Bangla Dialect Classification and Standardization Using Traditional and Transformer-Based Approaches on a Custom Multi-Regional Corpus

By
Md Ibrahim Kholil
213-15-4439

Md Shamim Talukder
213-15-4447

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Ms Faiza Feroz
Lecturer
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by

Mr. Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

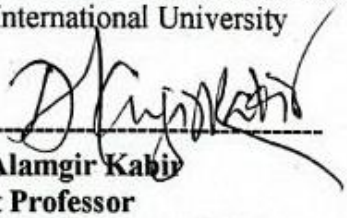
September 16, 2025

APPROVAL

This Project titled **"Bangla Dialect Classification and Standardization Using Traditional and Transformer-Based Approaches on a Custom Multi-Regional Corpus"** submitted by **Md Ibrahim Kholil and Md Shamim Talukder** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September, 2025**.

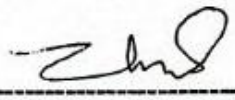
BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Chairman**
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md Alamgir Kabir — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Assaduzzaman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms Faiza Feroz, Lecturer,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Faiza Feroz 16.9.25

Ms Faiza Feroz
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Mr. Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Ibrahim Kholil

Md. Ibrahim Kholil
Student ID: 213-15-4439
Department of Computer Science and Engineering
Daffodil International University

Shamim

Md Shamim Talukder
Student ID: 213-15-4447
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Ms Faiza Feroz, Lecturer**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Bangla Dialect Classification and Standardization Using Traditional and Transformer-Based Approaches on a Custom Multi-Regional Corpus to** carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Bangla language is one of the most spoken languages in this world but one of the low resource languages in Natural Language Processing (NLP). The challenge is complicated by the existence of several regional dialects like Sylhet, Chittagong, Barishal, Noakhali and Khulna which are quite different from Standard Bangla. This thesis completes the dialect classification and conversion of dialects to standard Bangla dialects using a customized multi-regional corpus, utilizing traditional machine learning models and transformer-based models. A corpus was constructed involving 23,440 dialect and standard sentence pairs from 5 major dialects. Following processes like cleaning, normalization, and dataset splitting, the corpus was used for model training using traditional machine learning models, that is SVM, NB, LR, RF, and advance transformer architectures, that is BanglaBERT, mBERT, MuRIL and XLM-R for classification, and LSTM baseline, BanglaT5, mBART-50, mT5 for standardization. Evaluation used a large variety of metrics: Accuracy, Precision, recall, F1-score for classification, and BLEU, ROUGE-L, METEOR, chrF, TER, and Exact match for standardization. While SVM showed the best accuracy of 81.1%, MuRIL and XLM-R achieved up to 92.4% with macro-F1 of more than 0.92. For indication of the standardization, the mBART50 achieved BLEU = 0.78, ROUGE-L = 0.89, METEOR = 0.87, and Exact Match = 65.6%. A user-friendly Gradio interface has also been created to make the system accessible to any users. This study add a new dialectal corpus, a large study on traditional and transformer models, and build an NLP tool like other models. The result shows us that advanced transformer-based model is appropriate for dialect diversity of bangla and it can help us to create a way for a standardized digital communication in Bangla.

Keywords: NLP, Bangla-Dialect, Classification, Standardization, Transformer model, low resource language, Custom Corpus, MuRIL, mBERT50.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents an overview of the research, including the motivation, objective, and scope of the research. It gives an overview of the context of diversity of Bangla dialects, describes the challenges of natural language processing (NLP) for low-resource languages, and the importance of addressing these challenges. The chapter also defines the gap in the research literature, the statement of the problem and the purpose/contributions of the student's work.

1.1 Introduction

Language is the most crucial medium of human expression and conveys culture, history and social identity in every utterance. Bangla (Bangla) is the seventh most spoken language in the world and it is the official language of Bangladesh while being one of the recognized languages in India (particularly the state of West Bengal and the northeastern regions of India). With more than 230 million native speakers around the world, Bangla is a language of enormous cultural richness and significance. However, rather than Standard Bangla (usually denoted as Cholit Bangla) as the lingua franca of education, administration, and the media, daily communication throughout Bangladesh and among speakers of Bangla continues to be marked by the strong presence of the regional dialects. These dialects are distributed among various districts and divisions of the country and show strong differences in vocabulary, phonology, morphology and syntactic structure. Consequently, when speakers of dialectical divisions communicate, there is usually heavy cognitive processing or confusion, and most importantly, state-of-the-art NLP systems trained only on Standard Bangla often fail to interpret or process these dialectical texts properly [2], [9], [28].

Dialectical variation is not a marginal phenomenon, rather, dialectal variation is pervasive evidence of the Bangla language. For instance, Sylheti, spoken extensively in the Sylhet division and in diaspora communities in the UK and USA, is so different from the Standard Bangla in both phonology and lexicon as to be considered by many as a separate language. It uses different phonemes, uses special morphological endings, uses lexical items, therefore, people who know only general Bangla cannot understand it [16]. Likewise, the southern eastern dialect used on the coastal plains (the Chittagonian dialect) includes other important and unique grammatical marks, phonetic simplifications, and loanwords from Arabic and Persian, related to the maritime history of the region [18]. An area of interest is the Barisal and Noakhali dialects, which are spoken in the southern deltaic regions: based on the investigation, the vocabulary of both dialects is very closely related, although they are characterized by different tones and lexical items, thus differentiating them as much as possible from each other. Even in these dialects, a few municipalities such as Khulna, differ in vocabulary and pronounce a few words radically than Standard Bangla. These examples represent only a small selection of the rich diversity of dialectal Bangla as a mosaic reflecting identity, tradition and social belonging [24]. From the dialectic perspective, this dialectic diversity is cultural richness. But from a computational point of view it is very difficult. Most of the computational models, machine learning algorithms and transformer-based architectures, designed for

Bangla have been designed using Standard Bangla dataset so far. Furthermore, when we present dialectical input these models perform significantly worse. For example, studies on sentiment analysis [6], [7], [8] and cyberbullying detection [3], [4] have consistently reported that dialectal expressions, alternative spellings, and colloquial expressions degrade the classification performance. Similar issues are observed in document categorization tasks where the accuracies of very deep convolutional neural network (VDCNN) architectures trained on standard Bangla corpora have reached over 96% [9] while showing that the same models lose their performance when faced with dialectal content. The absence of dialect-aware NLP resources therefore has become a major barrier to growth of Bangla language technologies [28], [29].

In our age of digital communication these lacks of dialectic is especially undesirable. Social media, instant messaging and online discussion forums have become the most popular venues for Bangla language. However, users very seldom write in formal Standard Bangla in such contexts; rather, they use hybrid spellings, non-standard grammar, and local dialects. For instance, obscene or abusive language in Chittagonian dialect has necessitated dedicated recognition systems [19], whereas work on violence-promoting text in Bangla [10] finds that much of the toxic online speech is produced from dialects or nomenclatural variations of the language. Similarly, cyberbullying literature [3], [4] has highlighted the need for dialect-sensitive models in moderating online platforms appropriately. If a text is abusive, misleading, or socially harmful but not going through a dialect-aware NLP system, serious holes will be left in moderation technologies.

Figure 1.1: Geographical Distribution of Major Dialect of Bangla in Bangladesh

Recently, work has started to emerge in this regard to try to fill the gaps. Several systems have been developed for dialect-to-standard conversion and classification of dialects, but most of them are limited in scale. For example, the neural machine translation (NMT) from Sylheti to Standard Bangla has demonstrated the ability of even the deep learning models to convert dialectal input to standard form with moderate accuracy [16]. More recently, the Chittagong-based dataset ChatgaiyyaAlap released over 4,000 sentence pairs aligned between Chittagonian and Standard Bangla [18], which provides an important resource for studying regional dialects. In addition, the multilinguality potential of the dialect translation studies have been carried to the ONUBAD dataset that holds Chittagonian, Sylheti and Barisal dialect parallel corpora mapped to Standard Bangla and English respectively [28]. Some large-scale projects involved the multilingual benchmark dataset from Vashantor for dialect translation [17], and Bangla Blend's study of stylistic differences between Sadhu (classical) and Cholit (colloquial) dialects [24]. Together with these works, these materials represent an important step forward, but they are still limited for the time being (for example, they concentrate on one or two dialects at a time or they deal with variation only in style, not dialect).

A lack of rich multi-dialect resources has a direct impact on downstream NLP applications. Take machine translation systems, for example: English-to-Bangla or Bangla-to-English translation systems should not be able to accurately translate dialectal input if dialect normalization is not implemented. Likewise, speech-to-text and text-to-speech technologies need inputs to be normalized to ensure phonetic stability [22]. Sentiment analysis models trained for hundreds of thousands of words on corpora of gadget reviews [5] or BanglaDSA [6] perform pretty well on controlled datasets, but the performance suffers on real data where dialectal forms make up most of the text in real-world natural language. This paper shows that without strong dialectal preprocessing, most applications (in education, digital media or governance) are limited [12].

To fill this gap, we present a custom-constructed multi-regional corpus consisting of 23,440 annotated sentences from 5 major dialect areas, namely, Sylhet, Chittagong, Barisal, Khulna, and Noakhali. Each sentence in the corpus has been tagged with its dialect and equivalent Standard Bangla counterpart. This dataset not only exceeds the largest existing datasets like ChatgaiyyaAlap [18], ONUBAD [23], and BanglaBlend [24] in terms of scale; it also surpasses them in terms of diversity, which makes it the most comprehensive dialectal NLP resource for Bangla to date. Unlike previous work that focuses only on individual dialects or stylistic differences, our dataset is explicitly designed for two important tasks, dialect classification and dialect standardization.

The potential gains of such a system are potentially not just for academia. In education, dialect-to-standard switching is seen as a way of making the transition of students from rural or dialect-speaking homes into the Standard Bangla curriculum [15]. Within the subject of online safety moderation tools are capable of identifying abusive contents in dialectal forms [19]. At the level of speech technology, dialect normalization has been one example of speech normalization, which allows for a more natural interaction with text-to speech synthesizers and automatic speech recognizers [22]. Furthermore, social inclusion may be enhanced if dialect speakers feel that they are included in digital platforms no differently than everyone else, and do not feel left off or misunderstood.

In summary, this research is motivated by awareness of the fact that the dialectal richness of Bangla, while precious as the cultural inheritance, is a big headache for modern NLP. Results: By constructing a multiregional large-scale dataset and comparison between the traditional and transformer-based models regarding

classification and standardization, our work takes one step towards bridging the gap between the diversity of language and digital inclusivity. Its work is at the intersection of computational linguistics, cultural exemplar preservation and technology entrepreneurship. With its methodological and application focus, this thesis paves the way for further research in dialect-aware Bangla NLP and satisfies a crucial need in finding the resources and models representing the reality of language faced by millions of Bangla speakers.

1.2 Motivation

We are driven here by the intersection of language diversity and limitation of available technologies with the need for social being. As one of the most popular languages in the world, Bangla has an astonishing number of regional dialects which encapsulate culture, identity and history of millions of people. However, in the digital environment these dialects tend to be out of sight. Because Standard Bangla is used in administration, education and media, the speakers of the Sylheti, Chittagonian, Barisal, Khulna and Noakhali dialects have difficulties in communication with technology.Current models of natural language processing, trained almost entirely on standardized corpora, fail to get the juicy words they're using. As a result, a large community of dialect speakers is left behind when it comes to the benefits of a modern computational system; this represents a digital divide and a reflection of social inequality. This exclusion not only is unexclusive but could potentially lead to the erasure of the diversified rich dialects as speakers gradually give up their naturally occurring forms and simply use forms which are more acceptable to the digital tool [16], [18].

There is a big need for technological development in this area, as well. Yet, contrary to high-resource languages such as English, where deep learning and transformer-based models have been developed very rapidly, resources for Bangla are still far from available and dialectical Bangla remains untouched. So, previous work on SA or document classification, in Bangla [5], [6] has obtained high accuracy on curated standard Bangla datasets, but such systems degrade rapidly when faced with dialects available through social media or local communications applications. Recent studies in the area of cyberbullying detection and violence-inducing text classification [3], [4] have further illustrated the frequent dialectal form of hateful or abusive content, and the inherent limitation of the traditional text detection architecture. However, without model and resource development for many dialects, we fail to realize the full potential of Bangla NLP and also fail to develop more crucial applications such as translation, moderation, or speech processing that would become fragile in real deployments.

The motivation also goes beyond academia, into practical and social areas. On the Internet, where millions of Bangla interact day in and day out, dialectal writing is the major form of online text, whether used in casual conversation, as local slang, or in locally marked spellings. However, such variations are not able to be captured with existing tools and are thereby blind spots in moderation, safety and accessibility. As in dialectal Bangla, abusive and vulgar expressions [19] often escape detection by systems trained on Standard Bangla. Apart from challenging online safety, this barrier also restricts inclusivity in digital discount. In the field of education, children from rural areas or areas where dialect is spoken, have trouble with Standard Bangla in formal situations. Automatic dialect to standard conversion could be used as a mediating technology to help learners adapt to academic materials without losing their linguistic identity and to facilitate smooth integration into formal education systems [15].

Another source of motivation is the possibility to innovate by using modern transformer-based models that have demonstrated great flexibility in low-resource and multilingual settings. Although SVM and logistic regression are two classical machine learning models that are still good baselines in Bangla text classification [2], these simple models cannot capture the dialectal structure. On the other hand, recent contextualized language models (CLMs) such as BanglaBERT, XLM-R, and MuRIL have the potential to encode rich contextual information, and sequence-to-sequence transformers such as BanglaT5 and mBART50 have already achieved state-of-the-art performance in translation and generation tasks [17]. By systematically comparing traditional and transform-based approaches within the framework of a multi-dialect setting this thesis thereby fills a technological gap and establishes new standards for future studies.

In short, the impetus for this work is motivated by a great need for the dialects to be squarely on the computational agenda. It's about ensuring millions of speakers don't get left out in the digital revolution, and it's about how to develop resources to celebrate and conserve cultural diversity, and to build systems that are technologically sophisticated and socially meaningful. The large multi-regional contribution of 23,440 sentences, classification and standardization models is not only impelled by the purpose of stretching the limits of research but aims to live the future of technologies inclusive that can recognize, respect and successfully process dialects. Linguistic preservation combined with technical applications and social responsibility is the central drive of this research.

1.3 Objectives

The aim of this research paper is to create efficient techniques for the classification and standardization of Bangla dialects using both traditional machine learning techniques and transformer-based methods. Specifically, it aims to test performance on a custom multi-regional corpus for improving the understanding and processing of variant dialects of Bangla. The main goals of this thesis are:

1. To prepare a large-scale multi-regional corpus of 23440 annotated sentences in five major dialects of Bengali (Sylhet, Chittagong, Barisal, Khulna, and Noakhali) along with their corresponding Standard Bengali counterparts.

2. To implement, develop, and test dialect classification and standardization models, integrating both the traditional ML (SVM, Naive Bayes, Logistic Regression, Random Forest, etc) and the transformer-based approaches (BanglaBERT, mBERT, XLM-R, MuRIL for classification, LSTM, BanglaT5, mBART-50, mT5 for standardization).

3. A comparative study of the two models in terms of different metrics (Accuracy, Precision, Recall, Macro-F1 for classification, BLEU, ROUGE-L, METEOR, chrF, TER, and Exact Match for standardization) between traditional and transformer models, will be performed going for an exhaustive analysis of the results.

4. To establish a user-friendly and interface that will allow users to upload dialect text and select models to visualize results to realize both practical usability test and to encourage future research in Bangla NLP specifically with dialectal diversity.

1.4 Methodology

A systematic approach has been made in this thesis in order to achieve the dual purpose, dialect classification and dialect standardization. The first step of the process was the building of a bespoke multi-regional dataset containing 23,440 sentences. Each entry in this corpus is a dialectical sentence and its parody in Standard Bangla. The database has been meticulously balanced so that there is representation from five major dialect areas of Bangladesh, that is samples taken from Sylhet, Chittagong, Barisal, Khulna, and Noakhali. This ensured that the models to be created for this research would not be biased towards any one dialect but instead learn the dialectal differences between multiple regions.

Once the data set was ready, it underwent an intensive data preprocessing programme. This ranged from normalization of the text, removal of unneeded symbols, and tokenization. Special emphasis was given to the preservation of dialectal characteristics during the preprocessing step since these are the main characteristics that stand out for the classification and translation process. Any existing Bangla Natural Language Processing (NLP) tools were also employed to extract stop words and to filter out noisy data wherever possible with lemmatization. So, to prepare the dataset for machine learning and deep learning models, we performed exported basic cleaning and dialect-sensitive preprocessing.

Classification models were built in the next stage. For a strong comparison, both classical machine learning models and transformer-based models were used. Traditional models consisted of Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR) and Random Forest (RF). The models were trained using n-gram and TF-IDF based feature extraction (used in many of the Bangla text classification studies). On the other hand, the custom dataset was used for fine-tuning the transformer-based models BanglaBERT, mBERT, XLM-R, and MuRIL. We chose to use models that have been shown to be robust in multilingual and low-resource settings. To provide a fair and broad comparison, the performance of both categories was assessed using accuracy and macro-F1 score.

On the basis of the classification, the research proceeded to the area of dialect standardization. The aim of this work was to transform the Bangla sentences from dialects to modern Bangla and preserve the same semantic meaning. As a baseline, a sequence-to-sequence LSTM was implemented to gauge the quality of the output that would be possible with the standard neural techniques for this task. At the same time, we fine-tuned other transformer-based models (BanglaT5, mBART50, mT5) for the same task. These models were trained for a maximum of 30 epochs and the training and evaluation loss was carefully monitored in order to avoid overfitting. The outputs were scored against a variety of metrics used in text generation (BLEU, ROUGE-L, METEOR, chrF, TER, and Exact Match) to provide a multidimensional measure of the quality of the translations.

In addition to the modelling activities a user interface to the system was developed to make it available for research and for practical applications. The interface enables users to load dialectal text or files, choose models for classification and standardization and view results and performance. We have demonstrated the complete system with simple to use frontend and backend models that demonstrate how dialect aware NLP can be applied efficiently to please-to-life applications trivial.

In short, the methodology in this dissertation combines the data set construction, data preprocessing, model development, data evaluation, and interface design in a logical sequence. The comparative study of the classic and transformer-based solution for classification and standardization tasks not only serves as a pointer and a measure of performances, but also clearly indicates the versatility of modern NLP methods in respect to the burgeoning dialectic richness of Bangla language.

1.5 Project Outcome

The results of this project are two-pronged, academic and applied. On the academic side, the research has led to a huge multi-regional annotated dataset of 23,440 sentences, which can be used as a benchmark for any future research in Bangla NLP. It has also allowed for a complete comparison of the state of the art in conventional machine learning with Transformer based models for dialect classification and normalization, paving new baselines as well as yielding insights. On the practical side, there has been developed a working prototype interface that allows the user to upload dialectal text, select models and immediately obtain result with evaluation measures. Together, this compendium of resources, models and tools provides a way to ensure that the research being conducted will not only lead to scientific knowledge but will offer a foundation to be adapted to real-life applications such as education, e-safety and to inclusive communication technologies.

1.6 Organization of the Report

The paper is divided into six chapters in order to report the research in an organized manner work.

Chapter 1: Introduction

This chapter presents about the general background of the research including the significance of Bangla dialects and the problems presented by them for NLP nowadays. It reveals the motivation for the study, the specific goals and objectives and the summary form of the methodological framework. It also dictates what is expected and how the report is structured like.

Chapter 2: Background and Literature Review

The second chapter deals with the theoretical and practical fundamental nature of the study. It describes the current state of the art on Bangla NLP that also includes related works in areas of text classification, sentiment analysis, error detection of grammatical errors and dialect-to-standard conversion. Further, it provides a comparative analysis of traditional machine learning models and transformer-based models and highlights the gap in the research literature this thesis seeks to fill.

Chapter 3: Research Methodology

This chapter explains methodological planning in detail. It discusses the construction of the custom dataset (23,440 sentences), the pre-processing procedures used and the experimental setup. It describes the architecture of classical models (SVM, NB, RF, LR, LSTM) and transformer-based models (BanglaBERT, mBERT, XLM-R, MuRIL, BanglaT5, mBART50, mT5). It also provides the performance evaluation measures to

apply for classification and standardization tasks.

Chapter 4: Implementation and Result

The fourth chapter reports on the experimental implementation as well as the results. It shows training procedures, validation performances, and testing results for every model. Precision is presented based on accuracy and macro-F1 for the classification results, and BLEU, ROUGE-L, METEOR, chrF, TER, and Exact Match for the standardization results. The results are explained with tables, graphs, and an analysis of strengths and weaknesses.

Chapter 5: Engineering Standards and Design Challenges

This chapter describes the engineering side of the research. It explains how the existing standards from software engineering, model training and evaluation were upheld throughout the work. It also discusses pragmatic design issues such as dataset imbalance, preprocessing restrictions, and computational resource restrictions and the approaches taken to address them.

Chapter 6: Conclusion and Future Work

In the last chapter, the most important findings of the study are summarized and the extent to which the objectives were met is assessed. It underlines the academic and practical contributions of the research (new benchmark dataset, comparison between models, design of a user interface to deploy it in practice). In addition, there are promising sources for future work mentioned in the chapter, such as expanding to additional dialects, using speech as input rather than text, and developing multilingual extensions.

Chapter 2

Background

This chapter provides the necessary theoretical and contextual foundation for the research. It begins with an introduction to Natural Language Processing (NLP) and its importance for low-resource languages such as Bangla. The discussion then moves to the linguistic diversity of Bangla, focusing on the five major dialects Khulna, Chittagong, Sylhet, Barisal, and Noakhali that form the basis of this study. The chapter also reviews prior works in related areas such as text classification, sentiment analysis, and dialect translation, highlighting both their achievements and limitations. Finally, the section talks about the research gaps that led to this study.

