A Comparative Study of Sequential Models and Transformer-Based LLMs for Bangla Suspicious Political Comment Classification

Rasel Parvez[1], Md. Ikramul Hossain[1], Sadman Sadik Khan[1*], S. M. Aminul Haque[1], Sami Ahmed[2], and Abu Hurairah Rifat[2]

[1] Daffodil International University, Dhaka, Bangladesh
[2] Independent University, Bangladesh

{parvez15-5432,ikramul15-14091} @diu.edu.bd, sadman15-13696@diu.edu.bd*, aminul.cse@daffodilvarsity.edu.bd
{samiahmed19999, abuhurairah0619} @gmail.com

Abstract. The rapid worldwide spread of political discussions in Bangla calls for the automation of detection of suspicious or harmful content. This paper considers suspicious political comments in Bangla classification with both sequential-based and transformer-based deep learning names. The dataset is the Suspicious Bangla Text Dataset, which consists of 43,389 comments labeled as suspicious or non-suspicious. The dataset has undergone rigorous preprocessing: normalization, tokenization, and sequence padding. The two Recurrent Neural Networks, LSTM and Bi-LSTM, and two transformer models, BanglaBERT and mBERT, were trained with stratified 80-20 splits and tested. Experimentally, the results show that transformer-based models, especially BanglaBERT, got better than the sequential network by getting 93% overall accuracy, followed by the balanced precision, recall, and F1-score for both classes. These recent developments show how they were able to do contextual embedding and fine-tuning in low-resource languages by setting a benchmark framework with which suspicious political content in the Bangla language can be detected for safer online discourse and future NLP research in low-resource.

Keywords: Bangla NLP, Suspicious Comment Detection, Political Text Classification, BanglaBERT, LSTM, Bi-LSTM, Transformer Models, Low-Resource Languages, Deep Learning, Contextual Embeddings.

1 Introduction

The entire area of NLP is applied to analyze and interpret digital communication because in general, a machine can efficiently understand, process, and classify a huge volume of textual data. With the increasing social events happening online all throughout the world and where many political issues are discussed hotly in Bangla, there is thus an urgent need for automated detection systems

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_32

to detect dangerous or suspicious content. Such content sometimes spreads misinformation, targeted harassment, or threatening language. Hence, identifying such comments is necessary in order to maintain safer digital spaces and informed political discourse.

Being the seventh majorly spoken language worldwide, the Bengali language poses several unique challenges to NLP tasks. This is mainly because of its morphology being rich, syntactic structures being complex, the occurrence of informal spellings, and rampant code-mixing with English. Such language intricacies would often lead to great performance degradation of traditional machine learning algorithms. And so, recent studies focused on deep learning and transformer-based methods which make use of contextual embeddings to increase the accuracy and robustness of a classifier. The Suspicious Bangla Text Dataset having 43,389 labeled samples in two classes, that is to say, suspicious and non-suspicious, is a major contribution to research on the low-resource languages. Bangladesh datasets are hence combining data from various public sources: an expert-annotated DEATH row collected through research collaboration making it one of the largest and most diverse of all of the Bangla datasets for classification of political comments. An elaborate dataset is made possible for developing high-performance models of harmful political content detection, and at the same time forms a solid basis for future research in abusive content recognition and threat analysis.

Extraction of semantic and lexical contextual information from large corpora has set a stage for pre-trained language models such as BanglaBERT to change paradigms under NLP tasks for Bangla. For this project, the BanglaBERT was fine-tuned to classify political remarks into questionable versus not questionable. By incorporating a very potent preprocessing pipeline, tokenization, and hyperparameter tuning, this work achieves good accuracy in prediction without sacrificing computing efficiency.

The main contributions of this work are: 1. We establish an application of the model in real-world settings through text classification in political cases across low-resource-language scenarios; 2. We thoroughly evaluate the model across four different diagnostic metrics: fine accuracy, precision, recall, and F-measure, ensuring that the reader fully understands its performance; and 3. We provide a few suggestions for improving the fine-tuning of the model: more advanced attention mechanisms, cross-lingual embeddings, and application-specific pretraining on political discourse, where applicable.

