From Classical to Colloquial: Leveraging LLMs for Sadhu–Cholit Register Identification in Bangla

Rasel Parvez[1], Md Anwar Hossain[2], Showrov Azam[1], AKM Bahalul Haque[3], Sadman Sadik Khan[1*], and Sadekur Rahman[1]

[1] Daffodil International University, Dhaka, Bangladesh
[2] Maharishi International University,Fairfield, Iowa,USA
[3] Abo Akademi Universiy, Finland

{parvez15-5432, azam15-5843} @diu.edu.bd, sadman15-13696@diu.edu.bd*
anwarcse7@gmail.com, akmbahalul.haque@abo.fi
sadekur.cse@daffodilvarsity.edu.bd

Abstract. Posing as a diglossic and morphologically rich language, Bangla contains two major types of registers: Sadhu Bhasha, the classical type, and Cholit Bhasha, the colloquial. Identification of the registers can be beneficial for downstream applications involving NLP such as translation, OCR, and speech synthesis. The study involved developing a dataset balanced with 7350 Sadhu and Cholit sentences. The dataset was preprocessed by tokenization, normalization, and padding, then split 80–20 for training and testing. Four deep learning models, viz. LSTM, Bi-LSTM, BanglaBERT, and mBERT, were trained in identical settings, using Adam optimizers with a batch size of 32 for 10 epochs. Experimental results suggested that while sequential models did perform reasonably well, transformer models outperformed them substantially, with BanglaBERT attaining the highest accuracy of 95%. These results become the benchmark for Sadhu-Cholit classification and stress the importance of register sensitivity in the Bangla NLP.

Keywords: Bangla NLP, Sadhu Bhasha, Cholit Bhasha, Register Classification, Text Classification, Deep Learning, LSTM, BiLSTM, BanglaBERT, Multilingual BERT (mBERT).

1 Introduction

Natural Language Processing (NLP) has become the most disruptive AI field with potential to make a computer interpret, analyze, and generate human language. Consequently, the range of issues dealt with has expanded into very complicated enterprises such as sentiment analysis, authorship detection, and stylistic classification. Such tasks become a challenge for certain morphologically rich and diglossic languages like Bangla, for which indeed two language registers exist.Being spoken by more than 230 million people, Bangla poses a special instance

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_33

of diglossia with the coexistence of Sadhu Bhasha, which is classical, literary, and Cholit Bhasha, which is, on the contrary, modern and colloquial [1].The two are not interchangeable; they differ in grammar, syntax, and vocabulary, which at times makes computational systems misinterpret or misclassify the texts.

Automatic classification of Sadhu and Cholit registers carries a high degree of importance for various NLP applications. In machine translation, systems that are primarily trained on Cholit texts tend to fail when confronted with Sadhu constructions, and the consequence is that the translations suffer semantic distortion and poor fluency [2]. On account of a document digitization process with respect to classical literature, OCR has very high error rates, untrained as it is for registers; most such texts are Sadhu, with archaic syntaxes unfamiliar to current corpora [3]. The synthetic production of speech calls for distinct register identification, given that the Sadhu pronunciation is more formalized, Sanskritized phonology, while the opposite is true for Cholit, where simplified forms prevail [4]. Yet if another application of Sadhu-Cholit classification exists, it is, of course, to detect any remaining traces of the transition of Bangla from classical into modern form computationally for linguistic and cultural studies purposes [5].

Deep learning has therefore changed the paradigm of solving the problems of Bangla NLP. LSTM, BiLSTM, and other sequential neural network learning approaches have been shown to effectively model temporal dependences and long-range contextual information present in Bangla texts [6]. These models have been used successfully for multiple applications, including tense classification, sentiment analysis, and topic classification. Simultaneously, the launch of batch transforming architecture instituted another transition phase in the existence of Bangla NLP. BanglaBERT, a monolingual Bahngla large-scale pretrained model trained on multiple Bahngla corpora, outperformed all other models on various downstream tasks, highlighting the need for standalone models for less-resourced languages [7]. In contrast, mBERT, though trained on more than one hundred different languages, has turned out to be surprisingly effective for Bangla applications, especially in contexts where large-scale pretraining data are not available [8].

