International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 12 Issue 2 –JANUARY 2015.

COMPARATIVE STUDY OF FEATURE SELECTION AND CLASSIFICATION OF INDIAN ONLINE NEWS

Babu Renga Rajan.S[#1] Dr.K.Ramar[#2] Dr.K.G.Srinivasagan[#3]

[#1]*Professor, CSE & IT, PET Engg College, Vallioor, Tamilandu, India, it.sbr@petengg.ac.in*

[#2] *Professor, CSE & Principal, Einstein College of Engg.,Tirunelveli, Tamilandu, India*

[#3] *Professor, CSE (PG), National Engg. College, Kovilpatti, Tamilandu, India*

Abstract – Classification is plays major role in all areas of domains. Especially, in this era the data and information are largely influences the internet community. In the real time scenario, the user is in need of quite appropriate data / information in the internet pool. The main challenge is the data / information is stored in unstructured manner. In this paper, focus the various methods to classify the news in online mode with specific features. The performance of classifiers can be improved by reducing the number of features to be processed using four most effective feature selection methods viz. Document Frequency, Information Gain, Mutual Information, Chi-Square. K-NN, Decision Tree, Naive Bayes and Support Vector Machine classifiers are implemented and experimented with standard documents. The performance analysis is compared with respect to precision and recall. The classification results are 0.827, 0.903, 0.934, and 0.942 for precision and 0.756, 0.896, 0.929 and 0.936 for recall.

***Keywords:-** Feature Selection, Classification algorithms, Navie Bayes, Support vector machine*

I. INTRODUCTION

Due to the rapid growth of internet information a huge volume of unstructured data is available for the users. This information is useful to the users in plenty of ways. Document classification is one of the most important process of document analysis. If a system provides accurate and faster classification it would be useful to the internet community.

Text Classification (TC) is one type of classification tasks. It can automatically assign natural language texts based on their content to predefined classes or categories [1]. Classification has variety of applications like Web pages organized into category hierarchies, Journal articles indexed by subject categories, Patient records coded using international insurance categories, E-mail message filtering, News events tracked and filtered by topics, etc. Major problem of text categorization is the high dimensionality of the feature space. Many learning algorithms cannot handle such high dimensionality. Moreover most of these dimensions are not significant to text categorization. The process of selecting the representative features from the original feature space is called feature selection. At present the feature selection method is relied on statistical theory and machine learning. Few Feature selection methods are Document Frequency,-Inverse Document Frequency (TF-IDF), information gain, Gain Ratio, Gini Index, mutual information, Chi Square (CHI) and so on.

The rest of this paper is organized as follows. The next section provides a brief review of related works. Section III presents the proposed framework. The experimental results are discussed in section IV. Section V concludes the paper and put forward the directions of our future works.

II. RELATED WORKS

Several researchers handled feature selection process for Text classification with different forms like decision trees, naive-bayes, neural networks, and lately, support vector machines. Dash and Liu [2] gave a survey of feature selection methods for classification. Yang and Pedersen [3] compared state of the art five feature selection methods, such as document frequency (DF), information gain (IG), mutual information (MI), chi-square (CHI) and term strength (TS). The study results found IG and CHI to be the most effective. Although many techniques have been proposed, TC is a major area of research because of its effectiveness of present classifiers and still needs improvement. Classifier model is built describing a predefined set of classes and that model is used for predicting class labels of unlabelled sample [4]. Many researchers covered up the study of feature selection methods and classification algorithms in text classification. Some of them only had compared feature selection methods for a single classification algorithm [5],[6],[7]. Few of them undertook to compare multi feature selection methods and multi classification algorithms [8].[9],[10].

