International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 517 - 523

A Comparative Study of Text Classification Methods: An Experimental Approach

Rupali P.Patil
Department of Computer Science
S.S.V.P.S's Lk. Dr. P. R. Ghogrey Science College
Dhule, India
rupalihpatil@yahoo.com

R. P. Bhavsar
School of Computer Sciences
North Maharashtra University
Jalgaon, India
rpbhavsar@nmu.ac.in

B. V. Pawar
School of Computer Sciences
North Maharashtra University
Jalgaon, India
bvpawar@nmu.ac.in

***Abstract*—**Text classification is the process in which text document is assigned to one or more predefined categories based on the contents of document. This paper focuses on experimentation of our implementation of three popular machine learning algorithms and their performance comparative evaluation on sample English Text document categorization. Three well known classifiers namely Naïve Bayes (NB), Centroid Based (CB) and K-Nearest Neighbor (KNN) were implemented and tested on same dataset R-52 chosen from Reuters-21578 corpus. For performance evaluation classical metrics like precision, recall and micro and macro F1-measures were used. For statistical comparison of the three classifiers Randomized Block Design method with T-test was applied. The experimental result exhibited that Centroid based classifier out performed with 97% Micro F1 measure. NB and KNN also produce satisfactory performance on the test dataset, with 91% Micro F1 measure and 89% Micro F1 measure respectively.

Keywords*-Machine Learning;Naïve Bayes; K-Nearest Neighbor; Centroid Based; Text Classification*

*****

I. INTRODUCTION

This is an era of computer and Internet. In recent years numbers of users accessing computers are increased. Because of development in resources used for communication, data can be easily sent from one location to another. Thus online resources are increased. Therefore, today internet is the main source of information. As a result, information stored on the web is increasing rapidly. This information may be in the form of text, numerals, images, graphs, audio and video. It is continuously growing in size and complexity. One can use this huge data effectively; if and only if it is properly managed and organized according to our need. As most of this above information (above 80%) is stored as text [1], there is a problem of proper organization and management of this huge textual information. Classification is helpful in this direction. Text classification is the process in which text document is assigned to one or more predefined categories based on the contents of document. In general, let D be a set of documents, d_j be a document belongs to it and let $\{c_1, c_2, \ldots, c_n\}$ is set of all classes (text categories), then text classification assign document d_j to a single class or more than one classes. If a document is assigned to single class it is known as single class classification where as if it is assigned to more than one classes then it is known as multi class classification. Text classification has number of applications such as Email Classification [2],Topic Spotting on News Articles [3], Language identification[4] etc. In general, text classification plays an important role in information extraction and summarization, text retrieval and question-answering system.

Two major approaches described for text categorization are Rule based and Machine learning based approaches [5]. In Rule based approach rules are defined manually and document is classified according to these rules. This approach is suitable when document set is small. As rules are defined by human experts this approach is very accurate. But it totally depends on rules defined by human experts. Also if the document domain coverage is large then defining rules can be very tedious job. Moreover human experts may require writing more rules if document set increases. In Machine learning based approach text classifier is built automatically from a set of predefined classes. That is for construction of classifier, there is no need of human expert. Thus this approach saves human efforts, time and till provides comparable accuracy which is achieved by domain experts.

In machine learning generally two types of learning algorithms are found in the literature: supervised learning algorithms and unsupervised learning algorithms [6]. In this paper we have considered only supervised learning algorithms. Supervised learning means learning from examples. As humans learn from past practices, a computer system uses data to learn, which denote some "past practices" of an application area. Goal of supervised learning is to build a classification model on the basis of the data. The classes of new cases/instances can be predicted by using this model. In supervised learning, the data are labeled with pre-defined classes. In learning (training) phase the training data is used to learn a model. In testing phase, unseen test data is used to test the model and to measure the accuracy of model. If training examples are proper characteristic of the test data, good accuracy can be achieved on the test data. Many statistical and

machine learning techniques have been proposed for document classification such as Naïve Bayes [7], K Nearest Neighbor [8], Support Vector Machine (SVM) [9], Decision Tree (DT) [10], Neural Network (NN) [11] etc.

