Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 2, April 2021, pp. 742~749
ISSN: 2302-9285, DOI: 10.11591/eei.v10i2.2873

An in-depth exploration of Bangla blog post classification

Tanvirul Islam, Ashik Iqbal Prince, Md. Mehedee Zaman Khan, Md. Ismail Jabiullah, Md. Tarek Habib
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Oct 11, 2020
Revised Jan 28, 2021
Accepted Feb 15, 2021

Keywords:

Bangla blog
Bangla text classification
Bigram
Supervised machine learning
TF-IDF
Trigram
Unigram

ABSTRACT

Bangla blog is increasing rapidly in the era of information, and consequently, the blog has a diverse layout and categorization. In such an aptitude, automated blog post classification is a comparatively more efficient solution in order to organize Bangla blog posts in a standard way so that users can easily find their required articles of interest. In this research, nine supervised learning models which are support vector machine (SVM), multinomial naïve Bayes (MNB), multi-layer perceptron (MLP), k-nearest neighbours (k-NN), stochastic gradient descent (SGD), decision tree, perceptron, ridge classifier and random forest are utilized and compared for classification of Bangla blog post. Moreover, the performance on predicting blog posts against eight categories, three feature extraction techniques are applied, namely unigram TF-IDF (term frequency-inverse document frequency), bigram TF-IDF, and trigram TF-IDF. The majority of the classifiers show above 80% accuracy. Other performance evaluation metrics also show good results while comparing the selected classifiers.

Corresponding Author:

Tanvirul Islam
Department of Computer Science and Engineering
Daffodil International University
Mirpur Road, Dhaka, Bangladesh
Email: tanvirul15-6117@diu.edu.bd

1. INTRODUCTION

The modern world is rapidly moving towards a web-based platform for general co-operation in content making and utilization. Many people are interested in online resources for their daily choice of reading. Nowadays, the online blog is one of the most popular digital domains among people. Bangla text articles on digital platforms have increased for a few years. Bangla is the most widely spoken language in Bangladesh and the second most widely spoken in India [1]. Moreover, there was around 228 million native Bangla speakers worldwide, 7th most communicated language on the world [1]. In such an aptitude, it needs to organize Bangla texts in a standard way so that users can easily find their required articles of interest.

In the time of digitalization, quantities of Bangla online blogs are rising quickly and furthermore accessible in a great amount. A lot of blog readers like to read and analyze articles from various blog sources. Most of the time readers are interested only in the articles of their categories of interest. So, to access these text documents easily, we need to classify the topics of these texts by using some standard automatic topic classification methods.

Significant attention has been given by the researchers on blog post classification for languages like English and other languages [2, 3]. Some work has been done on Bangla news classification [4, 5], emotion detection from Bangla text [6] and malicious Bangla text detection [7]. However, there is no work done in a Bangla blog post classification.

Islam *et al.* introduced [4] a document categorization system where they used three supervised learning model i.e. SVM, naïve Bayes (NB) and SGD. Classification of Bangla newspaper documents in predefine twelve class was done [5] using SVM which gives promising accuracy. A new comprehensive dataset [8] is employed to classify Bangla articles from online news portals. They used two TF-IDF and Word2vec feature selection methods to train the five supervised learning models logistic regression, neural network, NB, random forest and AdaBoost. SGD classifier has been used [9] to categorize the Bangla document from BDNews24. It is observed that the proposed method performed better compared to SVM and NB classifier. Paper [10], introduced a voting classifier (VC) to help sentiment analysis for US airline companies based on SGD and logistic regression (LR) and used a soft voting mechanism to make the final prediction. In [11], the authors try to detect the sentiments from Bangla microblog posts to identify the overall polarity of texts as either negative or positive using SVM or maximum entropy (MaxEnt).