Though the emergence of NLP has had more development in Bangla, there are fewer studies concerning registers. Past research has focused on sentiment analysis, authorship detection, and domain-specific classification [9], [10]; however, instances of explicit Sadhu–Cholit classification are very scarce. Very few of these studies looked into stylistic variations between formal and colloquial Bangla, but they do little to capture the actual linguistic divergence between classical Sadhu and the very common Cholit [11]. Hence this study attempts to fill this gap by building a carefully annotated dataset of Sadhu and Cholit sentences and applying recurrent and transformer-based models for classification. The comparison of LSTM, BiLSTM, BanglaBERT, and mBERT aims at pro-

viding some strong baselines for Sadhu–Cholit identification along with offering insights into linguistic markers that weigh the classification results.

2 Literature Review

Bangla NLP research has vastly expanded in the past decade, prominent works being in text classification, style identification, or deep learning architectures. The advent of massive-scale pretrained models severely transformed Bangla NLP; however, Bhattacharjee et al. did come up with BanglaBERT-as-a-monolingual transformer model-that almost consistently performing better than multilingual ones for classification, sequence labeling, and sentiment analysis tasks [7]. The BUET NLP group further encouraged research in Bangla NLP by releasing open-source resources, going as far as including checkpoints of BanglaBERT and BanglaT5 for reproducibility.

Stylistic or register variation is a second key area of research. There are only a handful of researchers who actually looked into classifying formal and colloquial Bangla; transformer architectures have been used in the utmost precision for detecting stylistic variations [11]. This suggests that models can indeed learn register distinctions; hence their work directly applies to Sadhu-Cholit classification. Parallel to that, linguistics-based works like that of Sultana have classified colloquial Bangla verb morphology in-depth by describing distinctions in the usage of suffixes, tense marking, and syntactic freedom that can be exploited by computational models [9].

Nakib et al. prepared Sadhu-Cholit parallel corpora and systems for translating classical texts into modern ones [10]. The work provides, of course, datasets but also lists some lexical and syntactic markers that generally differentiate the two registers. Following on similar lines, Chatterjee's historical linguistic inquiry traces the slow ascendancy of Cholit over Sadhu in modern Bangla and enumerates the socio-political and cultural forces that had effect to bring about such a change [12]. Having this kind of knowledge makes it very easy to realize exactly why computational treatment of register classification is a very contemporary and relevant endeavor.

The information gained from several language registers can be used in the study of this area. BanglaT5 and BanglaNLG, working on Bangla text generation, illustrated that encoder–decoder pretrained models would learn syntactic cues when being fine-tuned [13]. Register-aware preprocessing proved to boost translation quality significantly in the work of Khan et al. [14] The recently built large-scale Bangla language models such as TituLLMs made apparent the problems involved in training on a mixed-style corpus, especially in the case of underrepresentation of Sadhu, thereby warranting the carving out of a dedicated classification effort [15].

Cross-lingual style classification for English and Hindi offers concessions, as Li et al. demonstrated how formal and informal registers can be detected on the basis of lexical, syntactic, or semantic features from methods that may be applicable to Bangla [16]. Register detection is further emphasized through practical applications. Rahman et al. have developed OCR post-processing systems, wherein register-sensitive models were able to reduce error rates in the digitization of archival texts [17]. Morphological analyzers and POS taggers developed for Bangla can be used to identify archaic affixes and sentence structures common in Sadhu [18].

One can argue that attention-based processes and transformers for their contextual understanding are prevailing in research-oriented Bangla news classification tasks. This argument is supported by our evaluation that compares four architectures-LSTM, Bi-LSTM, LSTM+Attention, and Bi-LSTM+Attention-on a balanced news dataset.

Another important one for classification is neural architectures. Considered the baseline for Sadhu–Cholit, due to the strong results obtained for Bangla sentiment analysis using BiLSTM with attention [19]. Multilingual BERT by Devlin et al. provided further proof of cross-lingual transfer learning lending support to low-resource languages such as Bangla [8]. Other efforts with LSTM and BiLSTM architectures were undertaken for Bangla newspaper headline classification and document categorization, further attesting to the near-ubiquitous efficacy of sequential-based models when it comes to individual classification problems [20], [21]. Emon et al. followed up on abusive content detection to assert that deep learning was a valid tool for modeling domain-sensitive registers [22] while Dhar et al. prove Bangla medical text classification with domain-specific features to enhance performance [23].

