Enhancing Bangla Document Classification Using a Hybrid Ensemble of Bangla-BERT and Bi-LSTM Models

Nafia Islam Naina[1], Khondoker Sabit Uz Zaman[1], Mahedi Hasan Emon[1], and Md. Sadekur Rahman[1*]

[1] Department of Computer Science and Engineering, Daffodil International University, Dhaka-1216, Bangladesh

{naina15-6065, zaman15-6154, emon15-6107}@s.diu.edu.bd, sadekur.cse@daffodilvarsity.edu.bd*

Abstract. With the tremendous increase of digital content in Bangla, the demand for automated document classification systems has been increasing. The complex morphology of the language, the absence of sufficient labeled documents, and the different writing styles make Bangla document classification hard to achieve. This research proposes a solution to overcome these problems by a new hybrid model that utilizes both Bi-LSTM and Bangla BERT models to complement each other in sequential dependency capturing and contextual embedding. To implement a robust Bangla document classification system, the Potrika dataset has been used in this work that consists of hundreds of thousands multi-category Bangla news articles. The data went through an extensive preprocessing pipeline which includes data cleaning, text normalization, tokenization, stop word removal, stemming, label encoding and more. From the entire dataset, 80% data was set aside for training and 20% for testing. Both models were trained individually. A soft voting ensemble was then performed on the outputs of both models to obtain the final predictions. The hybrid ensemble model is found to provide better performance than individual models with an accuracy of 97.16%, as shown through experimental results. Moreover, the proposed method outperforms all previously published Bangla document classification techniques in terms of accuracy and robustness. The results demonstrate the usefulness of ensemble methods in enhancing classification performance, which opens further research directions for Bangla NLP.

Keywords: Bangla Document Classification, Natural Language Processing, Deep Learning Models, Bangla BERT, Bi-LSTM, Hybrid Ensemble Model.

1 Introduction

Bangla is a South Asian language spoken by over 242 million people worldwide [16]. With the increasing amount of digital content in Bangla, such as news articles, blogs and social media posts the demand for automated document classification systems have grown. These systems play a crucial role in information retrieval, content management, and sentiment analysis. However, the syntax complexity of Bangla, along with the richness of the vocabulary and the size of inflection, makes

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_37

Bangla document classification a hard task by making the text representation and feature extraction more complex. Moreover, due to the limited size of well-annotated datasets, it was not possible to create a high-quality classification model. Such morphological complexity and semantic richness create challenges for categorical generalization of traditional models and indicates the need for advanced modeling approaches.

Several Bangla text classification tasks have been addressed using conventional machine learning approaches. Various algorithms have been utilized including Support Vector Machines (SVM), Random Forest, Naïve Bayes, and Logistic Regression with feature extraction methods such as TF-IDF or Count Vectorizer [5, 6, 13, 14]. These approaches are computationally inexpensive and interpretable but they do not focus on the sequential and contextual dependencies of words which are necessary to understand Bangla text, particularly when facing multi-class or multi-label setup. To alleviate these restrictions, approaches using deep learning models such as CNN, LSTM and Bi-LSTM have been used to automatically learn hierarchical representations directly from raw text. Although various embeddings have been used in the past for Bangla such as character-level embedding and pre-trained Word2Vec embeddings seems to work particularly well for handling morphological richness of Bangla as they capture subtle semantic and syntactic patterns which conventional models often miss [2, 3]. Additionally, deep learning models are capable of learning from large datasets, performing well with noisy text and automatically learning relevant features without the need for manual feature engineering, which makes them good for modern day NLP tasks in Bangla.

Hybrid and ensemble-based approaches are developed as powerful techniques to boost classification performance even further. Ensemble methods capitalize on the complementary strengths of various architectures and dilute individual model weaknesses by combining multiple models. For example, combining sequential models such as Bi-LSTM with transformer-based models like Bangla-BERT allows for long-range dependencies to be captured along with contextual semantic information. Instead of relying on the output prediction from an individual model, soft voting ensembles combine predicted probabilities from multiple models and provides more accurate and robust predictions compared to the individual models [12]. For training and evaluation of such models, there are large-scale datasets available, such as Potrika and BARD, which contain hundreds of thousands of articles from multiple categories [1, 11]. Moreover, earlier studies have shown that hybrid models achieve better performance than a single type of architecture when used on complex, real-world Bangla datasets [12].

This study develops a framework that combines sequential and contextual modeling in contrast to ensemble learning, to enhance classification performance and robustness. Our study makes the following contributions:

1. In our study, Bi-LSTM and Bangla BERT models are combined to jointly work on capturing the linguistic and the semantic feature of Bangla text.

2. We presented an ensemble method that consistently and convincingly outperforms the individual models used in this study.
3. Our proposed model has also surpassed other existing models, attaining an impressive accuracy of 97.16% on the Potrika dataset.

