Bangla News Headline Classification Comparison Among Different Classifiers

Khaled Ahmmed Anik, MD Zisanur Rahman, Md Humaion Kabir Mehedi,
Mohammed Julfikar Ali Mahbub and Annajiat Alim Rasel

Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka - 1212, Bangladesh
{*khaled.ahmmed.anik, md.zisanur.rahman, humaion.kabir.mehedi,*
mohammed.julfikar.ali.mahbub}*@g.bracu.ac.bd*
annajiat@gmail.com

***Abstract*—Document classification or categorization is the process of categorizing – or labeling – documents based on content. Bangla is a very enriched language in which text classification or categorizing is different from other languages. In this paper, we have compared some of the existing text classifiers to categorize Bangla Newspaper Headlines. Often readers find it difficult to select news categories in a densely gathered news-paper that is clustered with various other categorical news. This paper aims to find out a classifier that performs reasonably well in terms of categorizing news headlines. A collection of 136,811 Bangla news headlines from various newparer is used as dataset. The news headlines have been categorized into 5 categories. The Dataset has been pre-processed by NLP methods and for categorizing this dataset, existing available algorithms that are Recurrent Neural Networks (RNN), Random Forest, K-Nearest Neighbour (KNN) & deep Recurrent Neural Networks (DRNN) have been used and modeled. Among these classifiers, DRNN performed better with f1 score 84.**

***Index Terms*—Classification, NLP, RNN, Random Forrest, K-NN, D-RNN, classifiers, pre-processing**

I. Introduction

With the growing need for digitization, categorizing unstructured data has become quite necessary in order to classify each data. Moreover, news headlines often get clustered in the reader's perspective as all categories of news headlines are shown in a single instance. For this reason, document categorization has huge usability in different aspects of our life. In the process of text classification, some algorithms are run on a given document from which some different portions are assigned to a set of categories. This "set of categories" is represented as a flat hierarchical entity and each set of categories is unique which does not overlap with one another [1]. In text categorization, it is an important task to process the texts in a mode that the text classifiers can easily classify those. i.e., transformation into an suitable document representation [2]. Data processing is another crucial part of this process where data is brought to a minimal stage where it is ready to be used on a classifier. The actual goal of this study is to find out the estimation of which news headline belongs to which distinct category. For this study, a dataset of 136,811 news headlines has been gathered from different online news portals. To remove all the discrepancies, text processing has been applied that refines the contents of the dataset. Text processing includes tokenization, label encoding, removing low length data (removal of short headlines), Data cleaning (removal of stop words), and removing duplicate words of each category.

II. Related Works

Text categorization in Bangla Language has been here for a while since its beginning in the 2000's. In the beginning, there were few datasets to work with, but now the tables have turned, and a large number of datasets can be found in most of the places in which NLP algorithms can be implemented. M.Rahman *et al.* [3] have implemented BERT and ELECTRA transformers for Bangla Document categorization on three unique datasets and got satisfactory results. Their study shows that BERT outperformed ELECTRA for all provided datasets in precision. Other than that, the overall performance of ELECTRA was better than BERT and got the highest values of 94.18% and 96.39% for most cases.

H. Berger *et al.* performed another research [2] in which they compared text classification systems using N-gram frequency statistics. Additionally, three classification methods (PART, SMO, and NBm) were applied to n-gram character frequency data and word frequency-based document representation. We utilized a dataset of multilingual electronic communications made up of 1,811 emails that had been manually separated into layers. Additionally, the effect of information on the performance of classifiers was studied. Their research established the viability of multi-class email classification in SMO-related emails.

A. Sharfuddin *et al.* [4] demonstrated a method for sentiment analysis via the use of a deep recurrent neural network with BiLSTM implementation. Their dataset included 10,000 comments extracted from Facebook, of which 5,000 were favorable and the remainder were negative. The dataset was trained using their suggested technique, BiLSTM, which obtained 85.67 percent accuracy. Meanwhile, they trained the

dataset using SVM, Decision Tree Classifier, and Logistic Linear Regression, achieving 68.77 percent, 67.50 percent, and 60.94 percent accuracy, respectively.

A companion article [5] compares SVM, Naive Bayes, and Stochastic Gradient Descent (SGD) for Bengali Document Categorization. Several feature selection strategies, including Chi-square distribution and normalized TFIDF (term frequency-inverse document frequency) using a word analyzer, were employed to assess the above-mentioned classifier's effectiveness in predicting a document category. Following the use of several strategies SVM classifier achieved the maximum F1-score of 92.56 percent, while normalized TFIDF was used to pick features and CHI-square was combined with NB to get the lowest F1-score of 83.36 percent.