By addressing issues in Bangla political comment classification and linguistic and technical problems, this study sets up to better moderation agents and more accurate threat detection systems and brings more to NLP research in low-resource languages.

2 Literature Review

Research on Bangla abusive and hateful texts and suspicious text has been notorious lately, with transformers in the pretraining mode having the last say in treatment strategies. Bhattacharjee et al. introduced BanglaBERT, a monolingual BERT model pretrained on a large Bangla corpus, which got strong baselines across Bangla NLU tasks and inspired downstream fine-tuning for safety moderation and toxicity detection in Bangla social media [1]. On top of that, Jahan et al. developed BanglaHateBERT, a further-pretrained BERT model on large-scale abusive corpus data, released a balanced 15K hate-speech dataset, and reported consistent improvements over generic BanglaBERT in abusive language identification [2]. Das et al. collected one of the first fine-grained Bangla hate-speech datasets (7425 Facebook comments classified under seven different categories) and experimented with a CNN-based encoder-decoder classifier, opening the door to task-specific Bangla resources [3].

In addition to the monolingual studies, a number of works dealt with language transfer. Ranasinghe et al. evaluated multilingual transformers for—and, hence, in low-resource Indian languages—offensive language, showing that cross-lingual representations (mBERT, XLM-R) could be competitive for Bengali offensive detection when labeled data are scarce [4], and then further conducted a broader multilingual model evaluation for Indian languages with code-mixing challenges as well [5]. Gaikwad et al. explored cross-lingual offensive language detection (MOLD) and showed transferability from higher resource languages for Bengali tasks using available data sets such as TRAC for Bengali [6]. Mandl et al. gave an overview of the HASOC shared task series, which, besides spurring benchmarking activities for hate speech in multiple languages, also defined a standard evaluation protocol later reused for Bengali evaluation [7]. The work was updated by the FIRE 2022 emphasis on context-sensitive labeling as well [8].

Dataset creation continued side by side. UCI Bangla Hate Speech Detection Dataset contained an assortment of political, personal, geopolitical, religious, and gender-based abuse categories and became the most used benchmark to conduct research into Bangla toxicity [9]. The open-source projects like Bengali Hate Speech Dataset widened community access to labeled Bangla content for classification and lexical analyses [10]. The work by Das et al. further documented the design decisions of the dataset and the granularity in categories for Bangla hate speech on social platforms, thereby highlighting domain diversity and annotation problems [11].

Moderating Bangla brings its own challenges under transliteration and code-mixing scenarios. Raihan et al. have hemmed in offensive language phenomena in transliterated and code-mixed Bangla–English and have noted that a robust tokenization strategy combined with pretrained encoders can greatly increase robustness under noisy orthography [12]. More recently, using multi-label transliterated Bangla data sets beyond binary toxicity, the two datasets BanTH and

BANTH capture interlocking hate targets (e.g., gender, religion, origin) and poise strong encoder baselines and further pretrained transliterated encoders [13], [14].

Other safety issues in Bangla intersect with sentiment and aspect modeling. When performing Bengali sentiment analysis with transformer ensembles (mBERT, BanglaBERT, XLM-R), very high accuracy and F1 scores were obtained, indicating that domain-adapted encoders may transfer to toxicity identification given adequate fine-tuning [15]. Complementary work on large-scale Bangla sentiment corpora and hybrid feature learning contrasted their own approaches to preprocessing, normalization, and tokenization that also help in abusive content classification [16].

In this line of research, the detection of hate speech on the Bangla multimedia platform was investigated. Rezvan et al. put forward a multimodal Bengali meme dataset and experimented with all varieties of text–image fusion methods with BiLSTMs/ ConvNets and transformer encoders such as BanglaBERT, mBERT, and XLM-R, claiming that visual context is therefore key to disambiguating the toxicity of the text [17]. Akhter et al., meanwhile, designed a powerful hybrid ML model for detecting cyberbullying in Bangla, touching upon relevant topics such as feature engineering and class imbalance treatment important in the detection of suspicious political comments [18].

