Multi-class Short Text Classification Using Ensemble of Deep Learning Classifier

Chapter · October 2022

DOI: 10.1007/978-3-031-19958-5_45

Eftekhar Hossain
Chittagong University of Engineering & Technology
29 PUBLICATIONS **416** CITATIONS

Moshiul Hoque
Chittagong University of Engineering & Technology
181 PUBLICATIONS **1,403** CITATIONS

Mohammad Anisur Rahaman
Chittagong University of Engineering & Technology
16 PUBLICATIONS **220** CITATIONS

Multi-class Short Text Classification using Ensemble of Deep Learning Classifier

Miftahul Jannat[1], Eftekhar Hossain[1], Mohammed Moshiul Hoque[2]*, and Mohammad Anisur Rahaman[1]

[1] Department of Electronics and Telecommunication Engineering
[2] Department of Computer Science & Engineering
Chittagong University of Engineering and Technology
Chittagong-4349, Bangladesh
miftahuljannat1404@gmail.com, {eftekhar.hossain, moshiul_240, anisur.rahaman}@cuet.ac.bd

Abstract With the substantial outgrowth of e-commerce, social media usage and online news portals, a great wave has been observed in expressing views through short text. Most textual contents are unstructured and messy forms, which are impractical and cumbersome to organize or manipulate by human experts. Therefore, developing an automatic short text classification model concerning low-resource languages, including Bengali, is critical. Moreover, the crucial barrier to classifying short text in Bengali is the unavailability of text corpora, scarcity of linguistics tools, a limited number of words in the text, and a lack of dependencies between the words. This paper presents a short text classification model using the ensemble of four base deep learning classifiers (Neural Network (NN), Convolutional Neural Network (CNN), Bidirectional Long Short Term Memory (BiLSTM), and Bidirectional Gated Recurrent Unit (BiGRU)). Additionally, a corpus of around 0.13 million Bengali texts is developed for short text classification into six categories (e.g., international, national, sports, amusement, technology, and politics). The evaluation results on the developed corpus demonstrated that the proposed method outperformed all the baselines machine learning and deep learning models by obtaining the highest weighted f1-score of 84.4%.

Keywords: Natural language processing · Short text classification · Text corpus · Deep learning · Ensemble approach

1 Introduction

With the advent of the Internet, the generation of textual data is growing exponentially due to the extensive use of social media platforms, blogs, portals, and e-commerce sites. Most of the textual contents on the web are unstructured and messy forms, which is infeasible to analyze and organize manually. An intelligent text classifier can quickly and efficiently classify massive amounts of textual data into predefined categories at a lesser cost. The text classification primarily focuses on the texts those can be comprehended easily by observing the words' relational dependencies. However, a new genre of textual data, the *short text*, has emerged in recent years due to its many potential applications in online communication and advertisement. An intriguing and appealing short text (i.e., news headlines, product advertisement) is more powerful than a meaningful long text in gaining the attention of the readers or viewers. With the spiced-up headlines, people waste their time opening news that they do not even prefer reading. In such cases, automatic categorization of the news headlines is of utmost importance among NLP researchers. Although many works on long text classification have been done in many high-resource languages, short text classification is in the preliminary stage concerning resource-constrained languages, including Bengali. Moreover, over the years, different methods have been employed for short text classification, such as support vector machine (SVM) with rule-based features [10]. However, many recent studies have tried deep neural networks for short text classification and obtained promising results in some languages [11,1]. The existing Bengali short text classification techniques do not explore the ensembling of deep learning methods and their effectiveness in Bengali text classification. This work presents a framework that employs a deep learning-based ensemble approach to categorize the Bengali short texts to mitigate past study flaws. The key contributions of the work are illustrated in the following:

- Develop a Bengali short text corpus consisting of 0.13 million headlines into six categories: national, international, technology, amusement, politics, and sports.
- Propose a deep learning-based ensemble model with four base classifiers (e.g., NN, CNN, BiLSTM, BiGRU) to categorize the Bengali short texts by optimizing the hyperparameters.
- Investigate and compare the performance of the proposed method with baseline machine learning, deep learning, and existing techniques, thus setting up a remark that paves the way for future research.