Computational studies such as Syntactic verification carried out by Khan et al. [24] using n-gram models and Word prefix classification carried out by Islam et al. [25] establish some value for feature-level modeling in Bangla. While not focused explicitly on register, these studies provide linguistic feature extraction insights that can be translated into Sadhu-Cholit classification.

3 Methodology

The methodology employed in this study is a systematic, multi-phase approach that starts with data acquisition and ends with a comprehensive evaluation of the model. Each stage of the process, along with the steps to be performed, is duly pictorially represented in the table below. The major phases are delineated as follows in Fig. 1.

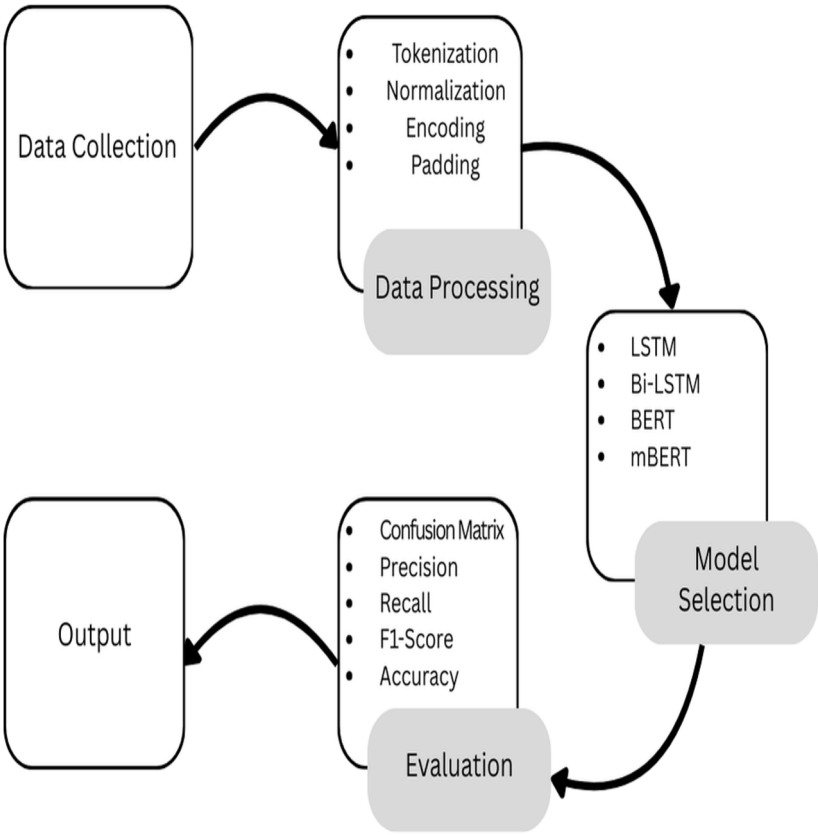

Fig. 1. Proposed Methodology

3.1 Data Collection

To perform the study, 7,350 Bangla sentences are provided with an equal split in two registers of Sadhu Bhasha (classical) and Cholit Bhasha (colloquial). Each class has 3,675 sentences, thus creating a balanced dataset for possible use in supervised classification. The data have been collected from multiple authentic sources, from classical literature, digitized archives, newspapers to contemporary online texts, covering both historical and modern registers of Bangla. For the purpose of replication and openness, the dataset was acquired from Mendeley Data[26]. The dataset is arranged into four columns: the original Bangla sentence, the English translation of the label, the Bangla label, and the class, wherein class values correspond to either Sadhu or Cholit. The sample dataset is presented in Fig. 2.

	Sentence	Labels(English)	Labels(Bangla)	Class
0	সেখানকার জানালা দিয়ে সমুদ্র দেখা যাইতেছিল	Saint	Sadhu(সাধু)	0
1	আমি কিছু দেখিতে পারিতেছি না	Saint	Sadhu(সাধু)	0
2	সকলেরই অনাবৃত দেহ সকলের সেই অনাবৃত বক্ষে আরশির...	Saint	Sadhu(সাধু)	0
3	মেয়েটি সেদিন ভিক্ষুককে সাহায্য করিয়াছিল	Saint	Sadhu(সাধু)	0
4	তুমি প্রশংসা কর না কর বৃদ্ধ বসিয়া তোমায় পুরা...	Saint	Sadhu(সাধু)	0
...	...	...	...	...
7345	রন্ধনপ্রণালী দোপেয়াজা জনপ্রিয়	Common	Cholito(চলিত)	1
7346	শেষে রহিম করিমকে বিপদে ফেলল	Common	Cholito(চলিত)	1
7347	হাফিজ কে তারাই বিপদে ফেলল	Common	Cholito(চলিত)	1