Perhaps a set of general reviews might strive to assemble methodological trends. A systematic review concerning hate-speech detection, enunciated by Jahan et al., discusses in particular deep-learning techniques and NLP pipelines, annotation schemas, imbalance handling, and explanability [19]. Modern online-hate-detection surveys emphasize certain pertinent issues concerning Bangla in the cross-lingual and culturally specific modeling sphere: issues of bias and an acute requirement of locale-specific resources [20].

3 Methodology

This procedure is an official four-step process. It begins with data collection and ends finally with the evaluation of the completed model upon completion for use. Below is a table that summarizes each step of the process and how operations will be carried out step by step. The major operations shall be described as follows in Fig. 1:

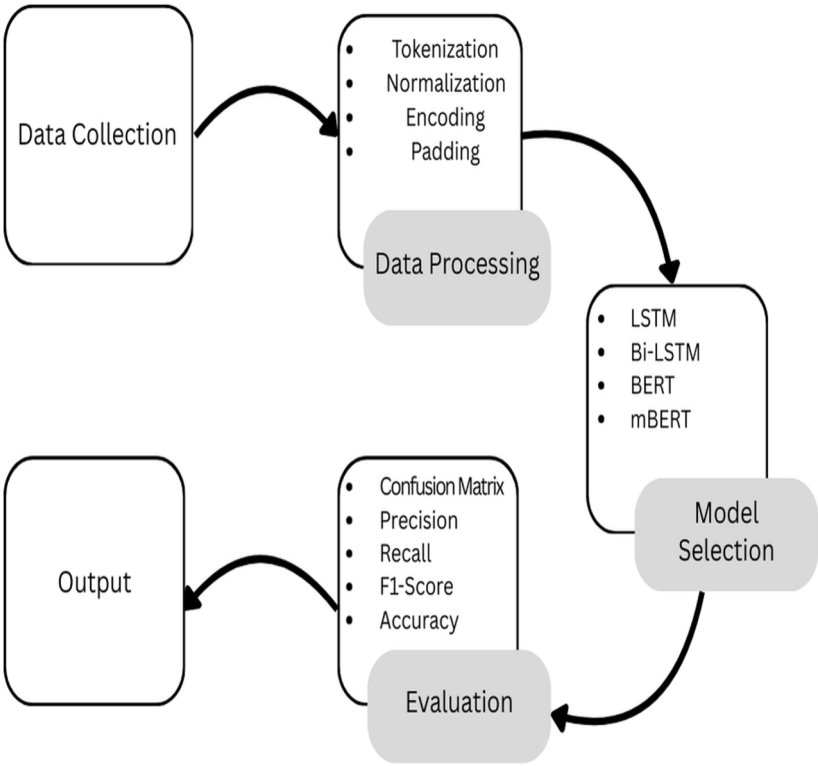

Fig. 1. Proposed Methodology

3.1 Data Collection

This study uses the Suspicious Bangla Text Dataset, which has 43,389 Bangla comments, out of which 23,874 were flagged as suspicious, and 19,515 had a non-suspicious tag. The data contained all possible source materials, social media, political discussion online forums in different regions, and comment sections of Bangla news portals in varying language styles. Apart from downloading samples from the public domain, additional manual annotation of samples was performed by native Bangla speakers to further improve quality and domain relevance. To that end, binary labeling was applied, while in cases where opinions conflicted, the majority vote held. This painstaking collection became a perfectly balanced dataset, representational of the language as it exists in political discourse, standing as an almost ideal classification case. The sample datatset and class distribution are provided in Fig. 2 and Fig. 3, respectively.