It is incorporated inverse class frequency (ICF) with term frequency (TF) and inverse document frequency (IDF) that provide enhanced performance for feature extraction [12]. Ankita *et al.* [13] also used term frequency-inverse document frequency-inverse class frequency (TF-IDF-ICF) for feature selection and evaluate that TF-IDF-ICF has performed better compare to TF-IDF. In [14], the authors inspect the use of cosine similarity and Euclidean distance as similarity measures on the vector space model based on TF-IDF. The acquire accuracy for cosine similarity and Euclidean is 95.80% and 95.20% respectively. Bangla web documents are categorized [15] using decision tree (C 4.5), k-NN, NB, and SVM while SVM was the most successful among these four methods. Paper [16] proposed a topic modeling-based approach to extractive automatic summarization. A general framework for short text classification by learning vector representations of both words and hidden topics together has been presented in [17]. Almost all of the preceding cited research deals with Bangla newspaper article classification and English blog article classification. Our main intent in this study is to classify the Bangla blog article based on supervised learning.

2. RESEARCH METHOD

The detection of multi-lingual unorganized and frequent words is a quite difficult task. We have employed three feature extraction technique i.e. unigram TF-IDF, bigram TF-IDF and trigram TF-IDF to train nine supervised learning model. Before implementing classification tasks, it needs to make the dataset accurate and valid by pre-processing which helps to train and test the data. On the pre-processing stage, Tokenization, stop word removal, stemming and feature extraction was performed on our dataset. The Data collection, pre-processing and feature extraction is detailed in the following subsection. After feature extraction, the dataset is split into train and test set and train datasets are employed to train the model. Finally, performance is evaluated to find the best classifiers and best feature extraction technique for this classification task. Figure 1 illustrates the block diagram of the entire process.

In performance analysis, accuracy and some other metrics are required to evaluate the performance of a classifier. A confusion matrix for a binary classification has the number of true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs). For the scenario of multiclass problem, the confusion matrix will be dimension n * n (n>2). A matrix of n rows and n columns, where is not possible to directly calculate TPs, TNs, FPs and FNs. The value of TPs, TNs, FPs and FNs for class I is calculated as [18]:

$$\mathrm{TP_i} = \mathrm{a_{ii.}} \tag{1}$$

$$\mathrm{FP_i} = \sum_{\mathrm{j=1,j\neq i}}^{\mathrm{n}} \mathrm{a_{ji}}. \tag{2}$$

$$\mathrm{FN_i} = \sum_{\mathrm{j=1,j\neq i}}^{\mathrm{n}} \mathrm{a_{ij}}. \tag{3}$$

$$\mathrm{TN_i} = \mathrm{um}_{\mathrm{j=1,j\neq i}}^{\mathrm{n}}\ ,\sum_{\mathrm{k=1,k\neq i}}^{\mathrm{n}} \mathrm{a_{jk}}. \tag{4}$$

To explore the effectiveness of the model, precision, recall and F_1 score are employed as evaluation metrics which are defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$F_1\,\text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

