A Multistage Feature Selection Model for Document Classification Using Information Gain and Rough Set

Mrs. Leena. H. Patil
Research Scholar, Dept. of Computer Sci. and Engg.
Sant Gadge Baba Amravati University
Amravati, India

Dr. Mohammed Atique
Associate Professor, Dept.of Computer Sci.and Engg.
Sant Gadge Baba Amravati University
Amravati, India

***Abstract*—Huge number of documents are increasing rapidly, therefore, to organize it in digitized form text categorization becomes an challenging issue. A major issue for text categorization is its large number of features. Most of the features are noisy, irrelevant and redundant, which may mislead the classifier. Hence, it is most important to reduce dimensionality of data to get smaller subset and provide the most gain in information. Feature selection techniques reduce the dimensionality of feature space. It also improves the overall accuracy and performance. Hence, to overcome the issues of text categorization feature selection is considered as an efficient technique. Therefore, we proposed a multistage feature selection model to improve the overall accuracy and performance of classification. In the first stage document preprocessing part is performed. Secondly, each term within the documents are ranked according to their importance for classification using the information gain. Thirdly rough set technique is applied to the terms which are ranked importantly and feature reduction is carried out. Finally a document classification is performed on the core features using Naive Bayes and KNN classifier. Experiments are carried out on three UCI datasets, Reuters 21578, Classic 04 and Newsgroup 20. Results show the better accuracy and performance of the proposed model.**

Keywords—Introduction; Document Preprocessing; Information Gain; Rough Set; Classifiers

I. INTRODUCTION

In the field of data mining, it has been observed that the data grow rapidly. With the rapid growth of data and the availability an increasing number of electronic documents, the task of classification becomes a key method [1]. Document preprocessing is an important parameter and feature selection is a common problem used in preprocessing for machine learning, data mining and pattern recognition [1][2]. Text categorization has always been a hot topic due to explosive growth of digital documents available. Due to huge development information acquirement and storage, tens, hundreds and even thousands of features [16] are acquired and stored in real world databases. Storing and processing relevant or irrelevant attributes becomes computationally very expensive and impractical [16]. A major problem of text categorization is its high dimensionality of features, due to which it misleads to the classifier [8]. The computational complexity of machine learning methods used for text categorization be increased and may bring about inefficient and results of low accuracy due to redundant or irrelevant terms in the feature space [6][14]. Mostly it is important to reduce dimensionality of the data to smaller set of features and relevant information for decreasing the cost in storing and reduction in the processing time [6], [13]. To overcome this, few attributes can be omitted, which will not seriously affect on classification accuracy. Many techniques in feature selection have been categorized namely filter and wrapper. The former employs to select attributes according to some significance measures such as consistency [4], information gain [3], distance [5], dependency [6] and others, later employs a learning algorithm to evaluate the attribute subsets. Rough set theory proposed [13] as a tool to organize conceptualize and analyze various types of data from knowledge discovery. It is useful in dealing with uncertainty and vague, knowledge information system. Rough set theory with attribute reduction offers a systematic framework for [15] distance based measures which attempt to retain the ability of original features for the objects from the universe. In a wide range of text categorization many feature selection methods are used [19]. Information gain is considered as the most effective method compared to other methods such as term strength, mutual information, X2 statistic, document frequency [24].

The task of text categorization is to classify the documents into predefined categories based on the contents of document [24]. Many methods have been applied to text categorization task on machine learning, such as KNN, Naive Bayes, C4.5 and SVM[14][15]. Several dimension reduction techniques like PCA, GA, IG [19] are carried out; still the problem of time complexity and text categorization[16][24] can be improved. Hence, in this we proposed multistage approaches: document preprocessing, feature selection and reduction technique which are used to reduce the high dimensionality of feature space. It removes the redundant and irrelevant attributes and thereby decreases the computational complexity of the machine learning process and increases the performance of classification. In the first stage documents are preprocessed with various steps. In the second stage, information gain is used to rank the importance of the features. In third stage Rough set approach is used to reduce the attributes. Finally, to evaluate the effectiveness of dimension reduction methods, experiments are conducted on Reuters-21,578, Classic 04 and

Newsgroup 20 dataset collection. For overall accuracy and performance the different classifiers like KNN and Naive Bayes are used. The results show that the proposed model is able to achieve high categorization effectiveness as measured by precision, recall and F-measure.

