Shadhu-Cholito Detection Across Scripts: A Comprehensive Approach to Banglish and Bengali Register Classification

Rafsan Hasan Pronay [1*], Anupam Singha [1, 2*] and Kingkar Prosad Ghosh[1, 3*]

[1]Department of Computer Science and Engineering, R. P. Shaha University, Narayanganj-1400, Bangladesh.
[2]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science & Technology, Chennai,India.
[3]Department of Computer Science and Engineering, Volgograd State Technical University,Volgograd,Russia.

*Corresponding author(s). E-mail(s): rafsanhasanpronoy00@gmail.com; anupumeos@gmail.com; kingkar@rpsu.edu.bd;

Abstract

The increasing usage of Banglish-a code-mixed variety of Bangla written in Roman script-presents significant challenges for NLP. This paper presents the first cross-script framework for identifying Bengali's two main language registers, Shadhu and Cholito, across both Bangla and Banglish text. A balanced benchmark dataset is developed, and a wide range of models is evaluated, including transformer architectures, such as MuRIL, XLM-RoBERTa, mBERT, and DistilBERT, a Bi-LSTM network, as well as traditional machine learning classifiers. Experiments demonstrate that MuRIL achieves the best performance on Bangla with 95.92% accuracy and 0.9591 macro F1-score, while mBERT performs best for Banglish, yielding 85.73% accuracy and 0.8573 macro F1-score. For the combined four-class corpus, the highest overall accuracy of 90.08% and 0.9001 macro F1-score was obtained by XLM-RoBERTa. Ensemble methods, weighted soft voting, and hard voting further develop robustness on transliterated and code-mixed data, 89.87% and 89.16% accuracy, respectively. These results set a strong benchmark for cross-script Shadhu-Cholito classification and form the basis for future applications, including sentiment analysis and machine translation tasks in the low-resource and mixed-script environments.

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_34

Keywords: Banglish; Bengali language registers; Shadhu-Cholito classification; Cross-script NLP; Ensemble methods; Low-resource languages

1 Introduction

Bangla, spoken by over 250 million people, is the most popular South Asian language [1]. Classically, it was written in the Bangla script, but its use with the expansion of internet communication has grown to a gigantic level. This has spurred the creation of Banglish, a blend which writes Bangla words using the Roman script, commonly disseminated with English sentences. While Banglish has become popular on mass media and informal platforms, its irregular spelling, code-switching, and informality pose severe challenges to Natural Language Processing (NLP) [2].

Parallel to this, the Bangla language itself has two registers: Shadhu (সাধু), the standard variant used in scholastic, literary, and formal usage, and Cholito(চলিত), the colloquial form prevalent in spontaneous usage. Shadhu Bangla largely takes classical forms (e.g., 'করিয়া', 'যাইতেছি') while Cholito takes simpler forms (e.g., 'করে', 'যাচ্ছি'), especially in verb conjugation and usage of pronouns. Accurate identification of such registers is of paramount significance in applications such as sentiment analysis, translation, and text categorization. But with the advent of Banglish, this becomes a bigger problem, and transliteration dissolves the line between formal and informal registers, such that even more advanced modeling strategies become difficult to demand.

While Banglish carries linguistic and cultural significance, Bengali NLP study has largely relied on monolingual Bangla or multilingual settings with few register-based variations. Precisely, Shadhu and Cholito categorization in both Bangla and Banglish are not heavily investigated [3]. Existing approaches do not generalize across script and style, hence the need for robust approaches capable of addressing this two-language challenge.

This research addresses the problem by establishing the first benchmark framework for cross-script Shadhu–Cholito classification. A well-balanced dataset for Bangla and Banglish registers is developed and a wide range of models is comprehensively evaluated, ranging from transformer-based models (e.g., XLM-RoBERTa, mBERT, BanglaBERT, MuRIL) to BiLSTM networks and TF-IDF–based classifiers (e.g., Logistic Regression, SVM). Ensemble learning techniques such as weighted soft voting and hard voting are employed to enhance robustness.

The key contributions of this research are the following:

- **Banglish Dataset:** Fresh, tagged dataset for Shadhu–Cholito identification in Banglish, contributing to low-resource language studies.
- **Comprehensive Benchmarking:** Thorough testing of transformer-based, recurrent, and traditional machine learning models for register classification in both Bangla and Banglish.
- **Ensemble Framework:** Demonstration of ensemble learning methods that improve classification accuracy and robustness across scripts.