2 Related Work

Significant research activities have been conducted on short text classification concerning high-resource languages (i.e., English, Chinese, and Turkish). Yin et al. [11] developed a short text classification system by utilizing FastText embedding, which obtained an f1-score of 81.3%. Lu et al. [7] proposed a BiLSTM and attention-based approach for classifying Chinese short text. Qiu et al. [9] developed a dataset of 12000 Chinese news headlines consisting of 18 classes. This work obtained the highest f1-score (0.783) with the neural bag-of-words method. Muhammad et al. [5] developed an Urdu news headlines classification system, creating a dataset of $111,859$ news headlines with four categories. Their evaluation achieved the highest accuracy of 94.27% with the Naive Bayes classifier. Irfan et al. [4] accumulated a corpus of five categories comprising 2800 Sindhi news headlines. Analyzing features with TF-IDF, this work found that multi-layer perceptron (MLP) and linear SVM provide better results. The scarcity of the benchmark corpora and the unavailability of NLP tools are the primary barrier to Bengali short text classification. Dhar et al. [2] developed a dataset of 474 Bengali news headlines into three categories and proposed an optimized machine learning method for short text classification. Khushbu et al. [6] proposed a neural network method for classifying Bengali short texts into 11 classes. This method was tested on a dataset consisting of 8000 news headlines and achieved 90% accuracy. Bhuiyan et al. [1] developed an LSTM architecture for classifying the Bengali texts from 4580 news headlines. Hossain et al. [3] proposed a deep convolutional and recurrent neural networks method for classifying Bengali web texts into 12 categories. This work investigated the effect of an ensemble classifier on short text classification on a large-scale dataset for superior performance.

3 Dataset Development

Availability of benchmark corpora is a prerequisite to developing any text classification system using deep learning. Due to the unavailability of the standard corpus, this work developed a larger corpus called *Bengali short text (BST) corpus* containing around 0.13 million news headlines. Four prominent Bangladeshi online news portals (i.e., 'Daily Ittefaq', 'Daily Inqilab', 'Daily Jugantor', 'Daily Jai Jai Din') have been considered. A crawler (Chrome web scrapper) is used to cull the news headlines of the most readable six news categories, namely 'International', 'National', 'Sports', 'Amusement', 'Technology', and 'Politics'. To verify the accurate assignment of news categories, we manually check the labels for a small portion (10%) of the dataset. Three postgraduate students are working on Bangla Language Processing (BLP) were employed to correct the labelling. The majority voting technique is applied to choose the initial label, and the labels are scrutinized by a BLP expert with many years of experience. The inspection outcome is measured by calculating the kappa score. We obtain a 0.92 score which is comparatively well.

Firstly, all the unwanted characters are excluded from the corpus, which includes punctuation marks, digits, symbols (!@#$%&), etc. We also eliminated headlines having less than two words as it would be challenging for a system to capture the contextual meaning from only two words.

4 Methodology

This work's primary objective is to categorize the Bengali short texts into six categories: *international*, *national*, *politics*, *sports*, *technology*, and *amusement*. We develop computational models using machine learning (ML) and deep learning (DL) based methods to accomplish the task. Figure 1 shows an abstract view of the text classification system.

Raw texts can not be directly fed to train the ML and DL models. Before that, features are needed to extract from these texts. This work utilized various

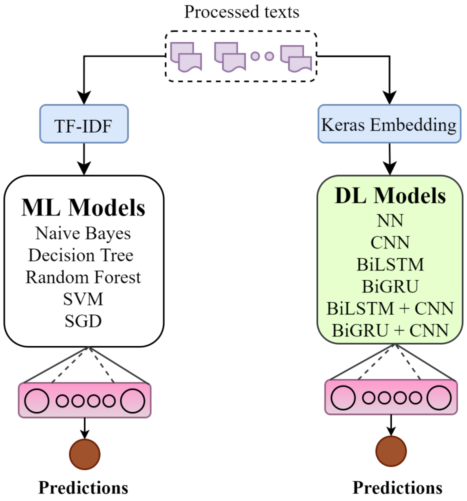

Fig. 1. Abstract process of Bengali short text categorization system

feature extraction techniques such as TF-IDF, and Word2Vec [8] to extract the relevant textual features.