II. PROPOSED MULTISTAGE MODEL

Figure 1 shows the outline of proposed multistage model for document classification. Preceding sections describes the different stages of proposed multistage model.

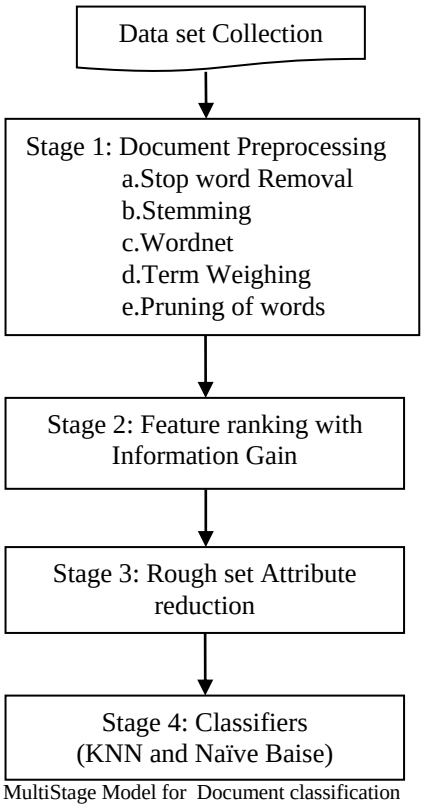

Fig. 1. Proposed MultiStage Model for Document classification

A. Document Preprocessing

To browse thousands of documents easily, document preprocessing becomes a most important tread. It articulates the required transformation processing to obtain the selected representation of documents. Thousands of words are present in a document set therefore; the aim of this is to reduce dimensionality to have better accuracy for classification [20]. Document preprocessing is divided into following stages:

- Stop words removal: Stop word List is used that contains the words to be expelled. The Stop word list is applied to remove terms that have a special meaning, but do not separate for topics.
- Word Stemming: Stemming algorithm such as porters is used to shrink a word to its stem or root form.
- WordNet: WordNet Senses Disambiguation is applied as an English Database.
- Global Unique words and frequent word set gets generated.

1) Stop-Word Removing

Stop-words remove the non-information behavior words from the text documents and reduce noisy data. To categorize large amount of word documents, stop word removal pay the similar advantages. Firstly, it could save an enormous amount of space. Secondly, it reduces the noise and keeps the core words, and makes later processing more effective and efficient.

2) Word Stemming

This process is used for transforming the words into their stem. In many languages the various syntactic forms of words are used and explicate. The most important technique called stemming is used for the reduction of words into their root. Many words from the English language can be reduced to their base form or stem, e.g. searches, searching belong to root stem search. An algorithm Porter Stemmer is apposite to stem documents. It is really a relatively accurate.

3) Wordnet

WordNet is a linguistic English Database developed in cognitive science laboratory [20][22]of Princeton University. It organizes words into a group called sysnets. Each of these contains a collection of synonymous words and corresponds to a concept. Therefore, WordNet is considered to be an English thesaurus database which maps their concepts. WordNet has four category noun, verbs, adjectives and adverbs. Using lexical database the WordNet approach measures the relatedness of terms from the words. WordNet as a dictionary covers some specific terms from every subject related to their terms. WordNet as a lexical database record all the stemmed words from the standard documents into their specific lexical categories.

4) Term weighing

Words are converted into terms, and thereby have to be measured the appearance of terms. Such processing is defined as term weighing. Therefore, each document depends on the term vector form they contained. The document vector is in the following format:

$$d = \{ w_1, \dots, w_i \dots\dots, w_{|T|} \},$$

Where w_i is the weight of the term with number i in the document, T is the term set and |T| is the cardinality of T.

The $tf - df$ is used for its weighing scheme to obtain the term vector T. $tf - df$ (term frequency x document frequency) is represented by $tfdf_{ij}$ and evaluated for the value calculated by dividing the term frequency (TF) by the document frequency (DF), where TF is the number of times a term t_j appears in a document d_i divided by the total frequencies of all terms in d_i , and DF is used to determine the number of documents containing term t_j divided by the total number of documents in the document set D.

$$tfdf_{ij} = \frac{TF}{DF} ,$$

where $TF = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$

and $DF = \frac{|\{d_i\, t_j \in d_i \in D\}|}{|D|}$

5) Pruning of words

The pruning process basically filters less frequent features in a document collection. The term vector is very high-dimensional and sparse. Also, it's seen that a number of elements in the term vector is ' '0' '. Hence, it is therefore required to prune the words those appear less than two times in the documents. This procedure shrinks the term vector dimension further.

