International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.75 (2015)

Text Document Classification using Rough Set theory and Multi-level Naïve Bayes

K.A.Vidhya
Research Scholar
Department of CSE,Anna University
avidhya06@gmail.com

T.V.Geetha
Professor, Department of CSE
Anna University
tv_g@hotmail.com

G.Aghila
Professor, Department of EEE
NIT Karikal,
aghilaa@gmail.com

Abstract— **Machine Learning algorithm plays a major role in designing an effective text document classification system. This work has been aimed at development of a text document classification system using Multi-Level Naïve Bayes and Rough set. The Rough Set Naïve Bayes system for document classification has been implemented and tested to improve the classification accuracy compared to the traditional Naïve Bayes, which utilizes "bag of words" approach in which, the classification of documents into the predefined categories is done by means of the probabilistic values. For improving the classification accuracy selecting the significant features itself becomes a vital task which is done through the Rough set theory, a mathematical tool utilizing Rough Set Attribute Reduction algorithm for reducing the number of features. In addition the proposed model utilizes Multilevel Naïve Bayes approach where the features from title, keyword and content are extracted for which a weightage factor is assigned to the words from the title and keyword and the word probabilities are calculated from the words obtainable in the content part. The precision rate on an average is high compared to the recall rate which proves that the system provides better text document classification with proposed Multi-level approach.**

Keywords: Supervised Learning, Text Mining, Naïve Bayes, Rough set theory, Feature Reduction.

I. INTRODUCTION

TEXT MINING [TM] is a fast growing area in research as in other areas like business process management and customer relationship management. Text document classification plays a vital role in text mining, information extraction and information retrieval. Text categorization is the process of classifying the document into predefined categories in which the proposed system has Multi-level hierarchy of predefined documents in-turn organized into different categories and each category represents a unique topic. Generally a Learner Classification system, categorizes the given new document according to the predefined categories available or defined earlier.

Document classification has always been an important application for information retrieval. It can improve the speed of information retrieval and aid in locating and obtaining the desired information rapidly and accurately. Document classification is the task of assigning a document to one or more pre-defined categories or classes based on its textual context. This problem has received extensive attention within the past few decades due to its potential applications in many fields [1], for example, information retrieval, collective filtering, spam filtering, and news stories classification [9].

The selection of the document feature and the process of classifying a new document based on the extracted document feature are two key issues for document classification algorithms. There are many machine learning algorithms that have been proposed for document classification in which the most popular methods include K-Nearest Neighbor (KNN), centroid classifier, naïve Bayes, decision tree, neural network, genetic algorithm, and support vector machine(SVM) [3], [7]. Most of the document classification method uses the vector space or bag-of-words model for representing document vectors. Each word in the documents corresponds to a feature in the vector representation. This leads to a dimensionality in thousands for a moderately sized document collection [3].

Recent researches on document classification have aimed to improve the accuracy of the classification in which, the large number of electronic documents scattered on the internet is characterized by the diversity in categories, slanted distribution, and difficulty in labeling. The handing out of such a large amount of information available on the internet is an immense defy for document categorization. Therefore, some classifiers have been designed for solving and preventing these problems out of which, hierarchical classification had engrossed significant interest [2] of research.

Linear categorization implies the categorization when the given categories are defined independently of one another. Most studies on document classification focus on this flat categorization. When the categories are assorted and the relationship between them is obscure, a hierarchical information organization method is required. The hierarchical classification system trains classifiers associated with the categories. The organization of a large number of categories as a tree aids the users to obtain information more rapidly and precisely.

Naïve Bayes supports supervised learning where the categories are predefined and learning model is built accordingly with that of the predefined categories [13]. Feature Selection addresses the issue of selecting the features that are highly informative among the available features and such selection itself is a problem which has been in research in the area of computational intelligence [7]. However the feature selector conserves the original meaning of the features after reduction unlike the other dimensionality reduction method [1].

Thus feature selection plays a vital role in selecting the number of features involved in document classification. Often there are many features involved in which the selection process is aimed at bringing unique set of features which helps in proper document classification.

In this paper a system has been proposed, implemented and tested aiming at selecting the number of features using quick reduct rough set algorithm and by using propositional selection of number of keywords from title, keywords and content then applying machine learning method to train the system using Naïve Bayes to achieve better classification accuracy.