	Text	binary_label
0	১ থেকে ১০০ এর মধ্যে আপনার প্রিয় নম্বর কি? এর উ...	0
1	ওএমজি ওএমজি ওএমজি হ্যাঁ হ্যাঁ এটি ... এটি নিখু...	0
2	এই ছুটির দিনটি একটি বোর্ডে পেরেক করা একজনকে উদ...	1
3	সেই রাজাকার বাহিনী আর জঙ্গী বাহিনী বঙ্গবন্ধু ক...	1
4	হ্যাঁ আমি বালিশগুলি লাইভ জার্নাল জারক রোটেশনগু...	1
...	...	...
43384	আমি ফ্রেঞ্চ শিখতে পছন্দ করি আপনার সম্পর্কে কেম...	0
43385	আপনি প্রিয় বোন একটি অদ্ভুত। এই তাপ স্তন্যপান।	1
43386	অভিশাপ। । । এখন আমি সত্যিই প্রাতঃরাশের খাবার চাই!	0
43387	আশা আছে নেক্সট টাইমে আমার যে কোন মাগী ধরে ফেলল...	1
43388	প্রাইভেট প্লেনে স্বামী-স্ত্রী শুধু দুজন যাচ্ছে...	0

Fig. 2. Sample of Dataset

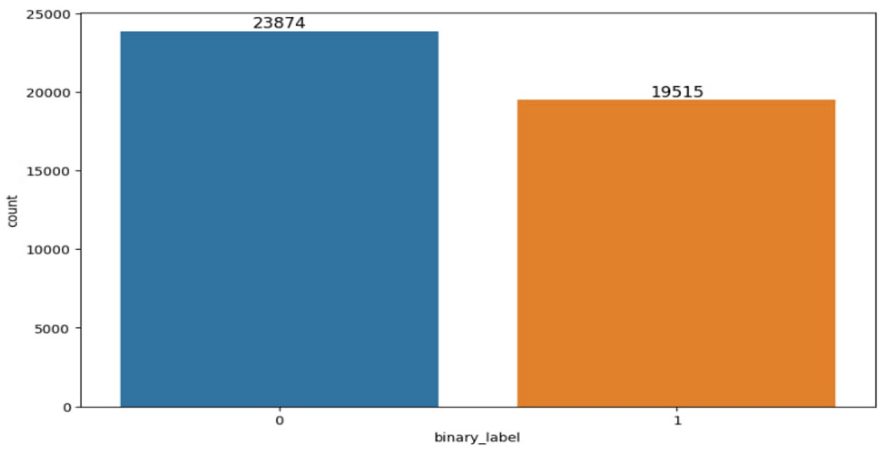

Fig. 3. Data Distribution

3.2 Dataset Preprocessing

The first point of evolution is noise removal stuff, which deals with everything from all kinds of punctuations, emojis, and special characters. Following that comes normalization to remove orthographic variations for coherence within the Bangla script. The text was then tokenized by BanglaBERT into subwords suitable for processing by the transformer-based model. Sequences were padded or truncated to a maximum length of 128 tokens such that input lengths are uniform for all of them. Non-Suspicious was encoded as 0 and Suspicious as 1. Thereafter,

an 80-20 stratified train and test split of the dataset was done without disturbing the class distribution. These flavorings gave the inputs a tabula rasa-a consistent format, with accurate labeling for a seamless train-test arrangement-and also gave the models a productivity boost during training.

3.3 Model Selection

When talking about suspicious comments, these many models are usually compared and tested. Among these were more classical recurrent models such as LSTM and Bi-LSTM, well-known for their capabilities to catch sequential dependencies in a given text. LSTM also solves the higher vanishing gradient problem, whereas Bi-LSTM views the same input from two directions and provides context.

Besides context embeddings, transformer-based models were also considered. A monolingual transformer BanglaBERT was fine-tuned to carry out a binary task after being pre-trained on a large Bangla corpus. This would guarantee that the cultural quirks of Bangla are captured in the model. But it was then decided to leverage mBERT, or Multilingual BERT, to see whether the multilingual embeddings can be transferred to the Bangla-specific task.