B. Feature Selection

Feature selection (FS) is a term given to the problem of selecting input attributes which are most predictive of a given outcome. The main objective of feature selection is to find the minimal feature subset from the problem domain which retain the suitable features, representing with high accuracy. Feature selection techniques are categorized into two forms: filter and wrapper. The filter method is used to select attributes according to their significance measures such as consistency [4], information gain [3], distance, dependency and others. The wrapper method employs a learning algorithm to evaluate the attribute subsets. The Significance measures are categorized into two: consistency based measure and distance based measure**.** From the original set of attributes the feature selection process chooses the subset of attributes. The main aim of feature selection recognizes the relevant features and [14][17] abolishes the irrelevant of dispensable features. The feature selection molded by two steps: First one rank the feature according to their importance and secondly, reduce the attributes. Hence, thereby decreases the computational time and increases the accuracy.

1) Feature Ranking with Information Gain

Many feature selection methods are successfully used for text categorization. [24] has compared five different feature selection methods like information gain, X2 statistic document frequency, term strength, and mutual information. They reported that information gain is the most effective method as compared with the other feature selection methods. [16] has presented that information has become one of the most popular approaches employed as a term importance criteria in the text document. [19] has presented that the information gain is based on information theory. [17] has proposed that before attribute reduction each term within the text are ranked depending on their importance for the classification. The terms are arranged in decreasing order using information gain. With this process for classification term of less importance are removed and terms of highest importance are identified, where attribute reduction methods are applied.

[9] has presented that a major problem of text classification is the high dimensionality of feature space and redundant terms. Therefore it is desirable to find some methods which can reduce attributes for improving the overall performance of classification. To solve such issue Information Gain was proposed which defines the expected reduction in entropy caused by partitioning the text according to the term. The Information Gain of term t is defined as:

$$Information\ Gain\ (t) = \sum_{i=1}^{|c|} P(C_i) log\ P(\ C_i) + P(t) \sum_{i=1}^{|c|} P(C_i|t) log P(C_i|t) + P(\bar{t}) * \sum_{i=1}^{|c|} P(C_i|\bar{t})\ log\ P(\ C_i|\bar{t})$$

Where C_i represents the ith category. $P(C_i)$ is the probability of the i[th] category. $P(t)$ and $P(\bar{t})$ are the probabilities that the term t appears or not in the documents respectively $P(C_i|t)$ is the conditional probability of the ith category given that term t appeared and $P(C_i|\bar{t})$ is the conditional probability of the ith category given that the term t does not appear. Before attribute reduction each term within the text are ranked depending on their importance for the classification. The terms are arranged in decreasing order using information gain. With this process for classification term of less importance are removed and terms of highest importance are identified, where the attribute reduction methods are applied. Feature selection is the dimensionality reduction technique where the dimension space is reduced by selecting the best features which represents the document and inputting it to the classifier

2) Feature Selection using Rough set

After document preprocessing the IG method is applied where the terms of high importance in document are acquired. Through IG the number of terms in the document is reduced but still the problem is high dimensionality of feature space for text categorization. Hence to reduce the dimensionality and time complexity used for text categorization and to increase the performance, the attribute reduction based on rough set is carried out. The purpose of this method is to minimize the information loss and maximize the reduction in dimensions. In 1982, Pawlak introduced the concept of Rough set theory [13][14]. The theory initially developed for a finite universe of discussion in which the knowledge base is a partition, obtained by any equivalence relation. In rough sets theory, the data is organized in a table called decision table. Rows of the decision table correspond to objects, and columns correspond to attributes. In the data set, a class label to indicate the class to which each row belongs. The class label is called as decision attribute, the rest of the attributes are the condition attributes. [13][14] has developed a mathematical tool of rough set theory.