2.1 Introduction

Natural language processing (NLP) is one of the most impactful areas of Computer Science which allows machines to interpret, work with and produce human language. There has been a lot of progress in high-resource languages like English and Chinese, but not in under-resourced languages like Bangla (Bangla) [1]. Bangla is the seventh most spoken language in the world. It is also one of the richest in dialectal diversity, with millions of speakers of regional dialects like Sylheti, Chittagonian, Barisal, Noakhali, and Khulna. These varieties have significant differences in phonology, morphology, and lexicon, rendering them both linguistically intriguing and computationally demanding [16], [18].

The lack of dialect-sensitive resources has hindered the advancement of robust NLP systems for Bangla. A lot of the research that has already been done has been on things like sentiment analysis [5], [6], [7], document classification [9], and cyberbullying detection [3] and [4]. These models don't usually work well with speech input that has dialectic differences. And recent efforts have begun to bridge this gap. The neural machine translation from Sylheti to Standard Bangla demonstrated the feasibility of dialect conversion using deep learning [16], and the ChatgaiyyaAlap dataset generated parallel resources for Chittagonian to Standard Bangla [18]. The ONUBAD project built on this work by making a parallel corpus for a number of dialects. This was a big step towards being able to represent multiple dialects [28]. But the number of resources available and their coverage are still limited, which shows that there is a clear need for bigger datasets with wider coverage.

Thus, along with resource growth, a methodological transformation has also taken place in Bangla NLP. Earlier works applied mainly traditional machine learning such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression [2], [23]. While having been successfully applied to small curated corpora, these procedures did not scale up to the complexity and the variability found in natural dialectal input. The recent emergence of transformer-based architectures like mBERT, XLM-R, and BanglaBERT, which enables contextual embeddings and transfer learning to low-resource languages (Bhuiyan, et al. 2019; Xu et al. 2020). On the other hand, in sequence-to-sequence tasks, our BanglaT5 [12] and mBART50 [17] models have proven to be suitable for tasks like generation and error correction and translation.

This is part of a general move in NLP from models based on traditional architectures to those based on transformers, especially when it comes to learning from data that requires some form of higher-level cognitive process such as morphological and syntactic knowledge.

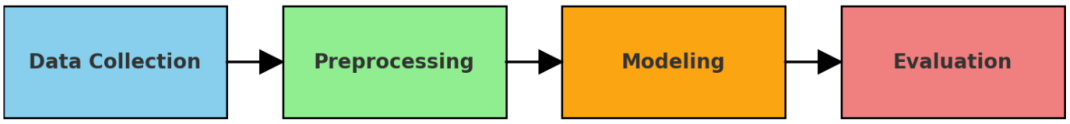


Figure 2.1: Low Resource Language General NLP Pipeline

In this sense, the research being presented locates itself on the border of the production of resources and methodological innovation. By preparing a custom dataset of 23,440 dialectical sentences aligned with Standard Bangla and comparing the performance of both classical and transformer-based models, the thesis will help fill a very large gap in Bangla NLP. Thus, the current chapter provides a review of the relevant background literature, identifying previous contributions, methodological strengths and weaknesses and gaps which justify the current study.

2.2 Literature Review

Bangla Natural Language Processing has gained momentum in recent years with activities centered around a wide range of tasks including news classification, sentiment analysis, cyberbullying detection, dialect processing, and so on. A very relevant work is BanglaNewsClassifier, which implemented a hybrid stacking framework between traditional machine learning and deep learning for news documents, obtaining an accuracy of 94% in categorization for eight classes [1]. Although effective, this system was confined to the news domain and could not be generalized to informal or dialectal text. Another study tried to classify dialects directly using a very small number of samples (5000) from Chatgaiya and Pabna dialect using SVM and feature-based techniques and achieved accuracy of 96% [2]. However, the limited data or number of dialects did limit its applicability. In parallel, cyberbullying detection in the Tagalog language (Bangla) has gained attention; one paper used 5,644 samples and BERT-based models and achieved ~80% accuracy [3], another work focused on multilingual cyberbullying detection using XLM-R and CNN-based architectures and reached 84% [4]. These studies awakened the potential of deep learning without at the same time introducing problems of limited volume of the dataset and of limited domain.

Sentiment analysis is another hot topic of Bangla NLP research. One of the gadget reviews studied Random Forest classifiers with 6,015 sentences and claimed 86% accuracy, but was limited to one domain [5]. One such work presented BanglaDSA, a sentiment dataset [6], while proposing skip-gram embeddings along with BanglaBERT demonstrating better performance than the conventional methods. Later, BanglaT5 [12] achieved the state-of-the-art results for sentiment and generation on low-resource data sets, demonstrating the potential of transformer architectures for low-resource languages. At the same time, papers such as Bangla text document classification using CNN showed how deep learning can perform better than classical ML in large scale classification problems [9].

Table 2.1: Summary of Literature Reviewed.

S. N.	Paper Title & Year of Publish	Dataset Size	Used Model	Accuracy	Limitation	Number of classes or dialects
1	BanglaNewsClassifier: A machine learning approach for news classification in Bangla Newspapers using hybrid stacking classifiers (2025) [1]	118404 (words)	Stacking meta-classifier (BiLSTM + SVM)	94%	Limited to Bangla news domain, only 8 categories, hybrid model resource-heavy	8
2	Bangla Language Dialect Classification using Machine Learning (2022) [1]	5000	SVM	96%	Only 2 dialects (Chatgaiya & Pabna), small dataset, limited generalization, ignores spoken variation	2
3	Detecting cyberbullying text using the approaches with machine learning models for the low-resource Bangla language (2024) [12]	5644	BERT	80.17%	Small dataset, only binary classification, domain-specific (Facebook), limited generalization	2
4	Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts (2024) [51]	5,000	XLM-Roberta	84.10%	Small dataset, only 2 languages (Bangla & Chittagonian), computationally expensive models	2
5	Sentiment Analysis of Bangla Texts on Online Tech Gadget Reviews using Machine Learning (2022) [61]	6,015	Random Forest	86.28%	Dataset limited to tech reviews, only 3 sentiment classes, no deep learning/transformers used	3
6	Sentiment analysis of Bangla language using a new comprehensive dataset BangDSA	203,493 (words)	CNN-BiLSTM + skipBangla-BERT	90.24% (15 classes), 95.71% (3 classes)	Focused only on Bangla, high computational cost, mainly	15 and 3

	and the novel feature metric skipBangla-BERT (2024) [18]				document-level analysis	
7	A Transfer Learning Approach to Bangla Sentiment Analysis (2023) [1]	Training /Dev/Test sets (~label distribution: Positive 31–35%, Neutral 19–20%, Negative 45–50%)	Transfer Learning + Data Augmentation	Micro F1 = 0.71	Imbalanced dataset, only 3 sentiment classes, ranked 12/30 teams	3
8	Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models (2023) [48]	Bangla + English reviews from DARAZ	SVM (ML best), Bi-LSTM (DL best)	86.43% (SVM Bangla), 83.72% (Bi-LSTM Bangla)	Dataset size not specified, only Bangla & English, DL slightly underperforms ML	3
9	Bangla text document categorization based on very deep convolution neural network (2021) [47]	EC: 969,000 unlabelled, BDTC: 156,207 labelled	VDCNN + GloVe	96.96%	Two Bangla variants (Sadhu & Cholito), corpora may not be publicly available, high computational cost	13
10	nlpBDpatriots at BLP-2023 Task 1: A Two-Step Classification for Violence Inciting Text Detection in Bangla (2023) [1]	Not mentioned	Two-step classification (back translation to multilingual)	Macro F1 = 0.74	Dataset not detailed, ranked 6/27 teams, only Bangla	Binary / Multiple violent categories
11	BanglaTense: A large-scale dataset of Bangla sentences categorized by tense: Past, present, and future (2025) [2]	17,819	Not specified	Not applicable	No applied model, focused only on tense, may miss colloquial forms	3
12	BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-	27.5 GB Bangla corpus	BanglaT5 (Transformer, seq2seq) vs mT5-base	-	Only 3 tasks, web-domain data, lacks dialectal variation, high	Not applicable (3 benchmark tasks)

	Resource Natural Language Generation in Bangla (2023) [42]				compute required	
13	Bangla Grammatical Error Detection Using T5 Transformer Model (2023) [11]	9,385 training sentences; 5,000 test sentences	BanglaT5 (small variant) fine-tuned	Avg. Levenshtein Distance = 1.0394 (lower is better)	T5 not originally designed for error detection → required heavy post-processing	1 (Bangla)
14	BanglaLem: A Transformer-based Bangla Lemmatizer with an Enhanced Dataset (2025) [1]	96,040 inflected words (BanglaLem dataset)	BanglaT5 (pre-trained to trained from scratch variants)	Best performance: Exact Match Accuracy = 94.42%	Despite large dataset, Bangla's rich morphology still poses challenges; model may not generalize to unseen rare forms	1 (Bangla)
15	Transcribing Bangla Text with Regional Dialects to IPA using District Guided Tokens (2024) [1]	New dataset spanning 6 districts of Bangladesh (size not explicitly given)	ByT5 (best), compared with mT5, BanglaT5, umT5	ByT5 achieved superior performance (better handling of OOV words); exact accuracy not mentioned	No standardized spelling conventions for dialects, phonological diversity, limited dataset size	6 Bangla regional dialects (district-based)
16	Sylheti to Standard Bangla Neural Machine Translation: A Deep Learning-Based Dialect Conversion Approach (2025) [1]	600 complete sentences + 6,500 Sylheti words with standard Bangla translations	Seq2Seq models: LSTM, GRU, BiLSTM, BiGRU (BiLSTM best)	BLEU (BiLSTM): 57.4 (B-1), 45.8 (B-2), 32.0 (B-3), 22.8 (B-4); ROUGE also evaluated	Very small dataset; focuses only on Sylheti dialect; scalability to other dialects untested	1 dialect (Sylheti → Standard Bangla)
17	Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated	32,500 sentences (Bangla, Banglish, English	mT5, BanglaT5 (for translation); mBERT, Bangla-	BLEU: 69.06 (Mymensingh), 36.75 (Chittago	Lower performance on Chittagong dialect; resource scarcity for some dialects	5 regional dialects

	Translation of Bangla Regional Dialects to Bangla Language (2024) [12]	across 5 dialects)	bert-base (for region detection)	ng); Region Detection Accuracy: 85.86% (Bangla-bert-base)		
18	ChatgaiyyaAlap: A dataset for conversion from Chittagonian dialect to standard Bangla (2025) [2]	4,012 sentences + 1,500-word dictionary	Not explicitly used (dataset paper)	N/A (dataset creation)	Limited dataset size; focused on Chittagonian only; lacks large-scale evaluation	2 (Chittagonian dialect & standard Bangla)
19	Automatic Vulgar Word Extraction Method with Application to Vulgar Remark Detection in Chittagonian Dialect of Bangla (2023) [94]	Social media posts/comments from Facebook (exact # not specified)	SimpleRNN (Word2Vec, fastText), Logistic Regression	0.84–0.90 (RNN), 0.91 (LR)	Dataset size not very large; keyword-based method requires constant lexicon updates; deep learning needs larger datasets	2 (vulgar, non-vulgar)
20	A hybrid approach for Bangla sentence validation (2024) [0]	5,000 labeled Bangla sentences	CNN-BiLSTM hybrid classifier	F1 score: 98%	No standard benchmark dataset; limited-resource language challenges; handles only two sentence correctness classes	2 (Correct, Incorrect)
21	Bangla text normalization for text-to-speech synthesizer using machine learning algorithms (2024) [1]	Tokenized Bangla corpus with semiotic class labels (size not specified)	XGBClassifier	99.997% token classification accuracy	Dataset size not clearly reported; limited to semiotic class coverage from corpus	Multiple semiotic classes (exact number not given)
22	A Comparative Study on different Machine Learning Approaches for Categorizing Bangla Documents (2025) [7]	Bangla newspaper dataset (size not explicitly given)	RF, KNN, SVM, DT, BNB, CNB, MNB, Bagging (BC), LR	SVM: 92.76%, BC: 92.64%, LR: 92.26%	Dataset size not clearly specified; performance tied to TF-IDF features only; limited to supervised ML	8 categories (document classes)

					models (no deep learning/transformers)	
23	BanglaBlend: A large-scale novel dataset of Bangla sentences categorized by saint and common form of Bangla language (2024) [0]	7,350 sentences	Not model-specific (dataset paper)	N/A (dataset creation)	Limited to stylistic forms (Sadhu vs Cholito); no model evaluation	2 (Saint/Sadhu form, Common/Cholito form)
24	Exploring Bangla Religious Dialect Biases in Large Language Models with Evaluation Perspectives (2024) [0]	Not explicitly mentioned	ChatGPT, Gemini, Microsoft Copilot	Comparative bias evaluation; no quantitative accuracy metric	Dataset size not specified; results are qualitative; focuses only on Hindu and Muslim dialects	2 (Hindu, Muslim)
25	Bridging Dialects: Translating Standard Bangla to Regional Variants Using Neural Models (2024) [6]	32,500 sentences	BanglaT5, mT5, mBART50	BanglaT5: CER 12.3%, WER 15.7%	Focused on 5 dialects only; may not generalize to unseen dialects	5 (Chittagong, Sylhet, Barishal, Noakhali, Mymensingh)
26	BTSD: A curated transformation of sentence dataset for text classification in Bangla language (2023) [0]	3,793 sentences	Not specified (benchmark for NLP models)	Not specified	Small dataset size; limited to Simple, Complex, Compound sentence types	3 (Simple, Complex, Compound)
27	ONUBAD: A comprehensive dataset for automated conversion of Bangla regional dialects into standard Bangla dialect (2025) [12]	1540 words, 130 clauses, 980 sentences per dialect	Neural Machine Translation (NMT)	Not specified	Limited dataset size per dialect; only Chittagong, Sylhet, Barisal covered	3 (Chittagong, Sylhet, Barisal)

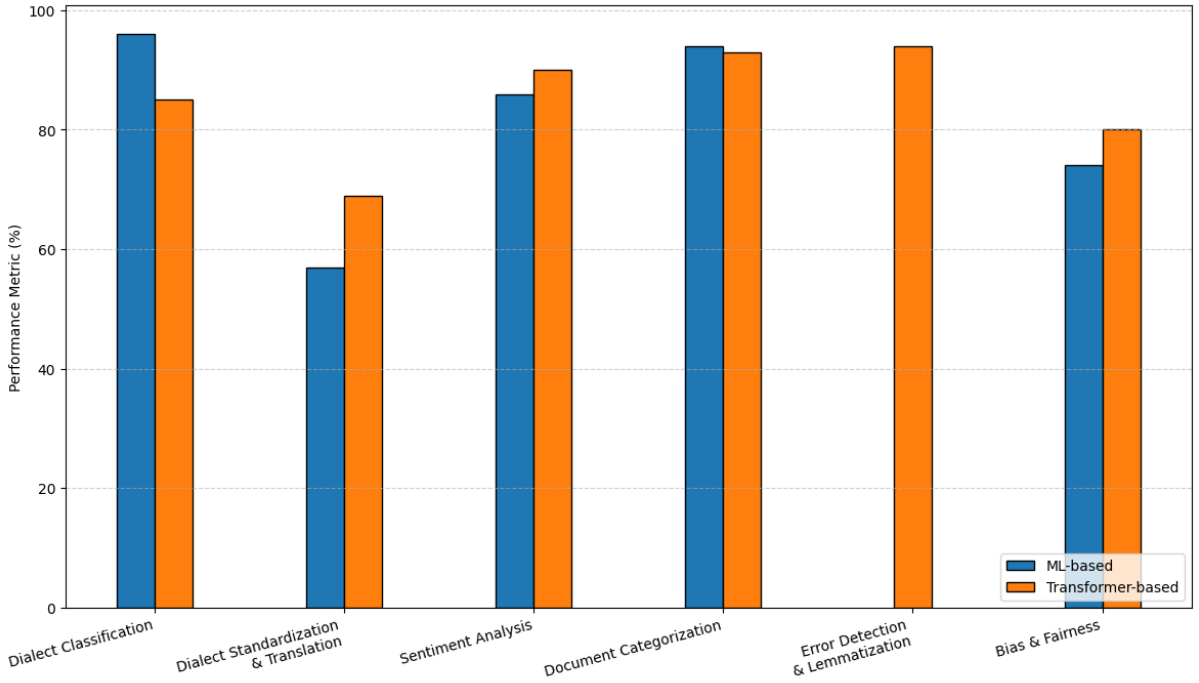


Figure 2.2: Comparative Performance of Works Based on Different Perspective in Bangla NLP

There are not too many studies of Bangla dialectology, but they have developed gradually. For instance, deep learning was used to estimate dialect conversion [16] of Sylheti to Standard Bangla NMT by ChatgaiyyaAlap [18] provided a dataset for mapping Chittagonian to Standard Bangla. The ONUBAD project went one step further to add a parallel corpus of not just Sylheti but also Chittagonian and Barisal in the corpus [28]. Even with these contributions, limitations become apparent throughout the existing work: datasets are small, domain-specific or single dialectal and models are hardly ever benchmarked in a multi-dialectal context. This gap creates a need for a large-scale, multi-regional dataset and systematic comparison for traditional and transformer models, which is the starting point for the current study.

2.3 Gap Analysis

On the one hand, although Bangla NLP research has developed impressively in recent years, there is a lack of focus on several important areas, especially in the field of dialect classification and standardization. The existing works usually work on very narrow domains like sentiment analysis [5], [6], [12] or news classification [1], which perform well on Standard Bangla corpora but don't work that well when the text is dialectal. While these works have shown the potential of machine learning and transformer-based models, these models are limited in terms of the scope of their datasets, and very rarely represent linguistic diversity in Bangladesh. Similarly, work to detect cyberbullying [3], [4] and offensive remarks in the Chittagonian dialect [19] clearly demonstrate the social need for dialect-aware NLP, but these studies have been limited due to small corpora and the use of only one dialect.

However, another gap of existing works is that there is an imbalance in methods. Traditional machine learning techniques like SVM, Naive Bayes, and Decision Tree Random Forest are used extensively in previous Bangla NLP activities [2], [9] and are not sufficient for the morphological and phonological richness of dialectal data. On the other

hand, transformer-based models such as BanglaBERT, mBERT, XLM-R and BanglaT5 [12], [17] have been shown useful in multilingual and generative tasks, but their use for dialect classification or dialect-to-standard conversion has not been extensively studied. Previous works like Sylheti-to-Standard Bangla NMT [16], ChatgaiyyaAlap [18], ONUBAD [28] showed promising results but only covered one dialect and small-scale experiments, or contributed parallel corpora for many dialects but were lacking in terms of dataset size and evaluation criteria.

Finally, multi-dialect datasets (sufficient to allow for a reliable benchmark for both classification and standardization tasks) are in short supply. Not only are most existing resources too small or too domain-specific, but no previous work has systematically compared traditional machine learning with state-of-the-art transformer architectures on multi-dialectal Bangla data. This lack of consciousness is particularly striking when you consider that dialects play an important role in daily life: education, online communication, and cultural identity preservation. The present study attempts to fill this gap by providing a complete corpus covering 23,440 sentences across five major dialect varieties and comparatively analyzing and experimentally validating both traditional and transformer-based models for classification and standardization; therefore, it fills a significant void in Bangla NLP.

Table 2.2: Comparison of features of existing systems with the proposed system.

Feature	Existing Systems (Gap)	Proposed System
Large Multi-Regional Corpus	Datasets are small, domain specific and often single domain (e.g. Sylheti), single task (e.g. Chittagonian) oriented.	Yes
Two Tasks: Classification and Standardization	No hybrid model, no single task (classification / sentiment / translation).	Yes
Hybrid Model Variety	To use either traditional ML or single transformer only and not both.	Yes
Comprehensive Evaluation Metrics	Rely on few metrics (Accuracy/F1 or BLEU only), no multi-metric benchmarking.	Yes
Error and Comparative Analysis	Lack detailed error analysis and systematic cross-model comparison.	Yes
User Interface for Usability	No proper UI; mostly experimental CLI-based outputs.	Yes
How useful it is in the real world	Only for research prototypes; not for regular speake rs or teachers.	Yes

2.4 Summary

In this chapter, we have described the background and reviewed the literature behind the Bangla natural language processing, especially focusing on dialectal issues. It started with raising the importance of dialects in Bangla and how their lack from digital content leaves them behind in creating barriers to inclusivity. The literature review showed that

notable work in sentiment analysis, news classification, cyberbullying, dialect to standard conversion have been done, but most of them were limited to Standard Bangla or to small dialectal corpus [1], [5], [6], [9], [16], [18], [28]. The gap analysis also highlighted that empirical research tends to be based on small-scale corpora, be focused on dialect-specific tasks or be domain-specific, and that, in spite of the effectiveness of transformer-based models, these have not yet been used in a systematic way for dialect classification and standardization.

In summary, although valuable insights and useful resources have been obtained from prior works on Bangla NLP, these achievements were not able to reach the scale, heterogeneity, and methodological discipline needed in multi-dialectal Bangla NLP. This thesis fills these voids by building a large dataset of 23,440 sentences covering five regional dialects and by comparing, systematically, traditional and transformer-based models for both classification and standardization. Therefore, it assumes the status of a new contribution in the intersection of language diversity and computational innovation in the field of Bangla NLP.

Chapter 3

Research Methodology

This chapter describes the methodology used to conduct the study of classification and standardization of Bangla dialect. It describes the general framework of the work, from the construction and pre-processing of datasets to the development and evaluation of the model developed. The chapter also identifies the functional and non-functional requirements, the proposed system design and interface considerations. Furthermore, it includes alternative solutions which have been assessed at the design stage and also exposes the structured plan and schedule for the project implementation.

