ISSN No:-2456-2165

# Survey on Text Classification

Leena Bhuskute
Department of Information Technology
Pillai College of Engineering (PCE),
New Panvel, India

Satishkumar Varma
Department of Information Technology
Pillai College of Engineering (PCE),
New Panvel, India

**Abstract:- Now a day there is rapid growth of the World Wide Web. The work of instinctive classification of documents is a main process for organizing the information and knowledge discovery. The classification of e-documents, online bulletin, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing procedures to develop significant information. Therefore, proper classification and information detection from these assets is a fundamental field for study. Text classification is significant study topic in the area of text mining, where the documents are classified with supervised information. In this paper various text representation schemes and learning classifiers such as Naïve Bayes and Decision Tree algorithms are described with illustration for predefined classes. This present approaches are compared and distinguished based on quality assurance parameters.**

***Keywords:-*** *Classifiers, Machine learning, Supervised learning, Text classification.*

## I. INTRODUCTION

The main aim of text categorization is to classify documents into a particular variety of predefined classes. The proposed methodology distributes the documents into sentences and classifies each sentence using keyword lists of every class and sentence similarity measure and then, it uses the categorized sentences for training. The proposed methodology shows an identical degree of performance, compared with the normal supervised learning methods. Therefore, this method is used in areas wherever affordable text categorization is required .It can also be used for making training documents with the rapid use of the Web and the availability of on-line text information has.

Text classification has been considered as a significant method used to manage and process an unlimited amount of documents in digital formats that are widespread and increasing continuously. In general, text classification plays a crucial role in information extraction and summarization, text retrieval, and question answering.

| Title | Text Type | Purpose |
|---|---|---|
| Guide to the inflation deflation | Magazine article | To explain |
| Today I lost my job | Journal entry | To describe |
| Where do we go from here | Editorial column in paper | To comment |
| Why we should nationalize all banks now | Speech | To argue |

Table 1:- Examples of Text Classification

## II. RELATED WORK

The previous works in this field have used numerous labeled training documents for supervised learning. One problem is that it's difficult to make the labeled training documents. While it's easy to gather the unlabeled documents, it's not really easy to manually categorize them for making training documents. The proposed methodology splits the documents into sentences and classifies each sentence using keyword lists of every class and sentence similarity measure. Therefore, this method may be employed in areas where affordable text categorization is required. It can also be used for making training document.

## III. LITERATURE SURVEY

In this paper we describe the relevant literature survey that uses various techniques for different Text classification methods.

| Authors | Observations | Future Scope |
|---|---|---|
| Niklas Lavesson et al. 2010 [1] | Machine Learning (ML) is closely related with pattern recognition.<br>ML is more mathematical and more successful It use Probability & statistic to design ML. | Correlation exists between previous information & study.<br>Self-assessment test used in this idea could also be too restricted. |
| B S Harish et al. 2010 [2] | Bag of word (BOW) is one among various methods used for representation of text.<br>Term weighting techniques are used to give correct weights to the term and enhance the performance of text classification.<br>SVM with term weighted VSM is assigned for Representation. | Lack of imposed structure of traditional database<br>Different algorithms perform different actions depending on information collection. |