Definition 1: (Decision table). A decision table is an ordered tuple $S = < U, A, V, f >$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite set of objects; $A = C \cup D$ is a finite set of attributes, where C is a set of condition attributes, D = {d} represents the decision attribute (or class label), $C \cap D = \emptyset$; $V = \bigcup_{a \in A} V_a$, where V_a denotes the domain of attribute a; $f: U \times A \rightarrow V$ is an information function which associates a unique value of each attribute with every object belonging to U, such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$.

Definition 2: (Indiscernibility relation). Given a decision table $S =< U, C \cup D, V, f >$, and an attribute set $B \subseteq$

$(C \cup D)$, B determines an indiscernibility relation IND(B) on U as follows:

$$IND(B) = \{ (x,y) \epsilon U \times U : for\ all\ a \epsilon B , f(x,a) = f(y,a)\}$$

The equivalence relation IND(B) partitions the set U into disjoint subsets, which is denoted by U/IND(B) (or U/B), where an element from IND(B) is called an equivalence class. For every object $x \in U$, let $[x]_B$ denote the equivalence class of relation IND(B) that contains element x, called the equivalence class of x under relation IND(B).

Definition 3:(Lower and Upper approximation). For the given S a subset of attribute $A \subseteq Q$ determines the approximation space. $AS = (U, IND(A))$ in S. For given $A \subseteq Q$ and $X \subseteq U$, the A-lower approximation $\underline{A}X$ of the set X in AS and the A-upper approximation $\overline{A}$X of the set X in AS are defined as follows:

$$\underline{A}X = \{ x \in U : [x]_A \subseteq X \} = \cup \{ Y \in A^* : Y \subseteq X\}$$

$$\overline{A}X = \{ x \in U : [x]_A \cap \mathrm{X} \neq \phi \} = \cup \{ \mathrm{Y} \in \mathrm{A}^* : \mathrm{Y} \cap \mathrm{X} \neq \phi$$

Rough set provides the concept to determines for a given information system the most important attributes. The main idea of the reduct is fundamental for rough set theory. An essential part of an information system is a reduct which is related to a subset of attributes. Another important part is a core. The reduct and core is an important concept of rough set theory which is generally used for feature selection and attribute reduction. Rough set theory determines the significance measures, degree of attributes and dependency.

Definition 4: (Positive Region). For the given information system $S =< U, Q, V, f >$ with the condition and decision attribute. $Q = C \cup D$ A$\subset$ C can be defined as A positive region $POS_A(D)$ in the relation $IND(D)$ as

$$POS_A(D) = \cup \{ \underline{A}X : X \in IND(D)\}$$

$\mathrm{POS_A(D)}$ contains all the objects in U. A positive region for any two subsets of attributes A, $B \in Q$ in the information system S. The subset of attributes $B \in Q$ defines the indiscrenibility relation IND (B) which defines the classification $B * (U/IND(B))$ with respect to subset A. Positive region of B is defined as

$$POS_A(B) = \cup_{X \in B} \underline{A}\ X$$

Definition 5: (Dependency): Positive region of B contains the entire object. The cardinality of positive region B defines a measure $\gamma_A(B)$ of dependency of the set of attributes B on A

$$\gamma_A(B) = \frac{Card\ (POS_A\ (B))}{Card\ (U)}$$

From the information system S a set of all attributes B depends on A in S, which is denoted as A→ B; iff satisfies the equivalence relation $IND(A) \subseteq IND(B)$. Two sets A and B are independent of S if neither $A \rightarrow B$ nor $B \rightarrow A$ hold. The dependency of set B to degree K to the set A in S is denoted as

$$A \overset{k}{\rightarrow} B, \qquad 0 \leq k \leq 1 if\ k = \gamma_A(B)$$

Definition 6: (Significance) Rough set defines a measure of significance or coefficient of significance of the attribute $a \epsilon A$ from set A with respect to classification $B * (U/IND(B))$ generated by set B.

$$\mu_{A,B}(a) = \frac{card\ (\ POS_A(B)) - \ Card\ (\ POS_{A-\{a\}}(B))}{Card\ U}$$

A significance of attribute a in the set $A \subset Q$ can be computed with respect to original classification Q*.