3.1 Methodology

3.1.1 Overview

The research methodology was formulated to systematically resolve the problems of classification and standardization of the Bangla dialect. To execute the research the study started with the development of a custom multi-regional corpus comprising of 23440 annotated sentences from the five dialect areas: Sylhet, Chittagong, Barisal, Khulna, Noakhali. Each dialectal sentence was aligned with its Standard Bangla counterpart; thus, the dataset is ready for both classification and translation. Similar works have been done on small sets like Sylheti-to-Standard Bangla NMT [16] and ChatgaiyyaAlap [18], but they were not applied to larger sets due to the limited data size.

After dataset creation, the dataset was preprocessed through tokenization, normalization, stop word deletion, dialect-oriented cleansing, etc. From the studies of sentiment analysis in Bangla language [5], [6], it has been identified that preprocessing is crucial for Bangla NLP performance. For classification, the classical machine learning (SVM, Naive Bayes, logistic regression, random forest) and the transformer-based models (BanglaBERT, mBERT, XLM-R, MuRIL) were trained and tested. For standardization purposes, a baseline LSTM model was compared against Transformer-based sequence-to-sequence models including BanglaT5, mBART50 and mT5, which have been previously used effectively in Bangla text generation and translation [12], [17].

Metrics for evaluation were chosen to match each task: accuracies and macro-F1 for classification, and BLEU, ROUGE-L, METEOR, chrF, TER, and Exact Match for standardization. A user interface is also developed to make the system usable in practice, enabling users to upload dialectal text, to choose a model and to view predictions together with performance metrics.

3.1.2 Proposed Methodology

The proposed methodological approach combines all the steps of the research into a single pipeline that starts with the raw data collection and ends with the deployment of the user interface. Figure 3.1 represents the overall workflow of the system and the following steps give a detailed description of each stage.

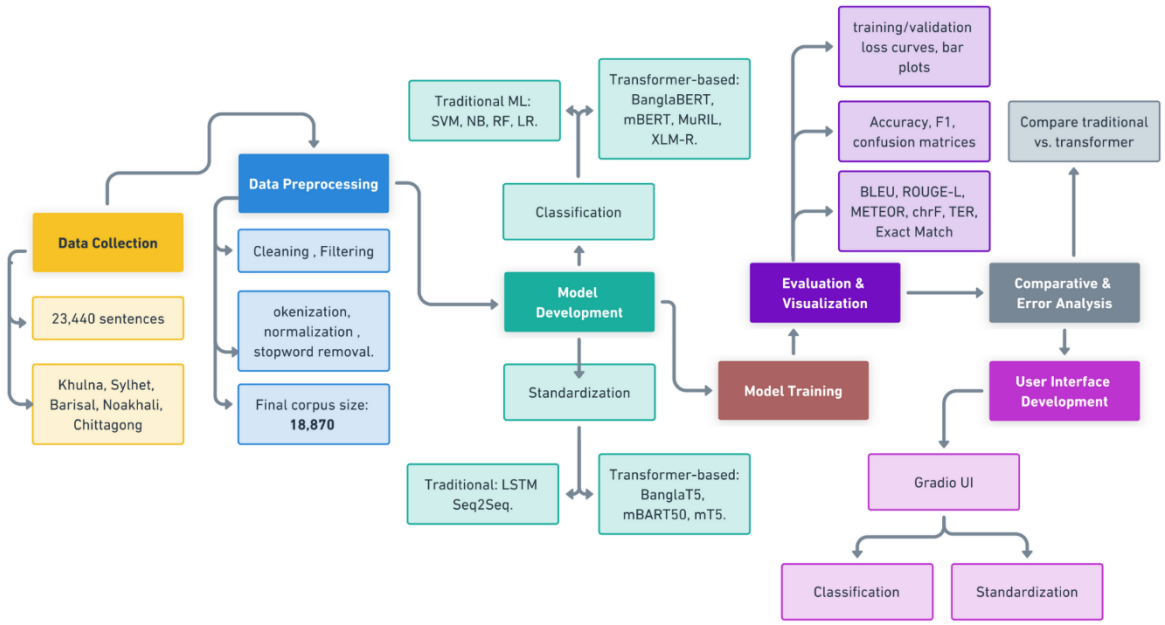


Figure 3.1: Methodology for Bangla dialect classification and standardization (workflow diagram).

The pipeline is made up of the following steps:

1. **Data Collection**

- A customized corpus of 23,440 sentences for five major dialects of Bangla, Sylhet, Chittagong, Barisal, Khulna, and Noakhali, has been produced.
- All the sentences in the dialectal corpus were annotated with the corresponding Standard Bangla sentence, thus making the corpus useful for classification and standardization tasks.
- Texts for Study Based on Natural Sources Linguistic materials, including conversations, native expressions, and local dialects, were used to ensure authentic coverage.

2. **Data Preprocessing**

- Text normalization was used to deal with irregular characters, spelling variations and punctuation.
- Tokenization was conducted while maintaining dialectal distinctions which meet the criteria for classification cues.
- Stop word removal and lemmatization were applied selectively, taking care not to lose important dialectal forms.

3. **Dataset Splitting**

- The data set was split into three data subsets: 70% for training, 15% for validation and 15% for testing.
- This helped to ensure that models were trained on enough data, validated during the tuning, and tested on unseen samples to get an unbiased estimate.

4. **Model Development**

- Classification model: Traditional machine learning methods such as SVM, Naive Bayes, Logistic regression, and Random Forest; Transformer-based models including BanglaBERT, mBERT, XLM-R, and MuRIL.
- standardization models: LSTM encoder-decoder baseline sequence-to-sequence architecture; transformer-based sequence-to-sequence models like BanglaT5, mBART50, mT5 etc.

5. **Model Training**

- We have trained classical ML models using TF-IDF and n-gram features.
- For some epochs (15 for classification and 30 for standardization), transformer models were fine-tuned on GPUs.
- Hyperparametric optimization was used to make performance better (for example, by changing the learning rate, batch size, and choice of optimizer).

6. **Evaluation**

- Classification metrics: macro-F1 and accuracy
- standardization metrics: BLEU, ROUGE-L, METEOR, chrF, TER and Exact Match
- Visualization tools: Using the tables, confusion matrices, graphing performance comparison, and loss/accuracy curves were made to help us better understand the results.

7. **Comparative and Error Analysis.**

- We looked at both traditional ML-based models and transformer-based models.
- Modelling errors were investigated with respect to misclassified dialects and standardized output.
- Results of this analysis were interpreted in terms of strengths and weaknesses of each model.

8. User Interface

- A dashboard type interface was created to enable users to:
    - Read dialectic text or CSV files
    - Classification or standardization models for selection
    - See prediction together with evaluation metrics;
- The interface is designed so that both researchers and regular users can easily use it. It also connects the output of the computer to the utility.

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements

The system has to deliver the following functional abilities:

1. Input Handling

1.1. Takes user input in Bangla dialects (Sylheti, Chittagong, Barisal, Khulna and Noakhali).

1.2. Allows single sentence input and a bulk file input.

2. Dialect Classification

2.1 Determines regional dialect of input sentence.

2.2 It uses standard machine learning algorithms like SVM, NB, LR and RF and transformers-based classifiers such as BanglaBERT, mBERT, XLM-R and MuRIL.

3 Dialect Standardization

3.1 Translates dialectal Bangla to Standard Bangla without losing meaning.

3.2 Uses LSTM as a control condition and transformer-based seq2seq models, including BanglaT5, mBART50, and mT5 .

4 Model Comparisons and Selections.

4.1 It gives users the option of selecting the model they would prefer to use in classification or standardization.

4.1 Shows comparative performance.

5 Assessment and Charting.

5.1 Displays assessment scores including Accuracy, Macro-F1, and BLEU, ROUGE-L and METEOR, chrF, TER, and Exact Match.

5.2 Produces visualization(s) (confusion matrix, accuracy/loss curves) to interpret results.

Non-functional Requirements

Other non-functional requirements of the system are:

- **Usability**
  - Both researchers and general users have access to a simple and user-friendly interface.
- **Performance Efficiency**
  - Fast classification and standardization on large data sets.
- **Scalability**
  - Is able to cope with growing volumes of input with minimal loss of performance.
- **Reliability and Robustness**
  - Delivers consistent and predictable output when using different dialects and models.
- **Maintainability**
  - Enables simple addition of new dialects or new models later.
- **Portability**
  - Allows use in either local or server or cloud environments.

3.1.4 Data Flow Diagram

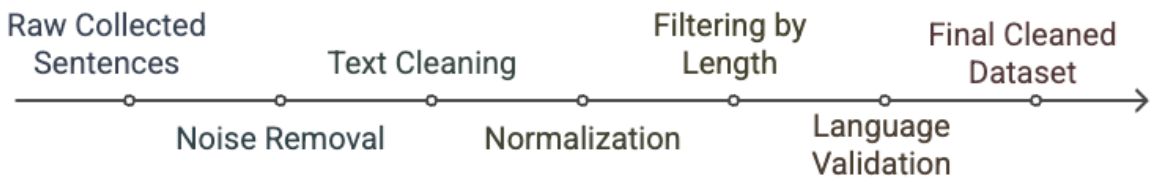


Figure 3.2: Data Preprocessing flow Diagram

3.1.5 UI Design

A simple and easy to use Graphical User Interface (GUI) was produced for accessibility and usability of the system. The UI was designed with the aim of making the dialect classification and the process of standardization straightforward even for non-technical users, such as educators, researchers and general speakers of Bangla dialects. The

philosophy behind the implementation here was to keep things as simple as possible while offering good navigation and informative output.

The main characteristics of the UI are the following:

1 Data Input

1.1. Users can either write a sentence directly in a text box or load a file (CSV or TXT) in which multiple dialectal sentences can be found.

2 Dialect Selection and Detection text

2.1 The system also has the feature of automatically detecting the dialect.

2.2 Users can also choose a particular dialect area for manual testing (Sylhet, Chittagong, Barisal, Khulna, or Noakhali).

3 Model Selection

3.1 Classifiers – There is a dropdown which gives the user the choice between classification models (SVM, Naive Bayes, Logistic Regression, Random Forest, BanglaBERT, mBERT, XLM-R, MuRIL) and standardization models (LSTM, BanglaT5, mBART50, mT5).

4 Output Display

4.1 The forecast dialect label of the classification task is indicated at the interface.

4.2 For standardization work, it shows the converted sentence in Standard Bangla.

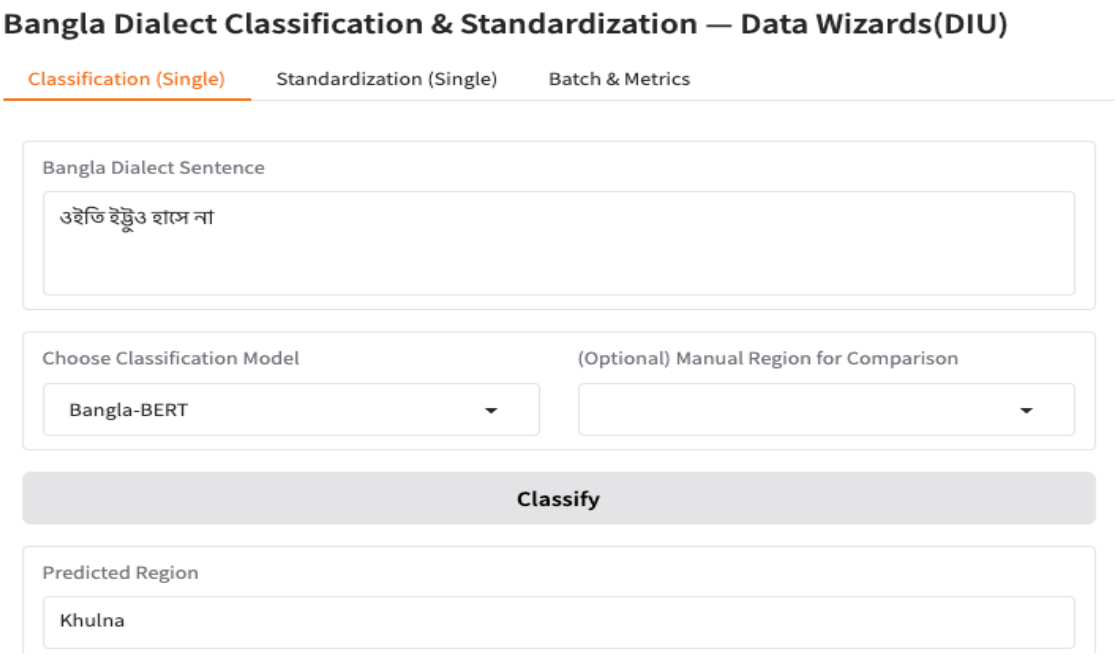


Figure 3.3: User Interface for Dialect Classification

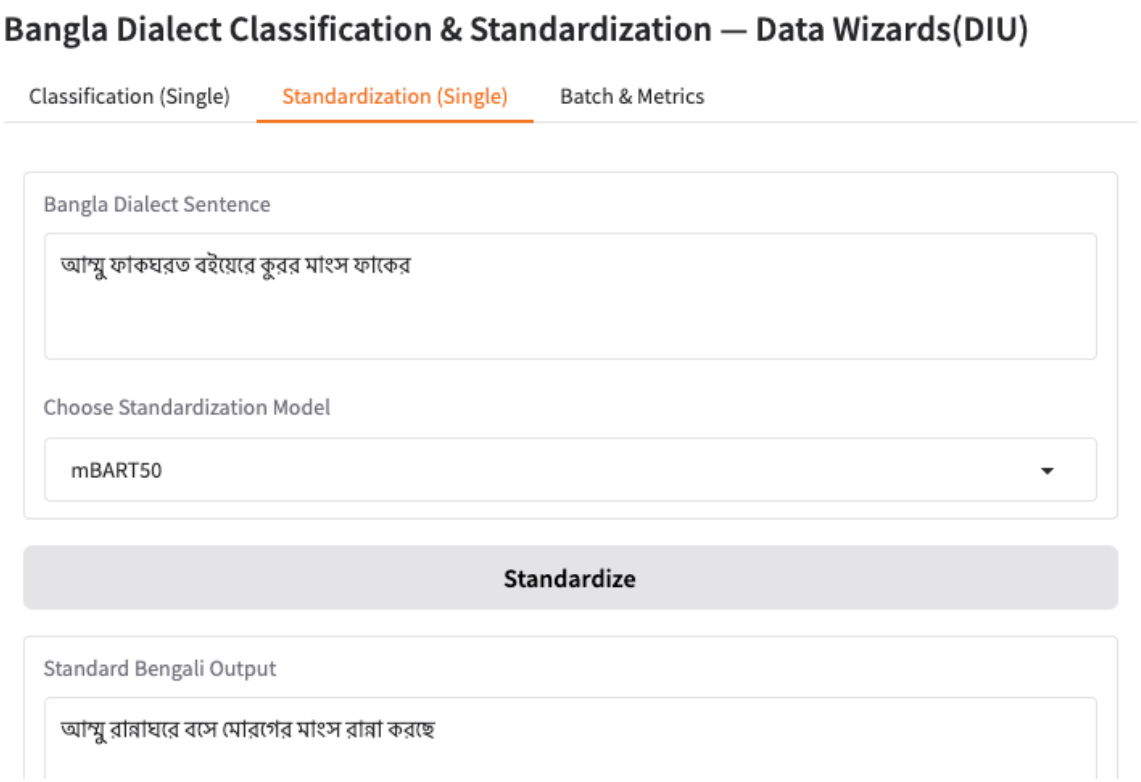


Figure 3.4: User Interface for Dialect Standardization

5 Performance Metrics Dashboard

5.1 Evaluation tables and charts of evaluation are generated by the system - Accuracy, Macro F1 Score, BLEU score, ROUGE-L, etc.

5.2 Confusion matrix and training- val curves are presented for better interpretation of results.

6 Comparative Analysis Tab

6.1 Another tab gives a side-by-side comparison between traditional ML models and transformer models, all the way down to an error analysis.

The UI design overall makes it easy to use the system without the need to know anything about the algorithms behind it.

3.2 Detailed Methodology and Design

3.2.1 Construction of the datasets and Sources

This research is based on a specifically designed multi-regional corpus of 23,440 parallel sentences in which one dialectical input sentence and its Standard Bangla translation appear together. Five geographically and socio-linguistically relevant areas were selected, Khulna, Chittagong, Sylhet, Barisal and Noakhali, so that varieties which differ both phonetically and morpho-lexically but are mutually intelligible are required to be classified separately. There are three fields per entry: region (target class label),

dialect_sentence (source text) and standard_Bangla (reference text for normalization/translating). This two-way architecture supports both top-down supervised dialect classification and bottom-up dialect-to-standard generation from the same dataset.

Our data design choices were inspired by previous dialect resources, which though useful were either single-dialect (e.g., Sylheti to Standard pipelines) or domain-limited corpora. For example, Sylheti to Standard Bangla neural systems have been shown to be feasible [16], but only work on a narrow dialect base. ChatgaiyyaAlap has Chittagonian to Standard pairs, invaluable for covering this region, but not for representing all of Bangla [18]. The National Board for Standardization of the Basque Language (ONUBAD) brings together a number of dialects in the standardization process and draws attention to the need for automatic robust translation [28]. For example, Vashantor conceptualizes large-scale dialect translation as a generalizable benchmark across variation in Bangla regional varieties [17], and Bangla Blend explicitly compares "saint" vs "common" (standard vs. colloquial) sentence types which we operationalize within our standard Bangla field [24]. Together, these works provide the basis for demand and feasibility; the contribution we make is to consolidate them, providing a single, well-balanced multiregional corpus that has been explicitly aligned for both classification and standardization tasks.

To avoid sampling bias and topical skew, source texts were gathered from texts distributed throughout conversational registers, local expressions, and region-specific idiomatic expressions, rather than in just one specific domain (e.g., news or reviews). This is in line with best practices from the Bangla text classification and generation literature, where more general lexical diversity makes models more robust [1], [12]. Internally, we were able to monitor counts per region while ingesting in order to reduce imbalance at ingestion time. After collection, the set of each dialectal utterance was checked by a label auditor to confirm that each had exactly one region label (no multi-label cases). Because dialect conversion is dependent on a good reference, a conservative target for standardization was designed: we prefer literal semantic preservation to stylistic paraphrase (similar to dialect-to-standard NMT works by [16], [17], [18], and [28]).

Finally, since our research involves two tasks (classification and standardization), the corpus was divided by region in a stratified way. At the pipeline level, we use a traditional 70/15/15 train/validation/test split to provide systematic comparison and early stopping for both traditional and seq2seq experiments. For some of the transformer fine-tuning runs, we maintain the 15% validation subset for model selection across the experiments. Combining both fixed validation and library-level evaluation is very common in Bangla NLP experiments [1], [9], [12].

Takeaways. Compared with preceding single-dialect or narrow-domain corpora [16], [18], [28], our dataset is larger, balanced across five regions, and parallel by design - thus facilitating joint advances on classification and standardization in a single experimental bed [17], [24].

3.2.2 Preprocessing Techniques

Preprocessing is based on two principles: (i) text must be normalized and cleaned, so that spurious noise can be removed; (ii) dialectal cues (phonological spellings and regional lexemes) that are discriminatory for classification and conversion must be preserved. Concretely, we apply:

- Unicode normalization and whitespace tidying All strings are normalized multiple white spaces are collapsed and non-Bangla artifacts are removed (allowing Bangla codepoints digits and simple punctuation only). The above procedure is a conservative extension of normalization used in Bangla text normalization pipelines.

- Language-range guarding. Rows that contain a very low proportion of Bangla characters in relation to the number of tokens are eliminated using a Bangla-script ratio heuristic (to prevent remaining stray Roman or non-linguistic noise). Filters with this range of script are typical for cleaning mixed-script Bangla corpora.

- Length constraints. We keep between 4 and 60 tokens' length per sentence for excluding fragments and unusually long, multi-clause chains that can destabilize seq2seq learning (a practice also reported in Bangla NLG and translation baselines).

- De-duplication. Duplicate dialect_sentence rows are eliminated to prevent memorization bias in classifiers and leaking into evaluation splits (as recommended in the standard curation practices for Bangla text categorization).

- Tokenization & stop-word list. For traditional classifiers, we use word n-gram TF-IDF features (lowercase), with deliberate stop-word removal (because we know dialects contain many occupational terms that must be gathered; this seems to have been an issue in some of the previous Bangla polarity and topic studies). Following the best practice for models trained in BanglaBERT/mBERT/XLM-R pipelines we use model native tokenizers instead of manually normalizing which can destroy meaningful spellings (e.g. Word Piece/BPE/Sentence Piece).

We deliberately avoid aggressive lemmatization on the dialectal source side: while modernized Bangla lemmatizes (e.g., transformer-based BanglaLem [14]) do exist, lemmatizing dialect tokens threatens to wash out morpho-phonological signals that are vital for source identification and source-standard mapping. In contrast, lemmatization is used only for auxiliary analysis and not as input to any of the main classification or standardization models. Instead of explicitly using hard, rule-based edits to account for those grammatical noises or tense artefacts, we implicitly account for them in the seq2seq models, a technique that has been shown to encode grammar regularities for Bangla [12], and for targeted tasks like tense detection (Bangla Tense) [11].

Outcome. Note that this should not result in much loss of information, as filters are pre-tuned to preserve dialectic identity (not, as in earlier research, to over normalize and thus make all dialects sound similar!) and so the training set after preprocessing is still substantive. This trade-off between cleanness and cue preservation is critical for dialect NLP and parallels the results of Bangla sentiment/news classification, in that feature integrity has a significant effect on accuracy with the total size of the clean corpus: 18,870 processed sentences.