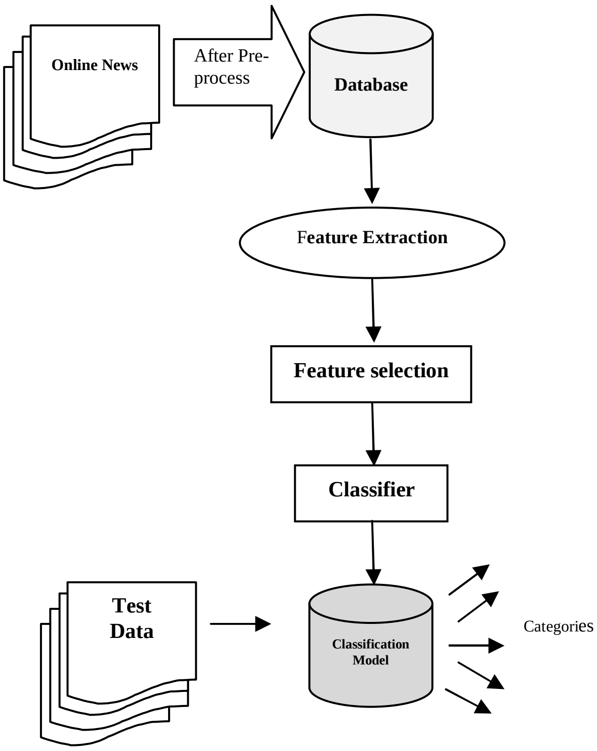

Figure 1: Framework of Classification

III. PROPOSED FRAMEWORK

The proposed framework for Document classification model is shown in Figure 1. The process is made up of three part Feature Extraction, Feature selection and Classification. The online news is gathered from websites in following categories, Business, Cricket, Entertainment, Hockey, Politics and Weather. The news items are collected and stored in databases after pre-process for further process. The data collected are distributed with six classes as shown in Figure 2.

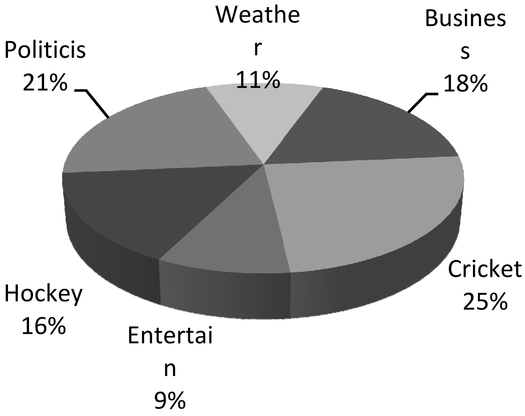

Figure 2 Distribution of News

A. Feature Extraction

The Feature extraction step is to transform the online data in the database into a format suitable for data mining by extracting and reducing terms from text. Natural language processing (NLP) technique is used to extract and reduce terms from online news collection. The following four processes of NLP can be applied

- Removal of HTML Tags,
- Tokenization,
- Stop Words Removal and
- Stemming

Removals of HTML Tags - All the online news collected from sources are converted into plain text after the removal of HTML tags.

Tokenization – Each document is treated as a string, and then partitioned into a list of tokens. This process of splitting terms from text is carried out by determining space between each term as a separator.

Stop Words Removal - In any language many words structure of language grammar. In English the words “a”, ”an”, ”the”, ”of” etc. are to be considered as stop words..The stop word list can be determined by the occurrence of such words [11], [12].

Stemming - The process to convert terms into their original form (root word) without prefixes and suffixes. For example “Learned” occurred 5 times, “Learning” occurred 3 times and “Learnt” occurred 2 times. They are converted to root form of word “Learn” and return the occurrence of 10 times. The stemming technique can reduce occurrence frequency of these words in documents by transforming them into one word as “Learn”. Porter Stemming algorithm is used to deal.

C. Feature Selection

Much research has been carried out on various feature selection algorithms. Feature Selection is to select the subset of features from original text documents. It can be achieved by keeping the informative words and remove the non-informative words. The informative words and non-informative words can be identified by the scores how it is helpful in classifying documents. Feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of the classifiers.

In this work, the following feature selection methods such as Information Gain, Mutual Information, Document Frequency and Chi-square are used on the collected online news data. The basic measures listed below are calculated for the feature selection process.

A - the number of documents in category, C_i, containing word/token, t.

B - the number of documents not in category, C_i, containing word/token, t.

C - the number of document in category, C_i,, not containing word/token, t.

D - the number of documents not in category, C_i, not containing word/token, t.

Document Frequency

Document frequency is a very simple feature selection method. Document frequency assumes that rare terms are “non-informative for category prediction, or non-influential in global performance" [3], and terms with higher document frequency are more informative for classification". Document frequency is calculated from A, B, C, D values as

$$DF = A + B \qquad - (1)$$

Only the terms that occur in a large number of documents are retained. Yang and Pedersen’s experiments showed that it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness [13].