The paper is organized in the following way: Section II for theoretical background, Section III for proposed system, Section IV for Evaluation and Results and Section V for Conclusion.

My sincere thanks to DST PURSE II research program which funded our research work and has been constantly supporting our research.

II. THEORETICAL BACKGROUND

Machine Learning is the study of methodologies for machine to learn about a system which can then be applied to the various tasks to design and implement the software. There are many tasks which are hard to handle like there exist no human experts, there exists human expert, the methodology keeps rambling, the problem should be personalized according to certain user [19], [25]. For example in certain manufacturing process there is a need to predict machine failures before they are to be analyzed by any sensor reading. Here a machine learning system has to be designed in order to study recorded data and subsequent machine failures along with the ability to learn prediction rules.

There are applications that need to be customized for each computer user separately. Consider, for example, a program needs to filter unwanted electronic mail messages in which different users will need different filters. It is unreasonable to expect each user to program his or her own rules, and it is infeasible to provide every user with a software engineer to keep the rules up-to-date. A machine learning system can learn which mail messages the user rejects and maintain the filtering rules automatically.

Machine learning addresses many of the same research questions as the fields of statistics [19], text data mining and market analysis with pscylogoical predictions, but with differences of emphasis. Statistics focuses on understanding the phenomena that have generated the data, often with the goal of testing different hypotheses about those phenomena. Text mining seeks to find patterns in the Text that are understandable by people. Psychological studies of human learning aspire to understand the mechanisms underlying the various learning behaviors exhibited by people like concept learning, skill acquisition and strategy change.

Text mining process includes two process [12] text refining which transforms free text into a statistical representation form in which data mining techniques can be applied and second is knowledge discovery which finds patterns or knowledge from the available data [3]. Classification of text documents have been used in various applications like filtering spam [14], e-mail categorization [15], e-mail monitoring and Ontology mapping [20] for yielding best results.

The task of the classifier is to use feature vectors to assign the represented category or class. Feature selection help us to focus the attention of a classification algorithm in those features that are the most relevant to predict the class. Although theoretically, if the full statistical distribution were known, using more set of features could only improve results but in practical learning scenarios it may be better to use a reduced set of representative features [15]. There are two main areas need to be taken into account in feature selection: accuracy and feature reduction. The other factor that needs to be considered is time complexity of the system but in this paper the focus is more towards accuracy and feature reduction. For every classification task, there are three main factors need to be considered: [15]. First task aims at feature selection and Feature reduction method where as the second task includes learning algorithm for classification and the third task involve classification of large dataset.

The concept of rough sets was proposed by Pawlak [8] as mathematical approach to handle imprecision, vagueness and uncertainty in data analysis. Rough set theory can be applied for feature reduction in database; given a dataset with vectorized attribute values it is possible to find a subset of a category.

The rough sets (RS) theory will have the difficulty in handling the real-valued attributes [12]. To solve this problem the attributes are quantified before [8] and a new dataset with numeric values are created but certain information might be lost during this process. In order to quantify large dataset many techniques of attribute reduction have been used to handle the data efficiently [12]. Of late there have been much interest in developing the methodologies that are capable of handling the data imprecision and vagueness and the many research has been done in the areas of fuzzy [17] and rough sets [17].

The Naïve Bayes classifier is generally used in text categorization (Lewis, 1998; Mitchell, 1997) due to its relatively good performance in large datasets and its ability to learn incrementally [10] which is a well known statistical method and has been successfully applied to classification tasks. The Naïve Bayesian classification method works even though there are some probability estimation errors and still gives reasonably accurate results [13]. Naïve Bayes method is a kind of simple classification method based on Bayes theory where "Naïve" refers to suppose of independent condition [19]. Although Naïve Bayes classification makes an unrealistic assumption that the values of the attributes are independent given the class of instance, this method is exceptionally successful with large datasets. The Naïve Bayes conditional probability is based on Bayes theorem for computing the conditional probability that given a document *d* and it belongs to category *c*

$$P(c/d) = \frac{P(d/c)P(c)}{P(d)} \quad (1)$$

Since Bayes theorem provides an effective way to calculate the posterior probability of each hypothesis given in the training data, this method is used as the basis for the naïve Bayes machine learning algorithm that calculates the probability for each possible hypothesis then outputs the most probable category. The two main design issues involved in applying to text classification problems are first to decide how to represent the arbitrary text document in terms of attribute values and second to calculate the probabilities required by the naïve Bayes classifier.