ISSN No:-2456-2165

| | | |
|---|---|---|
| Wen Zhang et al. 2011 [3] | LSI has good performance as compare to any other methods such as TF, IDF & Multi-words. LSI has favorable semantic, statistical quality and it can produce discriminative power for indexing | How to calculate the performances of indexing methods theoretically is a major difficulty. There are no idle measures to gauge two kind of quality that are semantic and statistical. |
| Mita K. Dalal et al. 2011 [4] | Decision tree algorithm is used as classifier. Naïve Bayesian approach was used as a preprocessor for dimensional property reduction followed by the SVM approach for text classification. The generic strategy for automatic text classification is used. | Decision tree classification does not assume independence among its feature. |
| P. Y. Pawar et. al. 2012 [5] | The document can be classified according to three ways such as unsupervised, supervised and semi supervised models. Efficiency of these models depends on the standard of the sample sets. | Supervised approaches must be trained on predefined positive & negative test samples. |
| Lam Hong Lee et al. 2010 [6] | Natural Language Processing (NLP), Data Mining and Machine Learning Strategies work together to distinguish and discover configurations in electronic text. Fuzzy correlation and Genetic algorithm are used. | The performance of a classification algorithm in text mining is noticeably influenced by the standard of data source. |
| Ildiko Pillan et al. 2014 [7] | Text level readability and sentence level readability and machine learning experiments for readability are done. SVM algorithm used. | Text readability measure mentioned has limitations when used on very short passages containing 100 words or less. |
| Aaditya Jain et al. 2016 [8] | This paper tells us about the popular classifiers used for text classification. | However the individual classifiers show limited applicability according to their respective domain and scope. |
| Sang-Bum Kim et al. 2006 [9] | This probability model would be freelance characteristic model so that the presence of one characteristic does not disturb another characteristic in classification tasks. | Dependability among these cannot be demonstrated by naïve Bayesian classifier. |
| Mahmud Alkoffash et al. 2006 [10] | The system sees the labels of the training documents but not those of the test set. Arabic text is used to classify. | Arabic text like morphemes that will generated from one root which can cause a poor performance in terms of both accuracy and time. |

Table 2:- Literature Survey in Detail

## IV.TEXT CLASSIFICATION METHODS

### *A. Machine Learning*

Machine learning is that the study of computer programs those progresses spontaneously through expertise. It is usually taken into consideration as branch of computer science and more precisely to subfield of Artificial Intelligence. Very similar to AI, machine learning may be a multidisciplinary exploration arena, drawing from work conducted in fields like measurements, arithmetic, environmental science, administration system, philosophy, technical model and psychosomatic science [1].

Although the territory of machine learning has been multidisciplinary in nature from its very initiation state to the present state of machine learning analysis is completely different from that of earlier analysis [1].

ISSN No:-2456-2165

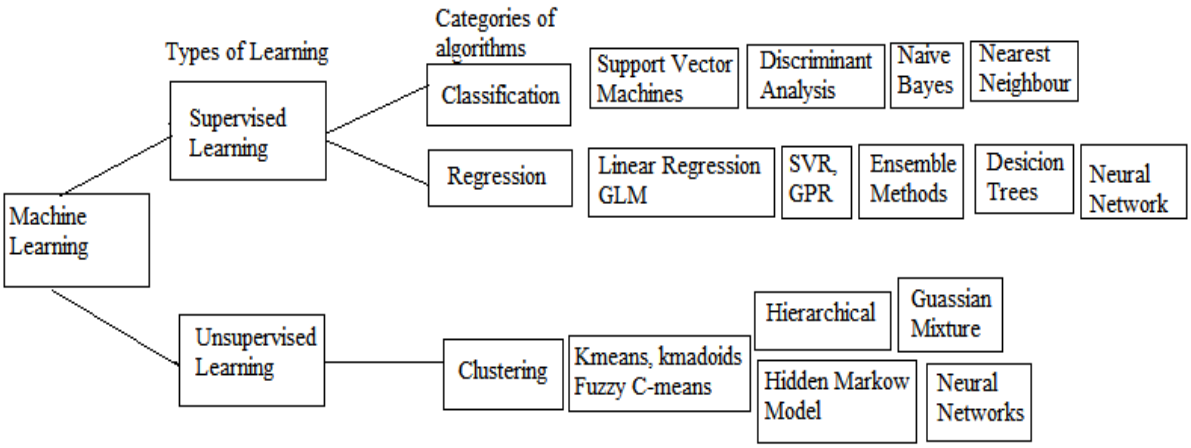


Fig 1:- Types of Learning Algorithm [6]

*B. Document Processing and Related Area*

In this step the document is preprocessed.

- *Text Clustering* – Make document collection with no external data.
- *Information Recovery* - Return a group of documents appropriate to a query.
- *Information Cleaning* - Filter documents that do not correspond to communication.
- *Information Mining* - Delete pieces of information. For example, names of people, ages and addresses in the text.
- *Text Categorization* - No query, communications, external details determine text topic.