2 Literature Review

Due to language complexity and the lack of a standard annotated corpus of Bangla, document categorization is still challenging for Bangla as a low resource language. Earlier attempts made in Bangla document classification primarily relied on either sophisticated deep learning models utilizing word embeddings by treating each approach independently or traditional type of machine learning algorithms which were dominant at that time with statistical features such as TF-IDF. However, these techniques are not robust and are limited to a specific text type.

Several studies have utilized the classical machine learning based approaches with feature extraction methods such as TF-IDF and Word2Vec to develop Bangla news classification systems. Tanvir Alam, M., Mofijul Islam, M. [1] introduced BARD, a Bangla news article categorization corpus on the largest Bangla dataset, consisting of 376226 articles in five categories. They used Random Forest, Neural Networks, and Logistic Regression models as well as the entire pair of TF-IDF and Word2Vec for generating features. In their study, the Word2Vec based neural network outperformed previous work, demonstrating the best effectiveness with 96% f1-score. Hossain, M.R., Hoque, M.M. [2] used Word2Vec embeddings and DCNNs for building a Bengali document classification system. They created their model based on large-scale dataset of 86,199 labeled documents, on 12 news categories, yielding up to 94.96% of accuracy. Similarly, Sharmin Yeasmin, Kuri, R., Rana, A., Md. Ashraf Uddin, Sala, M., Riaz, H. [5] proposed a Bangla news classification system, utilizing a multi-layer dense neural network and several machine learning methods, employing two datasets encompassing five news categories: a small dataset (being of 1,425 items from Prothom Alo) and a large dataset (consisting of 169,791 articles from Kaggle). They assessed models such as Naïve Bayes, SVM, Logistic Regression. With an accuracy of 92.63% on the smaller dataset and 95.50% on the bigger one, the neural network model produced the best results. Islam, T., Prince, A.I., Khan, Md.M.Z., Jabiullah, Md.I., Habib, Md.T. [6] constructed a dataset of 4,545 blog posts from different Bangla blogs in eight different categories. Nine machine learning models were assessed in their study utilizing TF-IDF features from unigram, bigram, and trigram. By attaining an accuracy of 87.4%, SVM with unigram TF-IDF demonstrated the greatest efficiency. Jakaria, A.J.M., Chowdhury, R.R., Konia, J.J., Roy, D., Meem, N.T.A. [14] presented a thorough comparison of nine traditional machine learning algorithms for classifying Bangla documents. They used a vast dataset of 4,37,870 cleaned documents from 32 different categories of Bangla newspapers. SVM performed better than all other models with 92.76% accuracy in their study.

Another set of studies investigated deep learning models focusing on CNN, LSTM, and hybrid systems to utilize the morphological and semantic richness of Bangla

text. Md. Mahbubur Rahman, Sadik, R., Al Amin Biswas [3] employed CNN and LSTM models in their character-level deep learning approach for classification of Bangla documents. To effectively handle the morphological complexity of Bangla, they encoded documents at the character-level as opposed to the traditional word-level encoding. They tested their models on 3 datasets: OSBC, Prothom Alo, and BARD and received the best accuracy of 95.42% on Prothom Alo with LSTM. Rabib, M., Sarkar, S., Rahman, M. [4] used both deep learning and traditional models to classify Bengali news in 12 categories on a balanced dataset provided by SUST, with 3,000 samples per class. Among all the techniques, CNN outperformed other models with an accuracy of 93.43%. A hybrid architecture of CNN and Bi-LSTM was presented by Kabir, H., Kazi Omar Faruk, Rahman, A., Nessa, I. [8] into Bangla articles of BARD dataset, containing around 76,000 articles distributed over 12 categories. Their model outperformed the traditional machine learning baselines on the same dataset, achieving up to 83.48% accuracy and 83.39% f1-score. Equivalently, Habib, A., Akter, A. [10] suggested categorizing Bangla newspaper articles into nine categories by implementing a hybrid deep learning technique that combines CNN and LSTM. They employed a large dataset containing approximately 400,000 items that were gathered via Kaggle. By attaining 89.51% precision, 92.43% recall and 92.57% f1-score, their model outscored conventional classifiers like SVM, Naïve Bayes, and logistic regression. Salehin, K., Alam, M.K., Nabi, Md.A., Ahmed, F., Ashraf, F.B. [9] utilized a massive dataset of 75,951 samples and two deep learning models along with five machine learning classifiers to classify Bangla articles into 12 categories. To enhance deep learning models, they implemented early stopping and utilized TF-IDF for feature extraction. In their study, LSTM performed most effectively with an accuracy and f1-score of 87%.