Similar to [2]M. Mansur *et al.* performed another N-gram-based text categorization research [6] in the Bangla newspaper corpus, in which the performance of text categorization grows progressively as n (from 1 to 3) increases, but the outcome varies when n is greater than 3. This demonstrates that more n-grams do not guarantee improved language modeling in Bangla N-gram-based text classification. This study aims to classify Bengali news headlines using different classifiers and put up a comparison among the used classifiers.

III. Methodology

The work process "Fig. 1" starts with data collection and pre-processing the dataset. Then we separate the dataset into training and testing parts. Besides, 4 different classifications: RNN, Random Forest, K-NN & D-RNN model have to be designed. Fit the dataset into the models and evaluate the accuracy of the models with an evaluation matrix.

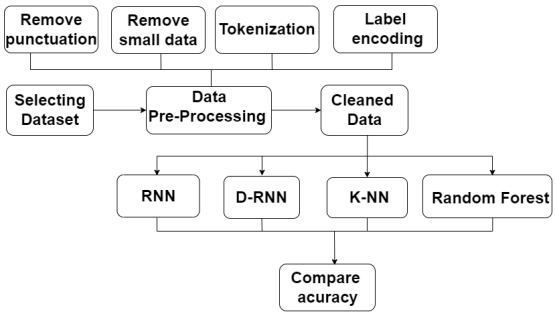

Fig. 1. Work Process

A. Data Collection:

Bangla datasets are not available and creating datasets is a time-consuming task. So, we collected a Bangla news heading dataset and some source code from GitHub for the study purpose [11]. Data contains headlines from different Bangla newspapers which helps to proceed further.

B. Dataset characteristics:

he collected dataset contains 136,812 rows and 3 columns "Fig. 2". The first row contains column headings, other rows contain a headline, category, and source newspaper name.

	headline	category	newspaper name
0	সীমানা পেরিয়েও হতাশ সীমান্ত	sports	Dainik Inqilab
1	সিৎসিপাসের কাছে হেরে বিদায় ফেদেরারের	sports	Dainik Inqilab
2	বন্দুক আর গো-রক্ষক দিয়ে দেশ চলে না : মমতা	International	Dainik Inqilab
3	নিদাহাস ট্রফিতে 'আন্ডারডগ' বাংলাদেশ	sports	Dainik Inqilab
4	ওবামার পরিচ্ছন্ন জ্বালানি পরিকল্পনা বাতিল করবে...	International	Dainik Inqilab

Fig. 2. Characteristics of the Dataset

Here, we worked with only headline and category, the category column is our target part, the goal of classification is to reach the correct target column based on headline features. International, sports, national, politics and IT are 6 different categories are holding 48,813, 33,277, 24,935, 16,491, 10,589, 2,806 news headlines "Fig.3" respectively.

Pre-processing is needed to train the dataset so that it can provide better accuracy.

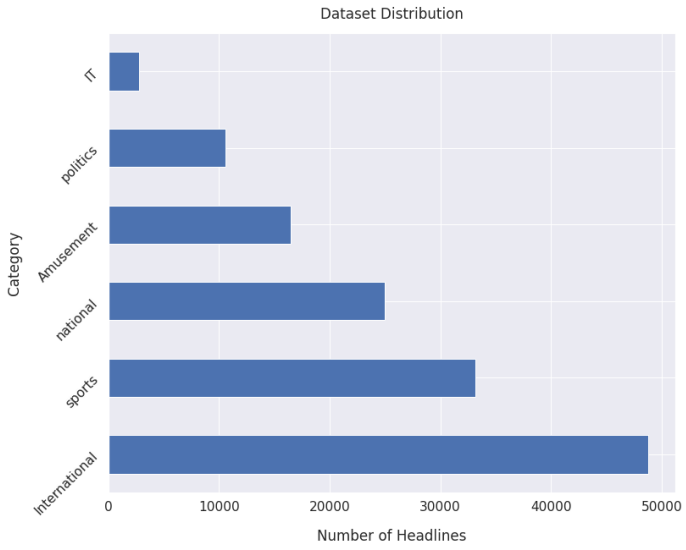

Fig. 3. Number of Headlines in each category

C. Data Pre-processing:

Before fitting into the chosen 4 classification models, we have to cut down unnecessary information and shape the data to get better accuracy from the models.