*C. Text Classification Block Diagram*

Document presentation is one of the many additional ways to reduce complexity and make document easier to handle. Shortly it is nothing but transformation of document from the complete text version to a document vector. Text representation is that the vital facet in document classification and denotes the mapping of documents into a compact type of its contents. A main representative of the text classification problem is that the particularly high dimensional property of text information.

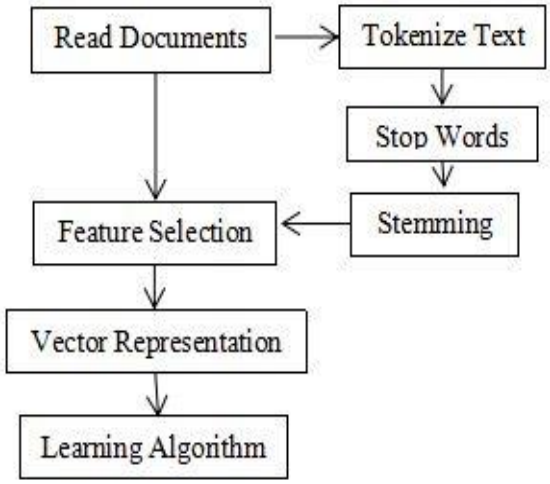


Fig 2:- Block Diagram of Text Classification

- *Read/Evaluate Documents*
- To read, evaluate and understand the document
- *Tokenize Text*
- Task of chopping it up into pieces, called tokens
- Text: "to sleep perchance to dream"
- Token: Examples: "to", "sleep", "perchance", "to", "dream"
- Type is that the category of all tokens containing identical character sequence. Examples: "to", "sleep", "perchance", "dream"
- Term could be a (perhaps normalized) type that is enclosed within the IR dictionary
- *Examples:* "sleep", "perchance", "dream"
- *Stop Words*
- Frequently occurring insignificant words need to be removed
- *Examples:* a, an, and, are, as, at, be, by, for, from, has, he, in, is, it.
- Stemming
- Removal of inflectional ending from words
- Process of conflating tokens to their root sort
- *Examples:* calculating to calculate, quantifying to quantify etc.
- *Examples:* Laughing, laugh, laughs, laughed.
- *Feature Selection*
- To form vector area
- Which improve the quality, strength and effectiveness and exactness of a text classifier.
- *Vector Representation*
- A document is a series of words
- So a document can be represented in vector format.
- *Learning Algorithm*
- Automatically constructs a classifier
- By learning the features of the categories
- Build a cluster of categorized documents
- Then uses the classifier to categorize documents into predefined categories.

## V.TEXT CLASSIFICATION TECHNIQUES

### A. NAIVE BAYES ALGORITHM

A naive Bayes algorithm is most familiar and real-world probabilistic classifier used in many applications. It considers that each one element of the examples are independent of every another element given the perspective of the category, that is nothing but independent hypothesis [9].

Naive Bayes algorithm, a family of classifiers that are supported by the popular Bayes theory, which is known for making simple but efficient models, especially in the areas of document segregation and disease prediction.. We can use this idea to train a naive Bayes spam filter and apply naive Bayes to song classification supported melodies.

*Example:* Let consider vehicle theft example where attributes are Color, Type, Origin and the subject is Stolen can be either yes or no.

| Colour | Type | Origin | Stolen |
|---|---|---|---|
| Blue | VAN | Domestic | Yes |
| Blue | VAN | Domestic | No |
| Blue | VAN | Domestic | Yes |
| Green | VAN | Domestic | No |
| Green | VAN | Imported | Yes |
| Green | MPV | Imported | No |
| Green | MPV | Imported | Yes |
| Green | MPV | Domestic | No |
| Blue | MPV | Imported | No |
| Blue | VAN | Imported | Yes |

Table 3:- Dataset for Naïve Bayes Classifier

Now we want to classify a Blue Domestic MPV. By using Naïve Bayes Classifier we can compute this. To do this we need to calculate the probabilities P(Blue|Yes), P(MPV|Yes), P(Domestic|Yes), P(Blue|No) , P(MPV|No) and P(Domestic|No) and multiply them by P(Yes) and P(No) respectively.