Information Gain

Information Gain is a measure of dependence between the feature and the class label. It is one of the most popular feature selection techniques as it is easy to compute and simple to interpret. It is commonly used as a term goodness criterion in machine learning [14], [15]. Information gain value measures the number of bits of information obtained for category prediction by knowing presence or absence of a term in a document. Information gain value is calculated as

$$\mathrm{IG} = \mathrm{A.logA} + \mathrm{B.logB} + \mathrm{C.logC} + \mathrm{D.logD} + (\mathrm{A}+\mathrm{B}).\log(\mathrm{A}+\mathrm{B}) – (\mathrm{C} + \mathrm{D}).\log(\mathrm{C}+\mathrm{D}) – (2)$$

Mutual Information

Informally, MI compares the probability of observing t and Ct together (the joint probability) with the probabilities of observing t and c independently (chance). Mutual information method assumes that the term with higher category ratio is more effective for classification" [1] Mutual information can be calculated as follows using our already calculated A,B, C, D values

$$\mathbf{MI} = \mathbf{log}\frac{\mathbf{A\ X\ N}}{(\mathbf{A}+\mathbf{C})(\mathbf{A}+\mathbf{B})} - (\mathbf{3})$$

Chi-Square

Chi-square [16] is used to assess two types of comparison: tests of goodness of fit and tests of independence. In feature selection it is used as a test of independence to assess whether the class label is independent of a particular feature. Chi square measures the lack of independence between a term, t, and the category, c. Chi square, can be calculated as follows,

$$\aleph^2 = \frac{\mathbf{N(AD - CB)^2}}{\mathbf{(A + C)(B + D)(A + B)(C + D)}} \text{ - } \mathbf{(4)}$$

N –Number of training documents.

D. Classification Algorithms

The following classification algorithms are employed to evaluate adapted the feature selection methods, k-Nearest Neighbor (kNN), Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM)

The k-nearest neighbour algorithm kNN [17] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. Calculate similarity between test document and each neighbour assign test document to the class which contains most of the neighbours.

Decision tree is also widely applied to document classification. Its tendency to base classifications on as few tests as possible can lead to poor performance on text classification. However, when there are a small number of structured attributes it showed better performance.

Based on Bayes principle, Naïve Bayes is used to calculate the characteristics of a new document using keywords and joint probability of document categories and estimate the probability of each class for a document.[18]

Support Vector Machine (SVM) [19] is a supervised classification algorithm. During training, this algorithm constructs a hyper-plane that maximally separates the positive and negative instances in the training set.

IV. Experiment Setup And Results

The data gathered was applied with all feature selection methods and its values have calculated and stored in database to verify the various feature selection methods functionality and how good it is going to reduce the dimensionality and improve the performance of a classifier.

A. Data Collection

Data was manually gathered with the help of some utilities, developed by us, that could help us gather data faster. 644 documents were manually collected and tagged with its class name. It is split into training set as well as test set in 10-cross fold. The online news data was collected from the websites of Indian news papers The Hindu [20], The Business Line [21] and Cricket web portal Cricinfo [22]. Then the files are pre-processed and stored in a database. Maximum number of words in a document is 1,585 and minimum number of words is 62.The density of the total dataset is 536.92 with range of 859 to 348.3.

B. Performance and Discussion

The gathered data needed preprocess and then feature selection methods applied and identified the best attributes based on its value received in each feature selection method and threshold. The total number of features in the initial set is reduced and the effectiveness of classifier is not reduced. After Tokenization the number of words in the whole dataset has 2,33,802 tokens or words. After the Stop Word removal process it contains 37.87% of words only. The stemming process reduced the feature space further. The Figure 3 depicted the number of features in all text documents to distinct word after stemming only 11.60%.

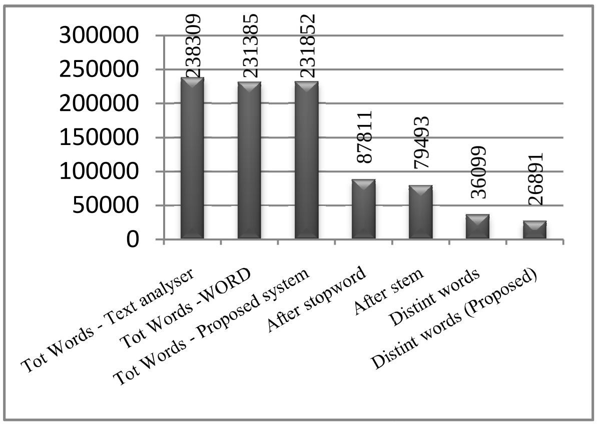

Figure 3 Feature Space after Pre-Process

All the feature selection methods are applied and its value is calculated for each feature set and stored in database. After finding out the optimum level of threshold the features above the threshold on each feature selection method was final feature sets. With this feature set classification process will be done on data set and performance measure was done.

The DF is ranging from 0 to 200 and used the cut off with features appeared at least in 10 documents in the Corpus. The accuracy is reduced by 5.9%.

