Performance Analysis of Supervised Machine Learning Approaches for Bengali Text Categorization

Ronald Tudu*, Shaibal Saha, Prasun Nandy Pritam, Rajesh Palit
Department of Electrical and Computer Engineering, North South University, Dhaka
Email: *ronald.tudu@northsouth.edu

***Abstract*—In this digital era, enormous amount of data are being generated everyday, and most of them are unstructured textual data. An automated text classifier helps to categorize the texts automatically into pre-defined categories. With the help of machine learning we can learn about the features of pre-categorized documents and predict document's category. Bengali language is one of the most spoken languages in the world. It has become essential to implement automated text categorization for Bengali language. Text categorization mostly uses data mining algorithms along with NLP tools, feature extraction and selection methods with vector space modeling. In this paper, we have measured the performance of Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Stochastic Gradient Descent (SGD) and Logistic Regression (LR) methods using an open source Bengali newspaper article corpus containing $84,906$ articles of 10 categories. The impact of the size of the training dataset on the accuracy of the classification was examined for different algorithms. We have documented the execution time to train the methods and discussed issues and challenges in Bengali text categorization. This paper can be used as a reference work for future researchers in Bengali text categorization.**

***Index Terms*—Text categorization, Machine learning, Bengali, Performance analysis**

I. Introduction

Automated text categorization is one of the emerging subject in text mining. The need for text categorization is increasing due to the massive textual digital data, which is growing in an exponential rate. It is estimated that the amount of unstructured data will be 40 zeta-byte by 2020. Machine learning approaches are used to categorize text with other several data mining algorithms. According to Islam *et al.* [9], the accuracy rate of classifiers increases as the size of training data increases. Text categorization is used in many applications such as content tagging, spam filtering and business intelligence.

Bengali is spoken in Bangladesh along with two Indian states West Bengal and Tripura. The number of Bengali speakers is approximately 200 million. There are about 80.83 million Internet users at the end of January 2018 [1], and 30 million of them are labeled as social media users (2018) [2] in Bangladesh. The need for text categorization for the Bengali language is appealing for researchers and scientists to analyze mass opinion, perspective, and detecting a subject of interest in a conversation. The amount of information available on the web is tremendous and increasing at an exponential rate. Lots of work have been done in sentiment analysis in different language especially in English and Indonesian to analyze cyber-bullying in text [19]. There are also many works done in Indian regional languages [16].

As a necessity of text mining is increasing, the researchers of Bangladeshi or some Indian researchers focused on developing applications using text mining. Thus a number of research work have been done in Bangla text mining too. Name entity recognition [3] is one of them where they categorized the name tag from articles. Another work used a supervised learning method for Bangla web document categorization. The main problem is all the researches in Bangla text mining are based on theory, but there is less application in this sector. There are numerous works on sentiment analysis in Bengali, some of the research work shows good results with good accuracy. Most of the research works use Romanized Bangla text using deep model recurrent model. There are some works where TF-IDF was mixed with SVM to see the performance of the model [9].

There are lots of data on the web in Bengali, it becomes very difficult to find data of interest in the web. When you want to post a question in a forum or a social media site it becomes necessary to categorize the text for the viewers to get maximum output. On the other hand, if you want you to post to be observed in a forum it becomes vital and categorizes the text and hash-tag the keywords. The problem is Bengali text categorization is not much done before in an application platform, but it is essential for Bengali people to categorize the status for people wellness. It is hard to summarize the most suitable summarize topic name for any types of writing. The main technical challenges and issues are working with unstructured data like texts. Pre-processing is the most difficult task in this project since there are very rare NLP tools to pre-process Bengali texts, like stemmer and selecting the best stop words. Working with a very big data set is also a challenge because we have to be careful about over-fitting.

In this paper, we have presented the performance analysis of four supervised text categorization techniques on Bengali newspaper articles. The accuracy rates of the classifiers on different size of datasets were examined and the variation of accuracy rates of the classifiers were observed with respect to size of the datasets. Confusion matrices for each classifier was built to see the miss-classification among the categories. We have included supervised learning classifiers, namely Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Stochastic Gradient Descent (SGD) and Logistic Regression (LR) methods in the experiments. According to Jindal *et al.,* [4], the most popular methods are Support Vector Machine (SVM) [5] followed by K-Nearest Neighbor (KNN) and Naive

978-1-7281-1390-6/19/$31.00 ©2019 IEEE
DOI 10.1109/APWConCSE.2018.00043

Bayes (NB). For vectorization of the text, TF-IDF and N-gram approaches were used in this analysis.