We can estimate these values by simply observing at above dataset.

$$P(Yes) = \frac{5}{10} = \frac{1}{2} \quad P(No) = \frac{5}{10} = \frac{1}{2}$$
$$P(Blue|Yes) = \frac{3}{5} \quad P(Blue|No) = \frac{2}{5}$$
$$P(MPV|Yes) = \frac{1}{4} \quad P(MPV|No) = \frac{3}{4}$$
$$P(Domestic|Yes) = \frac{2}{5} \quad P(Domestic|No) = \frac{3}{5}$$

According to Naïve Bayes Classifier

$$P(C|X) = P(x1|C) \times P(x2|C) \times \dots\dots\dots P(xn|C) \times P(C)$$

Values for Yes and No we have

$$Yes = P(X|Yes) \times P(Yes)$$
$$Yes = P(Blue|Yes) \times P(MPV|Yes) \times P(Domestic|Yes) \times P(Yes)$$
$$= \frac{3}{5} \times \frac{1}{4} \times \frac{2}{5} \times \frac{1}{2} = \frac{6}{200} = 0.03$$
$$No = P(X|No) \times P(No)$$
$$No = P(Blue|No) \times P(MPV|No) \times P(Domestic|No) \times P(No)$$
$$= \frac{2}{5} \times \frac{3}{4} \times \frac{3}{5} \times \frac{1}{2} = \frac{18}{200} = 0.09$$

Since 0.09 > 0.03, our example gets classified as 'No', and also there is no example of a Blue Domestic MPV in our dataset.

➢ *Advantages:*

- It Works fine with numeric and textual information.
- It is simple to develop.
- Perform computation very well.

➢ *Disadvantages:*

- The conditional thinking of freedom is violated by the details of the real world.
- Performance is worse when the elements are closely linked.

### B. DECISION TREE ALGORITHM

Decision tree is easy to understand as compare to other algorithms. It tries to solve problems using tree representation.

Decision tree can be used to generally represent decisions and decision making.

| Age | Income | Student | Cred_Sco | Buys_PC |
|---|---|---|---|---|
| Youth | High | No | Fair | No |
| Youth | High | No | Excellent | No |
| Mid_Age | High | No | Fair | Yes |
| Senior | Medium | No | Fair | Yes |
| Senior | Low | Yes | Fair | Yes |
| Senior | Low | Yes | Excellent | No |
| Mid_Age | Low | Yes | Excellent | Yes |
| Youth | Medium | No | Fair | No |
| Youth | Low | Yes | Fair | Yes |
| Senior | Medium | Yes | Fair | Yes |
| Youth | Medium | Yes | Excellent | Yes |
| Mid_Age | Medium | No | Excellent | Yes |
| Mid_Age | High | Yes | Fair | Yes |
| Senior | Medium | No | Excellent | No |

Table 4:- Dataset for Decision Tree

In this example, each element is discrete valued and above table presents a training set D.

Here class label element Buys_PC, has two discrete standards that is (Yes, No); therefore, there are two discrete classes (P, N).

Let assume, class P belongs to Yes and class N belongs to No.

Class P = Buys_PC = Yes = 9
Class N = Buys_PC = No = 5

Calculation of information gain of each element is required to find out splitting criteria for all these tuples.

*Step 1:* Determine expected information for categorizing tuple in D

$$Info\,(D) = -\sum_{I=1}^{C} P_i \log_2(P_i)$$

ISSN No:-2456-2165

$Info\ (D) = Info\ (P, N) = Info\ (9,5) = \frac{-9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$ = 0.94 bits

*Step 2:* Next, to work out on the probable information requirement for each element.