4.1 ML Classifiers

Five machine learning models namely Multinomial Naive Bayes (MNB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD) have been developed by utilizing the TF-IDF features. Various parameters are empirically tried to prepare the models. In the case of MNB, the additive smoothing parameter of the NB model is set to 1 and enabled the class prior probabilities option. Both DT and RF are implemented using 100 trees and to estimate the quality of the split in the tree 'gini' criterion was used. In the SVM, 'linear' kernel with 'l2' regularization is used and $1e^{-3}$ is chosen as the 'tolerance' value for the stopping criterion. For SGD, 'log' loss is used with the 'l2' penalty and the regularization coefficient 'alpha' is settled to 0.000028.

4.2 DL Classifiers

This work employed four DL architectures such as Neural Network (NN), Convolutional Neural Network (CNN), Bidirectional Long-short Term Memory (BiLSTM), and Bidirectional Gated Recurrent Unit(BiGRU) to categorize the Bengali news headlines. Apart from these, the combination of BiLSTM/BiGRU and CNN models is also explored. The preparation of the DL models for Bengali news headlines classification is illustrated in the following:

Table 1: Hyperparameter settings of DL models

Hyperparameters	**Hyperparameters Space**	**Optimum value**
Embedding Dimension	100, 128, 64, 32	64
Padding Length	17, 21	21
Padding Type	-	Post
Filter Size	128, 64, 32, 8, 10	128, 32
Kernel Size	5, 8, 10	5
Type of Pooling	’Global Average’,’max’	’Global Average’
LSTM units	32, 64, 100	32
GRU units	32, 64, 100	64
Batch Size	16, 32, 64, 100, 128	64
Number of Epochs	15, 20, 30	15
Optimizer	‘adam’,‘sgd’, ‘RMSprop’	‘adam’
Loss Function	-	Sparse categorical crossentropy

NN: The generated embedding matrix is transformed into a one-dimensional vector passing through a flattened layer. The flattened output is transferred to a shallow Neural Network (NN) consisting of one dense layer of 30 hidden neurons.
CNN: From the embedding layer, the embedding features are propagated into a one-layer CNN architecture having a convolution layer of 128 filters with a kernel size of 5. Features are further downsampled by applying to a Global average pooling layer. The output is then transferred to a dense layer of 30 neurons.
BiLSTM/BiGRU: We have applied the two variants of recurrent neural network (RNN) architecture (i.e., BiLSTM, BiGRU) to capture the long-range dependencies by utilizing information from both past and future. BiLSTM and BiGRU consist of one layer and have 32 and 64 units, respectively. The last time-step output of these architectures is passed to a fully connected layer of 24 hidden neurons.
BiLSTM/BiGRU+CNN: In the combined method, the BiLSTM/BiGRU and CNN are sequentially added with slight modifications in their architectures. The number of units in BiLSTM (32) and BiGRU (64) kept the same as described in the previous paragraph. The CNN consists of one convolution layer of 32 filters with kernel size 5. A global average pooling layer is used to extract the suitable sentence features. The extracted features are then passed to a softmax layer for classification.

Parameters for each of the model is selected by empirically tweaking several values. A summary of the hyperparameters for DL models is presented in Table 1.