In our study we have considered only supervised learning methods to learn our classifiers and estimate them on new test data set. Our study aims to compare three well known classification algorithms namely NB, KNN and Centroid Based. The performance of all three classifiers, on same data set is evaluated and compared by using same performance evaluation metrics.

The rest of the paper is planned as follows: Section II summarizes Literature Survey; while section III gives Methodology and theoretical description of NB, KNN and Centroid Based text classification algorithms used in this paper. Section IV describes Experimental setup and trials followed by Result and Discussion in section V. Section VI gives conclusion.

II. LITERATURE SURVEY

Since our aim is to compare NB, Centroid Based and KNN algorithms for English text we have considered the Reuters Standard English Dataset. Many works are already done in the area of text classification for classifying English text [12], [13]. In [14] Taeho Jo addresses two problems (high dimensionality and sparse distribution) of representing document using numerical vectors. He proposed a new neural network for text categorization called NTC (Neural Text Categorizer) which uses string vectors for document representation rather than numerical vector. For evaluating traditional (SVM, NB, KNN, Back Propagation) and proposed approach he has used three collections Newspage.com, 20NewsGroups and Reuters 21578. Experimental result shows that NTC is comparable with best traditional approach back propagation in terms of classification accuracy and the learning speed. In his study he has shown that NTC is more practical than others. Susan Dumais, John Platt, David Heckerman in [15] have compared effectiveness of five different inductive learning algorithms for text categorization namely Find Similarity, Naïve Bayes, Decision Tree, Bayesian Networks and Linear SVM, in terms of training speed, classification speed and classification accuracy. They have also compared training set size and alternative document representation. They have used new version of Reuters-21578 collection containing 75% of stories are for training purpose while remaining 25% are used for testing. For document representation they used tf*idf weighting for "Find Similar classifier" and for other classifiers they used binary representation. Mutual Information is used for feature selection. Their experimental results have showed that Support Vector Machines are more accurate with 92% accuracy on 10 most frequent categories and 87% accuracy with 118 categories. SVM is very fast to train and fast to evaluate. While Find Similar were lowest accurate with 64.6% accuracy for top 10 categories and 61.7% accuracy with all categories. But Find Similar is fastest learning method in all. In their experiment they showed that classification accuracy did not improve by using NLP derived phrases. Vidhya K. A, G. Aghila in [16] have proposed a hybrid text classification model based on Rough Set theory and Naïve Bayes Classifier. In their proposed model, for feature reduction Rough set theory is used and for classification of documents into the predefined categories Naïve Bayes theorem is used by means of the probabilistic values. The standard dataset Reuters-21578 and 20 Newsgroups are used. Instead of the traditional "bag of words approach, their model maintains a hierarchy of words. The proposed model improves the classification accuracy by overcoming the inaccuracy and ambiguity in data set. BaiRujiang and Liao Junhua in [17] have proposed a hybrid model RGSC-Rough Set and Genetic Algorithm for SVM classifier. They have used rough set theory to reduce the feature vector space and thus improve classification speed. To improve classification accuracy they present Genetic algorithm approach for feature selection and parameter optimization. They compared the proposed RGSC model with KNN and Decision Tree classifiers. For their experiment they used Reuters-21578 corpus. In their experiment for RGSC they got Average precision of 90.7%, Average Recall of 95.1% and Average F-measure of 92.5% which are greater than average precision, average recall and average F-measure of KNN and DT. Their experimental result showed that RGSC method is more effective than SVM and other traditional method. To improve the efficiency of basic EM method, Wen Han and Xiao Nan-feng in [18] have proposed an enhanced EM method, their approach is semi-supervised classification based on Naïve Bayesian. In their method they first reduce the feature space by applying DF*ICIF feature selection function and in subsequent iteration of EM method, using intermediate classifier unlabeled documents having maximum posterior category probability are transferred from unlabeled set to labeled collection. In enhanced EM numbers of iterations are less. Experimental results demonstrated that enhanced EM method obtains effective performance in terms of micro average accuracy and efficiency.