There was some heavy preprocessing of the text just to make it suitable for being fed into the model. All dirty things like punctuations, emojis, special characters, disturbing noises, etc., were cleaned from the text. The second step was normalization, dealing with spelling variants to keep the Bangla script uniform; then came tokenization with the BanglaBERT tokenizer that split the text into subword chunks considered suitable for transformer-based models. During this step, sequences were padded or truncated to 128 tokens to maintain uniform input length. Finally, stratified 80% train and 20% test splits were chosen for splitting up the dataset, making sure the class distributions were in both sets. Training could thus proceed faster and even more efficiently as a result of having clean standardized input from these two processing steps.

3.4 Model Training

Here, the models use LSTM and Bi-LSTM mixed with embedding and one or more hidden layers that comprehend sequential dependencies in Bangla text. One other instance might be the dropout at a probability of 0.5, given as an additional quick check to the issue of generalization. The models were trained over 10 epochs using the Adam optimizer with a 0.005 learning rate and min-batch size of 32. Also, binary cross-entropy loss was used for update optimization as this is a binary classification task.

The transformer-based models BanglaBERT and mBERT were fine-tuned on the

Suspicious Bangla Text Dataset. Input sequences were tokenized and padded to a maximum length of 128 tokens to maintain uniformity throughout the dataset. Besides, these models were trained using the same set of hyperparameter values used for training recurrent networks: batch size = 32, learning rate = 0.005, dropout of 0.5, and 10 epochs. The AdamW optimizer was applied to finely tune the model parameters during fine-tuning; the patience criterion of early stopping was employed, and the checkpoint saving strategy was used to save the best versions of the model.

3.5 Model Evaluation

Since all metrics observe the system's discriminatory power, and they exclude time consumption in training or deployment, the actual performances of BanglaBERT and mBERT are compared across several performance metrics. Summarized accuracy refers to accuracy across all samples, without sorting into registers. But in the case of the analyzed speech acts, assessment per register was very pertinent, according to the authors of. Since it is difficult to correctly diagnose Cholit versus Sadhu as very subtle levels and pronoun disambiguation are needed, results were obtained for each register separately.

Precision: Out of all the instances predicted to be positive, how many are actually true positive instances?

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

Recall: Describes the ability of a model to cover all relevant (actual positives) instances.

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

F1-Score: The harmonic mean of precision and recall; it balances false positives and false negatives.

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

Accuracy:TThe ratio of correctly classified predictions to the overall number of predictions made.

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

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

Confusion Matrix: This definitely is the table to give the number of different

classes.

ROC-AUC Curve: Its primary purpose is to evaluate the model's ability to distinguish between classes by using probability thresholds.

4 Result and Analysis

Table 1. Training and Validation Accuracy and Loss

	Training		**Validation**	
Model	**Accuracy**	**Loss**	**Accuracy**	**Loss**
LSTM	0.9937	0.0178	0.8655	0.6957
Bi-LSTM	0.9941	0.0172	0.8744	0.7217
BanglaBERT	0.9957	0.0127	0.9231	0.4497
mBERT	0.9924	0.0236	0.8956	0.5351

In Table 1, we describe the train and validation accuracies and loss for the four models-LSTM, Bi-LSTM, BanglaBERT, and mBERT. The training accuracy for all models ran really high between 99.24% and 99.57%, meaning they are pretty good at learning from the training data. On the validation performance, i.e., using unseen data for testing, the difference can be made visible. With its record-high validation accuracy of 92.31% and comparatively lesser validation loss, BanglaBERT can probably be said to be better adaptable with nuevo data. Bi-LSTM suffers high validation loss, implying some overfitting, while at the same time it does boast a good training accuracy of 99.41%. Therefore, the training and validation stages together testify to what was deduced from the table.