4.3 BSTxtC: Proposed Short Text Classifier

In choosing the models that will be used for the ensembling, we observe the developed models (described in Section 4.1 and 4.2) accuracy on the validation set. We found that the deep learning model’s performance varied slightly, whereas the machine learning model showed high variance. This work uses six DL models (NN, CNN, BiLSTM, BiGRU, BiLSTM + CNN, BiGRU + CNN)

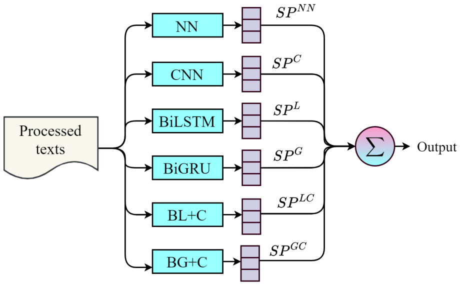

Fig. 2. Architecture of the proposed ensemble-based Bengali short text classifier. Here, BL+C, and BG+C represents the BiLSTM+CNN and BiGRU+CNN models respectively. And SP denotes the softmax probabilities

to develop the ensemble classifier (BSTxtC). This work employs an average ensemble technique which computes the average of the softmax probabilities of the participating models. Figure 2 shows the architecture of the ensemble model.

Let us consider a set of developed models, $M = \{NN, C, L, G, LC, GC\}$ where NN, C, L, G, LC, GC represents the neural network, CNN, BiLSTM, BiGRU, BiLSTM+CNN, and BiGRU+CNN respectively. We also have a 'm' number of test set instances. A model M_j classifies each instance m_i into one of the K predefined categories (i.e., International, National, Sports, Technology, Amusement, and Politics). For each m_i, a model M_j provides a softmax probability vector of size 'K', $SP^{M_j}[K]^{(i)}$. Thus models output becomes: $\left\langle SP^{NN}[]^{(1)}, SP^{NN}[]^{(2)}, ..., SP^{NN}[]^{(m)}\right\rangle$, $\left\langle SP^{C}[]^{(1)}, SP^{C}[]^{(2)}, ..., SP^{C}[]^{(m)}\right\rangle$,....., and $\left\langle SP^{GC}[]^{(1)}, SP^{GC}[]^{(2)}, ..., SP^{GC}[]^{(m)}\right\rangle$. Using these softmax probabilities, the proposed technique computes the output as described in Eq. (1).

$$AE_p = argmax\left(\frac{\forall_{i\epsilon(1,m)} \sum_{j=1}^{M} SP^{M_j}[K]^{(i)}}{M}\right) \tag{1}$$

here, AE_p denotes the vector of $m \times 1$, which contains the average ensemble method predictions.

5 Experiments

The experiment is conducted on google collaboratory. Pandas 1.1.4, python 3.6.9, scikit-learn 0.22.2 are used for processing and preparing the data and machine learning models. The Keras 2.4.0 library has been used in the backend with the tensorflow==2.3.0 framework to create the Deep learning models. Before starting the experiments, the developed corpus was split into two independent sets: train and test with a ratio of 4:1. For tweaking the DL models parameters, we chose 10% data of the training set for validation. We have applied random shuffling before splitting the dataset to remove any bias. Table 2 shows the class-wise split distribution of our dataset.

Table 2: Train and Test set distribution of BST corpus

Class Names	Train	Test
Technology	2123	536
Amusement	12522	3128
International	36491	8981
National	19532	5020
Politics	8136	2050
Sports	24481	6107
Total	**103285**	**25822**

5.1 Results

The developed models' performance on the BST corpus test set is compared based on their weighted f1-score (WF). The performance comparison of the models is reported in Table 3.

Table 3: Performance comparison on the test set of BST corpus. Here, accuracy, precision, recall, and weighted f1-scores are denoted as A, P, R, WF respectively.