Fig. 2. Sample of the Dataset

3.2 Dataset Preprocessing

In the process before model training, this dataset was preprocessed to ensure that it was consistent and compatible with deep learning architectures. First, tokenization was performed to segment the sentences into tokens suitable for the respective models. The normalization switches then followed to either avoid orthographic variation, unify the text encoding, or discard inconsistencies. Padding was used to maintain fixed input lengths for sequences shorter than required so that batch processing could be performed during training. An 80-20 split was done for training and testing with 5,880 sentences for training and 1,470 for testing. This split was done to ensure sample data for training and a fair evaluation, as seen in Fig. 2.

3.3 Model Selection

Four deep learning models were trained and compared against one another for the Sadhu–Cholito classification task: LSTM, Bi-LSTM, BanglaBERT, and mBERT. The rationale behind this selection was that models from both ends of the spectrum, i.e., sequential neural architectures and transformer-based pretrained models, should exist so that one can fairly compare the traditional sequence model with the contemporary language models for Bangla NLP. The first one was a unidirectional Long Short-Term Memory (LSTM) network. This architecture would parse the input sequences from left to right and thus capture long-distance dependencies across Bangla sentences. The very sequential nature of LSTM renders it a valuable tool for sentence-level classification where contextual flow is of utmost importance.

The other model-as it were-would be the Bidirectional LSTM, in which forward

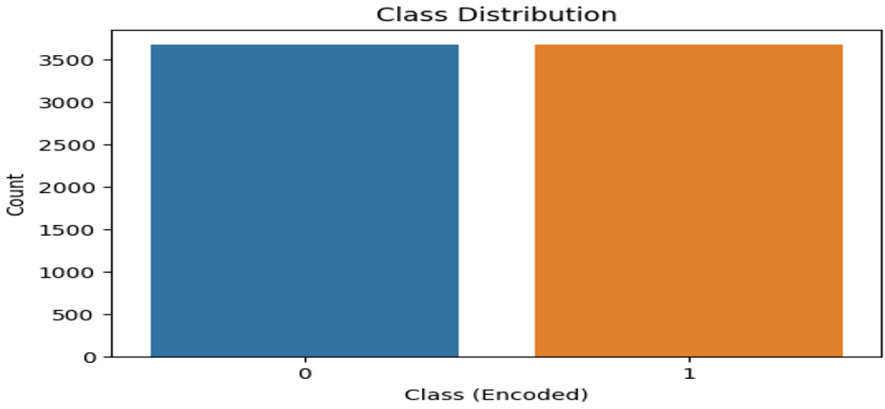

Fig. 3. Data Distribution

and backward sequences are processed in two parallel layers. In fact, these types of patterns become skilful to be differentiated into Sadhu and Cholit register: they are syntactic/semantic patterns observed by Bi-LSTMs under modeling of context from past and future simultaneously.

Our next model considered for training was BanglaBERT, a transformer model trained on a large corpus of Bangla. The fine-tuning of the Sadhu-Cholit dataset was done, considering an Adam optimizer with a batch size of 32 for 10 epochs. Since BanglaBERT has monolingual pretraining, it memorizes register-dependent linguistics cues such as morphology and lexical variation, which is utmost worthy from the perspective of the task.

Finally, mBERT was fine-tuned using the same setup as BanglaBERT. Although it was not trained only on Bangla, mBERT enjoyed the benefits of cross-lingual transfer, and the shared subword embeddings were utilized by several other languages as well. Adding this setting let us assess how well a general-purpose multilingual model would fare against a domain-specific monolingual one for regist.

3.4 Model Training

The optimizer was Adam for training the four model architectures: LSTM, BiLSTM, BanglaBERT, and mBERT, all of which were the same. All the models were trained in minibatches of size 32 for a maximum of 10 epochs. Other parameters such as the learning rate and dropout were adjusted for each model to achieve convergence within the stipulated number of epochs. Apart from customized adjustments, each model was trained with exactly the same conditions so that they could be compared for performance fairly. Training had to be per-

formed on a GPU-enabled server given the computations involved with recurrent and transformer-based models.