III. THEORETICAL BACKGROUND OF CLASSIFIERS

This section elaborates the important phases of Classification process, which by and large includes selection of classification method, text representation and dimensionality reduction and learning/training of classifier.

A. *Learning of Classifiers*

A Classifier is a model which classify new document based on the previous result of document classification. In this paper we have compared three supervised learning algorithms for document classification viz. Naïve Bayes, K-Nearest Neighbor and Centroid Based.

All the three classifiers require some labeled examples to train the classifiers in training phase, while in testing phase, unseen data is used to test and evaluate the classifiers. There is no need to write formal rules to train the classifier rather general subject knowledge is sufficient. Therefore it is easy to train the classifier and such inductive classifiers allow users to give category definition, which is important in some application.

B. *Text Representation and Dimensionality Reduction*

Since machine cannot understand the document in its raw form we have to represent it with some document representation model. For our case, we have used the popular vector space model. In our model we have used *tf*idf* weighting scheme for KNN and Centroid Based approach while tf is used for NB classifier. Here tf is term frequency and idf is inverse document frequency of a term t. Huge dimensionality is major issue for text classification. Hence we have to apply dimensionality reduction techniques. We have used *stopword removal* and *stemming* for dimensionality reduction. *Stopword* is a meaningless word like a, an, the;

whereas *stemming* is a process of removing suffixes and prefixes i.e. obtaining the root word (stem). As R-52 dataset used in our study is already stemmed and stopwords removed dataset.

C. Selection of Classifiers

1) Naïve Bayes: Naïve Bayesian is very fast and easy to implement so it is widely used in many practical systems. It is well known statistical method whose performance is relatively good for large datasets, so it is generally used in text classification problem [19],[20]. It is simple probabilistic classifier based on Bayes theorem. This classifier makes an independance assumption i.e. the values of the attributes are independent [21] given a class of instance this makes it suitable for large datasets.

Let C = {c_1, c_2, …,c_n} be set of predefined classes and d={w_1,w_2,…,w_m} be a document vector. We have to find conditional probability P(c_i|d) which is the probability of document d belong to category c_i. The document d will be assigned to category c_i which has maximum conditional probability p(c_i|d)

$$p(c_i|d) = \frac{p(c_i)(d|c_i)}{p(d)} \quad . \qquad (1)$$

The document is represented by words vector in vector space model. Therefore it is necessary to separately calculate the probability of all the words with remaining words and the resultant probability will get by multiplying them.

$$p(w_1, w_2, \dots, w_m) = p(w_1|w_2, w_3 \dots, w_m) * p(w_2|w_1, w_3 \dots, w_m) * p(w_m|w_1, w_2 \dots, w_{m-1})$$

The calculation can be simplified by strong assumption of NB algorithm that values of all the features are independent of each other in d when document d belongs to category ci then

$$p(c_i|d) = \frac{p(c_i) * \prod_{j=1}^{m} p(w_j|c_i)}{p(d)}$$

This assumption makes NB simple and fast algorithm and produce good result in most cases. In text classification we have to calculate p(w_j|c_i) of each word and p(c_i) of each category. p(d) is constant for all the given categories. Therefore above equation becomes

$$\max_{c_i \in C} p(c_i|d) = \max_{c_i \in C} p(c_i) \prod_{j=1}^{m} p(w_j|c_i)$$

and assign the document d to the category with maximum posterior probability.

2) K-Nearest Neighbor: It is a simple and widely used classifier [22],[23] for text classification. In training phase indexing is done and documents are represented in vector form. In testing phase, distance or similarity of each test document with each training document is calculated using distance measure like Euclidean Distance or similarity measure like Cosine Similarity. Then k-nearest neighbors (k=3 in our case) of test document are determined using these distances or similarities. Category of the majority of its nearest neighbors is assigned to the test document. In this paper, rather than using distances we have used Cosine similarity measure to find similarity which is calculated using following formula

$$\cos(d_1, d_2) = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_2}}{|\overrightarrow{d_1}| * |\overrightarrow{d_2}|}$$

Where d_1 and d_2 are documents' vectors.