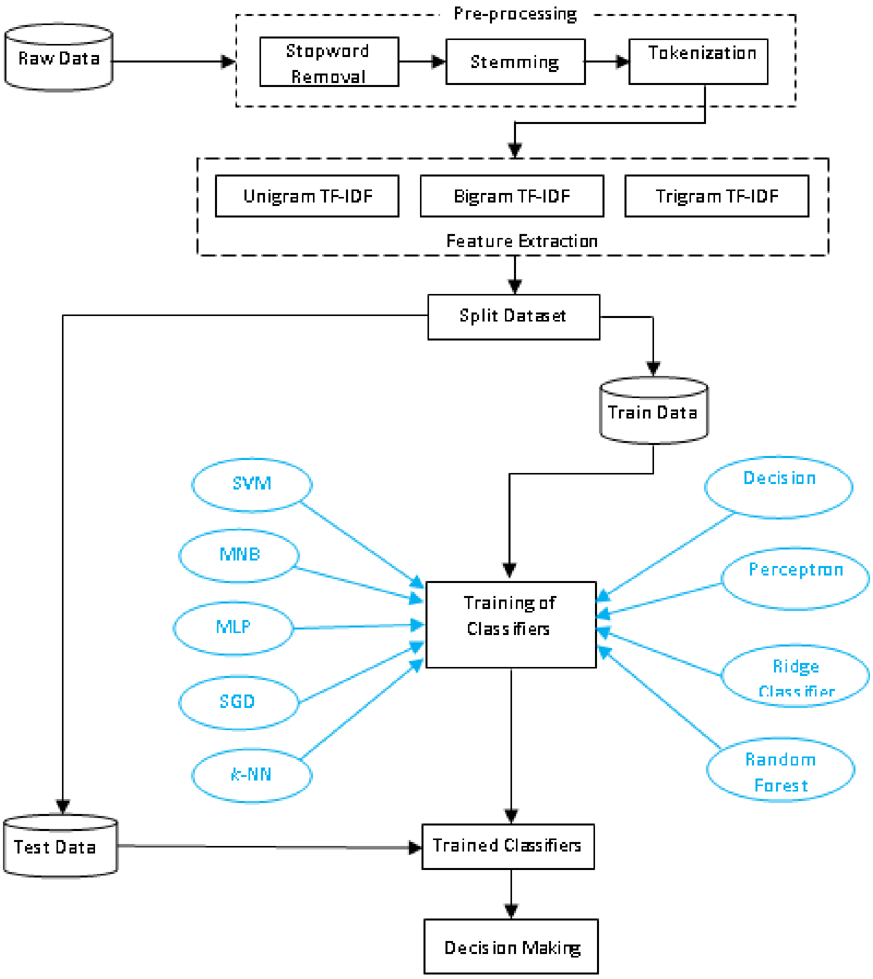

Figure 1. Block diagram of proposed model

2.1. Description of dataset

The Dataset used in this research is created by crawling few Bangla blog as like BDnews24blog [19], ebanglahealth [20], pavilion [21], and Vromon guide [22]. To automate the data collection process, a web crawler is developed using python. We considered only blog posts of eight predefined domain namely, finance, politics, technology, health, sports, entertainment, campus, and traveling. The distribution of dataset is given Table 1.

Table 1. Distribution of data

Class	Frequency
Finance	634
Politics	637
Technology	592
Health	492
Sports	406
Entertainment	638
Campus	631
Traveling	515
Total	4545

2.2. Data pre-processing

2.2.1. Tokenization

For better classification, it is necessary to go through tokenization and pre-processing before applying any feature extraction technique. Tokenization is the system that intends to break the content report into tokens delimited by a newline or tab or white spac. In this experiment, token is split based on some delimiter like punctuation, space, newline, emoji, and special character. All the special characters, punctuation, English and Bangla numerals, extra whitespace are eliminated in this stage. Word of other languages except Bangla and single-character words are also removed. Now the dataset is a bag of words. Figure 2 depicts the pseudocode of the tokenization process.

```
for i = 0 to i = length (dataset)
    list_of_word = dataset [i] split into (r"\.\s|\?\s|\!\s|\n")
    for j = 0 to j = length(list_of_word)
        marged_document = marged_document+ list_of_word [j] + " "
        dataset[i] = marged_document
    end
    marged_document = ""
end
```

Figure 2. Bangla_dataset_tokenization

2.2.2. Stop word removal

Stop words are the most repeatedly appeared words that don't help us find relevant information of the text documents. These words do not carry any significant information about the experiment. Some of the words are used only for grammatical rules. So, these non-significant words create huge noise which may damage the performance of text classification. The main reason is that they do not differentiate between relevant and non-relevant text. Stop word are removed from our dataset. A dataset of Bangla stop word [23] have been used for this process. The stop word dataset contains 390 stop words like “আমি”, “এবং”, “আরও”, and “এটা”. Figure 3 represents the pseudocode of the Stopword removal process.

```
for i = 0 to i = length(dataset)
    list_of_word = dataset [i] split into (r"\.\s|\?\s|\!\s|\n")
    for j = 0 to j = length(list_of_word)
        if  list_of_word [j] in stopword:
            continue
        marged_document = marged_document + list_of_word [j] + " "
        end
    dataset [i] = marged_document
    marged_document = " "
end
```

Figure 3. Bangla_stopword_removal

2.2.3. Stemming

It is essential to determine the root word of any word for an enhanced classification in case of a highly inflectional language like Bangla. The idea is to erase inflections from a word to obtain its stem word. There are several stemmers exist for Bangla language [24, 25]. A lightweight rule-based stemmer has been employed [24] in this research. Sample output of the stemmer used is given in Table 2.