Quick Reduct Algorithm is the most well-known algorithm for feature selection using Rough sets [12][13]. This is an incremental procedure, where it starts with an empty set and in each step a feature is added to the Reduct, in such way that dependency measure increases. The procedure stops when the dependency measure of the set of features being considered is equal to the dependency measure using all the conditional features. The algorithm attempts to calculate a reduct without exhaustively generating all possible subsets [13]. Its pseudo-code algorithm is given below:

QuickReduct(C,D)
C, the set of all conditional features;
D, the set of decision features.
(1) $R \leftarrow \{\ \}$
(2) do
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
(5) if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) Until $\gamma_R(D) = \gamma_C(D)$
(9) return R

The QUICKREDUCT algorithm attempts to calculate a reduct without fully generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset.

III. CLASSIFIER

The size of information grows rapidly, the problem arises of handling the data. It is infeasible to classify the data manually so automatic methods have been approached to reduce the time and effort for classification. Many document classifications have been built to categories the document according to their content. To improve the accuracy of classifier, researchers have worked on many ranking methods which select the term such as term frequency, chi squared, mutual information, and information gain. Still the problem arises is redundancy in the selected term. Redundant terms are equivalent to noise which causes a reduction in the accuracy of classifier. The classification accuracy changes according to the features being input to the classifier. If the features are of less redundant then the accuracy increases else it decreases. Feature selection algorithm with redundancy reduction for text classification, algorithm helps to decrease in redundant which improves the efficiency of the classifier.

A. Naive Bayes

Most Widely used classifier is the naive bayes. This classifier built the concept of probabilistic Classification where the probability is calculated for each document. it shows the belonging to the categories specified [10][12][13]. Many approaches using naive bayes classifier, multinomial naive bayes is used where the probability $P(C_j \mid d_i)$ of a document d_i belongs to the category C_j . C_j is calculated through the following equation :

$$P(C_j|d_i) = \frac{P(C_j) \prod_{k=1}^{|d|} P(w_{di,k}|C_j)}{\sum_{r=1}^{|c|} P(C_r) \prod_{k=1}^{|dj|} P(w_{d,k}|C_r|)}$$

where $|C|$ is the number of categories. $|d_i|$ be the length of document. $P(C_j)$ is probability of category is calculated according to the equation.

$$P(C_j) = \frac{1 + \sum_{i=1}^{|D|} P(C_j|d_i)}{|C| + |D|}$$

The probability of word gives that the category occurred $P\ (w_i \mid c_j)$ is calculated through the equation.c

$$P(w_i|c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_i, d_i)\ P\ (\ y_i = c_j|d_i)}{|v| + \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} N(\ w_{i,} d_i)\ P(y_i = c_j|d_i)}$$

where $|D|$ is the number of documents in the training set, $|v|$ is the number of words in the training set.

B. K-Nearest Neighbor

The KNN [3][17] algorithm is a well-known instance-based approach that has been widely applied to text categorization due to its simplicity and accuracy. To categorize an unknown document, the KNN classifier ranks the document's neighbors among the training documents and uses the class labels of the k most similar neighbors. Similarity between two documents may be measured by the Euclidean distance, cosine measure, etc. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. If a specific category is shared by more than one of the k-nearest neighbors, then the sum of the similarity scores of those neighbors is obtained from the weight of that particular shared category [2]. When classification is done by means of the KNN, the most important parameter affecting classification is k-nearest neighbor number. Usually, the optimal value of k is empirically determined. k value is determined so that it would give the least classification error.

IV. Experimental Evaluation

1) Performance Analysis

To evaluate the accuracy of text categorization results f-measure, precision and recall are used. These significance measures are mostly used to evaluate the accuracy of the result of classifiers for text categorization. The f-measure shows the combination of both precision and recall used in information retrieval. Precision is the proportion of correctly proposed document to the proposed document. Recall is the proportion of the correctly proposed documents to the test data that have to be proposed. In this paper F-measure, Precision, and Recall are not separated, they are computed for each class and average values of measures are used. Precision P and Recall R of each class are defined in equation below

$$\boldsymbol{P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

$$\boldsymbol{R} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

$$\boldsymbol{F} = \frac{\boldsymbol{2*P*R}}{\boldsymbol{P+R}}$$

Where TP, FP, and FN are true positive, false positive and false negative.