IG approach has values between -13,400 and 0. We choose the threshold value of 9400 and the lost 2 more true positives only. MI ranges from -5 to 4 and the threshold value at 0.5 is same TP.

Chi square approach value ranged from 0 to 340. We fix the cut off value at 30 where 520 features are used and achieved almost same accuracy and if we increase the cut off the accuracy decreases. The number of features reduced in each method and the accuracy of classifier is listed in the Table 1 and achieved the accuracy as in figure 4.

TABLE. 1
COMPARISON OF NUMBER OF FEATURES AFTER PRE-PROCESS

FS	Before	After	Reduced	TP before	TP After
DF	26891	1859	25032	437	411
GI	26891	6890	20001	520	518
MI	26891	9677	17214	564	564
CHI	26891	520	26371	610	603

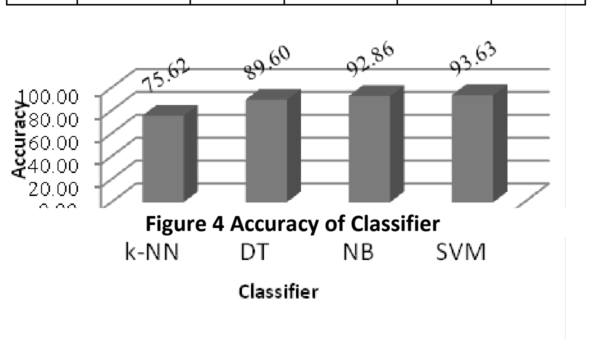

Figure 4 Accuracy of Classifier

Earlier SVM fetched 85.22% of accuracy and now it is improved to 93.63%. k-NN brought out 72.82% and after FS process it is 75.62%. Reduction of the feature set will improve the effectiveness of classifier with memory usage as well as computation time very much.

Second part of our experiments is done with the evaluation of classifiers [23]. 10 cross fold validation process of 638 documents is applied. Confusion matrix for TC is as in Table. 2

TABLE. 2
Confusion Matrix

		Predicted	
Actual		C	Not C
	C	True Positive	False Negative
	Not C	False Positive	True Negative

The simplest metric that can be used to evaluate a classifier, is accuracy measures the percentage of inputs in the test set that the classifier correctly labeled.

$$\text{Accuracy} \; = \frac{(\text{TP} \; + \; \text{TN})}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} . -(5)$$

Rarely if the classifier predicts more number of true negative values instead of true positive values, accuracy is value will not reflect the expectation of user. Use of Precision and Recall is better way to evaluate the classifiers.

$$\text{Precision (P)} \; = \frac{\text{TP}}{(\text{TP} \; + \; \text{FP})} - (6)$$

$$\text{Recall (R)of TC is} \; = \frac{\text{TP}}{(\text{TP} + \text{FN})} - (7)$$

$$\text{F1} \; = \frac{(2*\text{P}*\text{R})}{(\text{P+R})} \; - \; (8)$$

The 4 classifiers discussed earlier are deployed to classify the above documents with the best feature selection method CHI square method achieved better result compared to other classifiers. The memory space and execution time is optimized. The result is depicted in Table 3 and Figure.5 It is clearly evidence that Support Vector Machine has outplayed other classifiers.

TABLE . 3

Evaluation of Classifiers

	k-NN	DT	NB	SVM
Precision	0.827	0.903	0.934	0.942
Recall	0.756	0.896	0.929	0.936
F1	0.759	0.898	0.93	0.936

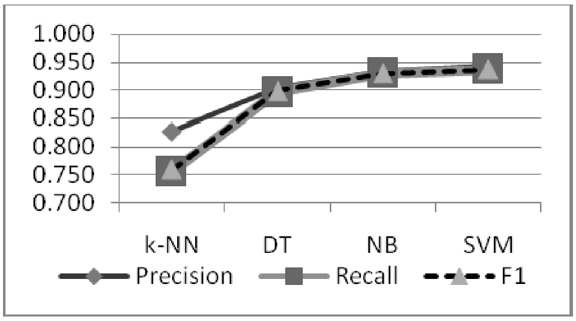

Figure 5 Clarifiers Accuracy

V. Conclusion

This Paper analyzes the four feature selection methods. The analysis imparted various threshold values for each feature selection method. It is confirmed that the Chi-square is the best feature selection method with the combination of SVM classifier bettered the k-NN, Naïve Bayes and Decision Tress classifiers in terms of its precision. SVM performed very well ahead of k-NN, DT and NB with 23.82% and 4.51% , 0.84 % respectively. If the more number of dataset is available then the test dataset can be provided to evaluate the test set separately instead of cross validation or split dataset.