Table 2. Sample output of used stemmer

Input word	Output word
মাসের	মাস
অফিসে	অফিস
থাকাটাই	থাকা
এমনটা	এমন
এসেই	এস
চাকরিটির	চাকরি

2.3. Feature extraction

The purpose of feature extraction method to reduce the dimensionality of the corpus by eliminating features that is inappropriate for the classification. Several statistical approaches can be used for feature extraction. TF-IDF have chosen in this experiment. TF-IDF considers two scores, TF and IDF. TF count the frequency of a term within the document, and IDF scale down the term that occurs frequently within multiple documents, as those are less important terms. To explore the proper feature extraction method three different string property e.g. unigram, bigram, and trigram are implemented in this experiment. Unigram features do not take into account the relevancy of words in a sentence. Bigram features consider the relevancy of two consecutive words within a sentence and trigram features include the relevancy between three consecutive words.

2.4. Training

An in-depth exploration has been performed to find a proper learning model for Bangla Blog Article classification. MNB, MLP, SVM, decision tree, random forest, SGD, ridge classifier, perceptron, and KNN has been considered among the available learning model from the literature review. Each of the datasets is split into an 80:20 ratio as a train set and test set. Three experiments have been conducted on this dataset. Firstly, unigram TF-IDF, then bigram TF-IDF and trigram TF-IDF are also employed as features extraction method with all the mentioned classifiers.

3. RESULTS AND DISCUSSION

Proposed System has been implemented on 4545 blog posts of eight distinct domains. Nine supervised learning models i.e. MNB, MLP, SVM, decision tree, random forest, SGD, ridge classifiers, perceptron and *k*-NN have been trained. Furthermore, the performance of each classifier has been experimented using three different string property unigram, bigram and trigram. First experiment was conducted using unigram TF-IDF features. Figure 4 illustrate the accuracy obtained by unigram TF-IDF features.

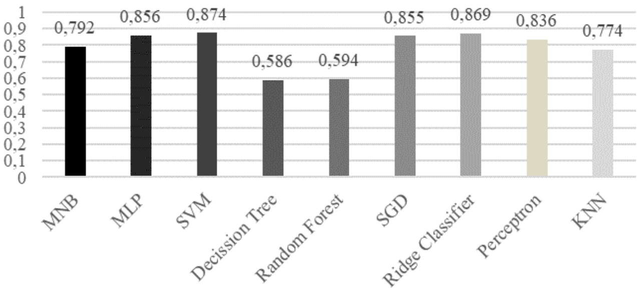

Figure 4. Accuracy using unigram TF-IDF features

SVM achieved the highest accuracy of 0.874 followed by Ridge classifier and MLP in the case of unigram TF-IDF features. SGD also has a decent performance. Decision tree and random forest has the poorest performance. Secondly, another experiment was conducted using bigram TF-IDF features. The accuracy achieved by the second experiment is in Figure 5.

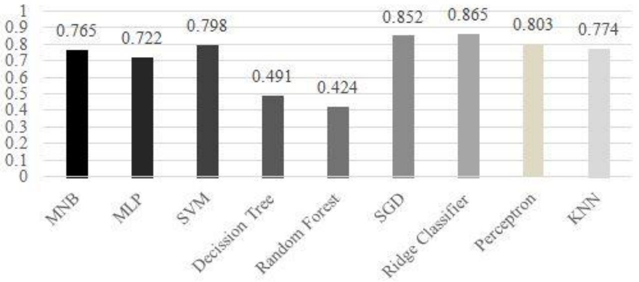

Figure 5. Accuracy using bigram TF-IDF features

In the second experiment, the accuracy of most of the classifiers decreased compares to the first experiment (unigram TF-IDF features). The performance gap between SVM with unigram TF-IDF features and bigram TF-IDF features are quite noticeable. In this case, Ridge classifier has the highest accuracy of 0.865. SGD is performing very close to it. Finally, the experiment has been conducted using trigram TF-IDF features which have the following accuracy given by Figure 6.