- **Cross-Script Evaluation:** Design of a testing platform assessing generalizability between Bangla, Banglish, and combined datasets.

To the best of current knowledge, no previous work has presented a Banglish Shadhu–Cholito dataset, cross-script evaluation across Bangla and Banglish, or a comprehensive benchmark involving classical ML, recurrent models, transformers, and ensembles for register classification. While existing works have looked only at monolingual Bengali or at sentiment analysis, this paper presents the first unified benchmark for Bangla–Banglish register detection, accompanied by a curated Banglish dataset and cross-script analysis.

2 Related Works

Research in Bengali NLP has grown rapidly in recent years, leading to advances in text classification, resource creation, and multilingual modeling. However, most prior studies remain focused on either monolingual Bengali or general multilingual corpora, with limited attention to *Banglish* and almost none to form-based distinctions such as *Shadhu* and *Cholito*. This section reviews relevant work in two main areas: Bengali text classification and code-mixed or multilingual analysis.

2.1 Bengali Text Classification and Resource Creation

Early progress in Bengali NLP has been driven by dataset building and classification models. Alam *et al.* [4] demonstrated that multilingual Transformer models could be efficiently transferred to Bengali text classification, achieving improvements in both domains. Ayman *et al.* [5] introduced *BanglaBlend*, a large-scale dataset that distinguishes between Shadhu and Cholito, which provided the first solid material for studying form-based distinctions. In following work by Ayman *et al.* Machine learning was used by [6] for the task, while Ria *et al.* [7] presented better architectures to better address low-resource tasks. Other neural approaches such as Bi-LSTM and LSTM have been employed for classifying registers, with Ayman *et al.* [8] showing their ability to learn contextual and semantic nuances. Parallel to this, Faisal et al. [9] created an emotion detection tool for Bangla and Banglish, addressing the challenge of code-mixing and language diversity. Ullah et al. [10] developed a multi-label classification model to automatically categorize the contextual themes of Bangla books. In parallel with this, Singha et al. [11] proposed an attention-based deep learning model for Bengali text summarization. These initiatives constitute a growing interest in Bangla classification tasks, but are either limited to monolingual Bangla or lacking in complete benchmarking on both scripts.

2.2 Code-Mixed and Multilingual Analysis

Due to increasing use of Banglish, different research studies have investigated code-mixed or multilingual settings. Sultana *et al.* [12] improved sentiment analysis of Bangla-English by cross-lingual word replacement and data augmentation and established the merit of the hybrid approach. Ahmed *et al.* [13] compared different

classifiers in Banglish sentiment analysis and concluded ensemble as a better methodology than sole classifiers. Similarly, Das *et al.* [14] employed CNNs to reduce colloquial Banglish and Twitter abbreviations. While these contributions are instructive regarding handling noisy and mixed-language text, their focus is less on register-sensitive classification than on sentiment analysis.

2.3 Summary

Overall, prior work has advanced Bengali NLP through datasets, deep learning, and multilingual modeling. Nonetheless, three gaps are of central concern: (i) work largely addresses sentiment or general classification, rather than register-detection specific; (ii) Banglish is underrepresented with limited datasets that reflect its transliterated and informal nature; and (iii) systematic benchmarking of Bangla and Banglish has not occurred. To address these lacunae, this paper introduces a new Banglish dataset, compares a broad variety of models, and offers the first benchmarking for cross-script *Shadhu–Cholito* classification.

3 Data Collection

The dataset utilized for this study comprises three datasets: Banglish Shadhucholito Dataset, Bangla Shadhucholito Dataset, and Merged Shadhucholito Dataset. These datasets were collected to analyze linguistic variation in Bangla and Banglish, namely Shadhu and Cholito, and their transliterated Banglish counterparts. Data collection was accomplished by sourcing, annotating, and preprocessing for a quality corpus to carry out linguistic analysis.

3.1 Data Sources

Two disparate datasets were utilized for the identification of Shadhu and Cholito registers in Bangla and Banglish in this study.

The Bengali sentences, formal Shadhu and informal CholitoBhasha, were collected from publicly available domains such as the BanglaBlend dataset [5]. The dataset provides a good foundation for the detection of formal and informal Bengali text, which is essential for proper training and evaluation of models.