1) Remove punctuation: We removed punctuation and stop words because the stop words and punctuation will mislead the system. We have a function where all rows of headlines will be sent and replaced with punctuation-free headlines "Fig. 4".

Original: ফ্লোরিডায় হামলাকারী 'মানসিকভাবে অসুস্থ': ট্রাম্প
Cleaned: ফ্লোরিডায় হামলাকারী মানসিকভাবে অসুস্থ ট্রাম্প

Original: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ: মাশরাফি
Cleaned: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ মাশরাফি

Fig. 4. 'Remove punctuation' function's result

2) Remove Small data: In the dataset, we have some small headlines which may not help our training part and it will increase the complexity of the prediction system. So, we removed the headlines "Fig. 5" with two or fewer word lengths.

After Cleaning:
Removed 4098 Small Headlines
Total Headlines: 132713

Fig. 5. Number of Headlines Removed and remained

3) Tokenization : Tokenization is the process of breaking down a big chunk of text into smaller pieces or words that help to give numerical identity or works as a feature of a particular category which helps to find out categories by matching word patterns. Here, we tokenized "Fig. 6" all the headlines of each category.

Class Name : IT

নতুন 167
ফেসবুক 165
ও 143
স্মার্টফোন 107
নিয়ে 95
থেকে 94
শুরু 86
ডিজিটাল 80
জন্য 79
মোবাইল 75
Total Number of Unique Words:57490

Fig. 6. Example of Tokenized word of IT category

4) Label Encoding: For the simplicity of the comparison with the category, we encoded the 6 categories with numerical unique values "Fig. 7", which will make further processing smooth. Here, we assign 6 different numeric values against categories.

D. Train the model:

Dataset has been split into an 80:20 ratio for training and test purposes I.

===== Label Encoding =====
Class Names:--> ['Amusement' 'IT' 'International' 'national' 'politics' 'sports']

Label Encoding ==> [0 1 2 3 4 5]

Fig. 7. Assigning numarical value to all Category

TABLE I
DATASET DISTRIBUTION

Set	Size
Full	132, 713
Training	95, 552
Test	13, 272
Validation	23, 889

The dataset is processed, cleaned, and ready for training. The same Dataset will be trained in RNN, Random Forest, K-NN & D-RNN models.

E. Random Forest:

A random forest's hyper-parameters are somewhat comparable to those of a decision tree or a bagging classifier. Fortunately, it is unnecessary to combine a decision tree with a bagging classifier since the random forest classifier class may be used instead. By using the algorithm's regressor, regression problems may be addressed using random forest.
While growing the trees, the random forest increases the model's unpredictability. When a node is split, it seeks the best feature from a random selection of characteristics rather than the key feature. We fit our train data into Random Forest, and it produces a 39.43 percent accuracy for test data.

F. KNN

The K nearest neighbors algorithms fall under the category of supervised learning and are most commonly used for classification and regression. This is a general algorithm that is also used to impute missing values and resample the dataset. After implementing KNN on the same dataset, we have achieved an accuracy of 32.35%.

G. RNN

RNN's developed from feedforward networks have comparable behavioral characteristics to the human brain. Simply said, recurrent neural networks are capable of making predictions from sequential data that other algorithms are unable to. Each word in the sequence is represented by an input vector that is a lookup of the word embedding matrix using a single hot encoded vector from the dictionary. Each keyword entered should have precise word embeddings. In this context, the term "word" refers to a token that can refer to either a word or a punctuation mark. After applying RNN to the processed dataset, we got an accuracy of 52.55%.

H. DRNN

A variety of applications nowadays are being implemented on Deep Recurrent Neural Network which is an efficient deep learning technique. The required depth of the DRNN cannot

be decided at first as variable length coding structures are designed to represent the DRNN's of particular depths. The activation function is critical to the success of DRNN's and should be utilized cautiously in these networks. On the basis of this insight, knowledge-based intersection and mutation operators are presented for carefully controlling the employment of activation functions in GA in order to optimize DRNN performance. 84.44% accuracy was obtained from the DRNN model.

IV. Result Comparison

The goal of this study is to find out the best performing classification model among RNN, Random Forest, K-NN & deep RNN II. DRNN is the best performing model in our test, and we got 84.44% test accuracy. In the training time, it goes through 7 epochs, and after each epoch, its accuracy gets increased and data loss decreases. The final epoch train data loss was 0.0845 "Fig. 8"."Fig. 2" shows how well it is performed by the diagonal line. This study did not get any other nearest good performance among the 4 selected models.