Approach	**Models**	**A(%)**	**P(%)**	**R(%)**	**WF(%)**
ML	MNB	78.93	80.0	78.9	76.8
	DT	66.54	66.1	66.5	66.2
	RF	75.27	75.0	75.3	74.6
	SVM	75.37	76.8	75.4	73.9
	SGD	78.96	79.1	79.0	77.9
DL	NN	83.20	83.1	83.0	83.0
	CNN	82.73	82.9	82.7	82.7
	BiLSTM	83.07	83.0	83.1	83.1
	BiGRU	83.47	83.5	83.5	83.4
	BiLSTM+CNN	82.21	82.1	82.2	82.1
	BiGRU+CNN	82.07	81.9	82.1	81.9
	(Proposed Method)	84.4	84.4	84.5	**84.4**

The outcomes exhibit that the SGD outperformed all the ML models with a weighted f1-score of 77.9%. Amongst the ML models, the DT achieves the minimum WF of 66.2% while the MNB, RF, and SVM achieve the WF of 76.8%, 74.6% and 73.9% respectively. On the other hand, in the case of the DL models, all of them showed outstanding performance compared to the ML models. BiGRU obtained the highest WF (83.4%) and accuracy (83.47%) amongst all the DL models. The NN and BiLSTM achieve almost similar WF of approximately 83%. Likewise, CNN scores a WF score of 82.7% whereas the combined models BiGRU + CNN and BiLSTM + CNN both obtained approximately 82% weighted f1-score. The results of the combined models are less than their counteractive individual performances. However, when all the DL models (NN, CNN, BiLSTM, BiGRU, BiLSTM+CNN, BiGRU+CNN) output probabilities are averaged, the overall prediction accuracy improves by ≈> 1% and attains the highest WF of 84.4%.

5.2 Error Analysis:

It is found that the proposed model (BSTxtC) performed best in Bengali short text classification compared to the other models. The confusion matrix is used to illustrate the quantitative errors done by the proposed method which is depicted in Figure 3. It is observed that 9181 instances of the ‘International’ category 8127 are correctly identified, whereas 1152 samples are misclassified. Out of the misclassified samples, maximum of 493 data are classified as the ‘National’ category. For the ‘Technology’ category, more than 40% (237 out of 536) instances are incorrectly identified where around 200 samples are misclassified as the ‘International’ and the ‘National’ class. However, the misclassification ratio is very low in ‘Sports’(≈ 7% (misclassified 409 out of 6107)) class compared to the ‘Technology’(≈ 45%), ‘Amusement’(≈ 13% (misclassified 393 out of 3125)), and ‘Politics’(≈ 28% (misclassified 568 out of 2050)) class. Moreover, in almost every class, the proposed model mostly misclassified the actual class as ‘International‘ or ‘National’ categories.

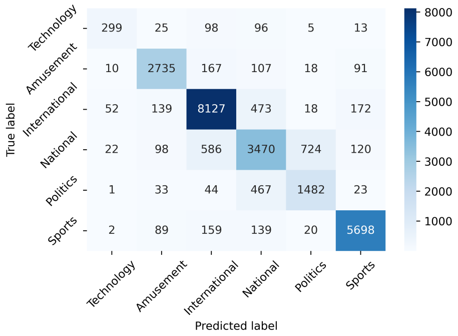

Fig. 3. Confusion matrix for proposed ensemble method

Table 4 shows some of the misclassified samples done by the proposed model.

It is noticed that due to the short text length, the model can not comprehend the actual context of the texts. For instance, in the case of the first example, the word **‘ফেসবুক’** (Facebook) is most common in the ‘Technology‘ category. As a result, the proposed model classified the text as from the ‘Technology‘ class. Similarly, in the second example, the word **‘আমাজন ’** (Amazon) frequently appears in the ‘International’ class, whereas the word **‘ডিক্যাপ্রিও’** (Dicaprio) in ‘Amusement’ class. However, due to the information scarcity, the proposed model gets confused and misclassifies the sample as text from the ‘International’ category. Analysis of the incorrect predictions revealed that it is difficult to

Table 4: Few sample headlines where the proposed ensemble model provides incorrect predictions

Headline	**Actual**	**Predicted**
'ফেসবুককে কোটি ডলার জরিমানা' (Crore dollars of fine has been imposed on Facebook)	International	Technology
'আমাজন রক্ষায় কোটি টাকা দিবেন ডিক্যাপ্রিও' (DiCaprio will donate crores to save Amazon)	Amusement	International
'রোগীর আত্মহত্যা' (Patient committed suicide)	National	International

identify the correct category of the shorter texts with familiar words frequently appearing in other categories.