The Banglish sentences used in this study were manually curated by the authors. They were produced via phonetic transliteration of Bengali sentences drawn from the Sahityapath (সাহিত্যপাঠ) textbook of the 2024–25 Higher Secondary Curriculum (11th–12th grade), specifically the Prose (গদ্য) section, which is part of the official curriculum. *Shadhu* (সাধু) examples were sourced from this textbook, whereas *Cholito* (চলিত) examples were collected from public Facebook posts and comments. The transliteration procedure preserves the linguistic properties of the original Bengali sentences while following conventions common in informal online communication. With regard to the textbook examples, this work are relying on fair use of publicly available educational material. Regarding social media data, this study only collected those posts that were publicly visible and anonymized all personally identifiable information. No

private or sensitive data were used. The data is only for research and benchmarking purposes, without any kind of commercial use; thus, this practice follows ethical standards commonly applied in NLP research.

3.2 Dataset Composition

The datasets contain sentences labeled by linguistic register and script:

- **Merged Shadhucholito Dataset**: Contains 14065 sentences, divided between Bangla and Banglish scripts. Each entry includes a sentence, label (Shadhu Bangla, Cholito Bangla, Shadhu Banglish, Cholito Banglish), script type (Bangla or Banglish), and numerical label (1 for Shadhu Bangla, 2 for Cholito Bangla, 3 for Shadhu Banglish, 4 for Cholito Banglish). Table 1 presents sample sentences from Merged Shadhucholito Dataset, illustrating the linguistic diversity and correspondence between Bangla and Banglish.

Table 1: Sample Data from Merged Shadhucholito Dataset

Sentence	Label	Script	Label
সেখানকার জানালা দিয়ে সমুদ্র দেখা যাইতেছিল	Shadhu Bangla	bangla	1
আমি গোসল করে খেতে যাব	Cholito Bangla	bangla	2
shekhankar janala diye shomudro dekha jaitechilo	Shadhu Banglish	banglish	3
ami gosol kore khete jabo	Cholito Banglish	banglish	4

- **Banglish Shadhucholito Dataset**: Contains 7010 Banglish sentences (3430 Shadhu Banglish, 3580 Cholito Banglish) for model training.
- **Bangla Shadhucholito Dataset**: Contains 7108 Bangla sentences (3475 Shadhu Bangla, 3633 Cholito Bangla).

3.3 Dataset Statistics and Visualization

Table 2 summarizes the statistics of Merged Shadhucholito Dataset.

Table 2: Statistics of Merged Shadhucholito Dataset

Category	Entries	Unique Words	Avg Length
Shadhu Bangla	3475	6047	8.75
Cholito Bangla	3633	6375	7.71
Shadhu Banglish	3430	9803	8.76
Cholito Banglish	3580	10176	7.92

Fig. 1 visualizes the sentence distribution, confirming balance across different categories.

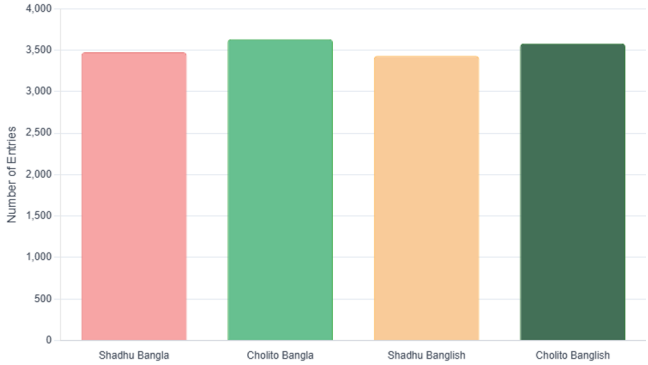

Figure 1: Distribution of Dataset Entries Across Shadhu and Cholito Categories.

Word clouds in Fig. 2 highlight frequent words in each category, generated using Python's `wordcloud` library with stop word removal and text normalization.

Figure 2: Word Cloud for ShadhuBangla/Banglish and Cholito Bangla/Banglish Sentences

4 Methodology

This work proposes a multi-model architecture for detecting *Shadhu* and *Cholito* forms in both Bangla and Banglish texts. The methodology comprises five primary stages: dataset curation, language-specific preprocessing, data stratification,

model evaluation, and ensemble-based aggregation. An overview is formalized in Algorithm 1.