3) Centroid Based Classifier (CBC): CBC is simple but effective document classification algorithm. In CBC, we calculate the centroid vector also called as prototype vector for each set of documents belonging to same class. If training data has k classes then total k centroid vectors $\{\overrightarrow{c_1}, \overrightarrow{c_2}, .., \overrightarrow{c_n}\}$ where $\overrightarrow{c_i}$ is the centroid of each class i are calculated using following two methods

a) Arithmetical Average Centroid (AAC): Most commonly used initialization method for centroid based classifier $\overrightarrow{c_i} = \frac{1}{|c_i|} \sum_{\vec{d} \in c_i} \vec{d}$ where centroid is the arithmetical average of all document vectors of class ci

b) Cumuli Geometric Centroid (CGC): $\overrightarrow{c_i} = \sum_{\vec{d} \in c_i} \vec{d}$ where each term will be given a summation weight.

Centroid of each category can be used to classify test document. To classify the test document d_t, we have to find similarity of document vector $\overrightarrow{d_t}$ with centroid vector $\overrightarrow{c_i}$ of each category $\{\overrightarrow{c_1}, \overrightarrow{c_2}, .. , \overrightarrow{c_k}\}$ using cosine similarity finally assign documents $\overrightarrow{d_t}$ to the class having most similarity value. That is d_t is assign to the class by using $\arg\max_{j=1..k} \left(cos\left(\overrightarrow{d_t}, \overrightarrow{c_i}\right)\right)$. The advantage of the CB classification algorithm is that it summarizes the characteristics of each class, in the form of concept vector. It's use has been demonstrated for text classification [24].

IV. Experimental Setup And Trials

We have designed an experiment to test the performance of NB, KNN and CB classifiers. The Experimental Set up is as given below

A. Experimental Environment:

For our experiment, we have implemented all three above mentioned classifiers in Java (jdk1.6.0). The experimental trials were performed on Pentium V with 2GB RAM and results were evaluated programmatically.

B. Data Set:

We have used Reuters-21578 (R-52) standard dataset publically available on web link [25], in which stopword removal and stemming is already performed. Total 52 categories are available for training and testing purpose with total 6532 training documents and total 2568 testing documents on 52 categories. Out of these, we have chosen 10 categories. Following table (TABLE I) summarizes categories and their distribution in the form of total number of documents.

TABLE I. DISTRIBUTION OF TRAINING AND TESTING DOCUMENTS

Category	Total documents	Training	Testing
Alum	**50**	**31**	**19**
Coffee	**112**	**90**	**22**
Cocoa	**61**	**46**	**15**
Copper	**44**	**31**	**13**
Cpi	**71**	**54**	**17**
Gnp	**73**	**58**	**15**
Gold	**90**	**70**	**20**
Grain	**51**	**41**	**10**
Jobs	**49**	**37**	**12**
Reserves	**49**	**37**	**12**
Total	**650**	**467**	**155**

C. The Performance Measure:

For quantifying result evaluation most commonly used performance measures like *Recall*, *Precision* and Micro and Macro *F1 measure* are used in our experiment. For text classification, *Precision* is the ratio of correct text documents to the total predicted text documents. *Recall* is the ratio of the correct text documents to the total text documents. It can be calculated as given below:

$$Precision = \frac{Tp}{Tp + Fp}$$

$$Recall = \frac{Tp}{Tp + Fn}$$

Where Tp: True Positive
Fp: False Positive
Fn: False Negative

F1 is calculated from precision and recall metrics. It is the harmonic mean of precision and recall and it is given by

F1= (2 * Precision * Recall) / (Precision + Recall)

F1-score can be computed on each individual category and then averaged over all categories; this is known as Macro averaging. This can be given in equation form as:

$$F1(macro - average) = \frac{\sum_1^A Fi}{A}$$

Where A is total number of categories. Each category has equal weight in Macro-averaged F1-measure.