In text classification problem the number of features can easily raise to hundreds and thousands. This poses a big hurdle in applying many sophisticated learning algorithms to text classification. Thus dimensionality reduction and feature reduction methods are included to reduce the number of features either by selecting the original features or transforming the features into new features as some functions of existing ones [12].

Despite its popularity, there has been some confusion in the document classification community about the "Naive Bayes [NB]" classifier because there are two *different* generative model in common use, both of which make the Naive Bayes assumption. One model specifies that a document is represented by a vector of binary attributes indicating which words occur and do not occur in the document. The number of times a word occurs in a document is not captured. When calculating the probability of a document, one multiplies the probability of all the attribute values, including the probability of non-occurrence for words that do not occur in the document. Here the document is considered to be the event," and the absence or presence of words to be attributes of the event. This describes a distribution based on a Multinomial model as follows.

Multinomial model:

In the multinomial model, a document is an ordered sequence of word events, drawn from the same vocabulary *V* 23. The assumption is that the lengths of documents are independent of class. There again make a similar Naive Bayes assumption: that the probability of each word event in a document is independent of the word's context and position in the document. Thus, each document *di* is drawn from a multinomial distribution of words with as many independent trials as the length of *di.* This yields the familiar "bag of words" representation for documents. Define *Nit* to be the count of the number of times word *wt* occurs in document *di.* Then, the probability of a document given its class from Equation 4 is simply the multinomial distribution:

$$P(d_i \mid c_j;\theta) = P(\mid d_i \mid) \mid d_i \mid ! \prod_{t=1}^{\mid V \mid} \frac{P(w_t \mid c_j;\theta) N_{it}}{N_{it}!} \quad (2)$$

Word Probability Estimate:

$$\theta_{wt|Cj} = P(w_t \mid c_j;\theta) = \frac{1+\sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is} P(c_j \mid d_i)} \quad (3)$$

Maximumlikehood Estimate:

$$P(c_j \mid d_i;\theta) = \frac{P(c_j \mid \theta)P(d_i \mid c_j;\theta)}{P(d_i \mid \theta)} \quad (4)$$

Working Mode:

In contrast to the multi-variate Bernoulli event model, the multinomial model captures word frequency information in documents. In the case of continuous inputs Xi, we can of course continue to use equations (4) and (5) as the basis for designing a Naive Bayes classifier. However, when the Xi is continuous we must choose some other way to represent the distributions P (XijY).

A. *Rough sets*

Basic concepts of crisp rough sets can be identified from [1], [2], [16]. From the existing literature, Rough sets are based on the information system that is a pair IS = (U, A), where U is a non-empty finite set of objects called the universe or corpus and A is a nonempty finite set of attributes.
With all the subsets of attributes S ⊆ A, there is an equivalence relationIND(S):

$$IND(S) = \{(x,y) \in U^2 \mid \forall a \in S, a(x) = a(y)\}$$

The partition of U generated by IND(S) is denoted

$$U / IND(S) = \otimes\{U / IND(\{a\}) \mid a \in S\} \quad (6)$$

$$A \otimes B = \{X \cap Y \mid \forall X \in A, \forall Y \in B, X \cap Y \neq \phi\}$$ and if $(x,y) \in IND(S)$ then *x* and *y* are indiscernible by attributes from *S.* For any selected subset of attributes S, there will be sets of objects that are indiscernible based on those attributes. These indistinguishable sets of objects therefore define an equivalence or indiscernibility relation, referred to as the S-*indiscernibility relation*. However, the target set X can be approximated using only the information contained within S by constructing the S-lower and S-upper approximations of X:

$$\underline{S}X = \{x \in U \mid [x]_s \subseteq X\} \quad (7)$$

$$\overline{S}X = \{x \in U \mid [x]_s \cap X \neq \phi\} \quad (8)$$

The S-lower approximation, or positive region, is the union of all equivalence classes in $[x]_s$ which are contained by (i.e., are subsets of) the target set. $POSs(Q) = \bigcup_{X \in U/Q} \underline{S}X$ (9)