5.3 Comparison with Existing Methods

To our knowledge, no benchmark dataset is publicly available on Bengali short text classification. Therefore, to justify the effectiveness of the proposed system, we have compared our approach with other quite similar tasks [1,6]. We applied existing techniques to our developed BST dataset and compared the performance. Table 5 exhibits the comparative analysis of the existing and proposed methods. The reported outcome reveals that our proposed method outperformed all the existing methods by obtaining a weighted f1-score gain of approximately 1.4%.

Table 5: Performance comparison with the existing methods of short text classification. Here, WF represents the weighted f1-score.

Techniques	WF(%) on BST corpus
Neural Network [6]	83.0
LSTM + Word Embedding [1]	83.1
Ensemble method **(Proposed)**	**84.40**

6 Conclusion

This paper presents a classification framework called BSTxtC that can classify Bengali short texts. We created a dataset of Bengali short texts (BST' to develop the system and empirically validate it. The BST dataset comprises approximately 0.13 million in news headlines comprising six categories (i.e., national, international, sports, politics, technology, and amusement) accumulated from various Bangladeshi online news portals. Various machine learning and deep learning models have been investigated on the BST corpus. Moreover, an ensemble-based classifier is proposed (BSTxtC) for the text classification after analyzing individual models' outcomes. The proposed approach can exploit the strength of the partaking models and thus lead to a boost in the overall system performance. The comparative analysis demonstrated that the BSTxtC method outdoes the machine learning, deep learning, and existing methods by obtaining the highest weighted f1-score of 84.4%. In the future, this work plan to increase the categories of the short texts and will investigate the performance by adopting the state of the art models such as attention and transformers.

References

1. Bhuiyan, M.R., Keya, M., Masum, A.K.M., Hossain, S.A., Abujar, S.: An approach for bengali news headline classification using lstm. In: Emerging Technologies in Data Mining and Information Security, pp. 299–308. Springer (2021)
2. Dhar, P., Abedin, M., et al.: Bengali news headline categorization using optimized machine learning pipeline. International Journal of Information Engineering & Electronic Business **13**(1) (2021)
3. Hossain, R., Hoque, M.M.: emantic meaning based bengali web text categorization using deep convolutional and recurrent neural networks (dcrnns). In: In: Misra R., Kesswani N., Rajarajan M., Bharadwaj V., Patel A. (eds) Internet of Things and Connected Technologies. ICIoTCT 2020. Advances in Intelligent Systems and Computing. pp. 494–505. Springer, Cham (2021)
4. Kandhro, I.A., Jumani, S.Z., Lashari, A.A., Nangraj, S.S., Lakhan, Q.A., Baig, M.T., Guriro, S.: Classification of sindhi headline news documents based on tf-idf text analysis scheme. Indian Journal of Science and Technology **12**, 33 (2019)
5. Khan, M.B.: Urdu news classification using application of machine learning algorithms on news headline. IJCSNS **21**(2), 229 (2021)
6. Khushbu, S.A., Masum, A.K.M., Abujar, S., Hossain, S.A.: Neural network based bengali news headline multi classification system: Selection of features describes comparative performance. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2020)
7. Lu, Z., Liu, W., Zhou, Y., Hu, X., Wang, B.: An effective approach for chinese news headline classification based on multi-representation mixed model with attention and ensemble learning. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 339–350. Springer (2017)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
9. Qiu, X., Gong, J., Huang, X.: Overview of the nlpcc 2017 shared task: Chinese news headline categorization. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 948–953. Springer (2017)
10. Silva, J., Coheur, L., Mendes, A.C., Wichert, A.: From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review **35**(2), 137–154 (2011)
11. Yin, Z., Tang, J., Ru, C., Luo, W., Luo, Z., Ma, X.: A semantic representation enhancement method for chinese news headline classification. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 318–328. Springer (2017)