F1-score calculated globally over all the test documents is called micro averaging. Each document has equal weight in Micro-averaged F1-measure. By taking the average of F1-measure values for each category i Micro-averaged F1-measure is obtained

$$F1_i = \frac{2*(Precision_i * Recall_i)}{Precision_i + Recall_i}$$

V. RESULTS AND DISCUSSION

Table II shows the performance of NB, CB and KNN classifiers for the above chosen 10 categories in terms of precision, recall and F1 measure metrics. Micro Average and Macro Average precision (P), Recall (R) and F1 measure (F1) are also depicted.

TABLE II. PERFORMANCE OF NB, CB AND KNN

Category	Naïve Bayes			Centroid Based			K-Nearest Neighbor		
	P (%)	*R (%)*	*F1 (%)*	*P (%)*	*R (%)*	*F1 (%)*	*P (%)*	*R (%)*	*F1 (%)*
Alum	100	74	85	100	89	94	100	63	77
Coffee	100	87	93	100	93	97	100	80	89
Cocoa	92	100	96	100	100	100	92	100	96
Copper	100	85	92	100	100	100	92	92	92
Cpi	100	82	90	100	94	97	100	88	94
Gnp	68	100	81	83	100	91	75	100	86
Gold	91	100	95	100	100	100	100	95	97
Grain	100	90	95	100	100	100	82	90	86
Jobs	100	92	96	100	92	96	91	83	87
Reserves	80	100	89	86	100	92	67	100	80
Micro Average	91	91	91	97	97	97	89	89	89
Macro Average	93	91	91	97	97	97	90	89	88

The same can be represented graphically in Figure 1, Figure 2 and Figure 3 below.

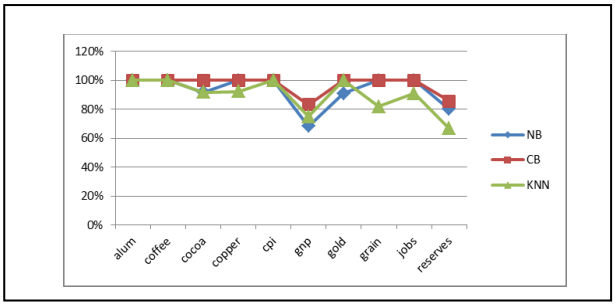

Figure 1. Performance of NB, CB and KNN in terms of Precision

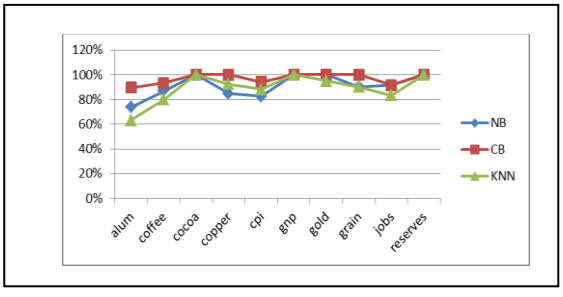

Figure 2. Performance of NB, CB and KNN in terms of Recall

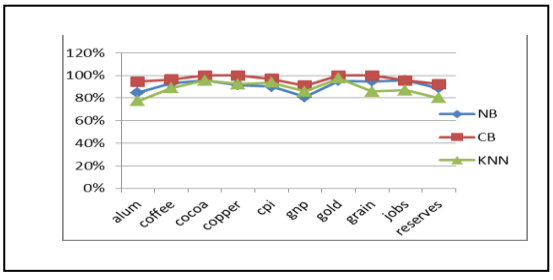

Figure 3. Performance of NB, CB and KNN in terms of F1Measure

In our experimental trials we have obtained Micro Average Precision of 91%, 97% and 89% for NB, CB and KNN respectively. While, Macro Average Precision of 93%, 97% and 90% was obtained for NB, CB and KNN respectively. The precision of each category for CB is higher than other two methods. This indicates that the CB method perform usually high precision. This is shown graphically in Figure 4.