3.5 Model Evaluation

The evaluators had to give the one unifying evaluative panoptic view about model performance. Hence, the usual metrics for classification were calculated, namely accuracy, precision, recall, and F1-score, given that each somehow complements the others and either serves in correctness in general or in class accuracy. Confusion matrices were also computed so that one could analyze the misclassifications occurring between the Sadhu and Cholit registers. This evaluation setting thus made it possible to directly compare the sequential models (LSTM, BiLSTM) with transformer-based models (BanglaBERT, mBERT) under exactly the same conditions.

Precision: How many of the predicted positive instances are in fact true positive instances.

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

Recall: How capable the model is of capturing all relevant (actual positive) 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: The overall proportion of instances that are correctly classified out of the given predictions.

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

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

Confusion Matrix: This is a summary table showing correct and incorrect predictions, for every class.

4 Result and Analysis

Table 1. Training and Validation Accuracy and Loss

	Training		Validation	
Model	**Accuracy**	**Loss**	**Accuracy**	**Loss**
LSTM	0.9770	0.0799	0.9167	0.2764
Bi-LSTM	0.9947	0.0218	0.9133	0.3420
BanglaBERT	0.9757	0.0604	0.9510	0.1609
mBERT	0.9753	0.0645	0.9320	0.2099

From Table. 1, one can observe the training and validation performances with respect to accuracy and loss for the four models: LSTM, Bi-LSTM, BanglaBERT, and mBERT. All models achieved high training accuracy, ranging from 97.33% to 99.47%, reflecting their strong ability to learn from the training data. However, the real distinction is visible in validation performance, which reveals the generalizability of the models. Among them, BanglaBERT demonstrates the best validation accuracy (95.10%) with a relatively low validation loss, indicating better adaptability to unseen data. On the other hand, Bi-LSTM, despite very high training accuracy (99.47%), suffers from higher validation loss, suggesting overfitting. The training and validation curves shown in Fig. 4 further provide visual evidence supporting the findings presented in the table.

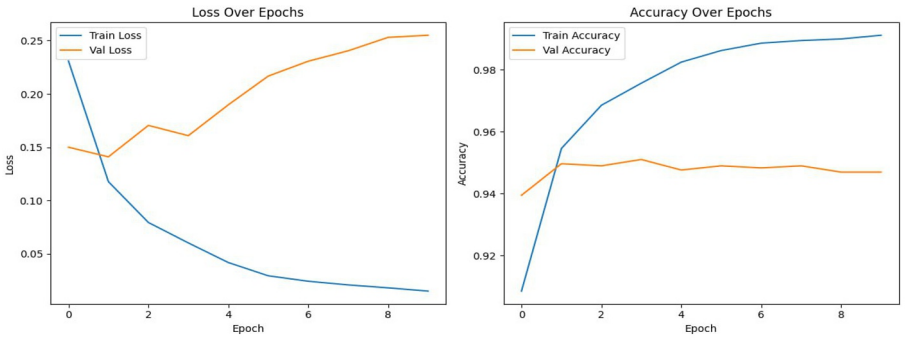

Fig. 4. Accuracy & Loss of Training & Validation for the Best performed Model(BanglaBERT)

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 epochs. 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	92%	0.92	0.92	0.92
Bi-LSTM	93%	0.93	0.93	0.93
BanglaBERT	95%	0.94	0.95	0.95
mBERT	93%	0.93	0.93	0.93

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 95% accuracy with precision, recall, and F1-score all at 0.95. This indicates that the model is very precise and balanced in classifying the two test sets of Sadhu and Cholit sentences without bias. Bi-LSTM and mBERT stand tied at 93% accuracy, balanced with a precision and recall of 0.93, thus reliably justifying the classification. The baseline LSTM yields accuracy of 92%, which, albeit slightly lower, still maintains consistent balance with all metrics. In all, this demonstrates consistent improvements in the performances of models from the traditional architectures toward transformer-based ones, while BanglaBERT proved to be the best for Bangla sentence classification within this work.