We have considered a dataset consisting of 84,906 Bengali newspaper articles of 10 categories. The details of the dataset is given in Section IV. The preliminary work was preprocessing and vectorizing the data. Secondly, we trained the classifiers with the datasets and evaluated the accuracy. The final step was to predict category which is the goal of text categorization.

The performance of different classifiers such as SVN, KNN, SGD, NB, and DT was investigated earlier, however, the authors in those works used smaller datasets. The authors in [9] concluded that by raising the size of training datasets, the accuracy of the classifier can be reached 100%. Although this trend is seen in machine learning approaches, in our experiment of Bengali text categorization we observed that using large training dataset overfits the classifier, and the accuracy reduces. The contribution of the authors in this paper can be documented as follows.

- We have compared performances of popular classifiers for Bengali text categorization;
- In the experiment a large sample dataset containing more than 84 thousand articles was used;
- It was observed that the classifiers overfit with the training data due to the large number features in big datasets. An effective stemmer is essential for Bengali language to reduce the number of redundant features;
- Our analysis concludes that using the knowledge of the confusion matrix a textual document can be tagged with multiple categories.

The paper is organized as the following order. Section II contains a discussion on the previous research works in this field. The methodology of conducting the experiments is given in Section III. The results are presented and analyzed in Section V. The discussions about the confusion matrix of all four classifier are also given in this section. Section VI discuss about future work and Section VII concludes this research work.

II. Related Work

While reviewing the literature of text categorization, we observe that many research works have been done for English language and other languages like Punjabi [6]. A few works have been done for Bengali language. Some prominent supervised techniques such as SVM, K-Nearest Neighbor (KNN) and Decision Tree (DT) are widely techniques and we discuss some of the works in this section.

The authors in [7] used a number of supervised learning methods for Bangla Web Document categorization. The paper automatically described the sort of category from a predefined set. They analyzed five categories with a dataset of one thousand records. The authors explored four classifiers named Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT). In this paper, the categories are Business, Sports, health, Technology, Education. About 90% of the total dataset are used as a training data. The total number of token was 22,218. The authors aimed to find the best classifier for Bengali web document among four classifiers and concluded that SVM gives highest accuracy.

The authors in [8] narrated Bengali document text categorization with stochastic Gradient Descent (SGD) technique. They designed seven experiments to compare their proposed method with others. The paper divided their methods into three categories. They conducted their experiments with 9,127 records of 9 categories. Approximately 60% datasets were used in the training datasets and the rest were used in test datasets. The paper also indicates the confusion matrix of different classifiers which authors used to compare their methods. The authors find that the value of F1 score (*i.e.* measure of test accuracy) is higher in SGD than the other approaches.

The Term Frequency-Inverse Document Frequency is used for calculating the frequency of a word in a document. Islam *et al.* in [9] used TFIDF weighted length normalization for feature selection after the pre-processing of the dataset is finished. The authors used 31,906 records of 12 categories. They examined the accuracy of the SVM approach with different number of datasets where the highest number was 31,906 and the lowest was 3,191 in the training dataset. In this paper the authors also try to calculate recall, precision and F1-measure of all 12 categories. In their paper, they also claimed that with the large number of dataset, they can touch the accuracy of 100%.

Chy, Seddiqui and Das [10] used a Bengali news classifier with Naive Bayes approach. They developed their own crawler to extract the news articles from web pages. Naive Bayes classifier is used in text classification because of its simplicity. In this paper the authors build their own steps to classify the news documents. The authors also used full text RSS, then pre-process their data with tokenization. The authors use TREC [11] evaluation technique to produce recall-precision graph.

The authors in [12] stated that they use 4,000 data in 8 categories where 500 data are available in each category. After applying several pre-processing methods, the number of tokens have been taken 9,57,623. The paper described the application of TF-IDF-ICF feature with dimension reduction technique. They also indicated the change in accuracy rates after applying reduction techniques.