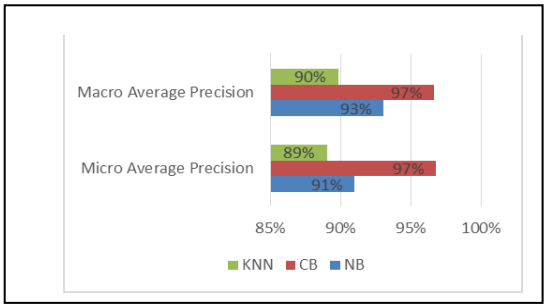

Figure 4. Performance of NB, CB and KNN in terms of Micro and Macro Average Precision

The Micro Average Recall of NB, CB and KNN are 91%, 97% and 89% respectively. The Macro Average Recall of NB, CB and KNN are 91%, 97% and 89% respectively. The recall of each category for CB is higher than other two methods. This indicates that the CB method perform normally high recall, which is shown in Figure 5.

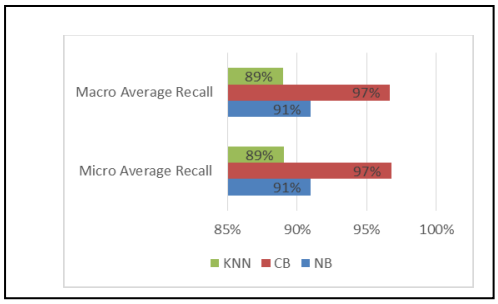

Figure 5. Performance of NB, CB and KNN in terms of Micro and Macro Average Recall

The Micro Average F1 of NB, CB and KNN are 91%, 97% and 89% respectively. The Macro Average F1 of NB, CB and KNN are 91%, 97% and 88% respectively. The F1 of each category for CB is higher than other two methods. This indicates that the CB method outperform NB and KNN as shown in Figure 6.

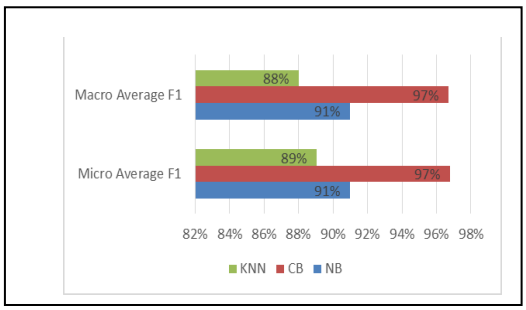

Figure 6. Performance of NB, CB and KNN in terms of Micro and Macro Average F1

A more accurate comparison of the different schemes can be obtained by looking at what extends the performance of a particular scheme is statistically different from that of another scheme. We have used Randomized Block Design method to compare the F1 measure obtained by different classifiers.

A. Randomized Block Design [26]:

In Randomized Block Design there is only one primary factor under consideration in the experiment. Similar test subjects are grouped into blocks. Each block is tested against all treatment levels of the primary factor at random order. This is intended to eliminate possible influence by other extraneous factor.

In our example we considered classifiers as treatments and classes as blocks. ANOVA table for Randomized Block Design by using Table III is as given below:

TABLE III. ANOVA TABLE USING RANDOMIZED BLOCK DESIGN

Source	DF	Sum of Square	Mean sum of square	F-Ratio	Table value
Treatment	**2**	**356.6**	**178.3**	**16.59**	**3.55455715**
Block	**9**	**530.7**	**58.97**	**5.49**	**2.45628115**
Error	**18**	**193.4**	**10.74**	**-**	
Total	**29**	**1080.7**	**-**	**-**	

(DF- Degree of Freedom)

H0: All classifiers do not differ significantly.

H1: At least one of the values differs from others.