The lower approximation is the complete set of objects in U/S that can be positively (i.e., unambiguously) classified as belonging to target set X. [2]. The S-upper approximation is the union of all equivalence classes in $[x]_s$ which have non-empty intersection with the target set .The upper approximation is the complete set of objects that in U/S that cannot be positively classified as belonging to the complement $NEGs(Q) = U - \bigcup_{X \in U/Q} \overline{S}X$ (10)

The boundary region, given by set difference $< \underline{S}X, \overline{S}X >$ consists of those objects that can neither be ruled in nor ruled out as members of the target set X.

$$BNDs(Q) = \bigcup_{X \in U/Q} \overline{S}X - \bigcup_{X \in U/Q} \underline{S}X \quad (11)$$

The vital part in data analysis is discovering dependencies. A *reduct* R_{min} is defined as a minimal subset *R* of the initial attribute set C such that for a given set of decision attributes *D*, $\gamma_R(D) = \gamma_C(D)$. From the literature [1], R is a minimal subset, if $\gamma_{R-\{a\}}(D) \neq \gamma_R(D)$ for all $a \in X$. Number of attributes can be removed from the subset without affecting the dependency degree. Hence, a minimal subset by this definition may not be the global minimum. A given dataset may have many reduct sets, and the collection of all reduct is denoted by

$$R_{all} = \{X \mid X \subseteq C, \gamma_X(D) = \gamma_C(D), \gamma_{X-\{a\}}(D) \neq \gamma_X(D) \forall a \in X\} \quad (13)$$

The intersection of all the sets in R_{all} is called the *core*, the elements of which are those features that cannot be eliminated without introducing more contradictions to the representation of the dataset [2]. For many tasks, a reduct of minimal cardinality is ideally searched for. That attempts to locate a single element of the reduct set $R_{min} \subseteq R_{all}$:

$$R_{\min} = \{X \mid X \in R_{all} \forall Y \in R_{all}, |X| \leq |Y|\} \quad (14)$$

The goal of RS- feature reduction is to discover reduct set of keyword. In this paper machine learning technique Naïve Bayes in the framework of the knowledge reduction approach based on rough set theory has been depicted. As far as accuracy and conciseness are concerned, the learning algorithms based on rough sets have significant pre-eminence.
The two main design issues involved in applying to text classification problems are first to decide how to represent the arbitrary text document in terms of attribute values and second to calculate the probabilities required by the Naïve Bayes classifier. In text classification problem the number of features can easily raise to hundreds and thousands which poses a big hurdle in applying many sophisticated learning algorithms to text classification. Thus dimensionality reduction and feature reduction methods are included in the RSBN system to reduce the number of features either by selecting the original features or transforming the features into new features as some functions of existing ones [16].

III. PROPOSED SYSTEM

The Naïve Bayes classifier is an efficient classifier model where the implementation of model is to train a system and classify it to the corresponding category which is much simpler when compared to other machine learning algorithms [19]. Figure 1 depicts the architecture of RSBN, which includes the following modules like pre-processing, rough set Feature Reduction, Multi-Level Naïve Bayes [MLNB] and followed by testing phase.
However, it has two main issues like first the classification accuracy is not optimal when compared to other classification algorithms and the second one is that to addresses the non-parameterized attributes. To overcome these issues the RSBN model depicts the enhanced Naïve Bayes learning model with rough set theory; a mathematical tool has been implemented. From architecture of RSBN, after the training dataset is provided to the system the document is pre-processed using the standard process like tokenization, stemming and stop-word removal.

Feature Selection:

From figure1 the feature selection method rough set theory has been used to define a set of keywords retrieved from the content words through which an information system is constructed based on the presence of the keyword in the category or not. The exact experimental results and evaluation is presented in the following section which mainly aims to reduce the number of attributes in the total corpus and the information system is maintained to make the decision on the classification of the text documents. The predefined knowledge defined here are the set of keywords fetched from ***DMOZ.org.***

Multi-Level Naïve Bayes

In Text classification normally "Bag of Words" approach has been used in which the semantic and syntactic relationship has been ignored as a consequence the system is not sufficiently robust. To overcome this issue the multi-level approach has been proposed aiming at a robust classifier model to classify the documents into the predefined category available. The following algorithm reported by us has been implemented and tested with the appropriate results for Naïve Bayes and quick rough set reduction and table I depicts the notation that has been used in the algorithm.