To solve multi-label classification, Akanda, W., Uddin, A. [7] utilized an extensive dataset of 416,289 articles from Prothom Alo with over 4,300 unique labels and six categories, using count vectorizer with the ML-KNN and the neural network models. They obtained 80.31% F1-score for the Economy category. A hybrid ensemble approach was proposed by Mahmud, T.A., Sultana, S., Mondal, A. [12] which combines machine learning (SVM, LSVC, Random Forest) and deep learning (LSTM, GRU) models on 20,000 Bangla news articles among 4 categories. The ensemble model attained 95.45% accuracy and outperformed individual classifiers. Ahmad, I., AlQurashi, F., Mehmood, R. [11] created the Potrika dataset which consists of over 664,880 articles collected from eight categories over six years from 6 Bangladeshi news portals. They used various feature extraction methods including TF-IDF, Doc2Vec and FastText and compared deep learning models against the traditional methods. The top score on manually labeled data was GRU with FastText 91.83% and on automatic data was KNN with Doc2Vec 75%, confirming the higher performance of deep learning, but stressing the tradeoff between the less costly automatic labeling and the very expensive training. On contrary, Md Gulzar Hussain, Sultana, B., Rahman, M., Md Rashidul Hasan [13] analyzed a smaller dataset of 12,500 articles from 12 categories from Prothom Alo to assess traditional machine learning methods. For extracting features, they employed CountVectorizer and TF-IDF. In their study, SVM performed better than Logistic Regression, with 84% accuracy as opposed to 81%. Moreover, 15. Sarkar, S., Hasan, M.N., Karmaker, S. [15] proposed a new dataset called BanglaNewsNet, gathering 7,245

tagged news articles from Prothom Alo. They contrasted large decoder-based models like GPT, Gemini, and DeepSeek with conventional sentence encoders like LaBSE, LASER, and BanglaTransformer. Using prompting-based classification and embedding similarity, the study discovered most large language models struggled with Bangla, even though they performed well in high-resource languages. Notably, the 200B-parameter LLM Gemini-1.5-Pro performed better than other LLMs and encoder-based techniques, obtaining 0.616 f1-score.

According to the above researches, both conventional machine and deep learning methods have significantly advanced the categorization of Bangla documents. Nevertheless, most current researches only use one model or one kind of feature representation, which restricts their capacity to generalize across various datasets and categories. Furthermore, although some recent research has tried hybrid or ensemble approaches, these have frequently been limited to a small number of classifiers or lack a systematic integration of both statistical and semantic representations. By putting up a hybrid ensemble-based framework for the classification of Bangla news articles, we seek to eliminate these limitations in our study.

3 Methodology

The methodology of this study includes all important stages, such as data collection, preprocessing, model selection, and model evaluation. Fig. 1 gives an overview of the entire workflow.

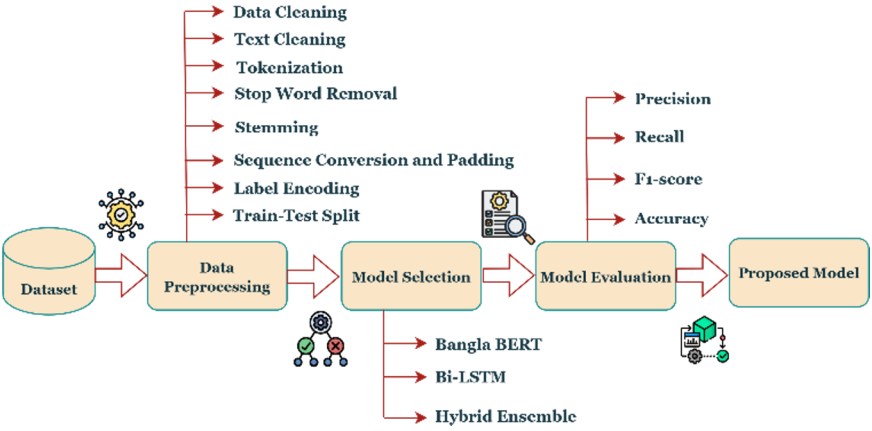

Fig. 1. Overall workflow of our study

3.1 Data Collection

We utilized the Potrika dataset, a widely used corpus of Bangla news articles for this study. The Potrika dataset contains samples of 8 different classes and consists of 329,110 rows and 2 columns, representing the text of the article and it's class, respectively. Table 1 presents the total number of articles in each category and Fig. 2 provides the dataset overview.

Table 1: Number of Articles per class.