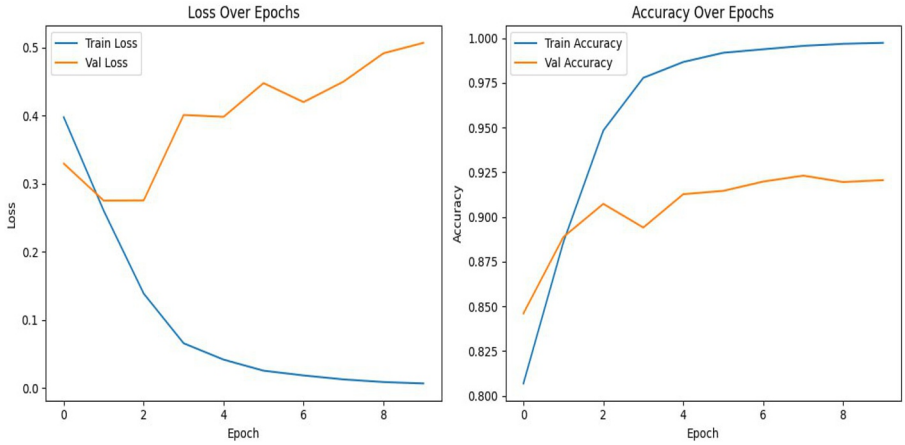

Fig. 4. Accuracy & Loss of Training & Validation for the Best performed Model(BanglaBERT)

The model shows consistent learning behavior, with training loss decreasing steadily and training accuracy approaching near-perfect levels. Meanwhile, validation accuracy remains stable around 94–95%, with validation loss showing gradual variation over (see Fig. 4). Both curves of loss and accuracy highlight smooth convergence throughout the training process. Considering these results together, the model maintains reliable performance in distinguishing between Sadhu and Cholit sentences, making it a strong candidate for Bangla sentence classification in this study.

Table 2. Comparison of Confusion Metrics for All Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	86%	0.86	0.86	0.86
Bi-LSTM	87%	0.87	0.87	0.87
BanglaBERT	92%	0.92	0.92	0.92
mBERT	90%	0.89	0.90	0.90

The Fig. 5 shows the confusion matrix for the BanglaBERT model and Table 2 presents the metric performances in terms of Accuracy, Precision, Recall, and F1-Score for the four models applied to the sentence classification task. The results reveal that BanglaBERT stands out as the best-performing model, achieving

92% accuracy with precision, recall, and F1-score all at 0.92. This indicates that the model is very precise and balanced in classifying the two test sets of Non-Suspicious and Suspicious sentences without bias.LSTM and Bi-LSTM has 86% and87% accuracy respectively, with a precision and recall of 0.86 and 0.87 , thus reliably justifying the classification. mBERT yields accuracy of 90%, BanglaBERT achieve the highest performance, each recording 92% accuracy with precision, recall, and F1-score of 0.92. It only shows how well these systems can amplify when targeting both benign and malignant sentences with equal judge capability.

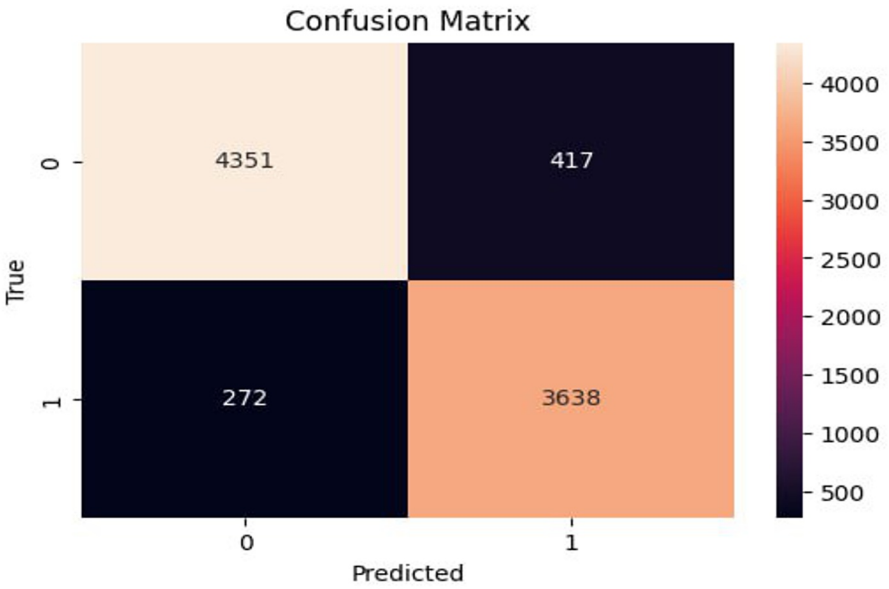

Fig. 5. Confusion matrix of the Best performed Model(BanglaBERT)