IV. EXPERIMENTS AND EVALUATION

The concise model of the distribution of class labels in terms of predictor features is built and tested for accuracy of the proposed system. The resulting RSBN classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

i) Input representation and Dataset Split –RSBN Model

The Standard benchmark 20-newsgroups dataset has been utilized to test the proposed system. The dataset has 20 pre-defined categories in which each category has about 1000 files in a specified category. The dataset is divided into two sets like 750 files towards training and 250 for test dataset. For training phase the conditional probability is calculated for individual keywords across the categories with ***tf-idf*** formula [15]. However in most of files the keyword fields are missing; in such cases the proposed system will generate the five to six keywords depending on the frequency of occurrence of keyword and priority is given to the keyword that occurs in quick reduct set.

20-Newsgroup dataset in which alt.atheism, computer, recreation, science and Politics form the top level of hierarchy and the rest of the group takes next level and third level accordingly. The predefined knowledge for each category provided as a set of keywords fetched from Dmoz.org which has about 276 keywords combined to all the categories available. These keywords when found in the content or title word given an extra weightage factor of .2 along with the respective weightage factor from the title and keyword weightage which has been discussed in the training phase.

ii) Rough Set Theory- Quick Reduct Algorithm:

The sparse matrix is formed in which set of condition attributes are the keywords from content and decision attribute that is the correct classification for the 20-categories available. The total number of keywords retrieved from the contents after pre-processing came around 63686. The quick reduct algorithm in step five is implemented and a set of 16 keywords are retrieved from the dependency degree γ equalizes to 1. Even though the process is time consuming as the reduct set has to be computed by combining the keywords across all the 63868 keywords the reduced sets considered to the important keywords for classification. For the keywords present a weightage factor is assigned for the documents and saved as title features and the weightage factor is calculated according to the equation15 which forms the first level. For the second level the features from keyword of the semi-structured documents are retrieved and the weightage factor is calculated using step2 of the value to be retrieved. $w_{ij} = t_{if} * \log(\frac{n}{n_i})$ Features from the content are retrieved and the conditional probability is calculated for the words using the following equation by taking the frequency of words in to account with the naïve independence assumption such that the word occurs in a particular category.

$$P(category / word) = \frac{P(word / category).P(category)}{P(word)}$$

At the end of the training phase the weigtage factor of each word in title and keyword along with the quick reduct keyword set and probability value of each keyword across the category based on the tf-idf formula is obtained. The following figure2 describes the distribution of keywords among the 20-categories the inference is that the internet keyword occurs more frequently in the science group compared to the other categories and the religion keyword occurs more frequently in alt.atheism. Using the equation 16 and equation 15 the probability value and the weightage factor are calculated. In the following figure 50,100…300 represent the frequency of occurrence of the keyword.