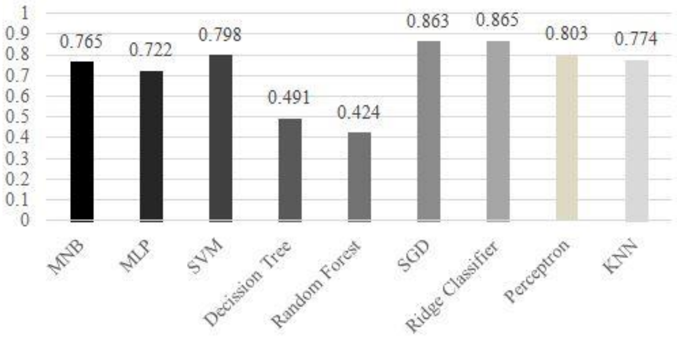

Figure 6. Accuracy using trigram TF-IDF features

In this study, the accuracy of SGD slightly increased compares 2nd experiment. The performance of other models did not have any change. Like bigram TF-IDF features, hare also Ridge classifier has the highest accuracy of 0.865. SGD is performing very close to it. The result indicates that SVM with unigram TF-IDF features have the highest accuracy among all the experiment. Ridge classifier with unigram TD-IDF features and SGD with bigram TF-IDF features are is second and third positions in terms of accuracy. Decision tree, random forest and MLP with trigram TF-IDF features are in the last three positions while MLP was performing quite better with unigram TF-IDF features. Precision, recall and F_1 score have been calculated for evaluating the performance of those models. Precision describes which portions of the Prediction (for a given class X) were actually correct. Besides recall describes which portions (for a given class X) was predicted correctly. F_1 score is the weighted mean of precision and recall. This score takes both into account. Precision, recall and F_1 score are given in the Table 3.

Table 3. Performance of difference approaches

Feature extraction technique	Learning model	Precision	Recall	F_1 Score	Accuracy
Unigram TF-IDF	MNB	0.83	0.79	0.78	0.79
	MLP	0.86	0.86	0.86	0.86
	SVM	0.88	0.87	0.88	0.87
	Decision Tree	0.59	0.58	0.59	0.59
	Random Forest	0.70	0.59	0.58	0.59
	SGD	0.86	0.86	0.86	0.85
	Ridge Classifier	0.87	0.87	0.87	0.87
	Perceptron	0.83	0.83	0.83	0.84
	k-NN	0.78	0.77	0.77	0.77
Bigram TF-IDF	MNB	0.83	0.79	0.78	0.77
	MLP	0.77	0.72	0.72	0.72
	SVM	0.85	0.80	0.80	0.80
	Decision Tree	0.60	0.49	0.51	0.49
	Random Forest	0.70	0.59	0.58	0.42
	SGD	0.86	0.86	0.86	0.85
	Ridge Classifier	0.87	0.87	0.87	0.87
	Perceptron	0.80	0.80	0.80	0.80
	k-NN	0.78	0.77	0.77	0.77
Trigram TF-IDF	MNB	0.83	0.79	0.78	0.77
	MLP	0.72	0.72	0.72	0.72
	SVM	0.85	0.80	0.80	0.80
	Decision Tree	0.60	0.49	0.51	0.49
	Random Forest	0.59	0.42	0.36	0.42
	SGD	0.86	0.86	0.85	0.86
	Ridge Classifier	0.87	0.86	0.86	0.87
	Perceptron	0.80	0.80	0.80	0.80
	k-NN	0.78	0.77	0.77	0.77

It can be observed from Table 3 that SVM with unigram features have the highest F_1 score. The performance of SVM has dropped in the case of bigram TF-IDF features and trigram TF-IDF features. Ridge classifier with unigram features and ridge classifier with bigram features are in the second position considering F_1 score. In terms of feature extraction, unigram has outperformed bigram and trigram. The reason behind this is data scarcity. As *N*-gram length increases, the occurrence of any given *N*-gram decreases. The sparser our data set, the worse we can model it. The number of events, i.e. *N*-grams during training period becomes progressively less as *N* increases. We have a comparatively very large amount of token types (i.e. the vocabulary of our dataset is very rich), but each of these types has a low frequency. For this, we have got better results with a lower order *N*-gram model. Moreover, our training data set is not so large, whereas a higher order of *N*-gram required a large number of data.

4. COMPARATIVE ANALYSIS OF RESULTS

Some relevant recent research results are compared below, to evaluate the efficiency of our proposed approach in classifying Bangla blog article. The study of literature exposes that all of the research reports are limited to the classification of Bangla news articles. Because of the unstructured nature of the blog article, classification of blog articles is difficult. Besides the diversity of content of blog article make the classification task more difficult. Furthermore, the lack of use of a common dataset of blog article makes it difficult to have an equitable comparison of the performance of different approaches. Even though such limitations, a review of numerical results related to Bangla blog article classification is attempted to evaluate the comparative merits of this work. The overview of all the approaches including our approach is given in Table 4.