4.1 Dataset Curation

Three distinct datasets are utilized:

- **Bangla Shadhucholito Dataset** ($\mathcal{D}_{\text{Bn}}$): Native Bangla script corpus labeled as *Shadhu Bangla* or *Cholito Bangla*.
- **Banglish Shadhucholito Dataset** ($\mathcal{D}_{\text{EnBn}}$): Manually constructed Roman-script transliterated Bangla, labeled as *Shadhu Banglish* or *Cholito Banglish*.
- **Merged Shadhucholito Dataset** ($\mathcal{D}_{\text{Merged}}$): Concatenation of $\mathcal{D}_{\text{Bn}}$ and $\mathcal{D}_{\text{EnBn}}$ for cross-script generalization.

Each sample is assigned one of the following labels:

$$\mathcal{Y} = \begin{array}{l}\{y_1 : \text{Sadhu-Bn},\ y_2 : \text{Common-Bn},\\ \quad y_3 : \text{Sadhu-EnBn},\ y_4 : \text{Common-EnBn}\}\end{array} \tag{1}$$

4.2 Language-Aware Preprocessing

Tailored preprocessing pipelines are developed to handle the script and linguistic nuances in Bengali and Banglish corpora.

4.2.1 Bengali Text Preprocessing

Bengali texts are processed using the following rules:

- Remove non-Bangla characters, emojis, numerals, and special symbols.
- Normalize Unicode representations using NFKC.
- Compress redundant whitespace and punctuations.

4.2.2 Banglish Preprocessing

For Roman-script Bangla, the cleaning strategy includes:

- Strip all emojis, digits, and special tokens.
- Normalize spacing and lowercase the text.

Table 3: Banglish Preprocessing Examples

Original Text	Processed Output
eiTa bhalo na!!!#$$ 123	eita bhalo na
ami kal Dhaka jacchi @5PM	ami kal dhaka jacchi
eta ekta Cholito banglish bakko!	eta ekta Cholito banglish bakko

4.3 Data Stratification

Each dataset ($\mathcal{D}_{\text{Bn}}, \mathcal{D}_{\text{EnBn}}, \mathcal{D}_{\text{Merged}}$) is partitioned via stratified sampling: Train 80%, Validation 10% and Test 10%. Stratification ensures proportional representation of all form-script labels.

4.4 Model Architecture and Justification

To capture diverse linguistic and semantic cues, a hybrid model suite is employed comprising three categories:

4.4.1 Transformer-Based Models

- **BanglaBERT**: Monolingual Bangla model fine-tuned for language-specific features [15].
- **mBERT**: Cross-lingual BERT trained on 104 languages; serves as a multilingual baseline [16].
- **MuRIL**: Designed for transliterated and code-switched Indic text; suited for Banglish [17]
- **DistilBERT**: Lightweight transformer with low latency and high inference speed[18].
- **XLM-RoBERTa**: Multilingual transformer pretrained on CommonCrawl; strong cross-lingual alignment[19].
- **ELECTRA**: Discriminator-based pretraining model with sample-efficient learning [20].
- **IndicBART**: FastText-based multilingual model optimized for low-resource Indic languages [21].

4.4.2 Recurrent Model

- **Bi-LSTM**: Sequence model that captures forward and backward dependencies in token sequences, especially effective for stylistic variations [22].

4.4.3 Shallow Lexical Models

- **TF-IDF + Logistic Regression**: Strong linear baseline on sparse word frequency vectors.
- **TF-IDF + SVM**: Margin-based model known for robust performance on short-text data.

4.5 Ensemble Aggregation

To aggregate the diverse capabilities of individual models, both probabilistic and voting-based ensembles are used.

4.5.1 Weighted Soft Voting

Predicted class probabilities $\mathbf{p}_i$ from each model are weighted and averaged:

$$\mathbf{p}_{\text{ensemble}} = \sum_{i=1}^{N} w_i \cdot \mathbf{p}_i \quad ; \quad \hat{y}_{\text{soft}} = \arg\max_c \mathbf{p}_{\text{ensemble}}[c]$$

4.5.2 Weighted Hard Voting

Predicted labels $\hat{y}_i$ are tallied using model-specific weights:

$$\hat{y}_{\text{hard}} = \arg\max_c \left(\sum_{i=1}^{N} w_i \cdot \mathbb{1}(\hat{y}_i = c) \right)$$

Here, $\mathbb{1}(\cdot)$ is the indicator function. Both strategies demonstrated superior macro-F1 and accuracy, particularly on $\mathcal{D}_{\text{Merged}}$.