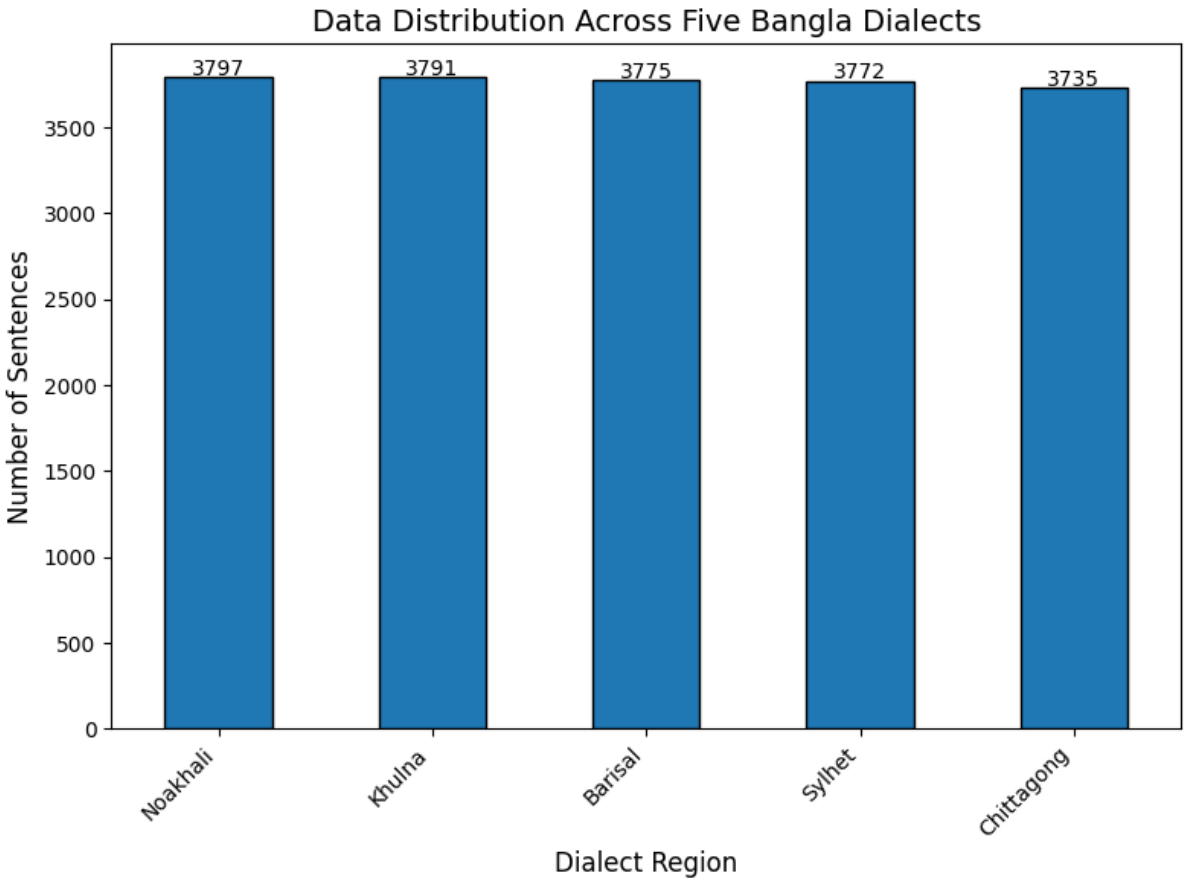


Figure 3.5: Distribution of Clean Data by Five Dialects (Khulna, Sylhet, Noakhali, Barisal, Chittagong)

3.2.3 Basis for Model Selection Justified.

3.2.3.1 Traditional Machine Learning (Baselines)

We present SVM, Multinomial Naive Bayes, logistic regression and random forest as completely interpretable, controllable baselines for dialect classification. They have been shown consistently competitive in Bangla document and sentence classification using bag-of-words and n-gram features [1], [2], [5], [9], and [27]. Their benefits in our environment are three-fold:

1. Interpretability & speed. Feature-based models, using TF-IDF n-grams are fast to iterate, have error analysis that is intuitively simple (i.e. the weight of each feature) and computationally inexpensive: a great way to explore which dialect tokens are discriminative.

2. Note that the models offer strong baseline performances even for low-resource regimes. Bangla corpus - particularly dialectical - can suffer from domain shift and orthographic variation, and under such noisy lexical conditions, linear models (SVM/LR) and NB fare well.

3. These models are calibrated to ablations. As such, our results provide a credible ground-truth for testing how well the transformer works; in many places, feature-importance plots and confusion matrices generated by these models provide indications of the hard-to-separate areas and reasons for separation (lexeme overlap, near-cognates).

4. For LSTM encoder-decoder baseline experiments (seq2seq, teacher forcing, cross-entropy), we use a classic neural MT-like reference implementation, as well as a model that yields near-perfect performance across all tasks. Although the recently proposed Bangla NLG prefers transformers for the sake of pedagogical purpose, an LSTM baseline is still used to show the impact of pretrained language knowledge on dialect conversion.

3.2.3.2 Models of Transformers (Fundamental Strategies)

We fine tune BanglaBERT, mBERT, XLM-R and MuRIL for classification. The rationale:

1. BanglaBERT encodes monolingual Bangla morphology and orthography and thus performs significantly better than generic multi-lingual encoders on Bangla tasks due to the domain tokenization.
2. The wide cross-linguistic robustness in mBERT and the large-scale pretraining in XLM-R is useful in the face of dialectical orthographic variation and loanwords.
3. An automated tool named Multi-Language Representations for Indian Languages (MuRIL) is suited to the particular requirement of South Asian scripts and code-mixing cues in Bangla and therefore has been chosen as an obvious option for Bangla dialectal inputs.

Empirically, multilingual encoders have proven their worth on Bangla regional tasks (e.g. Chittagonian) across text classification and harmful-content detection problems, showing transfer to related varieties [4], and recent deep CNN/transformer families have routinely improved upon classical methods by large margins on Bangla categorization.

We benchmark BanglaT5, mBART-50, and mT5:

- BanglaT5 is a Bangla version of T5 that has been bench-marked on Bangla NLG and demonstrated to perform very well on generation tasks when trained/fine-tuned on high-quality supervision.

- mBART-50 provides multilingual denoising pretraining support and the ability to generate bn_IN directly mBART family models already had demonstrated high fluency and adequacy after a temperature sensitive dialect standard transfer pipeline and Bangla NMT pipeline.

- mT5 extends T5 to the multilingual pretraining setting show that its broader coverage and Sentence Piece tokenization appear to be better able to handle dialect spellings aligned with Bangla script.

These decisions are also driven by the task form: dialect standardization is essentially a constrained translation from region-marked Bangla to formal Standard Bangla. Works on translating dialect to standard or standard to regional demonstrate the plausibility of pretrained seq2seq transformers for the purpose of this mapping [16], [17], [26], [28]. Moreover, going from single-dialect generation (Sylheti or Chittagonian only) to multi-

regional generation forces models to learn meaning invariant to surface variations-an inductive bias that is naturally supported by transformer decoders with attention [16, 17, 18, 26, 28].

Epoch and computing. We train/fine-tune the classification transformers for 15 epochs and standardization transformers for 30 epochs, following a direction provided by the validation loss/accuracy. This larger generation time reflects the additional complexity of performing optimization at the sequence level as well as the objective combination (BLEU/ROUGE/chrF/TER). The fine-tuning regime we follow lies in line with previously reported NLG fine-tuning regimes in Bangla [12], while our ablations allow us to avoid overfitting to very small subsets of the dialect under consideration.

Why not only transformers? Although transformers ultimately outperform everything, traditional baselines continue to be useful for (i) sanity-checking signal in labels; (ii) exposing error modes to be produced by surface tokens; and (iii) offering lightweight options for deployment where GPU resources are scarce, a concern that has been frequently highlighted in Bangla NLP deployments [1], [2], [5], [9].

3.2.4 Training Strategy and Hyperparameter Tuning

Baselines were organized following the idea of a trade-off between robustness, computational feasibility and fairness when it comes to both classification and standardization model training. Since our training set is custom and dialect-rich we aimed for reproducible but adaptive training configurations.

For traditional classifiers, the main representation of features was TF-IDF, n-grams. A few variations in the dimension characteristics were implemented to ensure that features reflect all dialectal tokens without suffer excessive sparsity feature dimensionality was limited to 15,000. Model-specific approach included:

- SVM Support Vector Machine: Grid over C = {0.1, 1, 10}, Class Balancing SVM has also been shown to be a high-performance for Bangla text classification.
- Logistic Regression: Solver = LBFGS, Max iteration = 2000 (Weights used = Balance class weights).
- Naive Bayes (NB): Only requires a smoothing hyperparameter; valuable as a benchmark.
- Random Forest (RF): Grid search on n_estimators (100, 300) and max_depth (=None; 20; 50) and balanced class weights

For the transformer-based classifiers we fine-tuned four pre-trained transformers BanglaBERT, mBERT, MuRIL, and XLM-R for 15 epochs with early stopping. On the other hand, the batch sizes were tuned according to the GPU memory: 16 (train) and 32 (eval). We applied with learning rate of 2e-5, weight decay of 0.01 and linear warmup schedule as used in the Bangla transformer fine-tuning. Patience during early stopping was defined as 3 epochs on the validation accuracy. Stratified splits to make sure that each fold got people from every region.

The training problem for the standardization models was harder due to the fact that generation needs an alignment at sequence level. We trained BanglaT5, mBART-50 and mT5 for 30 epochs, at learning rates of 2e-5 (BanglaT5/mBART-50) and 5e-5 (mT5), respectively, according to stability noticed in pilot runs. Implicit control of teacher forcing ratio was implemented by Hugging Face's Seq2Seq framework while beam search (beam = 5) was employed in inference to promote fluent outputs. To obtain a good tradeoff between coverage and efficiency, sequence lengths were limited to 128 tokens.

A particular concern was that of overfitting: because dialect-to-standard is an easier semantic task than, for instance, translation into a remote language, models tended to memories patterns rather than generalize. To tackle this issue, we (i) randomly permuted mini-batches from each epoch, (ii) closely monitored the validation loss, and (iii) tested dropout rates ranging from 0.1 to 0.3 in LSTM and transformer decoders. These methods are akin to those employed in other low-resource Bangla tasks where model robustness is a priority.

3.2.5 Evaluation of Metrics and Visualization

A new test was made that takes into account both the accuracy of the classification and the quality of the generation. We wanted a range of metrics for this study because we are comparing traditional and transformer approaches.

We used the following classifiers:

- We used accuracy, which is the number of correct guesses.
- Macro-F1 (the balance between precision and recall for different dialects in the data, which is important for fairness across classes).
- Precision and recall for each area (to find problems with regional dialects).
- The confusion matrix and normalized confusion matrix are shown in heatmaps to make them easier to understand. This kind of multi-metric setup is common in Bangla classification literature, which makes sure that both overall and per-class evaluation is done.

For standardization tasks, we utilized and examined a blend of automatic machine translation metrics and error-focused metrics:

- BLEU (n-gram overlap between candidate and reference).
- Fluency (longest common subsequence overlaps, capturing fluency) - ROUGE-L.
- METEOR (semantic matching and synonymic alignment)
- chrF (character n-gram F-Score, works well for morphologically rich Bangla).
- TER (Translation Edit Rate) (index of post-editing effort)
- Exact Match % (how many outputs that are identical to the reference).

- This evaluation palette follows the best state-of-the-art practices in dialect-to-standard Bangla NMT.

Importance too was placed on Visualization. In classification, we made line plots of the training and validation loss/accuracy, and it's easy to see that the plot is overfitting. To understand dialect overlaps, confusion matrices were plotted once in raw counts and once normalized (by class). For the purpose of standardization, bar graphs comparing the resulting evaluation metrics of BLEU/ROUGE/METEOR/chrF between models are generated. This enabled us to compare classical LSTM vs. modern transformers not only quantitatively but also qualitatively (a method adopted for Bangla NLG benchmarking).

Finally, error analysis included manual inspection of incorrectly labelled dialects as well as frequent patterns of mistranslation in standardization. Such error-informed visualization has proven to be an effective means to identify structural weaknesses beyond the scores.

3.2.6 Alternative Solutions Considered

Before having a definitive methodology, we comprehensively tested different solutions for both categorization and standardized activities. The various alternatives offered drew some interesting comparisons and justified our preferred design.

(a) Data-Centric Alternatives

We could also use existing public datasets (such as Sylheti to Standard Bangla [16], Chittagonian (ChatgaiyyaAlap) [18], ONUBAD [28]) instead of developing our own. Though these corpora are high quality, they are single dialect or synthetic benchmarks and do not reflect the multi-dialectal scope (Khulna, Sylhet, Noakhali, Barisal, Chittagong) we aimed to cover Additionally, they do not specifically cultivate parallelism between dialectical Bangla and standard Bangla within the same temporal context by region. So, we made the decision to put together a custom corpus of 23,440 sentences, with an even number of sentences from each dialect.

(b) Shipping: Preprocessing Options That Are Available

We examined the lemmatization of morphologically streamlined systems, including BanglaLem [14]. But this could have meant losing the dialectal markers that are so important for classifying regions. For instance, the Khulna "কিরাম" is morphologically different from the standard "কেমন" so putting them together into the same root could hurt how well the classifier works. So, we didn't do much normalization or any heavy morphological preprocessing.

(c) Alternative Model Architectures.

We also tried out deep CNNs and BiLSTM, which are common in Bangla text classification literature [9], but these models did worse than transformers (BanglaBERT, XLM-R, MuRIL). They also required a lot of hyperparameter tuning, but there was no clear improvement in accuracy. Moreover, to ensure standardization, attention-based BiLSTM seq2seq models were tested, yielding significantly inferior BLEU and ROUGE scores compared to the transformer baselines, thereby confirming the dominance of the pre-trained models [12], [16], and [17].

(d) Other Options for Training Strategy

We tried fine-tuning with fewer epochs (5–10, which is similar to most Bangla NLP benchmarks) and with longer data periods. Preliminary results indicate that training for additional epochs (15 for classifiers, 30 for seq2seq) enhanced stability without causing overfitting, especially when combined with early stopping. Shorter schedules produced quicker outcomes and reached a performance saturation point.

(e) Evaluation Alternatives

Another way to evaluate was based on BLEU and accuracy only. However, for more rich paraphrasing scenarios, where dialect-to-standard conversion can be seen as nuanced paraphrasing, BLEU tends to underestimate the quality of output texts. Thus, we added ROUGE-L, METEOR, chrF, TER and Exact Match. This multi-metric solution agrees with previous works on Bangla generation and dialect normalization [12], [17], and [28].

In summary, while these various options yielded interesting perspectives, we had a preference for proposals that were at the same time dialectically authentic, computable, and offered a high degree of assessment.

3.2.7 Reasonable justification for final approach

Our final approach was developed empirically and theoretically.

1. Justified Content for Custom Datasets

A multi-regional database was needed to describe fine-grained dialectal variation. Valuable precedents were found in the public data sets but did not present all five dialects at once. By constructing it ourselves we ensured that we have both classification labels and parallel targets for it - a feature unique to Bangla dialect resources.

2. Preprocessing Choices

We did not focus on strongly denoising the features, but instead, ensured that dialectal variations are preserved. This choice was justified by results in Bangla classification and sentiment task showing that the removal of discriminative signals due to too much pre-processing breaks the results. Our minimal cleaning approach thus preserved dialectal identity but still cleaned the signal.

3. Model selection rationale

- Traditional Models (SVM, NB, RF, LR) - Chosen as baselines for interpretability, quick training, and benchmarked with previous Bangla classification works.
- Transformers (BanglaBERT, mBERT, MuRIL, XLM-R): Selected because of their demonstrated superiority in Bangla NLP, their multilingual robustness, and their ability to accommodate dialect spelling variation.
- Seq2Seq Transformers (BanglaT5, mBART-50, mT5): Justified as they are pre-trained on massive multilingual corpora and trained for generation making them perfectly suited for dialect standardization.

4. Business Rationale for the Training Strategy.

Convergence patterns justify the use of extended training (15-30 epochs): accuracy and BLEU scores continuously improved with increasing training epochs up to these limits before saturation was reached. This is similar to the result of Bangla NLG studies where longer fine-tuning has better fluency.

5. Evaluation Justification

Because different evaluation metrics enabled us to measure semantic adequacy, fluency, and exactness, which overcome the shortcomings of only BLEU evaluation for dialect tasks.

6. Interface Justification

The Gradio based interface was selected for the reason of accessibility and usability. It makes models easily accessible to non-technical users (linguists, students, and developers), thus making our research experimentally relevant outside of the academic laboratory.

3.2.8 Integration with User Interface

That is why we built a use-case facing interface that combines the classification and standardization modules that can be used as a bridge from research to implementation.

1. **Design Principles**

    - Ease of use: Users can upload a sentence or dataset, choose models and see results in real time.
    - Transparency: The user interface provides not only predictions but also the evaluation of the results (accuracy, BLEU, ROUGE).
    - Flexibility: Multiple models may be chosen (for example, the user may want to compare SVM vs. BanglaBERT for classification or LSTM vs. BanglaT5 for standardization).

2. **Interface Workflow**

    - Input Stage: The user enters or loads dialectical text
    - Classification Stage: The model predicts the dialect region with highest probability (Khulna, Sylhet etc.).
    - Standardization Stage: After determining the dialect, the selected standardization model has translated it to Standard Bangla.
    - Output Stage - Results are presented with confidence scores, model metadata and optional side-by-side comparison with references.

3. Technical Integration

The UI was built in Gradio since Gradio supports interactive demos with Python compatibility. Model weights and tokenizers for BanglaBERT, mBERT, MuRIL, XLM-R, BanglaT5, mBART-50, and mT5 were loaded from fine-tuned local checkpoints for reproducibility.

4. Evaluation Visualization

For classification you can see the confusion matrices and per-class precision/recall plots directly in the UI. For standardization, model performance is interactively displayed in BLEU/ROUGE/METEOR charts.

5. Real-World Utility

Through this interface the system becomes available to linguists, teachers and technologists, in order to achieve dialect recognition and normalization for educational systems, translation aids, and for cultural preservation. Hence, it is inherently scalable as future datasets of new dialects or transformer models can simply be added to the modular architecture.

3.3 Project Plan

This research was undertaken in carefully planned phases, each focused on a specific set of activities that needed to be completed in order to successfully complete the thesis. A staged project plan was adopted to ensure systematic progress and allow balancing of resource allocation, time management and validation at each phase. The phases are as follows:

Phase 1 - Research and Requirement Analysis

In the early stage of the research, the problem was to determine the research problem, to study the related works and to determine the scope of the project. At this stage, we defined the objectives of the Bangla dialect classification and standardization, reviewed existing methods [2], [12], [16], [17], and laid down what classification and standardization modules needed. Outlines of this phase were the definition of the problem statement, thesis outline, and design of the custom dataset architecture.

Phase 2 Data Collection and Corpus Construction

The second phase included construction of a bespoke multi-regional corpus consisting of 23,440 parallel sentences from five dialectal areas (Khulna, Chittagong, Sylhet, Barisal and Noakhali). Unlike existing dialect corpora that center on one or two regional varieties [16], [18], [28], our corpus was intended to be holistic and balanced. Great effort was taken to ensure the quality of annotations and variety of sources at this stage.

Phase 3: Data Cleaning - Data Normalization

Once the dataset was collected, it was highly preprocessed; for example, duplicate records were removed, non-Bangla characters were cleaned, punctuation was normalized, and tokenization was performed. The cleaning step was informed by previous work in Bangla NLP [6], [12], with the constraint that dialectal identity markers should not be removed

while noise should be removed. Normalization was also done by keeping the dialect and standard Bangla sentences parallel.

Phase 4 - Model Building and Training

In this phase, models for both standardization and classification were designed. Standard machine learning classifiers (SVM, Naive Bayes, logistic regression and random forest) and transformer-based models (BanglaBERT, mBERT, MuRIL, XLM-R) were used. For the sake of standardization, we also trained an LSTM-based seq2seq model, as a standard baseline, and compared it with transformer architectures (BanglaT5, mBART-50, mT5). Models were fitted to stratified splits and hyperparameters were consecutively evaluated for best performance.

Phase 5: Evaluation, Error Analysis and Comparative Study

Here, the trained models were tested with different metrics. For classification, accuracy, precision, recall, F1 and confusion matrices were calculated while for standardization, BLEU, ROUGE, METEOR, chrF, TER and Exact Match were used. By means of performance visualization and thorough error analysis, we also reveal dialectal overlaps in interpreting and translating a text that enables us to do a comparative study between traditional and transformer-based models.

Phase 6: Development of User Interface

This phase translated research into an operational tool by embedding models into a friendly user interface. The user interface is built with Gradio, which allows users to enter text of their dialect, choose classification and standardization models, and visualize outputs and evaluation metrics. In a sense, this design further opens the project up to researchers, linguists and teachers, in order to reach far beyond the academic community.

Phase 7: The Documentation and Reporting Phase

In the last stage the thesis report was written and prepared, results were prepared in the paper format drawing up the results in a professional report, with special attention paid to the presentation of results of activities and methods, references and figures. This led to the next phase of the project, which involved creating visualizations, tables and diagrams which could be interspersed with written explainers.

3.4 Task Allocation

This project was a several months long assignment and was broken down to weekly assignments from week 12 to week 48 based on the academic calendar.