Table 4. Comparison with our proposed works

Work done	No of classes	Sample size	Source of dataset	Best performed feature selection technique	Best performed classifier	Accuracy
Proposed Work	8	4545	Bangla Blog	Unigram TF-IDF	SVM	87.4%
Dhar *et al.* [13]	8	4000	Online New Portal	TF-IDF-ICF	MNB	98.87%
Dhar *et al.* [12]	8	4000	Online New Portal	TF-IDF-ICF	MLP	97.65%
Mandal *et al.* [15]	5	1000	Online New Portal	TF-IDF	SVM	89.14%
Alam *et al.* [8]	5	376226	Online New Portal	TF-IDF	Neural Network	96%
Dhar *et al.* [14]	5	1000	Online New Portal	TF-IDF	Cosine Similarity	95.8%
Islam *et al.* [4]	12	31,000	Online New Portal	TF-IDE	SVM	*NA*

[1] *NA*: Not Mentioned

In the work [13], they classify news documents with a dataset of size 4000. MNB is giving promising result is there experiment. In Bangla web document classification [12] process, text documents of online news articles are classified where MLP achieved the highest accuracy. For developing a Bangla web document categorization system [15], different classifier has experimented where SVM with TF-IDF feature selection technique performed the best. News article classification using a comprehensive dataset [8] is done where neural network with TF-IDF feature selection technique has the most decent performance. Paper [14] introduced a text document classification approach where cosine similarity with TF-IDF feature selection technique achieved an accuracy of 95.8%.

Considering the discussed scenario and challenges in classification of blog articles, our obtained accuracy is more than 87% which is quite good. As we have specified earlier, the absence of the same dataset, dataset collected from a different domain and different source (Bangla blog instead of news portal) it is not equitable to explicit comparison of our work with others. Besides, it may not be improper to claim that our proposed approaches have enough distinguishing information to classify Bangla blog post.

5. CONCLUSION AND FUTURE SCOPE

The domain of the Bangla blog is rapidly expanding, so automated classification can be exceedingly beneficial. In this system, a dataset of 4545 instances from 8 distinct domains has been collected from various Bangla blogs. We have experimented with different learning models and different features selection approaches. After comparing performance, it is observed that SVM with unigram TF-IDF features achieved the highest accuracy and F_1 score of 0.87 and 0.88 respectively. SGD and ridge classifiers also perform closely to SVM. SGD and ridge classifier with unigram features obtained an accuracy of 0.85 and 0.86. Another observation is, for most of the cases unigram is performing better compare to bigram and trigram.

We expect that our dataset will help the Bangla NLP researcher to extend this Bangla blog article classification task. Besides this corpus can be employed in other Bangla NLP tasks.

ACKNOWLEDGEMENTS

The authors have received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

[1] C. Ferguson and M. Chowdhury, “The Phonemes of Bengali,” *Language*, vol. 36, no. 1, pp. 22-59, 1960. doi: 10.2307/410622.

[2] A. Sun, M. Suryanto, and Y. Liu, “Blog classification using tags: An empirical study,” *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pp. 307-316, 2007, doi: 10.1007/978-3-540-77094-740.

[3] K. Jun Lee, M. Lee, and W. Kim, “Blog classification using K-means,” *Proceedings of the 11th International Conference on Enterprise Information*, vol. 4, pp. 61-67, 2009, doi: 10.5220/0001949600610067.

[4] M. D. S. Islam, F. E. M. D. Jubayer, and S. I. Ahmed, “A comparative study on different types of approaches to Bengali document categorization,” *Proc. of 4th the Int. Conf. on Eng. Research Innov. and Educ.*, pp. 1-6, 2017.

[5] M. Islam, F. Jubayer, and S. Ahmed, “A support vector machine mixed with TF-IDF algorithm to categorize Bengali document,” *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 191-196, 2017, doi: 10.1109/ecace.2017.7912904.