Let's continue with the element Age. So that looks at the distribution of yes and no tuples of Buys_PC elements for each category of Age.

| Age | Buys_PC | | Total |
|---|---|---|---|
| | Yes | No | |
| Youth | 2 | 3 | 5 |
| Mid_Age | 4 | 0 | 4 |
| Senior | 3 | 2 | 5 |
| | | | 14 |

Table 5:- Frequency Table For Elements Buys_Pc & Age

*Step 3:* Find the probable information required to classify a tuple in D, if the tuples are separated according to Age is

$$Info_A(D) = \sum_{J=1}^{m} \frac{|D_j|}{|D|} \times Info(D_j)$$

$Info_{Age}\ (D) = \frac{5}{14} \times \left(\frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14} \times \left(\frac{-4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14} \times \left(\frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$ = 0.694 bits

*Step 4:* The gain in information due to such a separating,
Gain (A) = Info (D) – $Info_A$ (D)
Gain (A) states how much gained by branching on A.
Gain (Age) = Info (D) – $Info_{Age}$ (D)
= 0.94 – 0.694 = 0.246 bits.
Similarly, we can find out Gain (Income) = 0.029 bits, Gain (Student) = 0.151 bits, Gain (Cred_Sco) = 0.048 bits

*Step 5:* Element Age set as splitting element, because of high information gain among all other elements.

A branch having entropy 0 act as a leaf node. Branches with entropy more than 0 needs further splitting. Because of this here Senior, Youth, Cred_Sco, Student nodes are splatted further and Mid_Age node not splatted further, remain as leaf node with labelling "Yes".

Final Decision Tree returned by the algorithm looks like as shown in Fig.3.

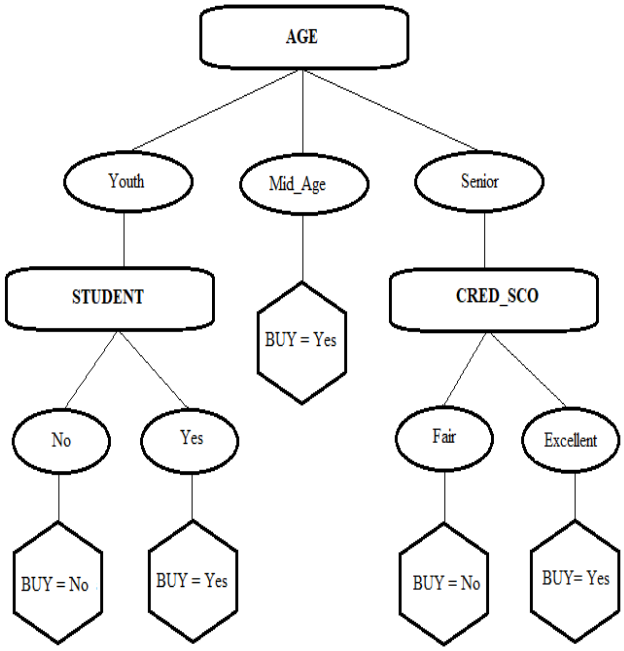


Fig 3:- Decision Tree

A decision tree can easily be converted to a group of rules by mapping from the root node to the leaf nodes one by one such as,

R1:-IF(Age=Senior)AND(Cred_Sco=FAIR) THEN Buy = No
R2:-IF(Age=Senior)AND(Cred_Sco=EXCELLENT) THEN Buy = Yes
R3:-IF(Age=Mid_Age) THEN Buy = Yes
R4:-IF(Age=Youth)AND(Student=No) THEN Buy = No
R5:-IF(Age=Youth)AND(Student=Yes) THEN Buy = Yes

These are the Decision Rules generated from above Decision tree.

➢ *Advantages:*
- It is simple to understand.
- Easily creates rules from decision tree.
- It minimizes problem complexity.

➢ *Disadvantages:*
- Training time is comparatively high.
- A document is just connected to one branch because of this once a mistake is happened at a higher level; whole sub tree is going to be wrong.
- It doesn't handle continuous variable well.
- It is going to suffer from over fitting some times.