Class	Number of Articles
National	41000
Sports	41000
International	41000
Entertainment	40772
Economy	40848
Education	40916
Politics	40179
Science & Technology	43395
Total	**329110**

	article	class
0	\n\nশ্রাবন্তীর 'শিকারি'র পর এবার আরেক ছবি 'বাদ...	Entertainment
1	টিভি অভিনেত্রী প্রত্যুষা ব্যানার্জিকে প্রায় মা...	Entertainment
2	বাস্তবের দুই বোন ডেকোটা ফ্যানিং আর এল ফ্যানিং ...	Entertainment
3	প্রবীণ চলচ্চিত্র অভিনেত্রী শবনম এখনও অভিনয় করত...	Entertainment
4	\nকরোনাভাইরাস আতঙ্কে কাঁপছে সারা বিশ্ব। গত ডিস...	Entertainment
...	...	...
329105	\n\n বিশ্বের বিভিন্ন প্রত্যন্ত অঞ্চলের মানুষের...	Science_Technology
329106	এটি মোট 142 ক্রেডিট সহ এয়ারস্পেস ইঞ্জিনিয়ারি...	Science_Technology
329107	\nঢাকা বিশ্ববিদ্যালয়, শাহাজালাল বিজ্ঞান ও প্রয...	Science_Technology
329108	\nআইফোন ৮ থেকে বাদ দেয়া হতে পারে টাচ আইডি, এমন...	Science_Technology
329109	হুয়াওয়ের লাভ ইন ফোকাস প্রতিযোগিতার বিজয়ীদের...	Science_Technology

Fig. 2. Potrika Dataset Overview

3.2 Dataset Preprocessing

After collecting the dataset, several preprocessing steps including data cleaning, text cleaning, label encoding are carried out to make the dataset clean and well-prepared for model training.

i) Data Cleaning: Rows containing missing values and duplicate articles were removed to avoid redundancy. Moreover, extremely short articles having less than 20 words were filtered out to provide sufficient context.

ii) Text Cleaning: All articles were cleaned for uniformity and noise reduction. Whitespace characters were replaced with a standard space. In this step, we compressed multiple spaces into a single space and removed all non-Bangla and non-alphanumeric characters as well as the leading and trailing spaces.

iii) Tokenization: Subword tokenization was performed using a pretrained Bangla-BERT tokenizer which separates text and gives out the token IDs while ensuring the subword meaningful units are captured. In case of Bi-LSTM, word level tokenization was done by bnltk Bangla tokenizer which splits sentence at word level to sequentially feed them for modeling.

iv) Stopword Removal and Stemming: In the preprocessing phase of Bi-LSTM, common Bangla stop words were removed to eliminate noise from the input data. The words are then processed using a Bangla stemmer, which reduces words to their root form and creates a better generalization of words. On contrary, these two steps were not performed for Bangla-BERT.

v) Sequence Conversion and Padding: In case of Bi-LSTM model, a Keras Tokenizer is utilized to transform tokenized words within integer strings and all strings were padded in regular order to obtain a similar size input. As for Bangla-BERT, input sequences were padded and truncated to an utmost range of 512 tokens.

vi) Label Encoding and Formatting: Categorical class labels were transformed into numerical values emoloying label encoding. For Bangla BERT, it has been augmented to convert the dataset into HuggingFace dataset objects and format the dataset into PyTorch tensors such as input_ids, attention_mask, and labels. Sequences and labels were transformed into numpy arrays for training Keras Model between Bi-LSTM.

viii) Data Splitting: The dataset was divided into 80% training and 20% testing sets to evaluate the model's efficacy properly. A stratified sample was used to preserve the proportion of the different classes in both sets.

3.3 Model Selection

To classify Bangla text documents with efficiency, we have utilized Bangla-BERT, Bi-LSTM and a hybrid ensemble model combining both architectures in this study.

i) Bangla-BERT: Bangla-BERT is a large scale pre-trained language model that captures deep contextual information with the help of a transformer network architecture. In contrast to the static embeddings of traditional approaches, BERT creates dynamic representations of words based on the context surrounding them, which is particularly important for understanding the nuances of Bangla sentence-level meanings. This ability to captures context and semantics makes it an ideal model for

Bangla text classification. We selected Bangla-BERT for this study as it gives strong contextual embeddings which enables a better multi-class article classification using the model.

ii) Bi-LSTM: Bi-LSTM is a recurrent neural network which reads both forward and backward sequences and captures dependencies from past and future tokens at the same time. By training bidirectionally, the model can understand the context of surrounding words better than an ordinary LSTM. As Bangla text classification models often require sequential, long-range dependencies, Bi-LSTM is an ideal approach, capturing both the context of a word as well it's semantic meaning. We selected Bi-LSTM for this study to utilize sequential data to achieve better classification performance.

iii) Hybrid Ensemble: The proposed hybrid ensemble model integrates the outputs of Bangla-BERT and Bi-LSTM to come up with the final prediction. Bangla-BERT provides richer contextual embeddings while Bi-LSTM capture sequential and syntactic information. The ensemble model combines these complementary features to enhance robustness and accuracy, and thus utilizes the strength of contextual representation along with sequential information.

3.4 Model Evaluation

This study uses several performance indicators such as precision, recall, f1-score, and accuracy to assess the overall performance of the model.