Table 3. Classification Report of BanglaBERT

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non Suspicious	0.94	0.91	0.93	4768
Suspicious	0.90	0.93	0.91	3910
Accuracy	-	-	**92**	**8678**
Macro Avg	0.92	0.92	0.92	8678
Weighted Avg	0.92	0.92	0.92	8678

Table 3 shows the class-wise evaluation metrics for the BanglaBERT model. The model performs well and consistently across the two classes. For the Non-Suspicious class, the precision is 0.94, the recall is 0.91, and the F1-score is 0.93, indicating that the model works extremely well in minimizing misclassification of non-suspicious cases. The Suspicious class, having 0.90 precision and 0.93 recall, attains an F1-score of 0.91, implying that the system performs well for detection of suspicious sentences with balanced precision and recall. The model's generalization strength is further highlighted by an overall accuracy of around 93% over 8,678 test samples. All of the macro and weighted average scores, such as precision, recall, and F1-score, were observed to have a rating of 0.92, indicating that the classifier handled the two classes in a balanced manner.

Table 4. ROC-AUC Scores of All Models Across Categories

Model	**Non-Suspicious**	**Suspicious**
LSTM	0.93	0.94
Bi-LSTM	0.97	0.95
BanglaBERT	0.98	0.96
mBERT	**0.99**	**1.00**

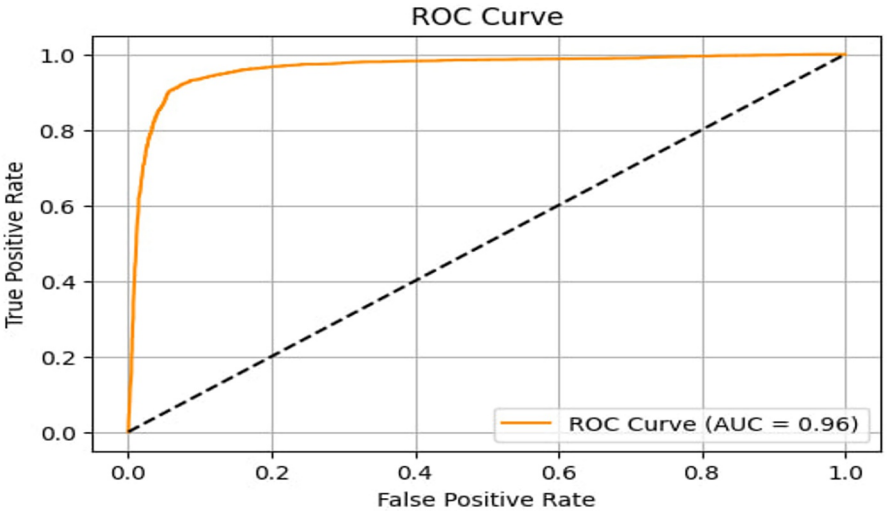

Fig. 6. ROC-AUC Curve of the Best performed Model(BanglaBERT)

Subsequently, looking at the performance measures of the models for the two categories of Non-Suspicious and Suspicious instances, it was observed that Table 4 reports the models to have consistently high AUCs for these two classes. The LSTM recorded AUC scores of 0.93 and 0.94 for the two classes, whereas the Bi-LSTM slightly improved with scores of 0.97 and 0.95. BanglaBERT further increased the scores with 0.98 and 0.96. Thing to note is that mBERT led with the highest scores of 0.99 in the Non-Suspicious category and a perfect score of 1.00 under the Suspicious category (see Fig. 6). Consequently, the model, mBERT in particular, discriminates very well between the two classes, hence giving considerable weight to the robustness and generalization capabilities of the proposed architectures.