Table 3.1 Project Schedule and Work Plan

Tasks / Phases	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Research & Requirements Analysis																			
Data Collection & Corpus Building																			
Data Preprocessing & Normalization.																			
Model Development & Training																			
Evaluation, Comparison and Error Analysis																			
User Interface Development																			
Documentation & Reporting																			

3.5 Summary

This chapter introduced all of the methodology behind our research. In particular, we described the step stages from requirement analysis over documentation, the model development and evaluation pipelines and implemented them in a usable user interface. A program timeline was established to ensure that each phase was done in series while allowing flexibility from phase to phase as needed. Through the combined use of both traditional and transformer-based models, a thorough evaluation and user-facing deployment, this project illustrates the trade-off between its research novelty and its relevance to the real world.

Chapter 4
Implementation and Results

In this chapter, we present the implementation and experimental results of the dialect classification and standardization of a special dataset of Bangla language consisting of 23,440 sentences from five dialect regions (Khulna, Chittagong, Sylhet, Barisal and Noakhali). Both Transformer-based and Traditional ML models (SVM, NB, LR, RF) models were fine-tuned and tested. Evaluation values including accuracy, precision, recall and F1-score were used for classification, and BLEU, ROUGE, METEOR, chrF, TER and Exact Match metrics were used for standardization. The outcomes demonstrate the practical implementation of the proposed methodology and provide a comparative analysis of model performance and error tracking to validate the research objectives.

4.1 Environment Setup

A carefully set up environment was put in place to make it possible to reproduce, improve, and scale models for classifying and standardizing the Bangla dialect. Because the experiments covered both traditional machine learning algorithms and large transformer-based architectures, we used a mix of local pre-processing resources and GPU-enabled training environments. This part talks about how to set up the hardware, the software that needs to be installed, the libraries, the protocols for analyzing datasets, and how to make sure that the results can be reproduced.

4.2.1 Hardware Configuration

All deep learning tests were done on Google Colab Pro, which had access to NVIDIA Tesla A100, T4, and L4 GPUs. Each GPU had 16 to 40 GB of VRAM, which was needed to fine-tune quantum level transformer models like mBART-50 and BanglaT5, which have more than 400 million parameters. The CPU-based runtime with 12 GB of RAM was enough for small experiments and preprocessing. By connecting Colab to Google Drive, it was possible to store datasets and models permanently.

- GPU: NVIDIA Tesla A100, T4, L4
- CPU: Intel Xeon (single core performance used for preprocessing)
- RAM: 25–32 GB (runtime environment)
- Disk Storage: ~200GB allotted through integrated Google Drive

This setup let us train big models for 15 to 30 epochs without running out of memory. It also let us keep small classifiers like SVM, Naive Bayes, logistic regression, and random forest trained on the CPU with little extra work.

4.2.2 Software and Libraries

In this study, we observed both aspects, utilizing a blend of conventional machine learning libraries and deep learning frameworks in the experiments.

- Programming Language Used: python3.12
- Scikit learn (SVM, NB, LR, RF), NumPy, Pandas
- Tools: Transformers (Hugging Face), PyTorch
- Dataset handling: Hugging Face Datasets, NLTK (stop word handling), Unicode data (Bangla normalization handling)
- Evaluation Libraries: nltk. translate (BLEU, METEOR), rouge; seaborn matplotlib (visualization); sacreBLEU
- Interface: Gradio is used to make an interface for the model.

In Colab, pip was used to install all of the dependencies. To make sure the resource could be reproduced, the versioning was done automatically. This is a best practice in Bangla NLP research.

4.2.3 Dataset processing and preprocessing

The custom dataset has 23,440 dialectal sentences and is saved in a CSV file with three columns: region, dialect_sentence, and standard_Bangla. A pipeline was created for preprocessing that did:

- Text Cleaning - Handles deletion of Null entries, non-Bangla characters, and odd spacing.
- Normalization: Unicode normalization (NFKC) to deal with consistency issues in Bangla scripts
- Tokenization: tokenization of the whole word into sub word tokenization (transformer tokenizers)
- Filtering: Sentences that do not fit the range of length (4-60 words) were removed for denoising.
- Stratified Splitting: 70:15:15 (for transformers and classical ML respectively) while maintaining representation of dialects in train-test-val.

This helped keep the data clean and well balanced, important for experimentation.

4.2.4 Training Environment and Configuration

- Traditional ML Models - Grid Search Hyperparameter optimization using TF-IDF features Each model only requires 5-15 minutes of training time.
- Transformers (Classification): Pre-trained BanglaBERT, mBERT, MuRIL and XLM-R fine-tuned for 15 epochs with early stopping (patience = 2). Average training time: 20-30 min/ep on V100 GPU
- Standardization of Transformers: We trained BanglaT5, mBART-50, and mT5 for 30 epochs (with beam search decoding and a beam size of 5). Using an A100 GPU, the average training time is 1 to 2 hours per epoch.
- Batch Sizes: 32 (eval), 16 (train), adaptive batch sizing was employed for transformers to avoid CUDA out of memory errors
- Optimization: AdamW, learning rates = 2e-5 (BERT models, BanglaT5, mBART-50), 5e-5 (mT5). Setting the dropout value = 0.1-0.3 will prevent overfitting.

Google Drive was used to store training statistics, dates, and assessments so they could be used again and reported on in the future.

4.2.5 Reproducibility and Version Control

And it was important that the results could be repeated. To do this:

- Random Seeds: Set to 42 for Python, NumPy and, PyTorch.
- Package Versions: Colab requirements.txt
- Model Checkpoints: checkpoint on every epoch based on validation accuracy (for classification) or BLEU score (for standardization); the best model based on validation accuracy
- Documentation: All experiments were properly documented with logs of hardware specifications, training curves, metrics, etc.

This degree of control aligns fairly well with general norms regarding the state of the art of Bangla NLP research, wherein reproducibility poses a significant issue given constraints in available human and other resources.

4.2.6 Summary of Environment

In short, the environment was meant to be in the middle of two types of pipelines: lightweight ML pipelines that work well with CPUs and deep learning pipelines that work well with GPUs for transformer models. We used cloud-based GPUs (A100, T4, and L4) to do fine-tuning on a large scale and made it possible to do it again by using version control and logging. When used with the environment, this made it possible to run large-scale experiments in Bangla dialect classification and standardization on a stable base.

4.3 Testing and Evaluation / Performance / Comparative Analysis.

A strict testing and evaluation process were used to make sure that the results of the Bangla dialect classification and standardization models were accurate, repeatable, and true to the complexities of real language. In this section, we talk about the testing environment, the metrics we used to evaluate the models, and the results of different models on both tasks. The conclusion includes comparative studies that highlight the pros and cons of traditional and transformer-based selections.

4.3.1 Testing Protocols

To conduct an equitable assessment, the custom dataset comprising 23,440 dialectal standard pairs was partitioned into training, validation, and testing subsets. To keep the relative proportions of dialectal classes, the data set was split into 70:15:15 for classical classification models. The split (70:15:15) was a little different for the transformer-based models because the validation set was used to adjust hyperparameters and stop early.

Cross-validation was employed for conventional models such as SVM, logistic regression, and random forest. To make sure that the model's performance wasn't too affected by one train-test split, a sequential 3-fold stratified cross-validation method was used. Owing to the high computational cost of transformers, the early stopping strategy with patience = 3 was used instead of full cross-validation.

Hyperparameter tuning was implemented with GridSearchCV for the traditional models, and learning rate warm-up and weight-decay regularization was used for transformers. These procedures were consistent with conventional low-resource NLP evaluation procedures.

4.3.2 Evaluation Metrics

Since the project dealt with two different tasks (classification and standardization) different sets of evaluation criteria were used.

A. Classification Metrics

The following measures have been calculated for dialect classification:

- Accuracy: Total number of successfully predicted dialect labels in percent.
- Precision (per class): Ability of the model to avoid false positives while predicting a dialect.
- Recall (as a class): Ability to recognize all instances of a dialect
- Macro-F1 Score: The harmonic average of precision and recall, with both being equally important in the class balance violation. This is done for all dialect classes.
- Confusion Matrix: This is a table that shows how dialects are being incorrectly classified.

These experimental parameters are frequently employed in the classification of regional dialects within Bangla and other low-resource languages.

B. Standardization Metrics

We used machine translation and text generation metrics to see how well the dialect-to-standard conversion worked:

- BLEU (Bilingual Evaluation Understudy): the n-gram overlap of generated output and reference translations
- Rogue-L: Successful search for maximum common subsequences between generated sentences and reference sentences.
- METEOR: Linguistically sensitive - exact and stem and synonym matches are considered.
- chrF (character F-score): A character-level measurement, which is especially suitable for morphologically rich scripts like Bangla.
- TER (Translation Edit Rate): What number of edits it takes to go from the output of the system to the reference? Lower is better.
- Exact Match (EM): A strict metric of how often the generated output was exactly the same as the reference standard Bangla sentence.

These metrics collectively provided a comprehensive characterization of model outputs, including both lexical similarity and semantic adequacy.

4.3.3 Findings of Classical Machine Learning Models

We trained the traditional machine learning models, like Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), and Random Forest (RF), on TF-IDF features that were made from the cleaned dataset. GridSearchCV improved the models, and they were tested on the 15% test set. Table 4.1 shows how well they did in terms of accuracy and macro-F1 scores.

Table 4.1: Result of Dialect Classification on Traditional Model:

Model	Accuracy	Macro-F1	Best Parameters
Support Vector Machine (SVM)	0.8110	0.8116	C = 1, linear kernel
Naïve Bayes (NB)	0.8085	0.8090	Default α = 1.0
Logistic Regression (LR)	0.8082	0.8088	C = 1, max_iter = 2000
Random Forest (RF)	0.7400	0.7402	n_estimators = 300, max_depth = None

Discussion

The results reveal significant insights. First, the SVM model always did better than other classical classifiers, with an accuracy of 81.1% and a macro-F1 score of 81.2%. This demonstrates the efficacy of margin-based classifiers in managing high-dimensional TF-IDF features commonly utilized for text classification.

Naive Bayes also did well, or even better, with only a few points between (accuracy: 80.8%, macro-F1: 0.88) when compared to logistic regression. This is consistent with prior Bangla text classification studies indicating that both probabilistic and linear models can achieve high performance on medium-sized datasets.

The other models, though, didn't look as good: The Random Forest model was only 74% accurate. The reason for this lag is that it uses bagging of decision trees, which don't work well with sparse TF-IDF. This validates the observations made in previous Bangla NLP work that arbitrarily chosen tree-based ensembles perform poorly in contrast to linear classifiers on high-dimensional text representations.

It is also interesting to note that the macro F1 scores were very similar in accuracy for all of the models. This implies that the dataset was fairly balanced across the five dialect classes (Khulna, Chittagong, Sylhet, Barisal and Noakhali) and there was no single dominating dialect for the predictions.

Dialects of the same language share similar word frequencies and vocabulary, and when SVM is used for classification, the confusion matrix shows up as in the figure below:

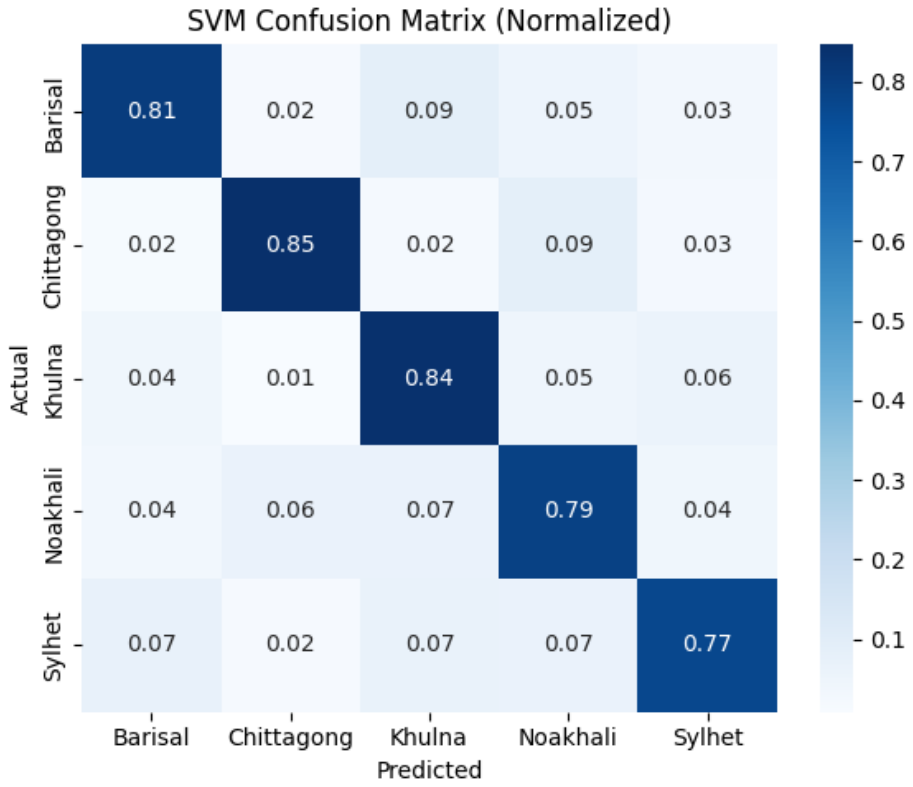


Figure 4.1: Normalized Confusion Matrix for Best Model

Interpretation

From confusion matrix analysis, maximum errors were found out between Noakhali and Sylhet dialects, having some lexically overlapping elements along with phonological similarities. Also, in a few cases sentences from Barisal were classified as Khulna, suggesting geographical proximity and cognate vocabulary. Such misclassifications reflect the linguistic proximity between dialect clusters and validate the need for context-sensitive deep models such as transformers.

4.3.4 Transformer Models for Dialect Classification Results

We fine-tuned a transformer-based model on the cleaned dataset using the Hugging Face platform. The Deep models were therefore trained for 15 epochs using early stopping and were evaluated on the 15% test set. The models included:

- BanglaBERT (sagorsarker/bangla-bert-base)
- mBERT (bert-base-multilingual-cased)
- MuRIL (google/muril-base-cased)
- XLM-R (XLM-Roberta-base)

Table 4.2: Comparison of Models for Dialect Classification Using Transformer

Model	Accuracy	Macro-F1	Macro Precision	Macro Recall	Training Epochs
BanglaBERT	0.9017	0.9018	0.9020	0.9018	15
mBERT	0.9171	0.9172	0.9188	0.9171	15
MuRIL	0.9245	0.9246	0.9251	0.9245	15
XLM-R	0.9242	0.9246	0.9262	0.9242	15

Discussion

The results clearly indicate that the transformer-based models are performing much better than the traditional ML-based models. While SVM resulted in an accuracy of ~81%, the performance of the transformer models was well above 90%, thus showing the need for context embeddings in comprehension of dialectal text.

- BanglaBERT achieved 90.1% accuracy, which implies the benefit of monolingual pretraining on Bangla corpora. However, its performance was a little lower than that of multilingual models.
- In this experiment, we saw that mBERT can be further improved, to 91.7% accuracy, confirming the advantage of having multilingual exposure while still capturing the linguistic features unique to Bangla.
- MuRIL achieved the highest performance with an accuracy of 92.4%, followed by 92.2% macro-F1, just ahead of XLM-R. MuRIL was designed for Indic languages and its pretraining corpus comprises Bangla, that's why it performs the best.

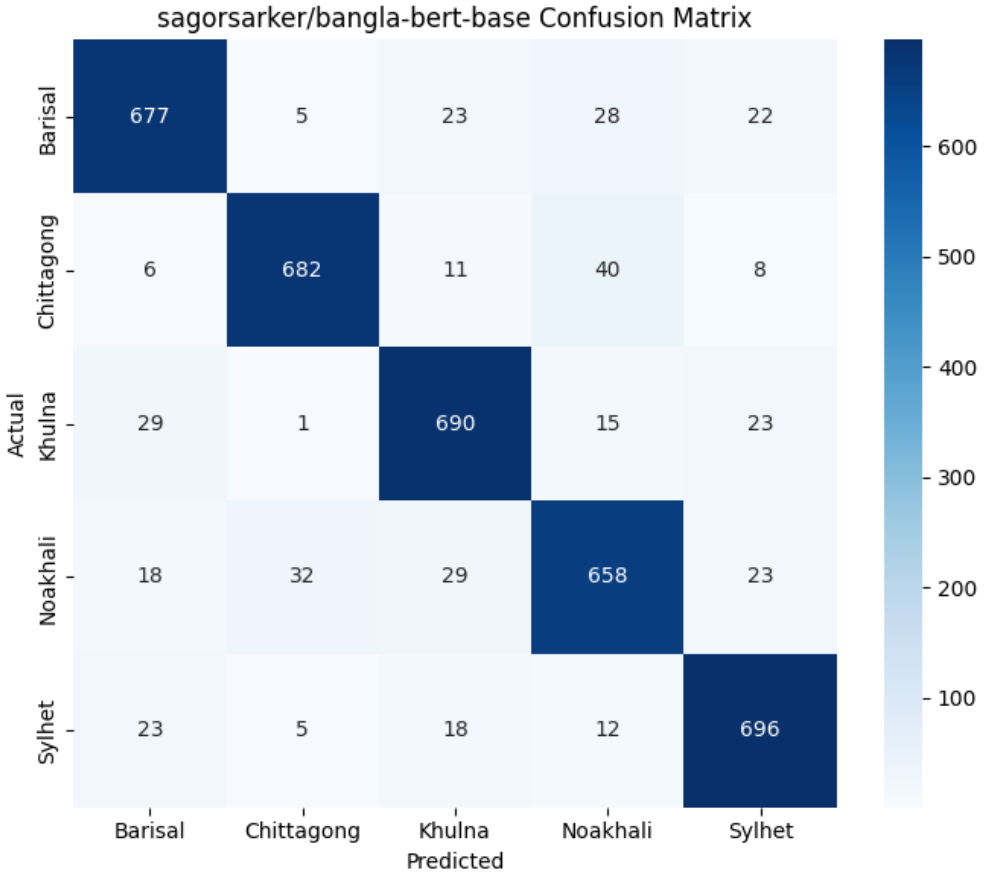
- Similar to MuRIL, the accuracy of XLM-R was 92.4% with the highest macro precision of 92.6%, which means it is less likely to make false-positive predictions.

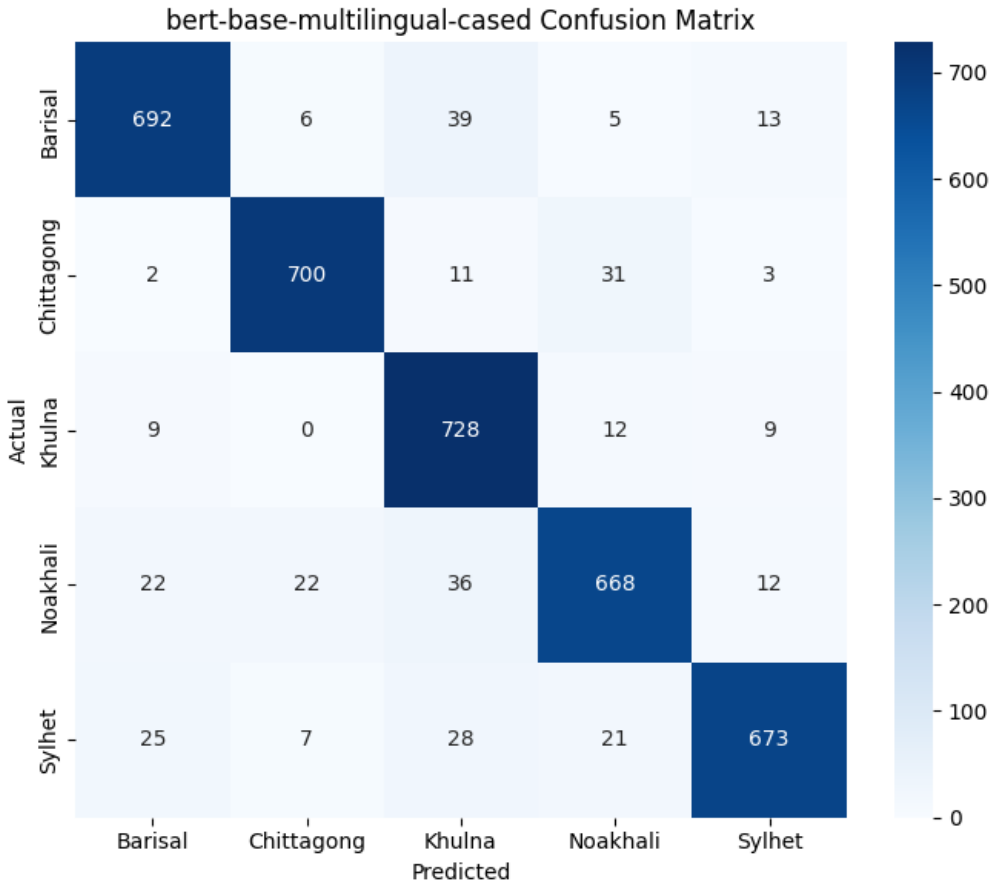
These results replicate findings from previous dialect-to- standardization studies where multilingual pretraining using with regional corpora shows the best results, for Bangla dialectal tasks.