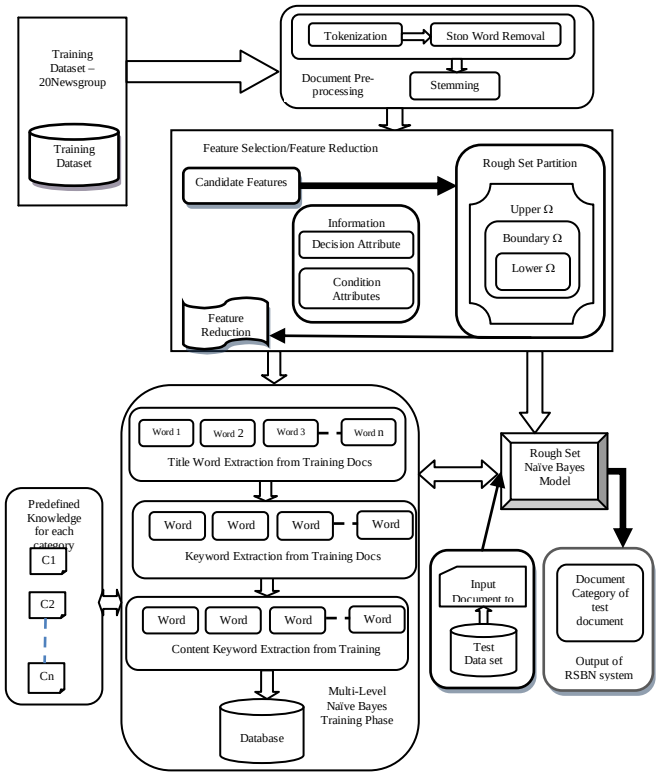


Fig1: Enhanced Version of Proposed System Architecture- RSNB from

TABLE I
Notation Explanation

Notation	Explanation
wi	weightage factor of word 'i'
tif	Term frequency of word 'i'
n	Total number of documents
Ni	Number of document in which the word 'i' has occurred
Pr(category/word)	Conditional probability of word with respect to category
Pr(category)	Probability of each individual category
P(tik|Cj)	Conditional probability that term 'i' occurs 'k' times in category Cj.
mk	Occurrence of word m {k=1 to n}
m	Number of unique words
|T|	Dictionary of words
R	Initial Null Set
x	subset of quick reduct set
C1,C2…Cn	N- Categories which are predefined

Procedure RSNB (Training Dataset, knowledge files, Test Dataset)

1. Retrieve the training documents from the dataset for individual category {C1,C2,....C10}

2. Calculate the weightage factor for title words, keywords using the following formulae $w_{ij} = t_{if} * \log(\frac{n}{n_i})$

3.Calculate word probabilities of the content words against the predefined category using Bayes theorem

$$P(category / word) = \frac{P(word / category).P(category)}{P(word)}$$

4. Annotate the individual document with keyword weightage, title weightage and content probabilities.

Quick Reduct Feature Reduction (From Literature by Pawlak)

QuickReduct (C, D) [From literature by Pawlak]

C, the set of all conditional attributes

D, the set of decision attributes

$R - \{\}$

do

$T \leftarrow R$

for each

if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$

$T \leftarrow R \cup \{x\}$

$R \leftarrow T$

until $\gamma_R(D) = \gamma_C(D)$

Return R.

RSNB Training Phase

5. Collect all the annotated documents with respect to each category and calculate probability for the entire training set using

$$P(t_{ik} \mid c_j) = \frac{1 + TF((t_{ik} \mid c_j).P(c_j))}{N=cj \mid T \mid + \sum_{s=1} TF(t_s, c_j)}$$

RSNB Testing Phase

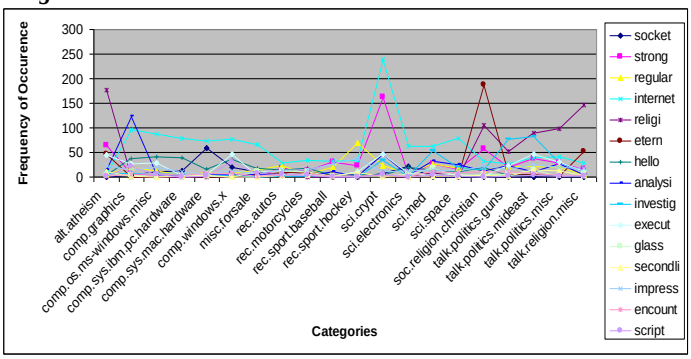
6. The most optimal category is assigned considering term frequencies belonging to individual in the test document by

$$ArgMax_{cj \in c} P(c_j \mid d_i) \approx ArgMax_{cj \in c} \prod_{k=1}^{N} P(t_{ik} \mid c_j).P(c_j)$$

i) Training Phase:

The following figure depicts the reduced number of keyword which exactly classifies the 20-categories and the inference is that certain keyword like cview has the maximum frequency of occurrence in category comp.graphics. The 20-Newsgroups dataset was considered in which the hierarchy of data is arranged and then values are retrieved. Once the training of the system is over with the training dataset the test document is used to the check the classification model formed by the proposed architecture. The above depicted algorithm ProcedureRSNB (RoughSetNaiveBayesAlgorithm) aims at classifying the document in to most optimal category. To evaluate the proposed classification system, precision, recall and F1-measure are used to evaluate the effectiveness of classification for the system. The table below shows an idea of how the proposed system will be evaluated with the precision, recall and F1-measure.