Jia and Mu [13] described a classification system with large corpus in Chinese language. The authors also used 6 categories and apply SVM techniques in Chinese language. Among those data, the authors use 50% as training data. They find a high score of F1 measurement in Chinese language. There are also some works done on N-gram based Bengali news text categorization [14].

After reviewing many paper on Bengali text classification, we found that there are no such work in large dataset as we used in this work. There are many technical challenges involved in large datasets. We observe those technical challenges and document in this paper.

III. Methodology

The process of Bengali text categorization has been described in this section. The flowchart of the methodology is given in Fig. 1. The whole process can be grossly divided into four steps as follows [4].

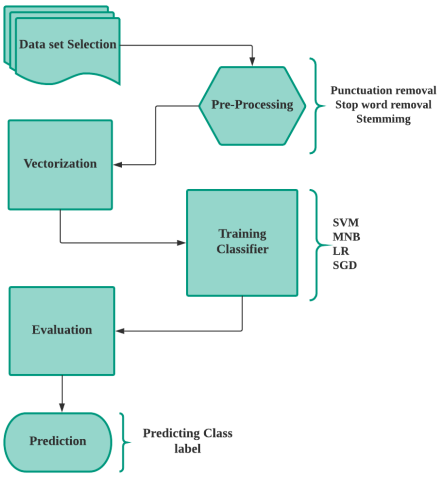

Fig. 1. Flow chart of methodology followed.

- Dataset selection
- Pre-processing
- Vectorization
- Classifier selection
- Evaluation

A. Dataset Selection

Dataset selection plays a very vital role in supervised machine learning approaches for categorization or classification. The quality and the quantity of data shapes the goal to predict a text to the right class. Newspaper articles are rich in information and easy to create a dataset for learning purpose. We have selected an open source Bengali corpus. This corpus is created from Bengali newspapers articles, and consists 12 categories, and among them we have used 10 in our work. The categories are Accident (AC), Crime (CR), Economics (EC), Education (ED), Entertainment (EN), Environment (EV), International (IN), Politics (PO), Science & Technology (ST), and Sports (SP).

The two other categories, Art and Opinion were not included as they are more prone to miss classification for the mixture of different content in these two categories. This dataset can be collected online at https://scdnlab.com/corpus/. The dataset is licensed under MIT and free to be used by anyone. This dataset is the result of a thesis conducted under the Department of Computer Science and Engineering in Shahjalal University of science and technology, Bangladesh.

B. Pre-processing

In this step, our goal is to remove noises from the data. As the texts are very unstructured way to represent information and contain noise, extraction of features from texts is very much challenging. There are some basic ways to pre-process text to preserve only relevant pieces of information. We divided the pre-processing task into three steps as follows.

- Punctuation removal
- Stop-words removal
- Stemming

1) Punctuation removal: All the punctuation are eliminated from the documents they play a very little or no role in categorization. They can be considered as noise. We have also removed Bengali numerals and special characters for the same reason. Some of the examples of the punctuation, numerals, and special characters that we removed are given below in Fig. 2.

০	১	২	৩	৪	৫	৬
৭	৮	৯	—	\	/	?
,	=	।	#	&	@	*
:	;	<	>	!	%	…

Fig. 2. List of removal symbols.

2) Stopword Removal: In computing, stop words are words which are filtered out before or after processing of natural language data (text). Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search. In our case, we try to filter out all the less contributing words in the Bengali language that helps the classifier to predict classes with less noise. We created a list of 501 stopwords for the Bengali language. The list contains most common Bengali words which include determiners, prepositions and coordinating conjunctions. The list is given below in Fig. 3.

অনেক	অনেকে	অনেকেই	অনেকেও	অন্তত	অথবা
অথচ	অর্থাৎ	অন্য	আজ	আজও	আছে
আপনার	আপনি	আবার	আবারো	আমরা	আমাকে
আমাদের	আমার	আমারো	আমি	আরও	আর
আগে	আগেই	আয়	অতএব	…	…

Fig. 3. List of stop words.

3) Stemming: Stemming is the process of reducing a word to its root form. Stemming is used for reducing different derivational or inflectional variants of the same word to increase effectiveness and efficiency of information retrieval [17]. It is an essential part in natural language understanding (NLU) and natural language processing (NLP). An example of stemming is given in Fig. 4 in the context of

text categorization. For this research, we used a light-weight stemmer.