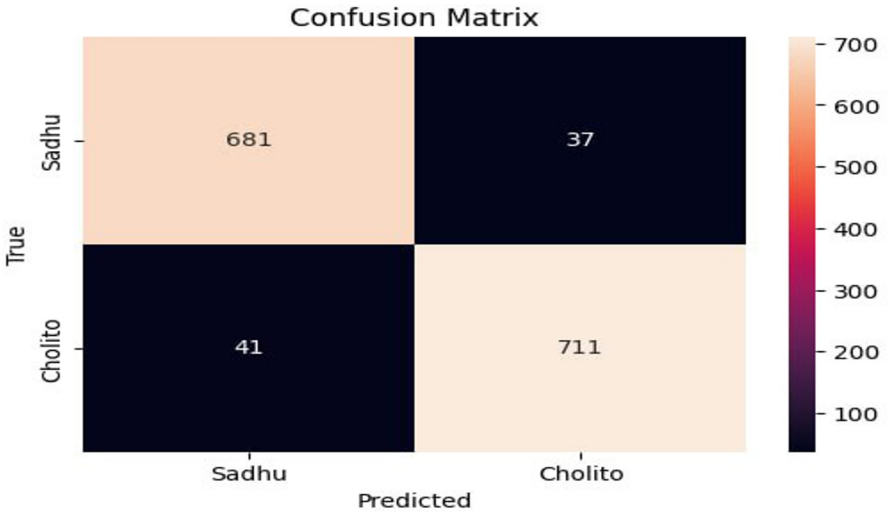

Fig. 5. Confusion matrix of the Best performed Model(BanglaBERT)

Table 3. Classification Report of BanglaBERT

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Sadhu	0.94	0.95	0.95	718
Cholit	0.95	0.95	0.95	752
Accuracy	-	-	**0.95**	**1470**
Macro Avg	0.95	0.95	0.95	1470
Weighted Avg	0.95	0.95	0.95	1470

The Fig. 5 shows the confusion matrix of BanglaBERT and Table 3 illustrates the class-wise evaluation metrics, namely precision, recall, F1-score, and support, for the final built model. The model exhibits strong and consistent performance across both classes. For the Sadhu class, precision is 0.94, recall is 0.95, and the F1-score is 0.95, reflecting that the model is highly effective in distinguishing Sadhu text forms with minimal misclassification. Similarly, the Cholit class achieves a precision, recall, and F1-score of 0.95 each, indicating equally reliable performance in identifying Cholit text forms.The overall accuracy reaches 95% over 1,470 test samples, demonstrating the ability of the model to generalize well beyond training data. Both macro average and weighted average metrics stand at 0.95 for precision, recall, and F1-score, confirming balanced behavior of the classifier. The closeness of these average values highlights that neither class

dominates the predictions, and the system performs equitably across all classes. Theoretically, the model is thus robust and well-suited for practical applications involving Bangla text classification between Sadhu and Cholit forms.

5 Conclusion

The Sadhu-Cholit categorization is considered to be an extremely abstract classification, which, in Bangladesh or in Bengal, with its diglossia, has as one of the primary computational barriers that stand in the way of style classification. We set up a balanced dataset with 7,350 sentences and set strong baselines for automatic identification of registers by evaluating recurrent and transformer-based architectures. The evaluation results stated that sequential modeling architectures of LSTMs and BiLSTMs could make use of contextual dependency information, whereas transformer-based architectures, powered by large scale pretrained Bangla corpora, could do even better, with BanglaBERT topping them all. BanglaBERT could deliver the highest accuracy of 95% ever recorded among all models tested and this strongly confirms the importance of domain specific pretrained models for register sensitive tasks.However, there are implications that go far further than classification accuracy in the present study.Automatic Sadhu-Cholit Identification engenders improvements in target applications, namely machine translation, OCR Post-Processing of classical texts, and speech synthesis, which depend on register-sensitive pronunciation and vocabulary.

Future research can build on this basis, augmenting the dataset with spoken transcripts and social media content to capture a broader range of register variation. Further studies can continue with hybrid approaches involving the attention-based recurrent model and transformer architectures; or larger pretrained multilingual and monolingual models fine-tuned for register awareness. All of these methods stand to bring about improvements in accuracy and robustness in classification, leading to higher Bangla NLP applications and more linguistic insights.