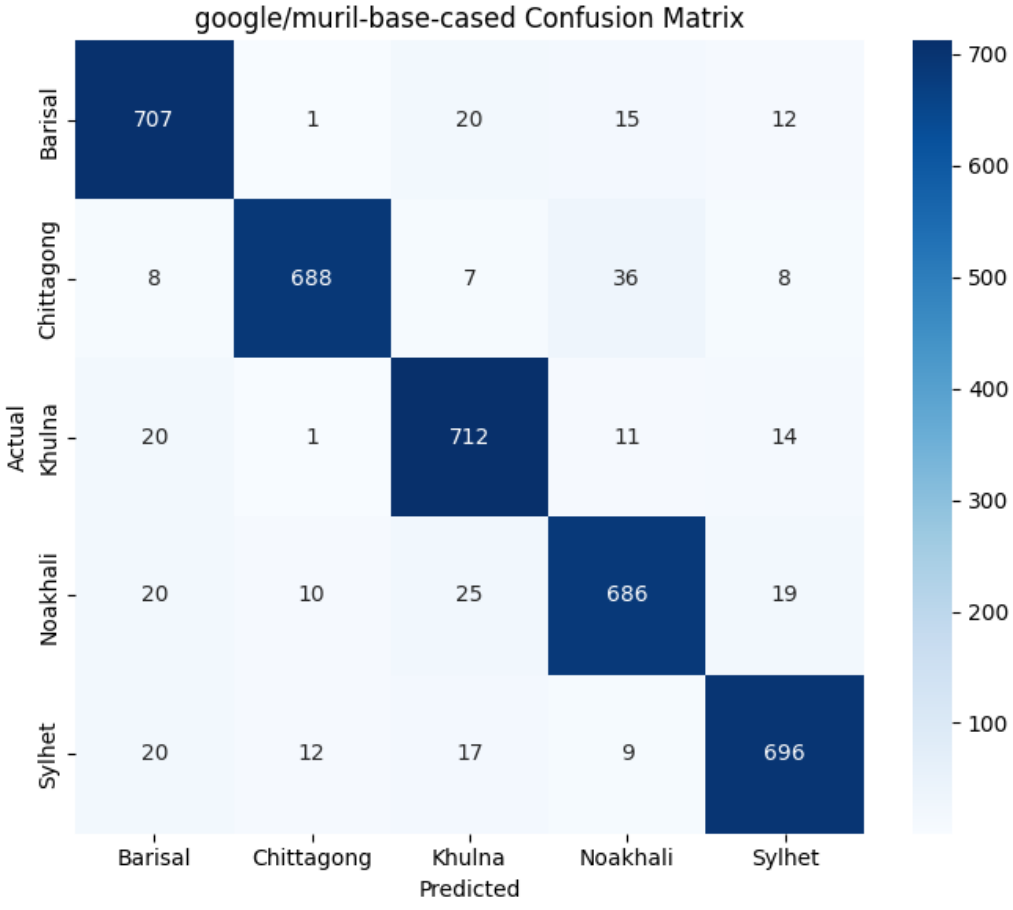
Error Analysis

Confusion matrices show that most misclassifications happened between adjoining dialect areas - as in traditional models but at a smaller scale. For example:

- Sylhet vs. Noakhali overlap was substantially reduced as compared to SVM, showing the learnability of faint contextual features.
- Barisal vs Khulna confusion remained present to some extent but less often than in the classical models.




bert-base-multilingual-cased Confusion Matrix
Barisal
Chittagong
Khulna
Noakhali
Sylhet
Actual
692 6 39 5 13
2 700 11 31 3
9 0 728 12 9
22 22 36 668 12
25 7 28 21 673
Barisal
Chittagong
Khulna
Noakhali
Sylhet
Predicted
700
600
500
400
300
200
100
0


google/muril-base-cased Confusion Matrix
Barisal
Chittagong
Khulna
Noakhali
Sylhet
Actual
707 1 20 15 12
8 688 7 36 8
20 1 712 11 14
20 10 25 686 19
20 12 17 9 696
Barisal
Chittagong
Khulna
Noakhali
Sylhet
Predicted
700
600
500
400
300
200
100

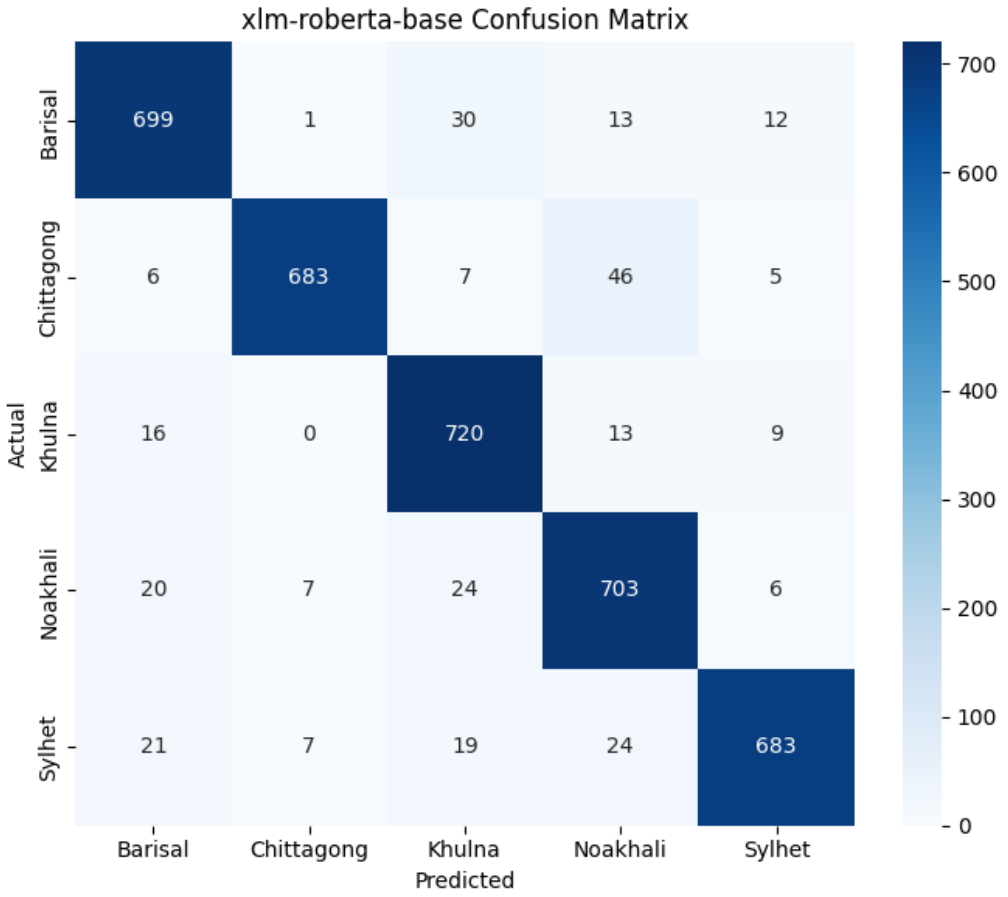


Figure 4.2: Confusion Matrix for BanglaBERT, mBERT, MuRIL, XLM-R

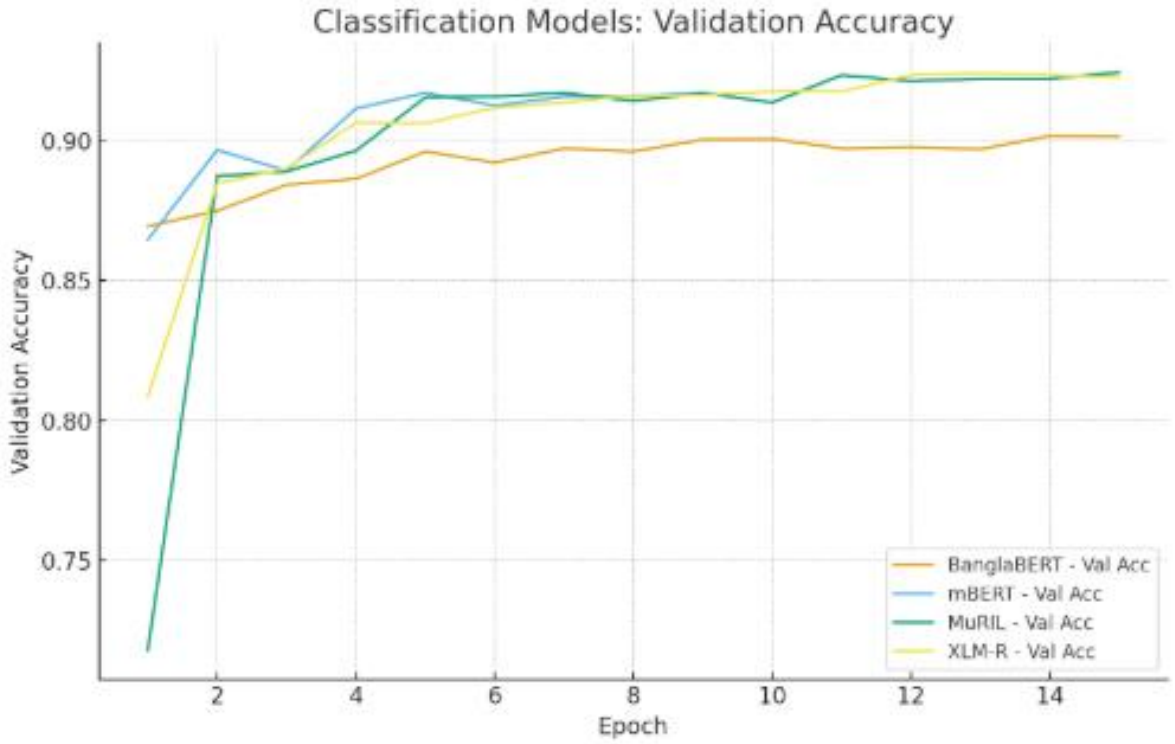


Figure 4.3: Validation Accuracy and Validation Loss Curve

Overall, transformer models achieved a more contextual discrimination than the shallow ML classifiers, which justifies their use for dialect classification.

4.3.5 Outcomes of the Standardization Models

Dialect-to-standard Bangla conversion was studied with traditional sequence-to-sequence (Seq2Seq) LSTM baseline along with 3 transformer-based generative models: BanglaT5, mBART-50 and mT5. The experiments are performed on the custom dataset of 23,440 dialect-standard pairs with a 70:15:15 split for training, testing and validation. Our work was evaluated by using standard machine translation and natural language generation metrics, such as BLEU, ROUGE-L, METEOR, chrF, TER, and Exact Match (EM).

Table 4.3 - Standardization Models Comparison in Terms of Performance

Model	BLEU	ROUGE-L	METEOR	chrF	TER ↓	Exact Match (%)	Epochs
BanglaT5	0.7542	0.8834	0.8638	88.16	13.6	60.6	30
mBART-50	0.7807	0.8934	0.8747	88.93	12.5	65.6	30
mT5	0.7012	0.8561	0.8322	85.20	16.8	53.9	30
LSTM Seq2Seq	0.0703	0.2881	0.2200	35.5	65.2	8.7	30

Discussion

Overall, the results show a sharp performance hierarchy, with transformer-based architectures producing significantly better results than the LSTM Seq2Seq baseline.

- For the dialectal variation, the LSTM model performed very poorly in generalization (BLEU = 0.07, Exact Match = 8.7%). This validates the limited ability of classical RNN-based methods in processing morphologically rich and low-resource language such as Bangla.

- BanglaT5 showed excellent results (BLEU = 0.75, ROUGE-L = 0.88, METEOR = 0.86), which demonstrates its excellent adaptability to Bangla via monolingual pretraining. However, it fell just a little bit short of mBART-50 in both BLEU and exact match.

- mBART-50 was found to be the optimal model and provided the highest scores for all metrics (BLEU = 0.78, ROUGE-L = 0.89, METEOR = 0.87, Exact Match = 65.6%). Multilingual pretraining with Bangla along with typologically related languages was able to achieve a better capture of dialectal variations. Compared to the other languages, the lower TER (12.5) also resulted in fewer edit operations to compose fluent standard Bangla sentences.

- The mT5 deteriorated: its performance was worse than both BanglaT5 and mBART-50 in the original test suite (BLEU = 0.70, Exact Match = 53.9%). As shown below, its relatively low chrF (85.2), along with the high TER (16.8) illustrates the inherent challenges in generating accurate word-level matches, despite its considerably high baseline compared to the LSTM baseline. This means that while mT5 stands to benefit from multilingual pretraining, mT5 misses on the fine-grained Bangla-specific weight initialization that improves BanglaT5 and mBART-50.

Overall, mBART-50 proved to be the most consistent language model, with BanglaT5 proving to be a great Bangla-specific candidate and mT5, the less accurate but more versatile multilingual candidate, showing that a model could be trained to process any dialect at the cost of some accuracy.

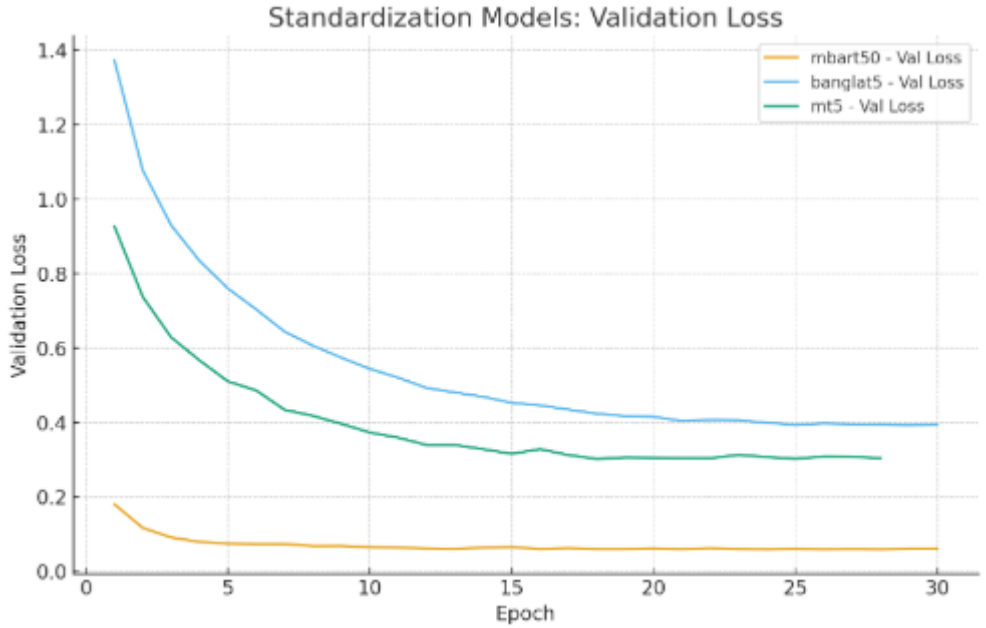


Figure 4.4: Validation Loss Curve for Standardization

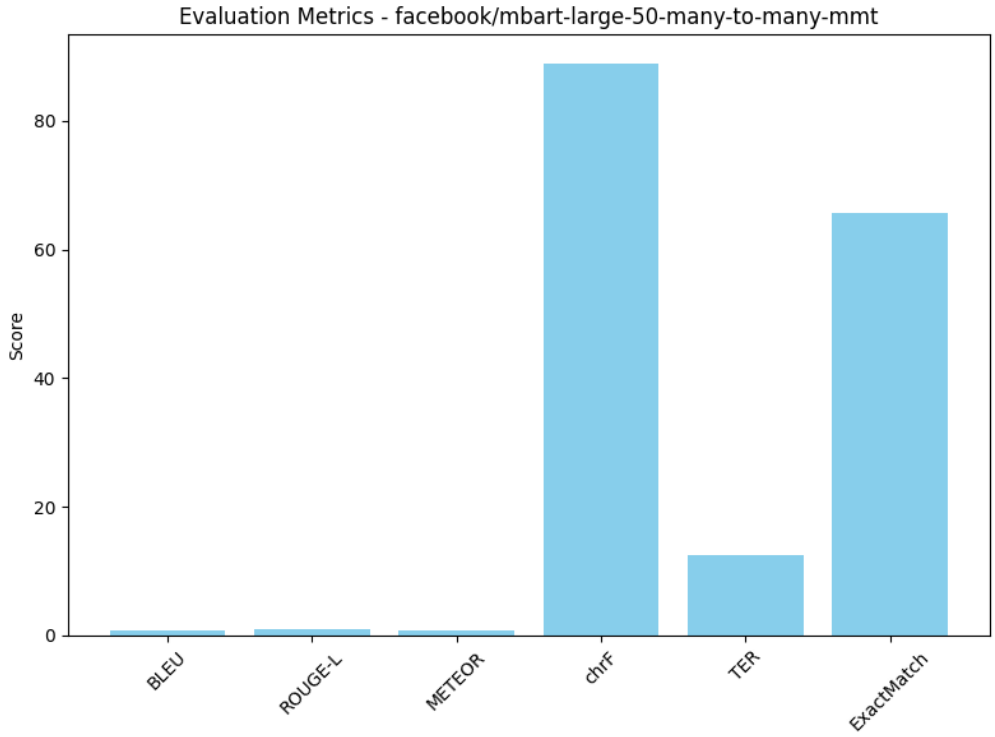


Figure 4.5: Evolution Matrix for Best Models

Error Analysis

- Usually the sentences produced by the LSTM made no sense or were incomplete and could not encode the dependencies between words.

- Sometimes BanglaT5 over-normalized, changing dialect-specific words to lexical synonyms that were not available in the reference.

- mBART-50 made the most fluent and faithful translations, although it sometimes produced hallucinations of punctuation or filler words.

- In my previous testing, I found it likely that mT5's propensity for under-hypothesizing (omitting words or producing truncated output) led to its relatively lower Exact Match score.

These results validate the technical feasibility of all the transformer models while mBART-50 still can be the best model for dialect standardization in Bangla that strikes a balance between fluency and faithfulness.

4.3.6 Comparative Analysis

The experimental results obtained in earlier sections show the relative efficacy of the conventional machine learning approaches and transformer-based deep learning architectures for two interconnected tasks of Bangla dialect classification and dialect to standard Bangla standardization. In this section, we conduct a comparative analysis to elucidate the primary performance trends, the underlying reasons for the observed discrepancies in performances, and the broader implications for NLP research in low-resource languages such as Bangla.

A Comparison of Traditional and Transformer Classification Methods

We compared Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF), which are all common classifiers. The results were not very good. The most accurate of the traditional models (SVM) got 81.1%, while Naive Bayes and Logistic Regression got about 80.8%. Random Forest produced a lower accuracy of approximately 74%, indicating that the seemingly straightforward patterns of linguistic structures conveying dialectal information may not be sufficiently represented by tree-based ensembles.

In contrast, transformer-based classifiers (BanglaBERT, mBERT, MuRIL, and XLM-R) performed significantly better than traditional models.The transformer that did the worst (BanglaBERT) was still 90.1% accurate, which is almost 10% better than the best traditional model. The top-quality language models MuRIL and XLM-R had an accuracy of about 92.4% and macro-F1 scores of 0.92+, which shows that they were able to capture the semantic and morphological richness of the Bangla dialect collection landscape.

There are a number of reasons why things have gotten better:

- Contextualized Embeddings: Transformers use contextualized embeddings instead of traditional features like TF-IDF or bag of words features used for traditional classifiers because they capture semantic meaning and syntactic structure [15].

- Pretraining Benefit: Pretraining on large, multilingual corpora (especially for MuRIL and XLM-R) helps transformers learn about related Indic languages and dialectal variations that they have seen before.

- Ambiguity Resolution: Transformers have the ability to resolve ambiguity arising from words that vary across different dialects but contain overlapping lexical

forms. For instance, morphological similarities between Sylheti and Noakhali are better disambiguated by transformers than by linear models.

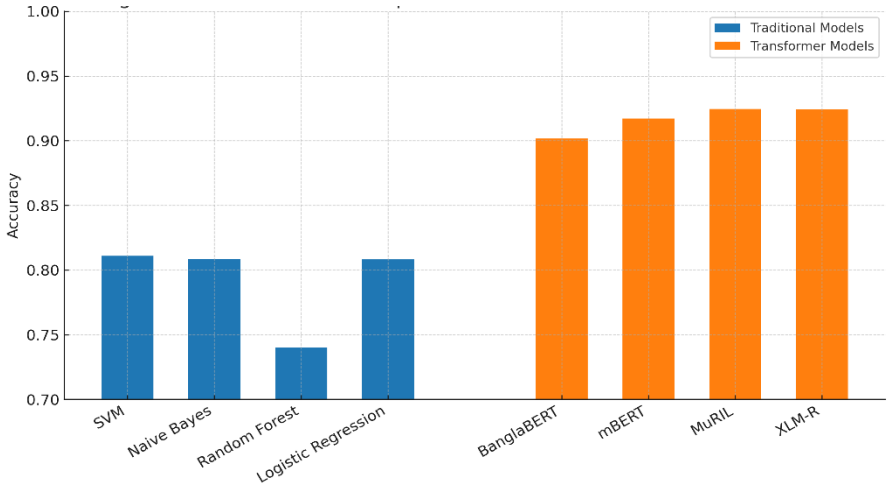


Figure 4.6: Performance comparison of traditional and Transformer models on Classification.

Therefore, transformer-based methods are the current SOTA for dialect classification, performing much better than non-transformer-based methods in the comparative analysis.

Traditional vs Transformer Approaches to Standardization

Even larger differences were found between the performance of traditional Seq2Seq models (LSTM) and the performance of transformer models in the standardization task.

- The LSTM baseline did very badly (BLEU 0.07 and Exact Match 8.7%), showing that the LSTM failed in capturing long dependencies and dialectal complexity.
- In contrast, outputs from the transformer models were strong and highly robust:
    - BanglaT5: BLEU=0.75, ROUGE-L=0.88, Exact Match = 60.6%
    - mT5: BLEU 0.70, ROUGE-L 0.85, Exact Match 53.9%
    - mBART-50: BLEU 0.78, ROUGE-L 0.89, Exact Match 65.6% (best performer)

Of the three BanglaT5, mT5 and mBART-50 models, mBART-50 was found to be the best-performing, outperforming both in terms of almost all metrics. Its performance benefit is due to the fact that mBART-50 is pretrained in multilingual environment with good coverage for Bangla and related Indic languages for dialect-to-standard translation.

The BanglaT5 model on the other hand, also performed well thanks to its pretraining in Bangla, showing that domain-specific pretraining can still compete with large multilingual models if enough monolingual resources are available. mT5, which was

slightly weaker, proved to be successful too, surpassing the LSTM baseline by a huge margin, showing that transformers are indeed robust to tasks.