Precision indicates the number of true positive divided by the total number of instances predicted as positives.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall is the fraction of positive instances that were correctly detected by the model.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

F1-score aggregates false positives and false negatives into a single value.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Accuracy is the percentage of correctly predicted cases (both positive and negative) among all the involved cases which indicates the overall correctness of the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Where, TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negative.

4 Result Analysis

4.1 Evaluating Models

In this section, the efficiency of the models employed in this study has been analyzed and compared with several metrics. Table 2 presents the performance analysis of the classifiers and a corresponding comparison bar chart is presented in Fig. 3.

Table 2: Performance analysis of the models.

Model	Training Accuracy	Testing Accuracy
Bangla-BERT	0.9734	0.9708
Bi-LSTM	0.9506	0.9470
Hybrid Ensemble	0.9744	0.9716

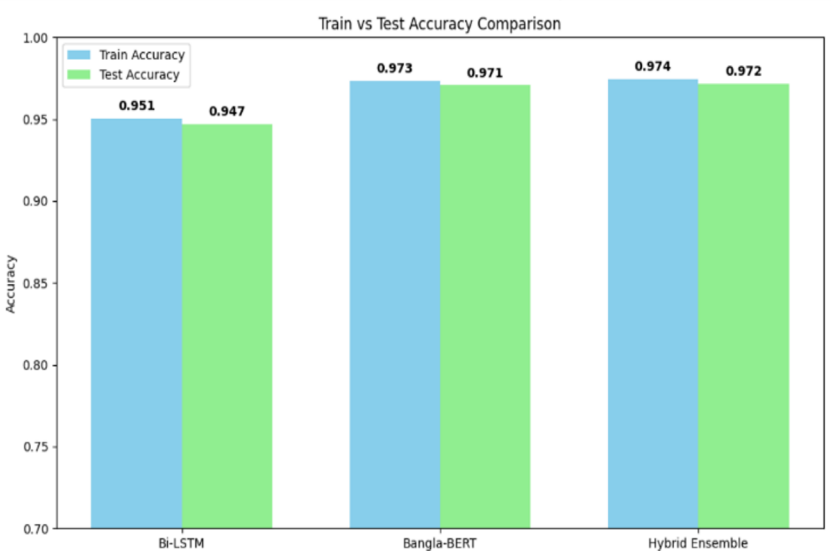

Fig. 3. Training vs Testing Accuracy Comparison

From the above results, we can see that the Hybrid Ensemble model obtained the highest performance with 97.44% training and 97.16% testing accuracy, thus outperforming Bangla BERT and Bi-LSTM. Bangla BERT also performed well with 97.08% testing accuracy, while Bi-LSTM obtained 94.70% testing accuracy. The results indicate that the hybrid ensemble model combining contextual embeddings from Bangla BERT with sequential modeling from Bi-LSTM yields a marginal yet statistically significant improvement in classification performance.

Table 3: Precision, Recall, F1-score, and Accuracy of the models.

Model	Precision	Recall	F1-Score	Accuracy
Bangla-BERT	97%	97%	97%	97.08%
Bi-LSTM	95%	95%	95%	94.70%
Hybrid Ensemble	97%	97%	97%	97.16%

The scores across all evaluation metrics of Bangla-BERT, Bi-LSTM and Hybrid Ensemble models are reported in Table 3. As shown in Table 3, Bangla-BERT attained 97.08% and Hybrid Ensemble model attained 97.16% accuracy. Both models obtained 97% score in other assessment metrics, surpassing Bi-LSTM model, which achieved 94.70% accuracy and 95% on other metrics. However, the overall best accuracy came from the Hybrid Ensemble model, which performed marginally better on testing accuracy compared to Bangla-BERT.

The confusion matrix, classification report and AUC-ROC curve of our best performing model Hybrid Ensemble are presented in Fig. 4, and Fig. 5 respectively.