Fig 2: The distribution of the keywords among the categories



tp (True Positive): The number of documents correctly classified to that class .tn (True Negative): The number of documents correctly rejected from that class. fp (False Positive): The number of documents incorrectly rejected from that class. fn (False Negative): The number of documents incorrectly classified to that class.

$$P : \Pr ecision = tp/(tp + fp) \quad (17)$$

Figure 3 shows the accuracy of the multi-level Naïve Bayes when compared with the traditional model yields better accuracy except for the three groups alt.atheism, comp.os.windows.misc and comp.windows.x is calculated for the individual categories using the equation 20 and the inference is that on average MLNB yields 94% accuracy.

$$R : \mathrm{Re}call = tp/(tp + fn) \quad (18)$$

$$F1Measure = 2.(P.R)/(P + R) \quad (19)$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (20)$$

ii) Testing Phase:

The testing database has 250 files for each category out which the classification is done and the results for the number of correctly classified, incorrectly classified and the precision recall rate are specified in the figure 3, 4 and 5. The precision and the recall rate of the given test data shows the system performance and accuracy of the proposed Multi Level Naïve Bayes along with quick reduct rough set reduction.

iii) Output of RSBN System:

In recent research, Naïve Bayes method has been combined with other techniques by which the accuracy of document classification has been improved. In the proposed system given a test document that will be classified to the category not only depending on the maximum probability occurring in the training dataset but also the weightage factor of the keywords and title words along with the quick reduct set depending on the training documents available. The formulas for precision, recall and F-measure is given in (17), (18), (19) in which for text document classification usually the precision rate should be higher compared to the recall rate while for

information retrieval system like search engines the recall rate should be high compared to the precision rate.

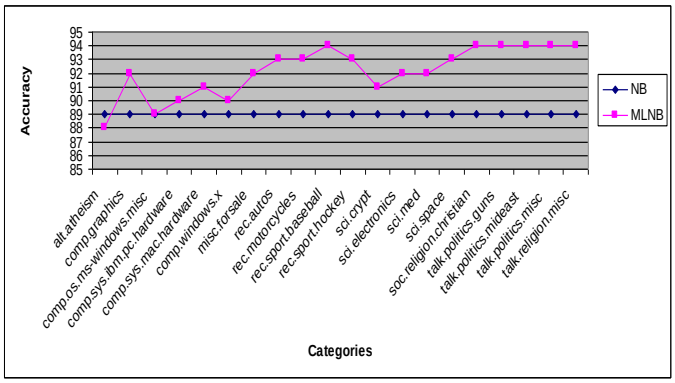


Fig 3: The accuracy of Naïve Bayes and Multi Level Naïve Bayes

The following figure illustrate the difference in F-measure of Naïve Bayes and Multi Level Naïve Bayes, which is calculated for the individual categories using the equation 19 and the inference is that on average MLNB yields average F-Measure of .95 when compared to NB which comes on an average .69.

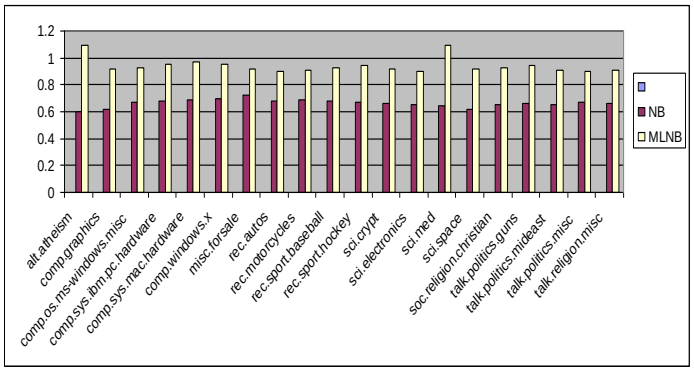


Fig 4: F-measure of Naïve Bayes and Multi Level Naïve Bayes

V. CONCLUSION

Machine learning approach for document classification has been in research for decades due to the explosive growth of digital documents and the learning method is required for documents available. The proposed model maintain a hierarchy of words rather than the traditional "bag of words" approach of Naïve Bayes and predefined knowledge is given by fetching the keywords from Dmoz.org for the relevant categories and the probability is calculated using the conditional probability of Naïve Bayes approach. The maximum likelihood is calculated based the number of documents available earlier in each category. For further improvisation of Naïve Bayes accuracy, of selecting proper features and reducing it through rough set quick reduct algorithm has been implemented and tested using the 20-Newsgroup standard dataset. From the experimental analysis it is obvious that the proposed method yields 94% accuracy in text document classification with the average F-measure rate of .95 which proves the system is better when compared to traditional Naïve Bayes approach. However the system has it own demerits like the runtime for the quick reduct algorithm is quite time intense which makes it critical when applied for online jobs. To overcome this issue the new approach for feature selection using particle swarm optimization and Naïve Bayes approach has been planned and testing is planned to be carried out using other standard dataset Reuters and webkdb.