Algorithm 1 Classification Workflow

Require: $\mathcal{D}_{\text{Bn}}, \mathcal{D}_{\text{EnBn}}$
Ensure: Predicted labels for $\mathcal{D}_{\text{Merged}}$
1: Preprocess each text in $\mathcal{D}_{\text{Bn}}$ and $\mathcal{D}_{\text{EnBn}}$
2: $\mathcal{D}_{\text{Merged}} \leftarrow \mathcal{D}_{\text{Bn}} \cup \mathcal{D}_{\text{EnBn}}$
3: **for** each dataset D in $\{\mathcal{D}_{\text{Bn}}, \mathcal{D}_{\text{EnBn}}, \mathcal{D}_{\text{Merged}}\}$ **do**
4: Stratify into Train (80%), Validation (10%), Test (10%)
5: **end for**
6: **for** each model M_i **do**
7: Train M_i on D_{train}, validate on D_{val}
8: Generate predictions $\hat{y}_i$ and probabilities $\mathbf{p}_i$ on D_{test}
9: **end for**
10: Combine outputs using weighted soft and hard voting
11: **return** Final ensemble prediction $\hat{y}_{\text{ensemble}}$

4.6 Experimental Setup

All the transformer models were fine-tuned with exactly the same configuration to ensure a fair comparison. The AdamW optimizer was employed with a linear learning-rate scheduler and a warmup ratio set to 0.1. Table 4 summarizes the hyperparameters used for all models.

5 Results and Discussion

This section compares the performance of transformer models and machine learning on three corpora: Bangla, Banglish, and combined Bangla and Banglish, for

Table 4: Unified training hyperparameters for all transformer models.

Hyperparameter	Value
Learning Rate	1e-4 (0.0001)
Batch Size	32
Epochs	25
Max Sequence Length	128
Optimizer	AdamW
Scheduler	Linear decay with warmup (0.1)

the classification of Shadhu and Cholito forms. Compared models are ELECTRA, XLM-RoBERTa, IndicBART, DistilBERT, MuRIL, mBERT, BanglaBERT, Bi-LSTM, Logistic Regression, SVM, and ensemble methods (Weighted average Soft Voting and Hard Voting). Performance metrics include accuracy and macro F1-score.

5.1 Performance Comparison on the Bangla Dataset

The Bangla dataset with Shadhu and Cholito Bangla texts was employed for testing model performance. The performance of all the tested models is presented in Table 5.

Table 5: Performance Metrics on the Bangla Dataset

Model	Accuracy (%)	Macro F1-Score
MuRIL	**95.92**	**0.9591**
XLM-RoBERTa	95.64	0.9563
DistilBERT	93.11	0.9309
mBERT	93.11	0.9309
Logistic Regression	87.90	0.8788
Bi-LSTM	87.76	0.8774
SVM	87.06	0.8704
IndicBART	86.36	0.8622
BanglaBERT	86.36	0.8634
ELECTRA	74.26	0.7426

MuRIL boasted the highest accuracy (95.92%) and macro F1-score (0.9591) due to its advanced multilingual training. XLM-RoBERTa (95.64%) and DistilBERT/m-BERT (93.11%) followed closely. ELECTRA registered the lowest accuracy (74.26%) and macro F1-score (0.7426).

Figure 3 displays the accuracies of evaluated models.

Figure 4 displays the confusion matrices of the ELECTRA and MuRIL models, showing their performance in predicting 'Cholito Bangla' and 'Shadhu Bangla' using different color schemes to highlight prediction accuracy.

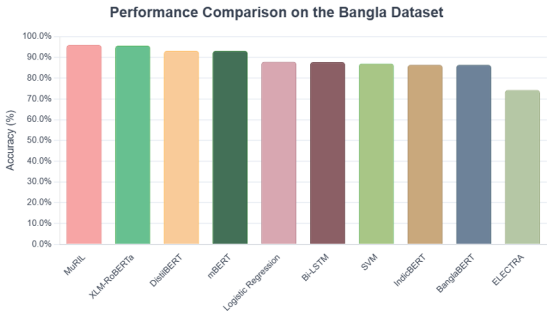

Figure 3: Comparison of model accuracies on Bangla dataset.