α=0.05

Test statistics $F0 = \frac{MST}{MSE}$ = 16.59462254 based on DF_1=2, DF_2=18

Where MST – Mean sum of square of treatment

MSE- Mean sum of square of Error

p-value :p(F>F0) <0.05 using table (Since (16.59462254 > 3.554557)

Since p-value < α=0.05 reject H0.

We used T-test to determine where the differences are:

Critical difference at 0.05 levels is 3.07976

TABLE IV. STATISTICAL COMPARISON OF DIFFERENT CLASSIFICATION ALGORITHMS USING T-TEST. THE VALUES GREATER THAN CRITICAL DIFFERENCE (3.07976) SHOWS THAT CLASSIFIERS IN ROWS ARE STATISTICALLY BETTER THAN THE CLASSIFIERS IN COLUMNS.

	NB	KNN
CB	**5.5**	**8.3**
NB		**2.8**

From this result we can say that centroid based classification algorithm outperforms all remaining algorithms, with NB being second and KNN being the last. Treatment means difference between NB and KNN i.e. 2.8 in this case, which is less than critical difference 3.07976. Therefore they do not differ significantly.

VI. CONCLUSION:

In this paper, we have reported our study on experimental evaluation of three well known classifiers NB, KNN and CB, with statistical significance test for English language text categorization on R-52 of Reuters-21578 already stemmed and stopword removed corpus. We have compared the performance of all the three classifiers. No feature selection was applied in the reported experimentation. The experimental result shows that all classification results on three classifiers are acceptable for sample English Language text. The performance of the Centroid Based classifier is best amongst all classifiers with

97% Micro Average F1 measure. Out of the three classifiers KNN has obtained lowest performance of 89% Micro Average F1 measure. Statistical comparison of different classification algorithms shows that Centroid Based algorithm significantly outperforms NB and KNN. Though the performance of CB is best, we have observed that the classification speed of NB is very fast among all three classifiers. We have also observed that KNN being lazy learning classifier, it is slowest among all.

ACKNOWLEDGMENT

We hereby acknowledge the financial and administrative support extended under SAP (DRS-I) scheme, UGC New Delhi at School of Computer Sciences, NMU, Jalgaon.

REFERENCES

[1] Vishal Gupta and Gurpreet S. Lehal,"A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, pp. 60-76, August 2009, doi:10.4304/jetwi.1.1.60-76

[2] IzzatAlsmadi, IkdamAlhami, "Clustering and classification of email contents", Journal of King Saud University - Computer and Information Sciences,Volume 27, Issue 1, pp. 46-57, January 2015, doi: 10.1016/j.jksuci.2014.03.014.

[3] Taeho C. Jo, Jerry H. Seo, Hyeon Kim, "Topic Spotting on News Articles with Topic Repository by Controlled Indexing", Chapter, Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents, Volume 1983 of the series Lecture Notes in Computer Science pp 386-391, 27 May 2002.

[4] HidayetTakç , TungaGüngör, "A high performance Centroid-Based classification approach for language identification", Pattern Recognition Letters, Volume 33, Issue 16, 1 December 2012, pp. 2077-2084, doi:10.1016/j.patrec.2012.06.01.

[5] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[6] Mohammed.abdul.wajeed, t. Adilakshmi, "Text Classification using Machine Learning", Journal of Theoretical and Applied Information Technology, 2005 - 2009 JATIT, ISSN: 2229-73673(1-2), pp. 233-237, 2012.

[7] Rish I., "An Empirical Study of the Naïve Bayes Classifier", www.cc.gatech.edu/~isbell/reading /papers/Rish.pdf.

[8] Taeho Jo*, (2008), "Inverted Index based modified version of KNN for text categorization", Journal of Information Processing Systems. Vol. 4. No. 1, 2008.