[6] S. Azmin and K. Dhar, “Emotion detection from Bangla text corpus using Naïve Bayes classifier,” *2019 4th Int. Conf. on Electrical Information and Comm. Tech. (EICT)*, pp. 1-5, 2019, doi: 10.1109/eict48899.2019.9068797.

[7] T. Islam, S. Latif, and N. Ahmed, “Using social networks to detect malicious Bangla text content,” *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-4, doi: 10.1109/ICASERT.2019.8934841.

[8] M. Tanvir Alam and M. Mofijul Islam, “BARD: Bangla article classification using a new comprehensive dataset,” *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, pp. 1-5, 2018, doi: 10.1109/ICBSLP.2018.8554382.

[9] F. Kabir, S. Siddique, M. R. A. Kotwal, and M. N. Huda, “Bangla text document categorization using stochastic gradient descent (SGD) classifier,” *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, Noida, pp. 1-4, 2015, doi: 10.1109/CCIP.2015.7100687.

[10] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, “tweets classification on the base of sentiments for US airline companies,” *Entropy*, vol. 21, no. 11, p. 1078, 2019, doi: 10.3390/e21111078.

[11] S. Chowdhury and W. Chowdhury, “Performing sentiment analysis in Bangla microblog posts,” *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1-6, 2014, doi: 10.1109/iciev.2014.6850712.

[12] A. Dhar, N. Dash, and K. Roy, “Categorization of Bangla web text documents based on TF-IDF-ICF text analysis scheme,” *Social Transformation –Digital Way*, pp. 477-484, 2018. doi: 10.1007/978-981-13-1343-139.

[13] A. Dhar, N. Dash, and K. Roy, “Classification of Bangla text documents based on inverse class frequency,” *2018 3rd International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pp. 1-6, 2018, doi: 10.1109/iot-siu.2018.8519866.

[14] A. Dhar, N. Dash, and K. Roy, “Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents,” *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall), Dehradun*, pp. 1-6, 2017, doi: 10.1109/ICACCAF.2017.8344721.

[15] K. Mandal and R. Sen, “Supervised learning methods for Bangla web document categorization,” *International Journal of Artificial Intelligence & Applications*, vol. 5, no. 5, pp. 93-105, 2014.

[16] Z. Wu, L. Lei, G. Li, H. Huang, C. Zheng, E. Chen, and G. Xu, “A topic modeling-based approach to novel document automatic summarization,” *Expert Systems with Applications*, vol. 84, pp. 12-23, 2017, doi: 10.1016/j.eswa.2017.04.054.

[17] H. Zhang and G. Zhong, “Improving short text classification by learning vector representations of both words and hidden topics,” Knowledge-Based Systems, vol. 102, pp. 76-86, 2016.

[18] B. P. Salmon, W. Kleynhans, C. P. Schwegmann, and J. C. Olivier, “Proper comparison among methods using a confusion matrix,” *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Milan, 2015, pp. 3057-3060, doi: 10.1109/IGARSS.2015.7326461.

[19] “Blog.bdnews24.com–pioneer blog for citizen journalism in Bangladesh,” [Online]. Available at: *https://blog.bdnews24.com/*. [Accessed: 14- Jul- 2020].

[20] “Bangla health tips–Bangla health tips, news and information,” [Online]. Available at: https://www.ebanglahealth.com. [Accessed: 14- Jul- 2020].

[21] "প্যাভিলিয়ন," [Online] Available at: *https://www.pavilion.com.bd* [Accessed: 14- Jul- 2020].

[22] “Vromon guide,” [Online] Available: *https://www.vromonguide.com.* [Accessed: 14- Jul- 2020].

[23] “Stopwords-iso/stopwords-bn,” [Online] Available at: *https://github.com/stopwords-iso/stopwords-bn.* [Accessed: 14- Jul- 2020].

[24] M. Mahmud, *et al*., “A rule based bengali stemmer,” *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2750-2756, 2014, doi: 10.1109/icacci.2014.6968484.

[25] T. Urmi,. Ismail, “A corpus based unsupervised Bangla word stemming using N-gram language model,” *2016 5th Int. Conf. on Informatics, Electronics and Vision (ICIEV)*, pp. 824-828, 2016, doi: 10.1109/iciev.2016.7760117.