"আমাদের দেশটি" ⇨ "আমি দেশ"

Fig. 4. Example of Bengali Stemming.

C. Vectorization

Vectorization is the process to represent documents in vector space. The process is to create mapping from term to term. It is called term because sometimes it can be arbitrary n-grams. In vectorization, each row represents a document and each column represents a term. Vectorization counts the number of term occurred in a document and represent it as a matrix. Vectorizing can be done in two ways: using vocabulary or feature hashing.

D. Classifier selection

Support Vector Machine (SVM) is a supervised learning method, which can be used in classification problems. In Support Vector Machine classification, the dataset is plotted in n-dimension where each feature represents a value of a coordinate. Then a logical hyper-plane is drawn that divides classes. Since textual data are linearly separable this algorithm works well in text categorization [18]. Among the popular SVM kernels like polynomial and RBF, we have used linear kernel as most of the text categorization problems are linearly separable [18]. Linear kernel has less parameters to optimize and works well with a lot of features. The performance of the classifiers based on other kernels do not increase with higher dimension space.

Naive Bayes is a popular probabilistic approach among classification problems. It basically works on independent conditional probability rather than the particular distribution of each feature. Multinomial Naive Bayes works on multinomial distribution. It works well for countable data such as word counts in texts.

In Stochastic Gradient Descent (SGD), a batch is the total number of examples uses to calculate the gradient in a single iteration. Stochastic indicates comprising each batch is chosen at random. Stochastic gradient descent uses only one single example iteration. Although given enough iterations SGD works but it is exceedingly noisy.

Logistic Regression (LR) is used when the dependent variable is dichotomous. Like other regression analysis it is used as a predictive analysis. It is used to describe data and to clarify the relationship between one dependent binary variable and one or more ordinal, nominal or interval independent variables.

E. Evaluation

The following metrics were used in analyzing the performance of the classifiers in our experiments.

- Recall: Recall refers to as sensitivity. Recalls are the fraction of relevant documents that are successfully retrieved. Recall is calculated by the number of true positive divided by the number of true positive and false negative.
- Precision: Precision related to positive predictive value. Precision is calculated by the number of true positive divided by the number of true positive and false positive. Precision is used to measure true positive accuracy of a document.
- F1 score: F1 is measured to find the accuracy rate of a classifier on a test dataset. F1 is the weighted average of precision and recall.
- Confusion Matrix: Confusion matrix is most unambiguous way to represent a prediction result of a classifier. Confusion matrix represents the number of miss-classified data between two categories.

IV. Experimental setup

As mentioned in Section III, a Bengali corpus [15] was built under a thesis work in the Department of Computer Science and Engineering, Shahjalal University of Science and Technology. This corpus is open for public use and was used in the experiments. There are 12 categories, however, 2 categories, Art and Opinion were excluded in this experiment.

1) Dataset Representation: Table. I shows the total number of samples datasets used in the experiments. From all samples, 90% data are used as training data and the rest are used as test data. The total sample size is 84,906, and four datasets were made consisting of 10,027, 42,370, 60,000 and 84,906 data. In every dataset, the percentage of training and test data remained the same.

Category	Samples	Training Set	Test Set
Crime (CR)	8565	7709	856
Economics (EC)	3445	3101	344
International (IN)	5151	4636	515
Sports (SP)	11888	10700	1188
Accident (AC)	6324	5692	632
Environment (EV)	4308	3878	430
Science and Technology (ST)	2901	2611	290
Entertainment (EN)	10093	9084	1009
Politics (PO)	20038	18035	2003
Education (ED)	12193	10974	1219

TABLE I
Training & Testing Dataset

2) Specification of the computer: A Windows 10 64 bit based desktop PC was used in the experimentation with Intel Core i7 3.6 Ghz 64 bit processor, 16 GB DDR4 RAM, 512 GB SSD storage, NVIDIA GeForce graphics card.

V. Results

In this section, the accuracy rates of SVM, MNB, SGD and LR classifiers are compared for different training and test datasets. Then the impact of the size of the training datasets on the classifiers is discussed. It is followed a comparison by the run-time or required time for training the classifiers. The confusion matrices for each classifier are then discussed to show miss-classifications among categories.