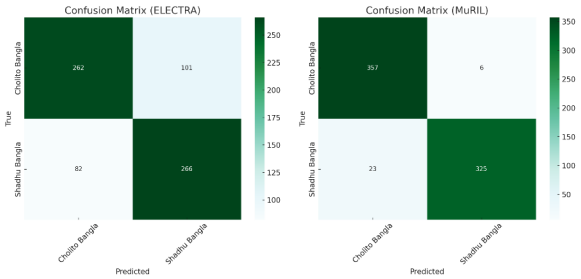

Figure 4: Confusion Matrices of ELECTRA and MuRIL

5.2 Performance Comparison on the Banglish Dataset

Banglish dataset, which includes Shadhu and Cholito Banglish text, was used to test model performance on transliterated text. Results are presented in Table 6.

Table 6: Performance Metrics on the Banglish Dataset

Model	Accuracy (%)	Macro F1-Score
mBERT	**85.73**	**0.8573**
Bi-LSTM	84.59	0.8457
MuRIL	84.17	0.8416
DistilBERT	84.17	0.8415
XLM-RoBERTa	83.88	0.8386
Logistic Regression	83.31	0.8328
SVM	82.31	0.8227
ELECTRA	81.17	0.8116
IndicBART	80.74	0.8067
BanglaBERT	55.63	0.5368

Highest accuracy (85.73%) and macro F1-score (0.8573) were achieved by mBERT, which showed better handling of transliterated features. Bi-LSTM (84.59%) and MuRIL/DistilBERT (84.17%) were at the second place. BanglaBERT recorded the least accuracy (55.63%) and macro F1-score (0.5368). Figure 5 represents model accuracies.

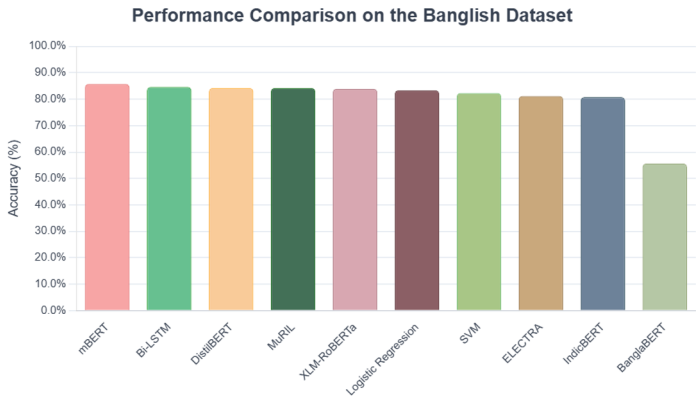

Figure 5: Comparison of model accuracies on Banglish dataset.

Figure 6 displays the confusion matrices of the mBERT and BanglaBERT models, showing each model's performance in predicting 'Cholito Banglish' and 'Shadhu Banglish' with varying levels of accuracy.

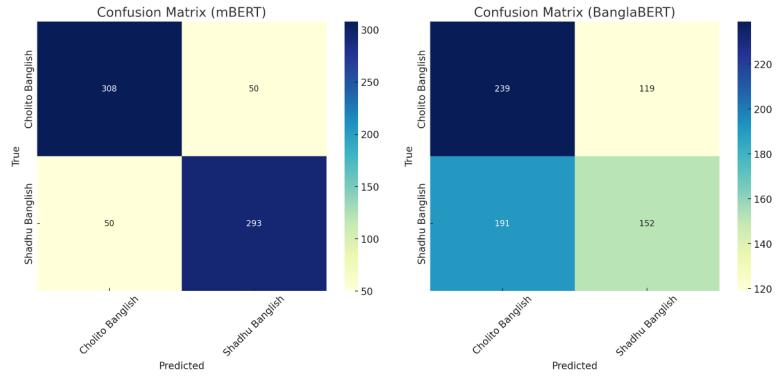

Figure 6: Confusion Matrices of mBERT and BanglaBERT

5.3 Performance Comparison on the Combined Dataset

The merged dataset of four classes (Shadhu Bangla, Cholito Bangla, ShadhuBanglish, Cholito Banglish) tested models' robustness against linguistic variations. Results are shown in Table 7. XLM-RoBERTa was best with an accuracy of 90.08% and a macro F1-score of 0.9001, benefiting from its cross-lingual ability. Weighted Soft Voting (89.87%) and Weighted Hard Voting (89.16%) followed. BanglaBERT was worst with an accuracy of 74.22% and a macro F1-score of 0.7233. Figure 7 graphs model accuracies..