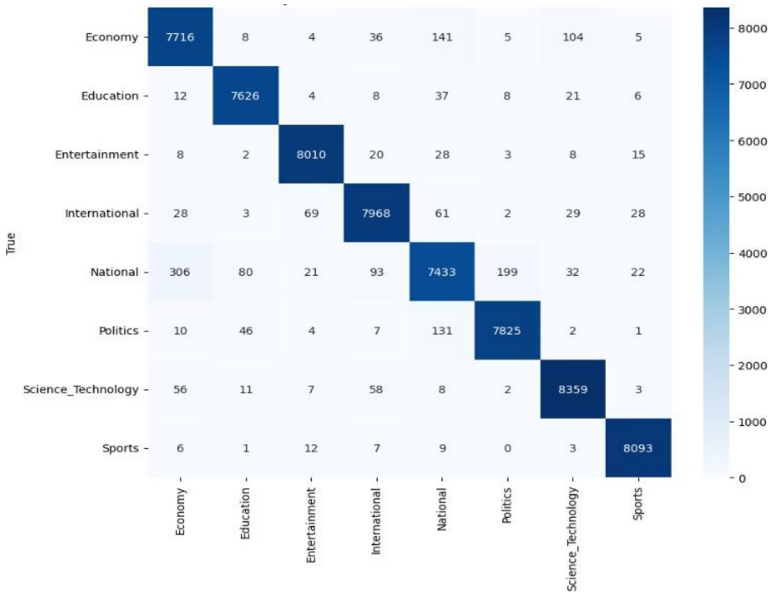

Fig. 4. Confusion Matrix of Hybrid Ensemble

The above confusion matrix highlights classification efficacy of the Hybrid Ensemble model among the eight categories. The diagonal values are relatively high in all categories which suggests effective predictive accuracy across the board. Examples include correct classification of 7,716 Economy, 7627 Education, 8010 Entertainment, 7968 International, 7433 National, 7825 Politics, 8359 Science & Technology and 8093 Sports articles. Other misclassifications are still relatively sparse, such as where an Economy article was mistakenly placed in National (141) or Science & Technology (104) and a National one categorized as Politics. However, these minor overlaps are trifling compared to the diagonal values which indicates that the model possesses extremely good classification performance.

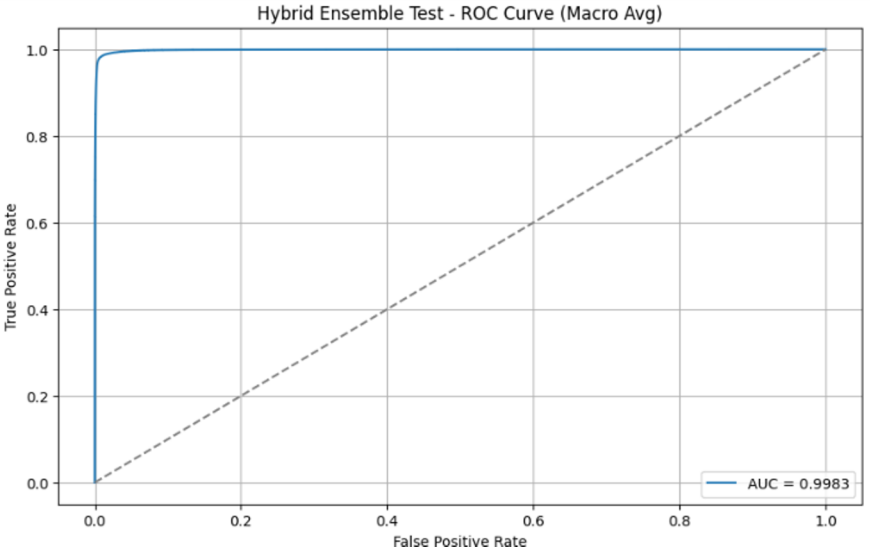

Fig. 5. AUC-ROC curve of Hybrid Ensemble

As shown in the ROC curve, the proposed Hybrid Ensemble model produces excellent classification performance. It can be noticed that the curve is almost close to the upper-left corner which means the model identifies the positive classes correctly with very minimal false positives. The AUC value of 0.9983 indicates great prediction skill of the model.

4.2 Comparison with other published models

Different kinds of algorithms have been widely used in different studies to efficiently classify Bangla documents. Several previous studies have reported competitive performance with other models. To evaluate the efficiency of our proposed solution, the performance comparison of our research work with past research work is shown in Table 4.

Table 4: Performance comparison with other published models.

Author(s)	Used Datasets	Used Algorithms	Best Algorithms (Accuracy)
Md. Mahbubur Rahman, Sadik, R., Al Amin Biswas [3]	OSBC, Prothom Alo, BARD	CNN, LSTM	LSTM (95.42% on Prothom Alo)
Rabib, M., Sarkar, S., Rahman, M. [4]	Public Bengali corpus from SUST	NB, Logistic Regression, Random Forest, Linear SVM, Bi-LSTM, CNN	CNN (93.43%)
Sharmin Yeasmin, Kuri, R., Rana, A., Md. Ashraf Uddin, Sala, M., Riaz, H. [5]	Prothom Alo, Kaggle dataset	NB, SVM, Logistic Regression, Multi-layer Dense Neural Network	Multi-layer Dense Neural Network (92.63% on Prothom Alo) 95.50% on Kaggle dataset)
Kabir, H., Kazi Omar Faruk, Rahman, A., Nessa, I., Benozir Zabin, Nahar, K., Iqbal, S., Md Sabbir Hossain, Annajiat Alim Rasel [8]	BARD	CNN-BiLSTM	CNN-BiLSTM (83.48%)
Salehin, K., Alam, M.K., Nabi, Md.A., Ahmed, F., Ashraf, F.B. [9]	Bangla news dataset	MLP, LSTM, SVM, Logistic Regression, Naïve Bayes, Random Forest	LSTM (87%)
Ahmad, I., AlQurashi, F., Mehmood, R. [11]	Potrika dataset	KNN GRU, LSTM, CNN	GRU (91.83% on manually labeled data), KNN (75% on automatically labeled data)
Md Gulzar Hussain, Sultana, B., Rahman, M., Md Rashidul Hasan [13]	Prothom Alo	SVM, Logistic Regression	SVM (84%)
Proposed Model	**Potrika dataset**	**Bangla-BERT, Bi-LSTM, Hybrid Ensemble**	**Hybrid Ensemble (97.16%)**