ISSN No:-2456-2165

## VI. COMPARISON OF ALGORITHMS

The comparison of two algorithms is described in short in below table

| Sr No. | Parameters | Decision Tree | Naïve Bayesian |
|---|---|---|---|
| 1. | Define | Decision tree when used for text classification it contains tree internal elements marked as terms, branches transient from them are marked by test on the weight. | Naive bias methodology is kind of segment classifier under known priori possibility and class conditional assumption. |
| 2. | Ease of use | simplicity in understanding and interpreting, | Easy for implementation and computation. |
| 3. | Cost | Low | Moderate |
| 4. | Application | Useful when modelling human decisions and behaviour. | It is mostly used in document classification |
| 5. | Speed | Fast | Slow |
| 6. | Space | More | Less |

Table 6:- Comparison of Algorithms

## VII. COMMON EVALUATION METRICS

Common evaluation metrics can be measured in various parameters, such as

- Accuracy
- Precision
- Recall

Accuracy is the most common performance measure that is nothing but proportion of correctly expected values to the all values. Precision is that the proportion of correctly expected positive values to the all expected positive values. Recall is that the proportion of correctly expected positive values to the all values in actual category.

The following four terms are used to describe these parameters

TP (True Positive):- Determined as a document being classified correctly as relating to a category.
FP (False Positive):- Determined as a document that is said to be related to the category incorrectly.
FN (False Negative):- Determined as a document that is not marked as related to a category but should be.
TN (True Negative):-Document that should not be marked as being in a particular category and are not.

| Accuracy | Precision | Recall |
|---|---|---|
| Accuracy can be defined as degree to that the results of a measuring, calculation and specification conform to the standard value. | Precision is nothing but refining a measurement, calculation and specification especially on the number of digits represented. | Recall means information retrieval is a statistical measure act as the fraction of all relevant material which can be returned by the search. |
| Accuracy = $\frac{TP+TN}{TP+FP+FN+TN}$ | Precision = $\frac{TP}{TP+FP}$ | Recall = $\frac{TP}{TP+FN}$ |

Table 7:- Common Evaluation Metrics

## VIII.CONCLUSION AND FUTURE SCOPE

This paper gives the detailed information regarding various text classification concepts. Different machine learning algorithms based on supervised and unsupervised learning perform their work differently with varying performance measures. This paper explains Naive Bayes and Decision Tree algorithms in detail how they classifies document with step by step procedures. In text classification how they allots one or more classifications to a document depending upon their content. Classifications are carefully chosen from previously known categories. The common evaluation parameters like Accuracy, Precision and Recall are generally used to quantify the performance of such algorithms. This study gives overall view about how certain algorithm will carry out text classification and by using common evaluation metrics performance of such algorithms can be evaluated. This tells that a lot of enhancement can still be done in text classification using data mining for obtaining better results and performance.

## REFERENCES

[1]. Niklas Lavesson, “Learning Machine Learning: A Case Study”, IEEE Transactions on Education, 2010, Vol No. 53, Issue No. 4, pp 672–676.
[2]. B S Harish, D S Guru, S Manjunath, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, 2010, pp 110-119.
[3]. Wen Zhang, Taketoshi Yoshida, Xijin Tang, "A comparative study of TFIDF, LSI and multi-words for text classification", Expert Systems with Applications, Volume No. 38, Issue No. 3, March 2011, pp 2758-2765.
[4]. Mita K. Dalal, Mukesh A. Zaver "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, Volume No. 28, Issue No.2, August 2011, pp 37-40.

[5]. Pratiksha Y. Pawar and S. H. Gawande "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Volume No. 2, Issue No. 4, August 2012, pp 423-436.
[6]. Lam Hong Lee, Aurangzeb Khan, Baharum Baharudin, Khairullah khan "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal Of Advances In Information Technology, Volume No. 1, Issue No. 1, February 2010, pp 4-20.
[7]. Ildikó Pilán, Elena Volodina, Richard Johansson "Rule-based and machine learning approaches for second language sentence-level readability", proceedings of the Conference: 9th Workshop on Innovative Use of NLP for Building Educational Applications, ACL 2014, USA, pp 174–184.
[8]. Aaditya Jain, Jyoti Mandowara "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification", International Journal of Computer Application, Volume No. 6, Issue No.2, April 2016, pp 126-129.
[9]. Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng "Some Effective Techniques for Naive Bayes Text Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Volume No. 18, Issue No. 11, November 2006, pp 1457-1466.
[10]. Mahmud Alkoffash, Adnan Alrabea, “Text Classification using KNN and NB”, International Journal of Computer Application (IJCA), Volume No. 24, 2006.