2) Results

The data used in the experiments are outlined in Table I, where the three datasets are used and downloaded from UCI machine learning databases. In the first stage preprocessing performed in four steps. Firstly stop words are removed, those which are useless for classification and may not be longer used. Stop words are removed according to the stop word list of 571 words. After stop word removal porter stemming algorithm is applied for stemming which reduce a words to its stem or root form. In the third step a wordNet, English thesaurus database is applied to have sense word. Lastly, the document vectors with tfidf weighing scheme is applied and the terms are extracted. All these process runs on Personal Computer with Windows XP and Intel® Core™ i7 CPU 2.66 GHz, 8.00 GB memory. The software used is MATLAB R2010b. The detail description about the preprocessed data is shown in Table II.

TABLE I. Data Set Description

Sr.No	Data Set	No. of Documents	Features	Classes
01	Reuters 21578	212	6539	04
02	Classic 4	54	1625	06
03	Newsgroup 20	52	1454	04

TABLE II. Preprocessed Document

Sr.No	Data Set	No. of Documents	Features Extracted
01	Reuters 21578	212	5677
02	Classic 4	54	1411
03	Newsgroup 20	52	976

The Classification KNN and Naive Bayes are applied on the whole dataset Reuters 21578, Classic04 and Newsgroup 20 and overall performances are examined. The results using KNN and Naive Bayes are summarized in Table III.

TABLE III. PERFORMANCE ANALYSIS ON THREE DATASET USING KNN AND NAIVE BAYES CLASSIFIER

Data set	No. of Features	KNN			Naive Bayes		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Reuters 21578	5677	82.79	93.45	87.79	79.78	83.90	81.78
Classic 04	1411	84.67	89.76	87.14	78.89	83.45	81.10
Newsgroup 20	976	81.98	91.89	86.64	80.97	87.56	84.13

The result shows that better accuracy is obtained by using KNN as compared to Naive Bayes Classifier.

In the Second Stage after the preprocessing, Information Gain is applied where the features are ranked and it reduces the dimension of feature space. In this the features are ranked individually and the classification is performed. The Classifier performance is examined with IG method and the results are shown in Table IV.

TABLE IV. PERFORMANCE ANALYSIS WITH IG USING KNN AND NAIVE BAYES CLASSIFIER

Date set	No. of Features	KNN			Naive Bayes		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
Reuters 21578	497	90.67	94.55	92.56	87.67	88.78	88.22
Classic 04	112	90.86	92.67	91.75	89.56	93.67	91.56
Newsgroup 20	43	88.97	96.78	92.71	81.78	92.76	86.92

In third stage after Information gain dimension reduction using rough set is examined. Using Rough set the attributes are reduct due to which dimension gets reduce and the feature space also decreases. Table V shows the classification performance of the feature ranking and the feature reduction performed by IG-RS method.

TABLE V. ERFORMANCE ANALYSIS WITH IG-RS USING KNN AND NAIVE BAYES CLASSIFIER

Date set	Features	KNN			Naive Bayes		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
Reuters 21578	3576	91.75	95.58	93.63	88.87	91.46	90.15
Classic 04	867	92.46	96.33	94.38	91.67	95.89	93.73
Newsgroup 20	292	90.64	97.91	94.13	83.08	94.25	88.32

The results show the better accuracy for KNN classifier as compared to Naive Bayes in terms of performance measures of classifier Precision, Recall and F-measure. With respect to the classifiers' performances, KNN Classifier shows higher performance than the Naive Bayes Classifier. Consequently, it is seen that a higher classifier performance is acquired with fewer features through hybrid methods.

V. CONCLUSION

A multistage feature selection model is proposed for document classification using information Gain and Rough set. Firstly, document preprocessing is carried out where the features are obtained through different steps like stopword removal, stemming, Wordnet, term weighing and pruning. On the original preprocessed document the classifier KNN and Naive Bayes are applied without dimension reduction and the classification performance are observed in terms of recall, precision and F-measures. Secondly, feature selection method Information Gain is applied in which features are ranked depending on their importance. Thirdly, rough set feature selection and attribute reduction is performed. Hence in proposed model features of less importance are ignored due to which dimensionality of feature space is reduced. Again computational time and complexity of the method are also reduced. At each stage classifiers performance are evaluated in term of precision, recall and f-measures. To analyze the effectiveness and accuracy of proposed model, experiments are performed using KNN and Naive Bayes classifier on Reuters 21578, Classic 04 and News Group 20.