The accuracy rates of all four supervised techniques are given in Fig. 5. The highest rate of accuracy is achieved for the

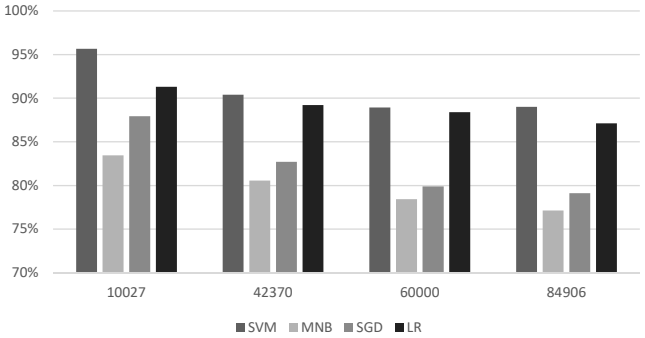

Fig. 5. Accuracy rates of SVM, MNB, SGD, LR classifiers

smallest dataset with 10,000 articles which is 93.3%. Support Vector Machine (SVM) algorithm shows higher accuracy rates in all four different sizes of datasets. In contrary, Multinomial Naive Bayes (MNB) shows the less accuracy rates with 75.16%. The accuracy rates vary 2% to 3% after applying dimension reduction. For example, SVM gives 87.5% accuracy rate without applying dimension reduction, but it gives 89.2% accuracy rate when dimension reduction is applied.

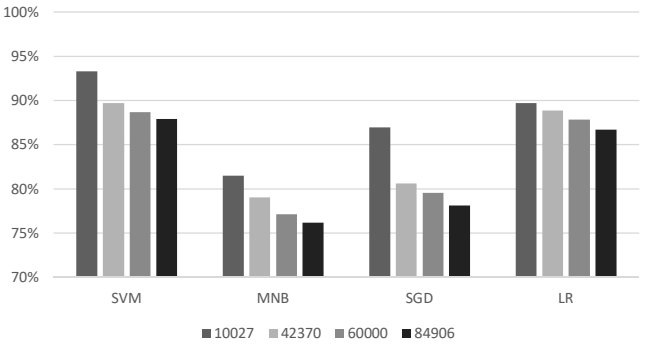

Fig. 6. Accuracy rates of classifiers with varying training data size.

Fig. 6 shows the decreasing rate of accuracy with respect to the increasing size of the training data for all four classifiers. The main reason for this decreasing accuracy rate pattern is over-fitting of data. Due to higher number of features, all four classifiers show less accuracy with big training data than the small training dataset. For example, SVM shows 93.3% with 10,027 training dataset, but gives 89.2% with 84,906 training dataset.

The execution times of all four classifiers for different datasets is shown in Fig. 7. Multinomial Naive Bayes (MNB) always shows the least run-time due to its simple form of classification. Multinomial Naive Bayes gives only 0.23 minute with full datasets where SGD and LR spends above 4 min. These processes were executed in a high-end desktop PC and the configuration is given in Section IV for reference.

The confusion matrix for Support Vector Machine (SVM) classifier is given in Table II. Compare to other confusion

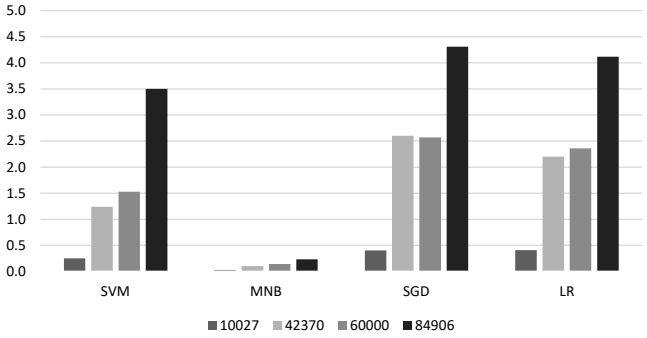

Fig. 7. Run-time Comparison of SVM, MNB, SGD, LR

	CR	EC	IN	SP	AC	EV	ST	EN	PO	ED
CR	748	1	7	9	16	14	4	15	58	13
EC	3	332	4	3	0	1	3	1	20	1
IN	3	4	518	5	4	1	3	6	5	7
SP	8	0	7	1091	4	12	1	17	11	20
AC	5	0	1	9	573	9	3	8	10	5
EV	8	2	3	6	11	296	0	11	42	29
ST	2	1	11	3	0	1	236	10	1	4
EN	13	1	5	18	4	7	11	881	24	26
PO	32	3	7	27	11	44	4	36	1773	103
ED	20	3	4	15	10	37	3	20	83	986

TABLE II
CONFUSION MATRIX FOR SVM

matrices, this matrix clearly shows less number of conflicts between categories. The highest conflict is between Politics and Education categories because there are overlapping features between them. Overall, the conflict in SVM classifier is negligible, and it is evident in the accuracy rate of SVM.