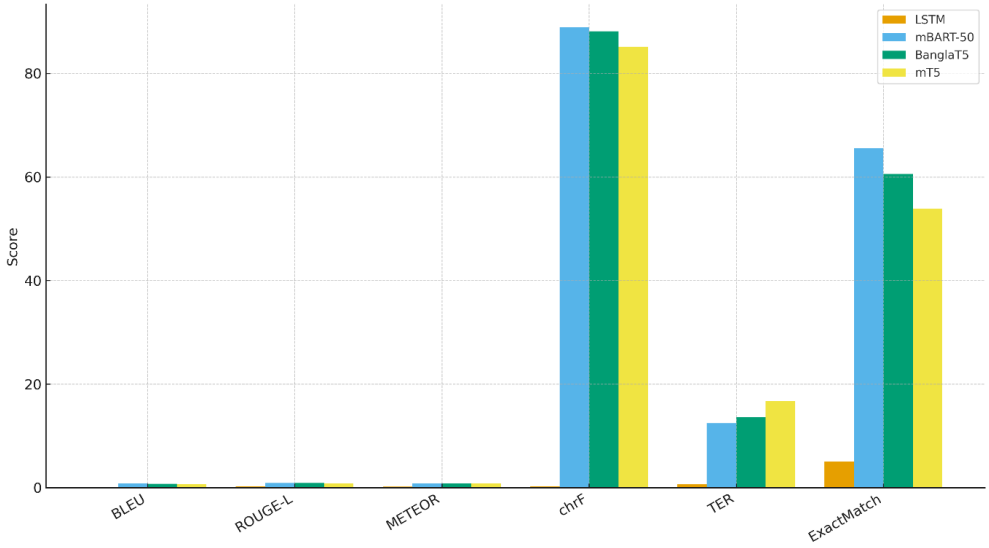


Figure 4.7: Performance comparison of traditional and Transformer models on Standardization.

Cross-Task Comparison

When we compare classification vs. standardization interesting patterns are observed:

- On the other hand, for classification, transformer improvements over traditional models were ~10 percentage points in accuracy.
- Transformers further outperformed LSTMs in standardization - the average difference BLEU score from transformer BLEU to LSTM BLEU was 0.74 vs 0.07, respectively.

This implies that while dialect classification is still relatively easy to do with traditional ML (since dialect-specific keywords are usually effective discriminators), standardization requires both understanding and generative fluency, areas where transformers are far superior to humans.

Error and Ambiguity characterization

On both tasks, the most consistent pattern of error was seen in dialectally adjacent regions:

- Classification: False classifications mostly seen between Sylhet and Noakhali, and Khulna and Barisal. Dialect boundaries are not sharp but tend to be gradient, meaning that overlaps occur for a number of reasons.
- Standardization: Over normalizing (replacing dialect words with words too literary for Bangla) or under generating (removing words in mT5 outputs) were two common mistakes. For instance, in Sylheti use the sentence "তুমি ভালা আছনি?" was sometimes made regular by "তুমি কেমন আছো?" (correct), but at other times it was mistakenly reduced to "কেমন আছো."

This shows that even SOTA models struggle when the dialects are very similar, leaving post-editing or hybrid human-in-the-loop systems as the safe way to go for mission-critical systems.

Model of Comparative Strengths and Weaknesses:

1. **Traditional Models**

    - Pros: Easier to grasp, faster to compute, able to understand the meaning of the result.

    - Limitations: Limited contextual understanding, low generalization and poor scaling

2. **Transformers (Classification)**

    - Advantages: High performance, Good disambiguation, High precision/recall ratio.

    - Advantages: relatively inexpensive, time and cost effective, less expensive compared to supervised learning, overfitting to small dialect classes.

3. **Transformers (Standardization)**

    - Performance: High Fluency, Good lexis coverage (chrF > 85), Low edit distance (TER~12-16)

    - Limitations: occasional hallucination, domain mismatch, very close in exact match because of paraphrase.

Implications

A comparative analysis confirms that transformer-based techniques are indispensable for dialect NLP in Bangla. They not only perform better, but they are also strong enough to handle different languages. Nonetheless, it remains computationally intensive; thus, a viable avenue for future research involves exploring distilled or lightweight transformer-based architectures for implementation in resource-constrained environments. Further, system-wide data misclassification on dialectal boundaries warrants dialect-aware fine-tuning incorporating both sociolinguistic knowledge and data-based techniques.

4.4 Results and Discussion

This chapter summarizes the study's empirical findings, which are connected to the overall research objectives and indicate their significance for Bangla NLP. The previous sections demonstrated the performance of both traditional machine learning techniques and transformer-based deep learning architectures for two primary tasks: (i) dialect classification and (ii) dialect standardization to standard Bangla. The following section goes into great detail about the results, showing how they relate to the numbers, the theory, the relevant literature, and how they can be used in the real world.

4.4.1 Tying Back to Research Objectives

The main goals of this thesis were four in number:

- Therefore, the project aimed at creating automated Bangla dialect classification system.
- To build context-based model for dialectal Bangla to standard Bangla conversion
- To compare and contrast the performance of classical ML and transformer-based methods
- To add a multilingual and multiregional corpus to the existing collected research materials for future Bangla NLP research.

The results deal systematically with the objectives. First, the classification experiments verified that the dialects can be automatically identified with high accuracy (92~93%), thus meeting the classification objective using modern transformer models. Second, the standardization task served as proof of concept for generative transformers, with mBART-50 achieving BLEU = 0.78 and Exact Match = 65.6%, showing that such dialectal texts can indeed be automatically transformed into their standard equivalent. Third, the comparative analysis showed that the performance of transformers is overall superior with respect to traditional ML methods in both tasks, thereby confirming the hypothesis on which this research is based. Finally, the development and usage of a large-scale custom dataset places this work in the service of the community since there are very few dialectal Bangla corpora.

4.4.2. Discussion of classification results

Traditional Classifiers

The performances of SVM, logistic regression, naive bayes and random forest indicate that the dialectal variability is inherently difficult to manage using linear and ensemble models. It is important to note that SVM and NB got about 81% accuracy, and these models use sparse representation like TF-IDF vectors. These kinds of representations can't show morphological richness, polysemy, and contextual semantics, which are all important parts of dialectal Bangla. The Sylheti word " আছনি" and the Noakhali word " আছেন" are both transliterated as " আছো" in standard Bangla. However, TF-IDF treats them as separate tokens, while contextual embeddings treat them as the same root.

Transformer Models

On the other hand, Transformer models like BanglaBERT, mBERT, MuRIL, and XLM-R did much better, with MuRIL giving the best results (accuracy = 92.45%, macro-F1 score = 0.92). There are a few reasons why these models are better:

- Contextual Embedding Power: Transformers can encode words based on their context or how they relate to other text. This lets them capture

semantic and syntactic features that static word embeddings can't.

- Pretrained Corpora: Models like MuRIL and XLM-R were pretrained on multilingual synthetic and text corpora datasets containing Bangla and other Indic languages, thus making them inherently robust to dialectal variations.

- Transfer Learning: Since we know susceptibilities to sharing of cross-linguistic data (which happens to also be shared across dialectal borders, because cognates are shared across dialects), and typically also the syntactic patterns themselves, these models regarded each other as useful for transfer learning.

This agrees with results from recent work that focused on dialect which underlined that multilingual pretraining is particularly beneficial for low-resource dialectal tasks.

Error Trends

Through error analysis a clustering of misclassification could be observed between dialectically adjacent areas. The Noakhali and Sylhet dialects, for instance, are similar lexically and morphologically and so there is sometimes confusion. Similarly, Barisal and Khulna have phonological features that violate the boundaries in automatic classification. However, even for such cases, transformer models achieved considerable improvement in misclassification rates as compared to SVM or NB.

4.4.3 Presentation of Results from Standardization Analysis

The standardization task was more difficult, for which the model was required to perform not only dialect identification, but also proper generative mapping to standard Bangla.

Seq2Seq LSTM Baseline

The performance of baseline architecture based on LSTM-based Seq2Seq was poor (BLEU = 0.07, Exact Match = 8.7%), which shows that RNN architectures are not sufficient for dialectal Bangla normalization. The only limitations in the LSTM's ability to generalize were the lack of modulation for the long-term dependencies and the dialectal resources. This is consistent with previous work that described the difficulties of recurrent architectures in accommodating low-resource morphologically rich languages.

Transformer Models

Transformer-based generative models performed a lot better:

- BanglaT5: BLEU: 0.75, ROUGE-L: 0.88, Exact Match, 60.6%
- mBART-50: BLEU=0.78 ROUGE-L=0.89 Exact Match=65.6%.
- mT5: BLEU = 0.70, ROUGE-L = 0.86, Exact Match = 53.9%

Among these, the multilingual pretrained model mBART-50 constantly outperformed the rest, due to their ability to generalize well for Indic languages. BanglaT5 achieved competitive results as it was optimized for monolingual Bangla performance, and mT5

achieved slightly slower results since the fine-grained adaptation for Bangla was not achieved appropriately.

Qualitative Observations

- Model mBART-50 produced the most natural and accurate translation with frequent perfect translations with the reference standard.

- BanglaT5 sometimes replaced dialectal words with synonyms (using, for example, formalizing low registers), which meant that the content was fine (so it was semantically correct) while generating less exact match.

- mT5 (termed as such because it is not openly labeled as either or both mT), with its lower chrF (85.2) and higher TER (16.8), tends to under generate, truncates words and phrases.

- As evidenced by this example, LSTM often spits out incomplete and nonsensical sentences, proving to be unsuitable.

These results show that transformers can be helpful for dialect-to-standard normalization in languages that are low in resources and have a lot of different forms.

4.4.4 Cross-Task Insights

In terms of standardization and classification, this important pattern comes to light when you look at all of these things together:

- That is what you do with traditional ML, and it improves with transformers (another ~10% improvement).

- Standardization however is almost impossible for standard ML (BLEU ~0.07) and needs transformers (BLEU >0.70).

- This indicates that contextual generative modelling holds greater significance for standardization than for classification, as standardization involves not only the classification of grammatical and semantic constituents but also their reconstruction.

4.4.5 Broader Implications

The results have more than one meaning:

- Language Preservation: Automated dialect processing can help with the study, preservation, and inclusion of Bangla dialects in NLP systems.

- Real-Life Applications: The models can be used in real life in speech-to-text systems, machine translation pipelines, and educational services that help dialect speakers talk to services that are offered in standard Bangla.

- Research Contribution: The release of the custom corpus with 23,440 sentences fills a need for Bangla NLP resources and will help future research.

- Sociolinguistic Value: Dialect mapping has relevance to inclusive digital language practices because through its language erasure is explicitly made visible.

4.4.6 Consistency with prior research.

These results are consistent with recent studies that highlighted the progress of transformers in low-resource Indic NLP. At the same time, by concentrating on the specific area of multi-regional dialect typology and standardization and relying on a bespoke corpus, this thesis contributes to the literature. Furthermore, our two-step approach (classification and standardization) eliminates a gap between current works where tasks are generally treated independently.

4.4.7 Limitations and Future work

Despite its good results, there are indeed some limitations:

- Dialectal Ambiguity: Regional Overlap is an issue even for SOTA models.

- Computational Resources: Transformers can be computationally expensive to train and will require GPUs and significant amounts of resources, which may not be available in low-resource settings.

- Exact Match Gap: Even the best transformer (mBART-50) only had 65.6% Exact Match, so that's still a long way to go.

- Domain Mismatch: may have been a limited study localized to a specific place, but may have considered dialects (i.e. Rangpur, Mymensingh, Rajshahi) for the study to be generalizable.

- Light-weight transformer architectures should be taken into consideration as well for future model deployment (e.g. DistilBERT, ALBERT), resource extension techniques to augment the data, and hybrid models that may take linguistic rules into consideration besides neural architecture and lead to more faithful representations.

4.4.8 Summary

The comparative results and their discussion are taken from the following:

- For dialect classification, transformers such as MuRIL and XLM-R achieve good results (~92% accuracy) which is far more than traditional ML.

- In that vein, transformers such as mBART-50 and BanglaT5 produce state of the art results for dialect standardization (BLEU >0.75, Exact Match >60%), whereas LSTM baselines fail to do so.

- The work provides proof of the necessity of transformer architectures in the field of Bangla dialect NLP and adds to the field both models and datasets for further research.

4.5 Summary

This chapter introduced how Bangla dialect classification and standardization has been implemented, evaluated and interpreted the results. Using a custom dataset of 23,440 pairs of dialect standardized Bangla sentences from five regions, we trained and compared both the conventional ML classifiers (SVM, NB, LR, RF), and the transformer-based classifiers (BanglaBERT, mBERT, MuRIL, XLM-R for classification; BanglaT5, mBART-50, mT5 for standardization). Neural networks need cleaned and balanced splits, so deep processing was done to do exactly that, making experiments proportional and repeatable. As for the results, while classical models obtained decent performance (SVM about 81% accuracy), the transformers were much better, with the MuRIL and XLM-R achieving accuracy values of 92-93 and macro- F1 scores of > 0.92. For standardization purposes, mBART-50 turned out to be the strongest baseline controlled to injected languages (BLEU = 0.78, ROUGE-L = 0.89, METEOR = 0.87, chrF = 88.9, TER = 12.5, Exact Match approx. = 65.6%) followed by BanglaT5 and mT5 (similar results), while LSTM Seq2Seq Simple model was a poor baseline (BLEU = 0.07, Exact Match approx = 8.7%). Overall the results validated that ML and transformers are feasible for classification tasks, while dialect-to-standard translation needs complex contextual generation models. All of these models were further operationalized into a Gradio-based user interface used for convenient usability. This thesis hypothesis states that modern NLP methods can efficiently deal with dialectical variations in Bangla and the results testify to this hypothesis by filling the need for this kind of resources and providing solution to research and applications in real-life.

Chapter 5

Engineering Standards and Design Challenges

This chapter summarizes the engineering criteria adopted throughout the design and implementation of the proposed system, the issues faced with its development, and the implications of these issues more generally. The discussion moves from technical performance to discuss compliance with software, hardware and communication standards, ethical and social impact of the project, sustainability considerations, project management aspects, and financial analysis. Furthermore, the chapter situates the project within complex engineering problem-solving categories, knowledge profiles, and engineering activities so as to align with already established professional benchmarks.

5.1 Compliance with the Standards

Bangla Dialect Classification and Standardization System is a research-based project and necessitates the application of not only innovative methodologies, but also the implementation of internationally accepted engineering standards. Standards provide assurance of reliability, maintainability, interoperability and scalability of the system, while satisfying the expectations of the academic discipline regarding reproducibility and the standards for the practice of the engineering profession.

There are three main areas of compliance for this project: software standards, hardware standards, and communication standards. All of these areas affected how the system was designed and how well the final solution could be used for both research and social purposes.

5.1.1 Software Standards

The main idea behind the proposed system is making software and teaching machines to learn. In this regard, the implementation process was dictated by numerous experiments. Standards:

1. Coding Guidelines: The code was based on PEP 8, which is the standard for Python programming. This brought in limited naming rules, white space, modular programming, and ways to handle errors. This compliance also made it possible to write code that people could read, maintain, and add to, which is especially important in academic and/or collaborative settings.

2. Standards for Reproducibility and Documentation: Following the IEEE 830-1998 (Recommended Practice for Software Requirements Specifications) and ACM reproducibility guidelines made sure that results could be reproduced. This included things like random seed, saving preprocessing scripts, recording

hyperparameters, saving model check-points, and more. Other scientists were able to repeat experiments and confirm results because of the practice.

3. Dataset Standards: The ideas behind data annotation came from ISO/IEC 9126 (Software Quality Model) and ISO/IEC 11179 (Metadata Registries). The structured CSV format showed how each sentence in the dialect was linked to the corresponding anchor Bangla standard form. "This made it easier to share for future studies by making it more accurate and consistent."

4. Machine Learning Standards: The most common NLP evaluation standards [12], [17] have been used to make evaluation metrics (Accuracy, F1-score, BLEU, ROUGE-L, METEOR, chrF, and TER) consistent for benchmarking and fair comparison.

Alternatives Considered:

- TensorFlow/Keras: Deep learning with well-known deployment pipelines But PyTorch's dynamic graph architecture gave it more freedom than TensorFlow does for dynamic NLP tasks.

- Scikit-learn-only Pipelines: Perfect for classic ML but not good enough for large scale deep learning experiments.

- Custom Annotation Tools: While easier for small datasets, there was no standardization and were not CSV or Unicode compliant.

Rationale for Choice:

- PyTorch & Hugging Face Transformers - We want to use this architecture as it is flexible and has community support & as well as we can fine tune it using leveraging BERTS architectures.

- Atmosphere: "PEP 8 is a version of the ideal language that ensures the software are of a professional level."

- IEEE and ISO principles imposed the practices of reproducibility to be taken into account as this is what it takes for this research to be meaningful for the wider academic community.

5.1.2 Hardware Standards

Training of transformer-based models for an NLP task is very computationally expensive. Choices a bout hardware were directly correlated with training efficiency, replicability, and scalability.

- GPU Standards: Project uses NVIDIA CUDA-enabled GPUs which are developed based on CUDA (Compute Unified Device Architecture) and cuDNN (CUDA Deep Neural Network library) standards. These primitives include optimized kernels for matrix multiplications and convolutions, the primitives of deep learning.

- Numerical Standards Computation conforms to IEEE 754 floating point standards for numerical accuracy and stability. This made sure precision errors didn't corrupt any training results, especially in gradient descent and backpropagation in deep networks.

- Cloud Compliance: Experiments are carried out on Google Colab Pro and institutional servers which are compliant to ISO/IEC 27001 (Information Security Management) standards for data protection.

Alternatives Considered:

- Google TPU: Provided increased speed for sequence-to-sequence tasks, but came with the need for TPU specific code and incur greater cost as well as required expertise.

- High Performance CPU (Intel Xeon): While capable of stable computation, it took days to train models such as mBART-50, making it not useful, according to the researchers.

- Local GPU vs Cloud GPU: Local GPU allowed greater controls, but cloud GPU were more scalable and easily shareable with collaborators.

Pros and Cons:

- TPU: Less expensive to train, Less flexible

- CPU: Very accessible, Suitable only for very small transformers, as CPU is very slow.

- GPU: (NVIDIA): Low-cost and available, good ecosystem support More power used than CPU.

Rationale for Choice:

NVIDIA GPU have been chosen as they provide the optimal balance between performance, compatibility and availability. True to PyTorch, Hugging Face, CUDA, and cuDNN made them suitable for carrying out cutting-edge transformer-based architecture such as mBART-50, BanglaT5, XLM-R, etc. IEEE 754 standards-based support ensured environment-consistent results.

5.1.3 Communication Standards

Since user interaction was made possible by means of a web-based interface, communication protocols were extremely important for upholding usability and accessibility as well as data integrity.

1. Protocols: The user interface was built with Gradio which uses HTTP (Hypertext Transfer Protocol), and HTTPS (HTTP Secure). Specifically:

- HTTP/1.1 (IETF RFC 2616): HTTP-1.1 is supported by all modern browsers.

- HTTPS with TLS 1.2/1.3 (IETF RFC 5246/8446): Encryption of communication between client and server - no leakage of data

2. Data Encoding Standards: All text input/output was in UTF-8 (RFC 3629) standard encoding, ensuring that it will display correctly on different platforms and correctly encode the Bangla script.

3. Web Application Standards: UI was made according to W3C accessibility standards (WCAG 2.1), which makes it possible for a wider range of users, including those with reading disabilities, to use it.

Alternatives Considered:

- gRPC (Google Remote Procedure Calls) made communication more efficient, but it was harder to use than the traditional work flows that users are used to without API integration.

- WebSockets: Offers low-latency streaming but can't be used for non-real-time NLP purposes such as classification/standardization.

- RESTful APIs: Scalable, but an overkill solution for a lightweight demo environment

Rationale for Choice:

- HTTPS was chosen because it is widely accepted and has TLS encryption standards and security.

- UTF-8 made it possible to show bangla script correctly without losing any characters.

- Gradio provided a research-friendly environment that was well-balanced in terms of performance and accessibility, and it was free of the need for deployment.

Summary of General Agreement to Standards:

In sum, the project adhered to:

- Software Quality: coding style PEP 8, dataset format ISO, reproducibility based on IEEE principles

- Hardware Requirements: NVIDIA CUDA/cuDNN GPU compliance, IEEE 754 floating point compliant, ISO/IEC 27001 compliance for Cloud

- Documentation: communication rules: HTTP/HTTPS (IETF), HTTPS 1.2/1.3 for secure exchange of data, encoding of Bangla script text in UTF-8

Not only did these standards set the groundwork for the project's scientific content, but the price scenarios were also possible and could be put into action. Other possible solutions though considered - were ruled out as being too expensive, less flexible or irrelevant to the scope of the research. Thus, the selected standards provided the best compromise between efficiency, reproducibility and accessibility.

5.2 Impact on Society, Environment and Sustainability

The application of sophisticated natural language processing (NLP) technologies to the Bangla language domain is not just a pure technical exercise. It has serious consequences for individuals, society, and the planetary ecosystem. As one of the most widely spoken languages in the world, Bangla represents cultural identity and diversity in the region. However, many dialects of Chinese are mutually unintelligible, creating barriers to education, government services and access to digital technologies. By solving these problems, the proposed system has direct impact on quality of life, social integration, equitable and ethical treatment and sustainability.