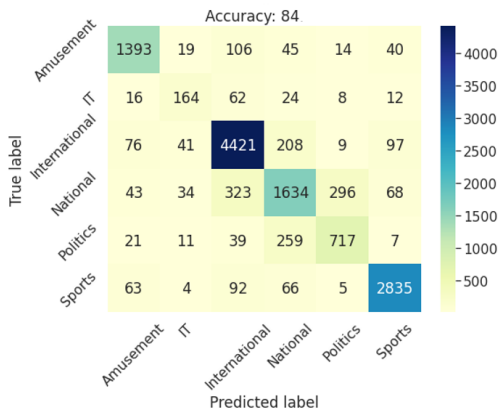

Fig. 8. DRNN accuracy Heat-Map

RNN model obtained 52.55 test accuracy which is not up to the mark comparing recent works on Bangla news classification with BNLP. RNN is the second-best performing model in this study case. On the other hand, KNN and Random Forest model data loss was very high and accuracy was 32.35 and 39.43 respectively.

TABLE II
Accuracy found in DRNN, RNN, KNN, Random Forest

Models	DRNN	RNN	KNN	Random Forest
Testing accuracy	84.44%	52.55%	32.35%	39.43%

This study found deep RNN as the best classification model among RNN, Random Forest, K-NN & deep RNN.

V. Conclusion

Four classification models: RNN, Random Forest, K-NN & deep RNN are used in this study for Bangla news classification from news headlines with a large collected dataset. This study shows that deep RNN models provide higher accuracy than the other 3 models based on pre-processing used in this study. Compared to other studies, results found for deep RNN are prominent and widely usable. RNN, KNN, Random Forest did not meet the expected standard. As we used 4 models to compare, some more popular classification models for Bangla such as BERT can be included in the future Though this study got some satisfactory results, the scope for improvement is there also. In the future, we will facilitate better categorization models and use more Bangla-friendly pre-processing.

Acknowledgment

We thank the members of the School of Data and Sciences of BRAC University for their generous guidance and support.

References

[1] Krendzelak, Milan & Jakab, Frantisek. (2015). Text categorization with machine learning and hierarchical structures. 1-5. 10.1109/ICETA.2015.7558486.

[2] Berger H., Merkl D. (2004) A Comparison of Text-Categorization Methods Applied to N-Gram Frequency Statistics. In: Webb G.I., Yu X. (eds) AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science, vol 3339. Springer, Berlin, Heidelberg.

[3] M. M. Rahman, M. Aktaruzzaman Pramanik, R. Sadik, M. Roy and P. Chakraborty, "Bangla Documents Classification using Transformer Based Deep Learning Models," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-5, doi: 10.1109/STI50764.2020.9350394.

[4] A. Aziz Sharfuddin, M. Nafis Tihami and M. Saiful Islam, "A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554396.

[5] Islam, M.S., Jubayer, F.E., & Ahmed, S.I. (2017). A Comparative Study on Different Types of Approaches to Bengali document Categorization. ArXiv, abs/1701.08694.

[6] Mansur, M., UzZaman, N., & Khan, M. (2006). Analysis of N-Gram based text categorization for Bangla in a newspaper magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Hossain, E., Chaudhary, N., Rifad, Z. H., & Hossain, B. M. (2020). Bangla-news-headlines-categorization. GitHub.

[8] Ahmed, M., Chakraborty, P., & Choudhury, T. (2022). Bangla Document Categorization Using Deep RNN Model with Attention Mechanism. In Cyber Intelligence and Information Retrieval (pp. 137-147). Springer, Singapore.

[9] Naqvi, R. A., Khan, M. A., Malik, N., Saqib, S., Alyas, T., & Hussain, D. (2020). Roman Urdu news headline classification empowered with machine learning. Computers, Materials & Continua, 65(2), 1221-1236.

[10] Alam, F., Hasan, A., Alam, T., Khan, A., Tajrin, J., Khan, N., & Chowdhury, S. A. (2021). A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models. arXiv preprint arXiv:2107.03844.

[11] Eftekhar-Hossain, Bangla-News-Headlines-Categorization, 2020, https://github.com/eftekhar-hossain/Bangla-News-Headlines-Categorization/blob/master/headlines.csv