Table 7: Performance Metrics on the Combined Dataset

Model	Accuracy(%)	Macro F1Score
XLM-RoBERTa	**90.08**	**0.9001**
Weighted Soft Voting	89.87	0.8983
Weighted Hard Voting	89.16	0.8911
MuRIL	89.09	0.8904
DistilBERT	88.17	0.8813
mBERT	85.76	0.8575
TF-IDF Logistic Regression	85.48	0.8546
TF-IDF SVM	84.63	0.8461
IndicBART	83.99	0.8394
Bi-LSTM	82.44	0.8239
ELECTRA	78.97	0.7897
BanglaBERT	74.22	0.7233

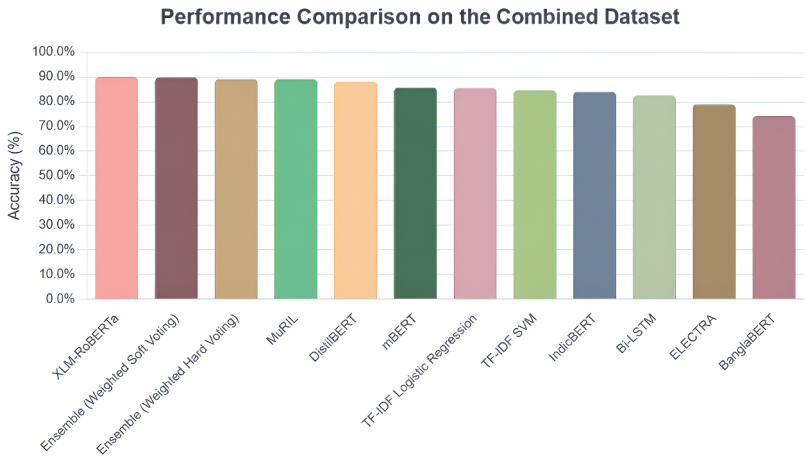

Figure 7: Comparison of model accuracies for Shadhu–Cholito classification on combined dataset.

5.4 Discussion

MuRIL (95.92% on Bangla), mBERT (85.73% on Banglish), and XLM-RoBERTa (90.08% on combined) outperformed other models due to their effectiveness in processing multilingual and transliterated text. Weighted Soft Voting (89.87%) and Weighted Hard Voting (89.16%) improved accuracy through ensemble methods. However, BanglaBERT (55.63% on Banglish) underperformed, likely due to its monolingual training, and ELECTRA (74.26% on Bangla) struggled with multi-class classification, particularly with transliterated Cholito sentences that resembled formal Bengali syntax.

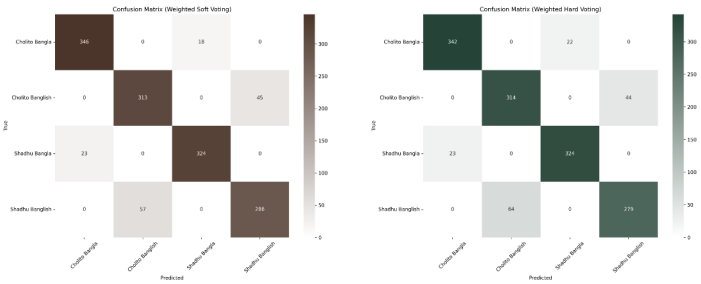

Figure 8: Confusion matrices for the Weighted Soft Voting and Hard Voting.

5.5 Error Analysis

Syntax Overlap: Simple Cholito sentences, like "ami jabo" (I will go), were misclassified as Shadhu due to their syntactic simplicity.

Transliteration Challenges: Variations in Romanized spelling caused overlap between formal and informal text, hindering accurate classification.

Code-Switching: ELECTRA faced difficulties with sentences containing both Bengali and English, leading to misclassifications.

Ensemble Limitations: While ensemble methods enhanced robustness, they did not fully resolve issues related to transliteration inconsistencies and code-switching.

Each transformer model was fine-tuned using 3 different random seeds: 13, 42, and 2025. The average results are reported as mean ± std. The variance across all models was small, std 0.4, confirming the stability of training.

6 Conclusion

This work presented an extensive analysis on *Shadhu* and *Cholito* form detection in Bangla and Banglish text with comparative analysis of several conventional machine learning, a recurrent neural network (Bi-LSTM), and transformer-based approaches. The experimental results demonstrated that the transformer models surpassed all the other baselines: MuRIL on B (95.92%), mBERT on Banglish (85.73%), and XLM-RoBERTa on the combined dataset (90.08%). Ensemble techniques yielded modest

improvements in robustness but did not outperform the top-performing standalone transformer models. The study underscores the challenge of Banglish processing, in which transliteration inconsistency and code-switching degrade the performance of monolingual models such as BanglaBERT.