References

1. A. H. Uddin, D. Bapery, and M. Arif, “Depression Analysis of Bangla Social Media Data using Gated Recurrent Neural Network,” ICASERT, 2019.
2. M. S. Rahman, S. A. Hossain, and M. J. Islam, “Improving Bangla Machine Translation by Handling Register-Specific Variations,” ICCCNT, 2021.
3. I. A. Azhar, S. Ahmed, M. S. Islam, and A. Khatun, “Identifying Author in Bengali Literature by Bi-LSTM with Attention Mechanism,” ICCIT, 2021.
4. UM. K. Hasan, S. A. Islam, M. S. Ejaz, M. M. Alam, N. Mahmud, and T. A. Rafin, “Classifying Bengali Newspaper Headlines with Advanced Deep Learning Models: LSTM, Bi-LSTM, and Bi-GRU Approaches,” AJRCS, vol. 16, no. 4, pp. 372–388, 2023.

5. Md. O. Faruqe, M. Jahan, A. Faisal, M. S. Islam, and R. Khan, "Bangla Hate Speech Detection Using Transformer-Based NLP and Deep Learning Techniques," ASIANCON, 2023.
6. A. Bhattacharjee, A. Rahman, M. S. Hossain, and A. K. Das, "BanglaBERT: Language Model Pretraining and Benchmarks," arXiv preprint arXiv:2201.02328, 2022.
7. BUET NLP Group, "BanglaBERT and BanglaT5 Model Releases," 2022.
8. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, pp. 4171–4186, 2019.
9. A. Sultana, "Verb Morphology in Colloquial Bangla," Journal of Linguistics, 2016.
10. N. Nakib, S. H. R. Tanim, and M. H. Tania, "Interpretation of Sadhu into Cholit Bhasha," Bangla NLP Workshop, 2020.
11. S. Sakib, M. A. Rahman, and F. Chowdhury, "Classifying Formal and Colloquial Bangla with Transformers," arXiv preprint arXiv:2303.11021, 2023.
12. K. Chatterjee, "Historical Evolution of Sadhu and Cholit Forms of Bangla," Bangla Linguistics Review, 2015.
13. A. Bhattacharjee, T. Hasan, and M. S. Rahman, "BanglaT5 and BanglaNLG Models for Text Generation," arXiv preprint arXiv:2210.09125, 2022.
14. N. Khan, M. F. Khan, M. M. Islam, and B. Sarke, "Improving Bengali-English Machine Translation with Register-Aware Preprocessing," GlobalNLP, 2021.
15. TituLLMs Consortium, "Training Large Bangla LLMs: Challenges and Opportunities," arXiv preprint arXiv:2402.01102, 2024.
16. Y. Li, A. Mukherjee, and P. Rosso, "Style Classification in English and Hindi Texts: A Comparative Study," ACL, 2020.
17. S. Rahman, A. Banerjee, and T. Mittra, "OCR Post-processing for Bangla Archives: A Neural Approach," IEEE Access, vol. 7, pp. 10293–10302, 2019.
18. S. Ahmed, R. Zaman, and A. Khatun, "Bangla Morphological Analysis and POS Tagging," COLING, pp. 235–245, 2018.
19. M. Hasan, S. Paul, and N. Ahmed, "Performance Analysis of LSTM and Bi-LSTM with Attention in Bangla Sentiment Analysis," ICCCNT, 2022
20. M. K. Hasan, T. A. Rafin, and N. Mahmud, "Deep Learning Approaches for Bengali Headline Classification," AJRCS, 2023.
21. M. Rahman, R. Sadik, and A. A. Biswas, "Bangla Document Classification using Character-Level Deep Learning," ISMSIT, 2020.
22. E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra, "A Deep Learning Approach to Detect Abusive Bengali Text," ICCIT, pp. 1–6, 2019.
23. A. Dhar, N. S. Dash, and K. Roy, "Categorization of Bangla Medical Text Documents Based on Hybrid Features," CCIS, pp. 181–192, 2019.
24. N. H. Khan, M. F. Khan, M. M. Islam, M. H. Rahman, and B. Sarke, "Verification of Bangla Sentence Structure using N-Gram Models," Global Journal of Computer Science and Technology, vol. 14, no. 3, pp. 1–7, 2014.
25. K. M. S. Islam, S. A. Khushbu, F. Yesmin, and M. Masum, "Bengali Words Classification by Prefix Using Machine Learning Classifiers," ICCCNT, pp. 1–5, 2021.
26. U. A. Ayman, C. Saha, and Z. Mawa, "BanglaBlend: A Large-Scale Novel Dataset of Bangla Sentences Categorized by Sadhu and Cholit Forms of Bengali Language," Mendeley Data, 2024. [Online]. Available: https://doi.org/10.17632/7rx9mk8v4m