In Table 4, the efficacy of our suggested model is contrasted with earlier research, indicating the significance of our Hybrid Ensemble approach. Previous researches have been conducted on different datasets using different methods, resulting in a very wide performance spectrum. As shown in Table 4, Md. Mahbubur Rahman, Sadik, R., Al Amin Biswas [3] computed CNN and LSTM and reported 95.42% accuracy with LSTM on Prothom Alo article dataset. Rabib, M., Sarkar, S., Rahman, M. [4] used public Bengali corpus by SUST and got 96.61% accuracy with CNN. Using multi-layer dense neural networks, Sharmin Yeasmin, Kuri, R., Rana, A., Md. Ashraf Uddin, Sala, M., Riaz, H. [5] obtained 95.50 % accuracy on Prothom Alo. Kabir, H., Kazi Omar Faruk, Rahman, A., Nessa, I., Benozir Zabin, Nahar, K., Iqbal, S., Md Sabbir Hossain, Annajiat Alim Rasel [8] employed CNN-BiLSTM hybrid model and obtained 83.48% accuracy on BARD dataset. Using multiple ML and DL models on Bangla news dataset, a maximum of 87% was reported by Salehin, K., Alam, M.K., Nabi, Md.A., Ahmed, F., Ashraf, F.B. [9]. Ahmad, I., AlQurashi, F., Mehmood, R. [11] created the large-scale Potrika dataset and achieved 91.83% accuracy using GRU on manually labeled data. They also obtained 75% accuracy on automatically labeled data using KNN. Moreover, Md Gulzar Hussain, Sultana, B., Rahman, M., Md Rashidul Hasan [13] reached an average accuracy of 84% over a dataset of Prothom Alo articles using SVM and Logistic Regression.

Our proposed model, using BanglaBERT and Bi-LSTM in a Hybrid Ensemble, outperformed all previous work, obtaining 97.16% accuracy on the same Potrika dataset [11]. The findings indicate that the combination of pre-trained contextual embeddings with sequential models can successfully extract semantic as well as sequential features in Bangla text and provide significant improvements in classification performance over standalone deep learning or classical machine learning model.

4.3 Limitations

Although our proposed hybrid ensemble model exhibited strong efficacy, several limitations persist that may impact the generalizability and reproducibility of the results. First, the study does not include comparisons with a full set of existing Bangla text classification baselines, which makes the claim that it outperforms all previous methods less strong. Furthermore, the training configuration including hyperparameters, tuning strategy, model stability across multiple runs, and computational overhead is not thoroughly examined that complicates efforts for other researchers to replicate the precise conditions for optimal performance of the hybrid system. Moreover, separate preprocessing was done for Bi-LSTM and Bangla-BERT. Preprocessing steps such as stemming and stop word removal were performed for Bi-LSTM, but it was not used for Bangla-BERT. The evaluation depends more on accuracy, which may not fully exhibit how well each category is performing when class imbalance exists. Lastly, the lack of an extended error analysis makes it harder to find out about class-specific weaknesses or model biases.

5 Conclusion

This study proposes a wide and systematic approach of Bangla news classification by utilizing advanced deep learning techniques. Our proposed Hybrid Ensemble model that combines Bangla-BERT and Bi-LSTM, achieved the best results among all the proposed models as it is proficient to extract both contextual and sequential features of Bangla text. In contrast to the individual Bangla-BERT and Bi-LSTM models in our study, the hybrid model shows the best results, achieving the highest training and testing accuracy among the models. Moreover, it surpassed all other previously published models in terms of accuracy. The results emphasize the combination of pre-trained contextual embeddings with sequential learning as an ensemble framework that outperforms Bangla-BERT and Bi-LSTM models in terms of generalization and robustness. In addition, this study draws attention to the need for extensive data preprocessing, appropriate model selection, and ensemble approaches in delivering high-performing text classification. The proposed hybrid model is also able to outperform individual deep learning models and existing approaches by a reasonable margin which introduces new levels of standards in reliable Bangla news classification, thus creating possibilities towards a more intelligent and adaptable Bangla NLP applications.