From the experimental results, hence it is concluded that A Multistage feature selection model for document classification using Information Gain and Rough set is efficient to reduce the dimensionality of feature space.

Future scope of the work for text categorization is to reduce the dimensionality, computational time and complexity by developing different reduction algorithm. Also the classification performance can be improved by developing different hybrid model which will be more useful for document clustering.

REFERENCES

[1] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang, "An integration of WordNet and fuzzy association rule mining for multi-label document clustering", Data & Knowledge Engineering 69, 1208–1226, 2010.

[2] F. Beil, M. Ester, X. Xu, "Frequent term-based text clustering", Proc. of Int'l Conf. on knowledge Discovery and Data Mining KDD '02 pp. 436–442, 2002

[3] C.K. Lee, G.G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization", information Processing Manage 42 155–165, 2006

[4] M. Dash, H. Liu, "Consistency-based search in feature selection", Artificial Intelligence 151 155–176, 2003.

[5] R. Jensen, Q. Shen, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches", IEEE transactions on Knowledge and Data Engineering 16 (12) 1457– 471, 2004.

[6] Qinghua Hu, Daren Yu, Jinfu Liu, Congxin Wu, "Neighborhood rough set based heterogeneous feature subset selection" Information Sciences 178, 3577–3594, 2008.

[7] W. Pedrycz, G. Vukovich, "Feature analysis through information granulation and fuzzy sets", Pattern Recognition 35 825–834, 2002

[8] Q. Shen, R. Jensen, "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring", Pattern Recognition 37, 1351–1363, 2004

[9] R.B. Bhatt, M. Gopal, "On fuzzy-rough sets approach to feature selection", Pattern Recognition Letters 26, 965–975, 2005

[10] R.B. Bhatt, M. Gopal, "On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters 26 1632–1640, 2005.

[11] D.G. Chen, C.Z. Wang, Q.H. Hu, "A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets", Information Sciences 177, 3500–3518, 2007

[12] Z. Pawlak, "Rough Sets, Theoretical Aspects of Reasoning About Data", Kluwer Academic Publishers, Dordrecht, 1991.

[13] Z. Pawlak, A. Skowron, "Rough Sets: Some Extensions", Information Sciences 177, 28–40, 2007

[14] Yuhua Qian, Jiye Liang, Witold Pedrycz, Chuangyin Dang, Positive approximation: An accelerator for attribute reduction in rough set theory", Artificial Intelligence 174, 597–618, 2010.

[15] Yuhua Qiana, Jiye Lianga, Witold Pedrycz, Chuangyin Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework", Recognition 44, 1658–1670. 2011

[16] C.S. Yang, L. Shu, "Attribute reduction algorithm of incomplete decision table based on tolerance relation", Computer Technology and Development 16 (9), 68–69 72, 2006

[17] Y.H. Qian, J.Y. Liang, F. Wang, "A new method for measuring the uncertainty in incomplete information systems", Fuzziness and Knowledge-Based Systems 17 (6), 855–880, 2009.

[18] J.Y. Liang, K.S. Chin, C.Y. Dang, C.M. Yam Richid, "A new method for measuring uncertainty and fuzziness in rough set theory", International Journal of General Systems 31 (4) 331–342, 2002

[19] G.Y. Wang, H. Yu, D.C. Yang, "Decision table reduction based on conditional information entropy", Chinese Journal of Computer 25 (7) 759–66, 2002.

[20] C.L.Chen, F.S.C. Tseng, T. Liang, "An integration of fuzzy association rules and WordNet for document clustering", Proc. of the 3th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 147–159, 2009.

[21] Porter, M. F. "An algorithm for suffix stripping. Program", 14(3), 130 - 137, 1980.

[22] George A. Miller –"WordNet: A Lexical Database for English".

[23] H. Wang, "Nearest neighbors by neighborhood counting", IEEE Transactions on PAMI 28 942–953, 2006

[24] Harun Uguz, "A Two stage Feature selection method for text categorization by using Information gain, Principal Component analysis and Genetic Algorithm", Knowledge Based System 24,1024-1032, 2011.