REFERENCES

[1] *Miller, T.W. & Nguyen H.T. (2005). Data and Text Mining: A Business Applications Approach, Pearson Prentice Hall*

[2] *Dash, M. & Liu, H. (1997) "Feature Selection for Classification". Intelligent Data Analysis, Vol.1, no.3, pp. 131-156.*

[3] *Yang, Y. & Pederson, J.O. (1997) "A comparative Study on Feature Selection in Text Categorization", In Proceedings of the 14th International Conference on Machine Learning, pp. 412-420.*

[4] *Jiawei Han & Micheline Kamber (2012), Data Mining Concepts and Techniques, Morgan Kaufmann.*

[5] *Li, S., Xia, R., Zong, C. & Huang, C.R. (2009). A framework of feature selection methods for text categorization. Proc. Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing, pp.692-700.*

[6] *Pal, M. & Foody, G.M. (2010). Feature Selection for Classification of Hyperspectral Data by SVM. IEEE Transactions on Geoscience and Remote Sensing, Vol. 48, No. 5, pp.2297-2307.*

[7] *Meng, J. & Lin, H. (2010). A two-stage feature selection method for text categorization, Proc. 7th Int.Conf. on Fuzzy Systems and Knowledge Discovery,pp.1492-1496.*

[8] *Brank, J., Mladenić, D., Grobelnik, M. & Milić-Frayling, N. (2008). Feature Selection for the Classification of Large Document Collections. Journal of Universal Computer Science, Vol. 14, No. 10, pp.1562-159.*

[9] *Yang, Y. & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization.Proc. 14th Int. Conf. on Machine Learning, pp.412- 420.*

[10] *Cai, D.M., Gokhale, M. & Theiler, J. (2007). Comparison of feature selection and classification algorithms in identifying malicious executables. Journal of Computational Statistics & Data Analysis,Vol. 51, issue 6, pp.3156-3172.*

[11} *Ho. T.K, Fast Identification of Stop Words for Font Learning and Keyword Spotting. In Proceedings of the Fifth International Conference on Document Analysis and Recognition (pp. pp. 333-336). : IEEE Computer Society, 1999*

[12] *Wilbur, J. & Sirotkin, K. The automatic identification of stop words. Journal of Information Science, 18, pp. 45-55, 1992*

[13] *F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1-47. 2002.*

[14] *J.R. Quinlan. Induction of Decision Trees. Machine Learning, 1(1): pp.81-106, 1986 IG*

[15] *T. Mitchell. Machine Learning. McCraw Hill, 1996*

[16] *H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In J.F. Vassilopoulos,editor, Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, November 5-8, 1995, pages 388{391, Herndon, Virginia, 1995. IEEE Computer Society.*

[17] *Tam, V., Santoso, A., & Setiono, R. , "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", 16th International Conference on Pattern Recognition, 2002-4,235–238.*

[18] *Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification", IEEE, TKDE, Vol. 18, No. 11, , Pp-1457- 1466 ,November 2006.*

[19] *T. Joachims, Learning to Classify Text using Support Vector Machines, Kluwer, 2002.*

[20] *www.thehindu.com*

[21] *www.thebusinesslinenewspapers.com*

[22] *www.espncricinfo.com*

[23] *Marina Sokolova, Guy Lapalme A systematic analysis of performance measures for classification tasks, Information Processing and Management 45 (2009) 427–437*

Babu Renga Rajan He is working as Professor CSE & IT, PET Engg. College. He is Fellow member in Institution of Engineers, Life member of Indian Society for Technical Education, .His research interests include Data, Text mining, Natural Language Processing and information retrieval.

Dr.K.Ramar is presently working as a Principal in Einstein College of Engineering, Tirunelveli, Tamilnadu.. He is a Fellow member in the Institution of Engineers. Life member in Computer Society of India,, Indian Society for Technical Education, and Executive member in System Society of India..He has published more than 100 papers in international and national journals. His research interests include Pattern Recognition, Image Processing, Computer Networks, Fuzzy Logic based Systems and Data Mining.

Dr.K.G.Srinivasagan is presently working as Professor CSE - PG in National Engg. College, Kovilpatti, Tamilnadu. He is Life member of Computer Society of India, Indian Society for Technical Education. His research interests include Pattern Recognition, Image Processing, Data Mining, Natural Language Processing.