	CR	EC	IN	SP	AC	EV	ST	EN	PO	ED
CR	551	5	5	18	46	9	11	51	36	26
EC	1	239	2	2	1	3	3	2	16	8
IN	15	8	423	9	17	8	5	10	10	23
SP	5	2	29	1019	5	16	3	19	14	40
AC	96	0	4	23	486	14	5	10	24	19
EV	5	3	2	3	17	157	0	5	14	19
ST	3	10	12	3	2	0	198	10	1	17
EN	21	9	22	27	7	14	13	771	16	36
PO	113	51	54	58	35	129	15	83	1795	263
ED	32	20	14	24	17	72	15	44	101	743

TABLE III
CONFUSION MATRIX FOR MNB

Table III indicates the confusion matrix of Multinomial Naive Bayes (MNB) classifier. MNB classifier has many major conflicts between two categories. Politics has overlapped with Education category 263 articles. There are many categories which overlap with each other in MNB classifier. MNB also has higher conflict rate between Accident and Crime or Politics and Crime categories.

Table IV refers to the confusion matrix of Stochastic Gradient Descent (SGD) classifier. In some particular categories, SGD confusion matrix shows no conflict between two categories. Similar to Table II & Table III, SGD also shows conflict between Politics and Education, but the rate is higher than the other two. Table V indicates the confusion matrix of

	CR	EC	IN	SP	AC	EV	ST	EN	PO	ED
CR	556	0	40	12	20	7	10	40	10	14
EC	3	229	4	2	1	3	4	0	6	1
IN	6	1	231	1	2	3	5	2	0	3
SP	15	9	54	1094	19	33	18	38	22	64
AC	44	3	10	10	525	16	5	3	11	9
EV	1	2	1	0	6	117	1	0	3	1
ST	0	1	5	0	0	0	149	0	0	1
EN	13	16	21	21	4	11	30	803	6	21
PO	178	77	182	43	50	183	4	36	1948	265
ED	26	9	19	3	6	49	3	20	21	785

TABLE IV
CONFUSION MATRIX FOR SGD

	CR	EC	IN	SP	AC	EV	ST	EN	PO	ED
CR	739	0	9	7	18	14	9	25	34	16
EC	2	315	3	3	0	4	3	0	19	1
IN	5	4	495	1	7	4	6	5	6	6
SP	6	0	16	1102	4	13	2	12	13	30
AC	8	0	1	9	562	10	5	5	10	8
EV	2	3	3	2	11	273	0	2	13	17
ST	0	4	8	2	0	2	208	5	0	4
EN	16	1	12	22	4	7	20	878	14	23
PO	45	14	13	32	18	53	11	47	1861	106
ED	19	6	7	6	9	42	4	26	57	983

TABLE V
CONFUSION MATRIX FOR LR

Logistic Regression (LR) based classifier. Logistic regression is basically used for statistical analysis. So the conflicts among the categories in LR is low which allows better accuracy rates than the MNB & SGD classifiers.

VI. FUTURE WORK

During the experiment, we observed that the accuracy of the categorization tools was decreasing with the increase of training data size. This is one of the most interesting observation we found while categorizing a lot of Bengali textual data. It was happening due to over-fitting of the model and that can be mitigated by reducing number of features in the training dataset. Another reason for over-fitting was the overlapping of features which creates ambiguity and a reason for miss classification. In Bengali, a word can be present in a lot of different form and those words need to be reduced to their root words to avoid redundant features. This task should be done during pre-processing phase and termed as stemming. There is a lack of effective stemmer for Bengali language and it can be considered as the future work of this paper. A text can also fall into multiple categories. After analyzing the confusion matrix, meaningful insights can be extracted and documents can be double categorized.