REFERENCES

[1] A McCallum, K.Nigam. A comparison of event models for naïve Bayes text classification. AAAA-98 workshop on Learning for text Categorization, 2004.

[2] B. Kamens, Bayesian Filtering: Beyond Binary Classification. Fog Creek Software, 2005.

[3] Dino Isa, Lam Hong Kee, V.P. kallimani and R.Rajkumar "Text Document Pre-processing with Bayes formula for classification using SVM" IEEE Transactions on Knowledge and Data Engineering -2008.

[4] Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim "Exploiting noun phrases and semantic relationships for text document clustering " Information Sciences. www.elsevier.com/locate/ins.2008

[5] Hamel, L.; Nahar, N.; Poptsova, M.S.; Zhaxybayeva, O.; Gogarten, J.P.;" Unsupervised Learning in Spectral Genome Analysis" IEEE FBIT CNF 2007.

[6] Han X., Zu G., Ohyama W., Wakabayashi T., Kimura F., Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination, LNCS, Volume 3309, Jan 2004,pp. 463-468.

[7] Huan Liu, Senior Member, IEEE, and Lei Yu, Student Member, IEEE "Toward Integrating Feature Selection Algorithms for Classification and Clustering" IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, april 2005..

[8] J. Komorowski, Z. Pawlak, L. Polkowski, and A. Skowron, "Rough sets: A tutorial," in Rough-Fuzzy Hybridization: A New Trend in Decision Making, 1999.

[9] Pachghare, V.K.; Kulkarni, P.; Nikam, D.M. "Intrusion Detection system using Self Organizing Maps" International Conference on Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. 22-24 July 2009 Page(s):1 – 5.

[10] Sang-Bum kim, Kyong-soo Han, Hae-Chang Rim, Sung Hyon Myaeng "Some Effective techniques for Naïve Bayes Text Classification" IEEE Transactions on Knowledge and Data Engineering -2006.

[11] S.J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle, "A Case-Based Technique for Tracking Concept Drift in Spam Filtering,"J. Knowledge Based Systems, vol. 18, nos. 4-5, pp. 187-195, 2004.

[12] S.J. Delany, P. Cunningham, and L. Coyle, "An Assessment of Case-Based Reasoning for Spam Filtering," Artificial Intelligence J., vol. 24, nos. 3-4, pp. 359-378, 2005.

[13] Tao Jiang, Ah-Hwee Tan, Senior Member, IEEE, and Ke Wang"Mining Generalized Associations of Semantic Relations from Textual Web Content" IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, February 2007.

[14] Tseng, V.S.; Ja-Hwung Su; Hao-Hua Ku; Bo-Wen Wang;" Intelligent Concept-Oriented and Content-Based Image Retrieval by using data mining and query decomposition techniques" IEEE International Conference on Multimedia and Expo. June 23 2008-April 26 2008 Page(s):1273 – 1276

[15] YI-Hsing Chang, Hsiu-Yi Huang, "An automatic document classification Based on Naïve Bayes Classifier and Ontology".Proceedings of the seventh International conference on Machine Learning and Cybernetics,2008.

[16] Vishal Gupta , Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging Technologies in Web Intelligence, VOL 1, No.1, August 2009.

[17] Vidhya.K.A ,G. Aghila , "Hybrid Model for Text Document Classification", "The 2nd International conference on computer and Automation Engineering", 2010

[18] Vidhya.K.A, G. Aghila ," A Survey of Naïve Bayes Machine Learning approach in Text Document Classification", International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010

[19] Libiao Zhang, Yuefeng Li, Chao Sun, Wanvimol Nadee, "Rough Set Based Approach to Text Classification", IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT),2013.

[20] Leena Patil, Mohammed Atique ,"An Improved feature selection based on neighborhood positive approximation rough set in document classification", International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 5, No. 1, pp. 13-30, 2015.