6 Future Work

Future research can enhance this study by integrating a more robust benchmarking framework that encompasses supplementary Bangla NLP models, multilingual transformers, and standardized machine-learning baselines to facilitate statistically valid comparisons. Using hyperparameter optimization techniques like Bayesian search or population-based training can help make training more stable. It is noticed that the model's variance changes when different training seeds are used. Some preprocessing steps could change how well contextual embedding works. Future studies might examine additional pipelines specifically designed for BERT-based architectures. These could include lemmatization, normalizing at the sub-word level, or even training without preprocessing. To address the class imbalance, focal loss technique can also be used. Moreover, more advanced techniques such as stacking, meta-learners, or adapter-based hybridization can be added to the hybrid ensemble framework, that can make it work better and use less computing power.

References

1. Tanvir Alam, M., Mofijul Islam, M.: BARD: Bangla Article Classification using a New Comprehensive Dataset. 2018 International Conference on Bangla Speech and Language Processing (ICBSLP). (2018). https://doi.org/10.1109/ICBSLP.2018.8554382.

2. Hossain, M.R., Hoque, M.M.: Automatic Bengali Document Categorization Based on Deep Convolution Nets. Advances in Intelligent Systems and Computing. 513–525 (2019). https://doi.org/10.1007/978-981-13-5953-8_43.
3. Md. Mahbubur Rahman, Sadik, R., Al Amin Biswas: Bangla Document Classification using Character Level Deep Learning. 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). (2020). https://doi.org/10.1109/ismsit50672.2020.9254416.
4. Rabib, M., Sarkar, S., Rahman, M.: Different Machine Learning Based Approaches of Baseline and Deep Learning Models for Bengali News Categorization. International Journal of Computer Applications. 176, 10–16 (2020). https://doi.org/10.5120/ijca2020920107.
5. Sharmin Yeasmin, Kuri, R., Rana, A., Md. Ashraf Uddin, Sala, M., Riaz, H.: Multi-category Bangla News Classification Using Machine Learning Classifiers and Multi-layer Dense Neural Network. International Journal of Advanced Computer Science and Applications. 12, (2021). https://doi.org/10.14569/ijacsa.2021.0120588.
6. Islam, T., Prince, A.I., Khan, Md.M.Z., Jabiullah, Md.I., Habib, Md.T.: An in-depth exploration of Bangla blog post classification. Bulletin of Electrical Engineering and Informatics. 10, 742–749 (2021). https://doi.org/10.11591/eei.v10i2.2873.
7. Akanda, W., Uddin, A.: Multi-Label Bengali article classification using ML-KNN algorithm and Neural Network. 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). 466–471 (2021). https://doi.org/10.1109/icict4sd50815.2021.9396882.
8. Kabir, H., Kazi Omar Faruk, Rahman, A., Nessa, I., Benozir Zabin, Nahar, K., Iqbal, S., Md Sabbir Hossain, Annajiat Alim Rasel: Automatic Bangla Article Content Categorization Using a Hybrid Deep Learning Model. (2022). https://doi.org/10.1109/r10-htc54060.2022.9930103.
9. Salehin, K., Alam, M.K., Nabi, Md.A., Ahmed, F., Ashraf, F.B.: A Comparative Study of Different Text Classification Approaches for Bangla News Classification. 2021 24th International Conference on Computer and Information Technology (ICCIT). 1–6 (2021). https://doi.org/10.1109/iccit54785.2021.9689843.
10. Habib, A., Akter, A.: Deep Learning Bangla Text Classification Using Recurrent Neural Network. International Journal of Research in Advanced Engineering and Technology 8(1), 10–16 (2022).
11. Ahmad, I., AlQurashi, F., Mehmood, R.: Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language. arXiv:2210.10903 [cs.AI] (2022). https://doi.org/10.48550/arXiv.2210.10903.
12. Mahmud, T.A., Sultana, S., Mondal, A.: A New Technique to Classification of Bengali News Grounded on ML and DL Models. International Journal of Computer Applications. 185, 15–21 (2023). https://doi.org/10.5120/ijca2023922897.
13. Md Gulzar Hussain, Sultana, B., Rahman, M., Md Rashidul Hasan: Comparison analysis of Bangla news articles classification using support vector machine and logistic regression. Telkomnika. 21, 584–584 (2023). https://doi.org/10.12928/telkomnika.v21i3.23416.
14. Jakaria, A.J.M., Chowdhury, R.R., Konia, J.J., Roy, D., Meem, N.T.A.: A Comparative Study on different Machine Learning Approaches for Categorizing Bangla Documents. International Journal of Computer Applications. 186, 32–39 (2025). https://doi.org/10.5120/ijca2025924391.
15. Sarkar, S., Hasan, M.N., Karmaker, S.: Zero-Shot Multi-Label Classification of Bangla Documents: Large Decoders Vs. Classic Encoders. arXiv:2503.02993 [cs.CL] (2025). https://doi.org/10.48550/arXiv.2503.02993
16. Wikipedia Contributors: Bengali Language. Wikipedia (2025). https://en.wikipedia.org/wiki/Bengali_language.