VII. CONCLUSIONS

Text categorization is getting importance as there has been tremendous advancement in the field of machine learning. Thus there is a strong necessity to examine the existing strategies for Bengali language. This research is mainly categorization of the Bengali newspaper articles to find the topic of large scale document. In the context of social networks, categorization helps people to profile people according to their interest. Bengali text classification application is new research area for Bengali language. So, it will be very helpful for Bengali literature or web to analyze data and get information from web. In this paper, We analyzed four algorithms, and document their performances. Only supervised learning method were considered in this paper. Among all the 4 algorithms, Support Vector Machine gives the highest prediction accuracy. On average 87.5% accuracy was achieved from SVM. Thus SVM was the best algorithm among all with the higher accuracy rate and with average training time. There were some limitations of resources the authors face while conducting this research work. In the pre-processing phase, a good stemmer was very much necessary for reducing the number of features and to avoid over-fitting of the training dataset. The findings in this research work will provide information and assist future researchers on this area.

REFERENCES

[1] "Bangladesh Telecommunication Regulatory Commission," [Online]. Available: http://www.btrc.gov.bd/ content/internet- subscribers-bangladesh- january-2018. [Accessed 25 10 2018].

[2] "The Financial Express," [Online]. Available: https://thefinancialexpress.com.bd/sci-tech/ social-media-users-30-million-in-bangladesh-report-1521797895/. [Accessed 25 10 2018].

[3] A. Senapati, A. Das and U. Garain, "Named -Entity Recognition in Bengali," in Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation, 2013.

[4] R. Jindal, R. Malhotra and A. Jain, "Techniques for text classification: Literature review and current trends," 2015.

[5] C. Cortes and V. Vapnik, "Support Vector Networks."Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[6] V. Gupta and G. Lehal, "Automatic Punjabi Text Extractive Summarization System," in Proceedings of COLING 2012, 2012.

[7] A. K. Mandal and R. Sen, "Supervised Learning Methods for Bangla Web Document categorization," International Journal of Artificial Intelligence & Applications, vol. 5(5), pp. 93-105, 2014.

[8] F. Kabir, S. Siddique, M. R. A. Sabbir and M. N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier," 2015 International Conference on Cognitive Computing and Information Processing(CCIP), 2015.

[9] M. S. Islam, F. E. M. Jubayer and S. I. Ahmed, "A support vector machine mixed with TFIDF algorithm to categorize Bengali document," 2017 International Conference on Electrical, 2017.

[10] A. N. Chy, M. H. Seddiqui and S. Das, "Bangla news classification using naive Bayes classifier," 2014.

[11] E. M. Voorhees and D. K. Harman, "TREC: Experiment and Evaluation in Information Retrieval," in MIT Press Cambridge, 2005.

[12] A. Dhar, N. S. Dash and K. Roy, "Categorization of Bangla Web Text Documents Based on TF-IDF-ICF Text Analysis Scheme," 52nd Annual Convention of the Computer Society of India, 2018.

[13] Z. Jia and J. Mu, "Web Text Categorization for Large-scale Corpus," 2010 International Confrence on Computer Application and System Modeling (ICCASM 2010), 2010.

[14] M. Mansur, "Analysis of N-Gram based text categorization for Bangla in a newspaper corpus," BRAC University, 2006.

[15] Md. Saiful Islam, Md. Abu Shahriar Ratul, Md. Yusuf Khan, "An Open Source Bengali Corpus", Shahjalal University of Science and Technology.

[16] M. Hanumanthappa and . M. Swamy, "A Detailed Study on Indian Languages Text Mining," International Journal of Computer Science and Mobile Computing, vol. 3, no. 11, pp. 54-60, November 2014.

[17] Debasis Ganguly, Johannes Leveling, Gareth Jones, "Bengali (Bangla) information retrieval", Technical Challenges and Design Issues in Bangla Language Processing. pp. 273-301, 2013.

[18] Thorsten Joachims, "Text categorization with Support Vector Machines: learning with many relevant features", In Proceedings of the 10th European Conference on Machine Learning (ECML'98), 1998.

[19] Hariani and Imama Riadi, "Detection of Cyberbullying On Social Media Using Data Mining Techniques," International Journal of Computer Science and Information Security, pp. 244-250, 2017.