[9] Gharib T. F., Habib M. B., Fayed Z. T., (2009), "Arabic Text Classification Using Support Vector Machines", wwwhome.cs.utwente.nl/~badiehm/PDF/ISCA2009.pdf

[10] Su J. and Zhang H.,(2006), "A Fast Decision Tree Learning Algorithm", American Association for Artificial Intelligence, AAAI'06 Proceedings of the 21st national conference on Artificial intelligence, Vol. 1, 2006, pp. 500-505, ISBN: 978-1-57735-281-5

[11] Fouzi Harrage, Abdul Malik Salman, Al-Salman, Mohammed BeMohammed, "A Comparative Study of Neural Networks Architectures on Arabic Text Categorization using Feature Extraction", Publisher:IEEE, 2010 International Conference on Machine and Web Intelligence (ICMWI), 3-5 Oct. 2010, pg 102- 107, ISBN:978-1-4244-8608-3, DOI10.1109/ICMW.2010.5648051

[12] Ms. K. Umamaheswari, Dr. S. Sumathi, Ms. V. Aparna, Ms. A. Arthi, " Text Classification Using Enhanced Naïve Bayes With Genetic Algorithm", International Journal Of Computer Applications In Xiu-Li Pang, Engineering, Technology And Sciences (IJ-CA-ETS) Issn: 0974-3596, Volume 1, Issue 2, April '09 – September '09, Page: 263.

[13] Abhishek Sanwaliya, Kripa Shanker and Subhas C. Misra, "Categorization of News Articles: A Model based on Discriminative Term Extraction method" IEEE, Second International Conference on Advances in Databases, Knowledge, and Data Applications, 978-0-7695-3981-2/10, IEEE Computer Society, 2010, DOI 10.1109/DBKDA.2010.18

[14] Taeho Jo, April (2010), "NTC (Neural Text Categorizer): Neural Network for Text Categorization", International Journal of Information Studies Vol. 2 Issue 2.

[15] Susan Dumais, John Platt, David Heckerman,MehranSahami "Inductive Learning Algorithms and Representations for Text Categorization*",robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf*

[16] Vidhya.K.A, G.Aghila, "Hybrid Text Mining Model for Document Classification", Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on (Volume:1), 26-28 Feb. 2010, pp.210-214, DOI:10.1109/ICCAE.2010.5451965

[17] BaiRujiang and Liao Junhua, "A Hybrid Document Classification based on SVM and Rough Sets", International e-Conference on Advanced Science and Technology , pp. 18-23, 7-9 Mar. 2009, IEEE Computer Society, 978-0-7695-3672-9/09 IEEE DOI 10.1109/AST.2009.14.

[18] Wen Han and Xiao Nan-feng, (2011) "An Enhanced EM Method of Semi-supervised Classification Based on Naive Bayesian", Eighth IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 2, P. 987 – 991, 26-28 July 2011, DOI:978-1-61284-180-9

[19] Kavi Narayana Murthy, "Automatic Categorization of Telugu News Articles", doi: 202.41.85.68.

[20] Kourdi Mohmed EL, Benasid Amine, Rachidi Tajje-eddine, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", Semitic '04 Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages pp. 51-58, Association for Computational Linguistics Stroudsburg, PA, USA ©2004.

[21] Huan Liu and Lei Yu (2005),"Towards Integrating Feature Selection Algorithm for Classification and Clustering" IEEE Transactions on Knowledge Engineering Vol. 17 No. 4, April 2005.

[22] DR. Riyad Al-Shalabi, Dr. GhassanKanaanManaf H. Gharaibeh, (2006), "Arabic Text Categorization Using KNN Algorithm".*www.uop.edu.jo/download/research/members/CSIT2006/.../pg20.pdf*

[23] K Raghuveer and KaviNarayana Murthy, "Text Categorization in Indian Languages using Machine Learning Approaches", IICAI 2007: 1864-1883

[24] Eui-Hong(Sam) Han and George Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results", Principles of Data Mining and Knowledge Discovery, pp. 424-431, 2000.

[25] www.cs.umb.edu/`smimarog/textmining/datasets/SomeTextDatasets.html

[26] http://www.r-tutor.com/elementary-statistics/analysis-variance/randomized-block-design