5 Conclusion

Given the focus of this study on detecting and classifying political suspicious comments in Bangla, the authors compare sequential models (LSTM, BiLSTM) to transformer-based architectures (BanglaBERT, mBERT). Having the advantage of the high-quality Suspicious Bangla Text Dataset and a strict preprocessing pipeline involving normalization, tokenization, and sequence padding, these models accept clean, standardized input for training. From the experimental results, it has been found that transformer-powered models, in particular BanglaBERT, better capture the context and semantics of Bangla political comments than recurrent networks. Combined, metrics ranging from accuracy, precision, recall, F1-score to confusion matrices have demonstrated that these models were reliable and could distinguish suspicious comments from non-suspicious ones.Thus, the present work acts as a methodological framework and benchmark for performance-based suspicious political comment classification in Bangla and is thus a small step toward low-resource language NLP research.

References

1. A. Bhattacharjee et al., “BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Bangla NLP,” arXiv:2101.00204, 2021.
2. M. S. Jahan, M. Haque, N. Arhab, and M. Oussalah, “BanglaHateBERT: BERT for Abusive Language Detection in Bengali,” in Proc. 2nd Workshop on Resources and Techniques for User Information in Abusive Language Analysis (REST-UP), 2022.
3. A. K. Das, A. Hussain, and M. Mandal, “Bangla hate speech detection on social media using deep learning,” Journal of Intelligent Systems, vol. 31, no. 1, 2021.
4. T. Ranasinghe et al., “Multilingual Offensive Language Identification for Low-Resource Languages,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., 2021.
5. T. Ranasinghe, M. Zampieri, and A. E. Orasan, “An Evaluation of Multilingual Offensive Language Identification for Indian Languages,” Information, vol. 12, no. 8, 2021.
6. S. S. Gaikwad et al., “Cross-lingual Offensive Language Identification for Low-Resource Languages,” in RANLP, 2021.

7. T. Mandl et al., “Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Content Identification,” arXiv:2108.05927, 2021.
8. P. Chaturvedi et al., “Overview of the HASOC Subtrack at FIRE 2022,” in FIRE 2022, 2022.
9. UCI Machine Learning Repository, “Bengali Hate Speech Detection Dataset,” 2022.
10. rezacsedu, “Bengali-Hate-Speech-Dataset,” GitHub repository, accessed 2025.
11. A. K. Das et al., “Hate speech detection in Bangla social media: dataset and modeling,” Journal of Intelligent Systems, 2021.
12. M. N. Raihan et al., “Offensive Language Identification in Transliterated and Code-Mixed Bangla,” in Bangla Language Processing Workshop, 2023.
13. S. Banerjee et al., “BanTH: A Multi-label Hate Speech Detection Dataset for Transliterated Bangla,” arXiv:2410.13281, 2024.
14. F. Haider et al., “BANTH: A Multi-label Hate Speech Detection Dataset for Transliterated Bangla,” Findings of NAACL, 2025.
15. M. Z. Islam et al., “Exploring transformer models in Bengali sentiment analysis,” Journal of King Saud University – Computer and Information Sciences, 2024.
16. M. S. Hossain et al., “Sentiment analysis of Bangla language using a comprehensive dataset and hybrid learning,” Journal of King Saud University – Computer and Information Sciences, 2024.
17. R. Rezvan et al., “Multimodal Hate Speech Detection from Bengali Memes and Text,” arXiv:2204.10196, 2022.
18. A. Akhter et al., “A robust hybrid ML model for Bengali cyberbullying detection,” Journal of King Saud University – Computer and Information Sciences, 2023.
19. M. S. Jahan et al., “A systematic review of hate speech automatic detection,” Neurocomputing, 2023.
20. S. Kumar et al., “A Survey on Automatic Online Hate Speech Detection in Low-Resource Languages,” arXiv:2411.19017, 2024.