There still exist certain limitations, such as this work did not perform cross-script train-test evaluation owing to resource constraints (e.g., training on Bengla and testing on Banglish), and identify this as an important future direction for analyzing script-level generalization. Given that there exist variations in typing habits between users, transliteration and spell variants are not adequately treated, and domain generalization remains a challenge. Future work will be focused on augmenting the Banglish dataset with naturally occurring social media text and expanding the proposed framework for downstream tasks such as sentiment analysis, machine translation, and conversational agents.

References

[1] Wikipedia contributors: Bengali language — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Bengali_language. Accessed: 2025-09-20 (2025)

[2] Tahereen, T.: Banglish: Code-switching and contact induced language change in a spoken variety of bangla (2016)

[3] Sen, *et al.*: Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. IEEE Access **10**, 38999–39044 (2022) https://doi.org/10.1109/ACCESS.2022.3165563

[4] Alam, T., et al.: Bangla Text Classification using Transformers (2020). https://arxiv.org/abs/2011.04446

[5] Ayman, U., et al.: Banglablend: A large-scale nobel dataset of bangla sentences categorized by saint and common form of bangla language. Data in Brief **58** https://doi.org/10.1016/j.dib.2024.111240

[6] Ayman, U., Rahim, A., *et al.*: Bengali text classification: Distinguishing saintly and common forms using machine learning model. In: 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), pp. 1–7 (2024). https://doi.org/10.1109/COMPAS60761.2024.10796448

[7] Ria, *et al.*: Toward an enhanced bengali text classification using saint and common form, pp. 1–5 (2020). https://doi.org/10.1109/ICCCNT49239.2020.9225358

[8] Ayman, U., *et al.*: Bengali text classification using bi-lstm and lstm: Differentiating between saint and common forms of bengali text, pp. 2440–2445 (2024). https://doi.org/10.1109/ICCIT64611.2024.11022005

[9] Faisal, M.R., *et al.*: Bengali banglish: A monolingual dataset for emotion

detection in linguistically diverse contexts. Data in Brief **55**, 110760 (2024) https://doi.org/10.1016/j.dib.2024.110760

[10] Ullah, M.S., *et al.*: Classifying bangla book's context: A multi-label approach. In: 2023 26th International Conference on Computer and Information Technology (ICCIT), pp. 1–5 (2023). https://doi.org/10.1109/ICCIT60459.2023.10441414

[11] Singha, A., *et al.*: Bengali text summarization with attention-based deep learning. In: 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), pp. 1–5 (2023). https://doi.org/10.1109/ASIANCON58793.2023.10270772

[12] Sultana, B., *et al.*: Enhancing bangla-english code-mixed sentiment analysis with cross-lingual word replacement and data augmentation, pp. 652–657 (2024). https://doi.org/10.1109/ICEEICT62016.2024.10534454

[13] Ahmed, *et al.*: Sentiment analysis for banglish text using machine learning approach, pp. 3378–3383 (2024). https://doi.org/10.1109/ICCIT64611.2024.11022100

[14] Das, *et al.*: A deep learning study on understanding banglish and abbreviated words used in social media. (2021). https://doi.org/10.1109/ICICCS51141.2021.9432339

[15] Bhattacharjee, A., et al.: BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla (2022). https://arxiv.org/abs/2101.00204

[16] Pires, T., et al.: How multilingual is Multilingual BERT? (2019). https://arxiv.org/abs/1906.01502

[17] Khanuja, S., et al.: MuRIL: Multilingual Representations for Indian Languages (2021). https://arxiv.org/abs/2103.10730

[18] Sanh, V., et al.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020). https://arxiv.org/abs/1910.01108

[19] Conneau, A., et al.: Unsupervised Cross-lingual Representation Learning at Scale (2020). https://arxiv.org/abs/1911.02116

[20] Clark, K., et al.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (2020). https://arxiv.org/abs/2003.10555

[21] Dabre, *et al.*: Indicbart: A pre-trained model for indic natural language generation. (2022). https://doi.org/10.18653/v1/2022.findings-acl.145 . http://dx.doi.org/10.18653/v1/2022.findings-acl.145

[22] Huang, Z., et al.: Bidirectional LSTM-CRF Models for Sequence Tagging (2015). https://arxiv.org/abs/1508.01991