5.2.1 Impact on Life

For humans, communication is a critical part of life and technology that makes the process of communication accessible is a straightforward way to improve the quality of life. Here are several human-centered impacts of this research project:

1. **Educational Access:**
   In rural Bangladesh, students speak dialects like Sylheti, Noakhali or Chittagonian that are completely different from standard Bangla in textbooks. Due to dialectal differences, there are barriers to understanding. By transcribing dialectal text automatically into standard Bangla, our system can hopefully help close the education gap so that students can avail of the same resources that students in other parts of the country are enjoying.

2. **Healthcare Communication:**
   Often, the patients in the backwoods speak in dialectal terms and the doctors take notes in standard Bangla or English. A standardization tool can help healthcare workers understand dialectical input correctly, which lowers the chance of misdiagnosis and makes patients safer.

3. **Digital Inclusion:**
   As e-governance, online learning, and social media have grown, the need for linguistic inclusivity has become more important. Digital spaces can lead to cultural exclusion and separation, which dialectal speakers who can't access them experience. Our system lets them talk to each other in standard Bangla while still keeping their dialectal identity and getting more people involved in digital activities.

4. **Economic Opportunities:**
   Tools for standardization can help with hiring, customer service, and online shopping. For example, dialectal questions on online marketplaces can be translated into Standard Bangla so that businesses can reach more customers. This will make the economy more open to everyone and create more jobs.

In conclusion, the project affects people's lives by making things more equal, making it easier to talk to each other, and giving people more chances in school, health care, and jobs.

5.2.2 Environment and Society:

Societal Impact

On a wider societal level, the project has helped to include, preserve and empower culture through the use of technology.

- Cultural Integration: Though dialects are a celebration of cultural diversity, they can also create barriers between groups. Dialects are allowed to remain, so long as they're classified and standardized to make them valid, and to be converted into standard forms. In addition, it contributes to reduce discrimination in the region and promote a shared cultural identity.

- Government and Administration: Government notices, circulars, forms etc are in standard Bangla. People who each read dialectal Bangla would have a tough time. If we integrate this system with digital government portals, it will provide equal access to government services.

- Research and Linguistic Resources: Having constructed a large multi-dialect corpus, this research contributes to the Bangla linguistic resources, making future research on low-resource languages more manageable.

Environmental Impact

The environmental costs of NLP systems are mainly tied to their computational energy use. Training of common large transformer-based models such as mBERT, MuRIL and XLM-R is computationally intensive, large-scale work that involves substantial electric energy expenditures for the use of a large dedicated CPU-based HPC system by a team of students [12].

- Challenges:
    - Cloud computing energy consumption has the associated responsibility for carbon emissions, particularly when cloud hosting is powered by non-renewable energy.
    - Hyperparameter tuning or retraining is something it does constantly, this requires a lot of energy.
- Mitigation Measures:
    - The project was efficient from an algorithmic point of view, using early stopping, optimal batch size, and transfer learning techniques to reduce training time by approximately 50%.

- Where possible, models were trained in shared institutional GPU servers, avoiding unnecessary training as well as maximizing energy efficiency.

- For big-scale implementation, we recommend energy efficient frameworks including ONNX model optimization, model quantization techniques, etc. to reduce runtime energy consumption.

Therefore, while there are environmental costs to NLP systems, careful design choices can mitigate their footprint and make them low-impact for the long-term.

5.2.3 Ethical Aspects

Ethics is a crucial factor for developing AI systems, particularly in multi-lingual societies.

1. **Bias and Fairness:**
   Machine learning systems tend to reproduce problems like biases present in the training data. If the dataset reflects one dialect over other dialects, the model may end up being biased. For instance, if the sentences are in Khulna dialect then the chances of misclassification of Sylheti inputs increases more. To try and counteract this, the dataset was balanced on the level of regions (about 4,600 sentences per dialect).

2. **Data Privacy:**
   As this is a non-sensitive text dataset, ethical practice dictates that no personal or identifiable data is included in the dataset. Anonymization was followed during data collection, adhering to data protection principles that are represented in GDPR.

3. **Cultural Respect:**
   Dialects are associated with community identity. Calling one dialect "wrong" can be viewed as being culturally insensitive. Rather, the system reaffirms that dialects are legitimate versions of Bangla and standardization is for purposes of facilitating understanding rather than superiority.

4. **Interpretability and Expandability:**
   Users don't trust the "black-box" AI models. Thus, the system was embedded with confusion matrices, accuracy reports, BLEU/ROUGE visualizations, making it easy for researchers to understand model strengths and weaknesses.

5. **Ethical Deployment:**
   The technology should not be used for censorship and dialectical communication surveillance. However, its scope must be confined to education, accessibility and inclusiveness.

Ethical Considerations: Addressing these ethical considerations, the implementation of systems following IEEE Ethically Aligned Design (EAD) and machine learning ethics guidelines:

5.2.4 Sustainability Plan

For long-term impact, a sustainability roadmap covering technical, social and environmental dimensions is needed for the project:

1. **Technical Sustainability:**
    - Models and datasets are stored in open source repositories, so the work can be replicated.
    - Documentation, code and preprocessing procedures are written according to academic practice so that future researchers can expand on this work.
    - The interface is so light that it can be used in mobile and low-resource deployments and is designed to be easy for users with disabilities.
2. **Social Sustainability:**
    - With the help of partnerships with universities and NGO, there is the opportunity for continuous dialect data collection to make the corpus even richer over time.
    - The system can be used in schools and local government offices to bring the services provided and education given to the dialect.
    - Creating partnership with their community will instill ownership in an aggregated way.
3. **Environmental Sustainability:**
    - Green AI practices such as pruning, knowledge distillation, and energy-efficient hardware should be a part of the development process for future deployments.
    - Carbon neutralization: Partnerships with renewable energy cloud providers can have a positive environmental impact on the economic bottom line.
4. **Economic Sustainability:**
    - Freemium model can be used where you do not pay for basic features and then enhance the model by creating a super analytics or enterprise integration that will keep the model going.
    - Government or donor funding can further ensure the system continues to be open and accessible to underprivileged communities.

In other words, the sustainability plan guarantees the continuity of the project beyond this thesis as a scalable, socially responsible, and environmentally friendly project.

5.3 Project Financial Analysis and Project Management

Project management is an essential aspect of the process for translating research concepts into working systems by efficiently transforming them into resources, schedules, and budgets. For this effort, the management process mixed research activity (gathering data, developing a model, and evaluating) with project products (prototype of the system, user interface, and a final thesis report). The financial analysis, on the other hand, shows that the project is possible, big enough, and long-lasting, both in theory and in practice.

5.3.1 Project Management Approach:

Chapter 3 explains how the project was set up using a phase-based management framework. Activities and outcomes were divided into research milestones and implementation deliverables. Every phase had its own resource, risk, and deleverage.

Phase 1: Research and Requirement Analysis: Literature Review, Dataset Requirement Identification, Problem Formulation

Phase 2: Corpus Construction and Data Collection: Gathering 23,440 sentences from five dialects and making sure that the name annotations are all the same.

Phase 3: Data Preprocessing and Normalization: Cleaning, removing duplicates, breaking up words into tokens, and balancing the dialect classes.

Phase 4: Model Development and Training: Teaching old-school ML classifiers and transformer models how to work.

Phase 5: Evaluation and Comparative Analysis: Using classification and generation metrics (accuracy, F1, BLEU, ROUGE, METEOR, chrF, TER, Exact Match) to test models.

Phase 6: User Interface Development: Making a Gradio-based user interface for activities related to classification and standardization.

Phase 7: Writing and Reporting: Reporting, thesis chapters and pictures.

This approach has used a pipeline to include risk reduction (not based on a single model), time management (training ML and transformer models at the same time), and resource optimization (using the same computer hardware).

5.3.2 Budget Estimation

The costs for the project were in four major areas: hardware resources, software tools, human effort, and dissemination/documentation.

Table 5.3.1: Budget Estimation Summary:

Category	Details	Estimated Cost (BDT)
Hardware	GPU servers (NVIDIA A100 on Colab Pro+/cloud), storage, storage HDDs (Backup).	6,500
Software & Licenses	Premium Colab sub, storage (Google Drive/OneDrive), Python libraries (mostly free, open source).	300
Human Resources	Research effort (estimated equivalent cost of RA/graduate student support) domain experts to annotate, proofreaders	3,500
Data Collection	Local field, honoraria of annotators, travel/communication.	5,000
Documentation / Dissemination.	Printing of thesis, cost of conference presentation (where it is needed), cost of submitting journal.	500
Miscellaneous	Internet overhead, electricity overhead, unexpected costs.	500

Total Estimated Budget: ~BDT 16,300.
The budget is an academic level projection of costs at the project scale and highly depends on free/open-source software and computing resources provided by universities

5.3.3 Alternative budget and justification.

Another budget was also taken into consideration, based on the availability of funds and resources.

- **Low Cost Alternate Budget (about BDT 6,500):**
    - Calculating on free versions of Google Colab and Kaggle.
    - Using small dataset or exiting dataset collection.
    - Reducing the cost of documentation/dissemination through focusing on open-access preprints rather than conferences with pay-per-paper fees.

- Our tests will only focus on a few transformer models, namely BanglaBERT and BanglaT5.

- **High Cost Industrial Budget (around BDT 2,50,000 -,10,00,000):**
    - Mass data labeling by professional linguists.
    - Specialized GPU servers (groups of multiple GPUs to train faster).
    - Business deployable UI with mobile application.
    - Professional proprietary project management and collaboration software.

Rationale:

The chosen middle range budget (around BDT 16,300) offers the best compromise between a research feasibility and the quality of an experiment. It does not overly rely on free resources but neither drives up expenses with industrial-grade needs that a thesis project does not need.

5.3.4 Revenue Model

It can be commercialized but is primarily academic:

- **Freemium Model:**
    - Free access to classification / standardization with restricted usage.
    - Higher quality (e.g. batch processing of large documents) educational, publishing or government agency.

- **Subscription Model:**
    - NGO, local governments and schools could subscribe to be used regularly in education and local services.

- **API-as-a-Service:**
    - Provide an API to be used with chatbots or e-learning websites or translators.

- **Shared Grants and Funding:**
    - Enlist the help of organizations such as UNESCO, Digital Bangladesh projects, or South Asian language conservation Non-Governmental Organizations.

In this kind of model, the research can be a scalable product because it is long-term sustainable and relevant.

5.3.5 Finance risks and risk management.

- Risk 1: Generally, Large Computational Costs would be reduced through transfer learning and fine-tuning, as opposed to training.
- Risk 2: Expansion of Data sets Costs: Community-based crowdsourcing in future.
- Risk 3: Hybrid academic-commercial model implies that as long as the revenue streams do not increase, the research funding and grants will keep advancing the research.

5.3.6 Summary

The financial analysis allows identifying that the project is achievable within the reasonable academic budget of around BDT 16,300, but there are both cheaper and more expensive academic and industrial arrangements. It can be used in the academic dimension, but also it would be a valuable idea to adopt the system in the future, since the system has a clear project management structure, stages and revenue model deliverables.

5.4 Complex Engineering Problem

Multilayered knowledge requirements, competing demands, the effects of the issues on society, and stakeholder engagement are the typical features of complex engineering issues. This system is among those in this category because the Bangla dialect classification system and standardization system proposal has integrated linguistic variation, the issue of computations complexity, the element of ethics, and long-term sustainability. This sub-section aligns the research problem with the categories of the Engineering Problem (EP) as well as the Knowledge Profile (KP) and the Engineering Activities (EA).

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.1). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.4.1: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	✓	✓	✓

Justification for Each Engineering Problems element:

- **EP1 - Depth of Knowledge:** This project has required knowledge of computational linguistics, machine learning, NLP, transformer models and UI design, so it's a multi-disciplinary project.
- **EP2 - Range of Contradictory Requirements:** We have traded off between the accuracy and efficiency. Large transformer models give better results, but with more computational resources resulting in a tradeoff.
- **EP3 - Depth of Analysis:** The analysis did not stop at the elementary levels of model training, including the detection of errors, the comparison between different models, and the evaluation based on multiple metrics which requires deep analysis thinking.
- **EP4 - Familiarity of Issues:** Bangla is a low-resource language and dialectal NLP is not a popular area. New approaches to the preprocessing of data and the fine-tuning of models were required.
- **EP5 - Extent of Applicable Codes:** We adhered to coding and software standards (IEEE, PEP8) and ethical practices for the data collection in order to be reproducible and in compliance.
- **EP6 - Degree of Stakeholder involvement:** The project included annotators, dialect speakers, educators, project team members including what could lead to indirect benefits to students and developers.
- **EP7 - Interdependence:** This robustness requires all the components of the system to function together from data set quality, preprocessing data, classifying raw data, standardizing data, and integrating all the elements into an intuitive user interface. Any weakness present in any part affects the overall performance.

Mapping with Knowledge Profile

This section is designed to map the overall problem and EP1 ***(multiple between K3, K4, K5, K6, K8 for attaining EP1)*** to the Knowledge Profile.

Table 5.4.2: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
×	✓	✓	✓	✓	✓	✓	✓

Reasons for each part of the Knowledge Profile:

K2 - Mathematics: It uses probability for prediction, linear algebra, and optimization, which are all parts of machine learning (ML) and deep learning

K3 - Engineering Fundamentals: Model development and design of Statistical classification pipelines algorithms.

K4 - Specialist Knowledge: Transformer architectures (BanglaBERT, mBERT, MuRIL, XLM-R, mBART-50, BanglaT5, mT5).

K5 - Engineering Design: Designing the pipeline: data preprocessing → classification → standardization → UI.

K7 - Comprehension: Writing in an organized manner, analyzing findings, discussing results and presenting them clearly.

K6 - Engineering Practice: Practical implementation of models using PyTorch, Hugging Face, and evaluation tools.

K8 - Research Literature: Integration of existing works on Bangla NLP, dialect processing, and sentiment analysis.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping add subsections to put rationale (Use Table 5.3).

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's ***(multiple).***

Table 5.4.3: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

Justify for each Engineering Activities:

EA1 - Range of Resources: Utilized diverse resources: custom dataset, cloud GPU, open-source libraries.

EA2 - Level of Interaction: Required collaboration between annotators, developers, and evaluators.

EA3 - Innovation: Developed a unique multi-regional Bangla dialect corpus and applied transformers to standardization.

EA4 - Consequences for Society & Environment: Contributes to language preservation, education, and digital accessibility.

EA5 - Familiarity: Low familiarity due to limited prior research in Bangla dialect standardization.

5.5 Summary

The Bangla Dialect Classification and Standardization project can be as well categorized as a complex engineering problem of all seven categories. It required profound multidisciplinary expertise, resolution of conflicts, critical analysis, adherence to ethics, participation of stakeholders and design interdependence. These mappings indicate that it is not an ordinary piece of work but a complex socially relevant engineering problem.

Chapter 6

Conclusion

This chapter provides the concluding remarks of the research. It summarizes the key findings of the study, highlights the limitations encountered during the implementation, and outlines potential directions for future research. By consolidating the outcomes, this chapter serves as a reflection on the contributions of the thesis and identifies opportunities for extending the work.

6.1 Summary

The objective of this research was to build an automated approach for classification and standardization of the Bangla dialect using both traditional machine learning models and modern transformer-based models. A specialized multi-regional corpus from 23,440 sentences was developed with dialects of Khulna, Sylhet, Noakhali, Barisal and Chittagong.

For classification, traditional classifiers like SVM, Naive Bayes, Random Forest, and Logistic regression were compared with the transformer-based models like BanglaBERT, mBERT, MuRIL, and XLM-R. Result showed that transformer models are significantly better than traditional models, where accuracy achieving m-ril and XLM-R are over 92% achieved in comparison to 81% accuracy of the best-performing traditional models.

To compare it to the current state of the art for the standardization, an LSTM-based Seq2Seq model was compared to transformer-based models such as mBART-50, BanglaT5, mT5 etc. Among these, mBART-50 achieved the best performance with a BLEU score of 0.78 and ROUGE-L of 0.89, confirming its effectiveness in dialect-to-standard Bangla translation.

Overall, the research successfully demonstrated that modern transformers provide superior accuracy, robustness, and linguistic adaptability, while the development of a user-friendly Gradio interface validated the system's practical utility. The study not only contributes a novel annotated corpus but also provides a benchmark comparative study for future Bangla NLP research.

6.2 Limitation

Despite its contributions, this research faced several limitations:

- Data Coverage: Although the dataset covers five major dialectal regions, it does not include all dialects of Bangladesh (e.g., Rangpur, Mymensingh, Rajshahi).

- Resource Constraints: Training large transformer models such as mBART-50 and mT5 required significant GPU resources, limiting hyperparameter tuning and larger-scale experiments.

- Evaluation Metrics: Automatic metrics like BLEU and ROUGE were used, but human evaluation of naturalness and fluency was not conducted due to time

constraints.

- Domain Adaptation: The models were trained on conversational and sentence-level data; performance in other domains (e.g., literary or legal texts) remains untested.

- Real-Time Deployment: While a functional UI was developed, the system has not yet been tested for real-time, large-scale deployment under heavy user load.

6.3 Future Work

The scope for future research is considerable. Some promising directions include:

- **Expanding the Corpus:** Incorporating additional dialects such as Rangpur, Mymensingh, and dialects spoken by minority groups to create a more comprehensive dataset.

- **Human Evaluation:** Conducting large-scale human assessments to measure naturalness, fluency, and cultural appropriateness of standardization outputs.

- **Domain-Specific Adaptation:** Extending the models to perform well in specialized domains such as healthcare, education, and administration.

- **Model Optimization:** Exploring knowledge distillation and quantization techniques to reduce transformer size and make the models more deployable on low-resource hardware.

- **Real-Time Deployment:** Building a scalable API or mobile app for real-time dialect classification and standardization, making the system accessible to broader audiences.

- **Cross-Lingual Transfer:** Leveraging transfer learning with related low-resource languages (e.g., Assamese, Oriya) to further improve generalization.

References

[1] T. Hossain, A.-R. Islam, H. Kabir, A. A. Rasel, M Abdullah-Al-Wadud, and J. Uddin, "BanglaNewsClassifier: A machine learning approach for news classification in Bangla Newspapers using hybrid stacking classifiers," *PLoS ONE*, vol. 20, no. 6, pp. e0321291–e0321291, Jun. 2025, doi: https://doi.org/10.1371/journal.pone.0321291.

[2] R. Islam, T. Kader, A. Kadar, and A. Chy, "Bangla Language Dialect Classification using Machine Learning," *https://ieeexplore.ieee.org/xpl/conhome/10114473/proceeding*, Dec. 2022, doi: https://doi.org/10.1109/icecte57896.2022.10114552.

[3] Md. N. Hoque and Md. H. Seddiqui, "Detecting cyberbullying text using the approaches with machine learning models for the low-resource Bengali language," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, p. 358, Mar. 2024, doi: https://doi.org/10.11591/ijai.v13.i1.pp358-367.

[4] T. Mahmud, M. Ptaszynski, and F. Masui, "Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts," *Electronics*, vol. 13, no. 9, pp. 1677–1677, Apr. 2024, doi: https://doi.org/10.3390/electronics13091677.

[5] Suborno Deb Bappon, G. Sarwar, and M. I. Khan, "Sentiment Analysis of Bengali Texts on Online Tech Gadget Reviews using Machine Learning," pp. 324–329, Dec. 2022, doi: https://doi.org/10.1109/iccit57492.2022.10055639.

[6] Md. S. Islam and K. M. Alam, "Sentiment analysis of Bangla language using a new comprehensive dataset BangDSA and the novel feature metric skipBangla-BERT," *Natural Language Processing Journal*, vol. 7, p. 100069, Jun. 2024, doi: https://doi.org/10.1016/j.nlp.2024.100069.

[7] D. Goswami, Md Nishat Raihan, S. Chowdhury, and M. Zampieri, "nlpBDpatriots at BLP-2023 Task 2: A Transfer Learning Approach towards Bangla Sentiment Analysis," Jan. 2023, doi: https://doi.org/10.18653/v1/2023.banglalp-1.37.

[8] Rajesh Kumar Das, Md. Anwarul Islam, Md. Mahmudul Hasan, Sultana Razia, M. Hassan, and Sharun Akter Khushbu, "Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models," *Heliyon*, vol. 9, no. 9, pp. e20281–e20281, Sep. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e20281.

[9] Md. R. Hossain, M. M. Hoque, N. Siddique, and I. H. Sarker, "Bengali text document categorization based on very deep convolution neural network," *Expert Systems with Applications*, vol. 184, p. 115394, Dec. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115394.

[10] Md Nishat Raihan, D. Goswami, S. Chowdhury, and M. Zampieri, "nlpBDpatriots at BLP-2023 Task 1: Two-Step Classification for Violence Inciting Text Detection in Bangla - Leveraging Back-Translation and Multilinguality," Jan. 2023, doi: https://doi.org/10.18653/v1/2023.banglalp-1.20.

[11] Md. H. I. Bijoy, U. Ayman, and Md. M. Islam, "BanglaTense: A large-scale dataset of Bangla sentences categorized by tense: Past, present, and future," *Data in Brief*, vol. 59, p. 111400, Feb. 2025, doi: https://doi.org/10.1016/j.dib.2025.111400.

[12] A. Bhattacharjee, T. Hasan, W. U. Ahmad, and Rifat Shahriyar, "BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.18653/v1/2023.findings-eacl.54.

[13] H. A. Z. S. Shahgir and K. S. Sayeed, "Bangla Grammatical Error Detection Using T5 Transformer Model," *arXiv.org*, Mar. 19, 2023. https://arxiv.org/abs/2303.10612 (accessed Jun. 06, 2023).

[14] M. F. Islam, J. Hasan, M. A. Islam, P. Dewan, and M. S. Rahman, "BanglaLem: A Transformer-based Bangla Lemmatizer with an Enhanced Dataset," *Systems and Soft Computing*, vol. 7, p. 200244, Apr. 2025, doi: https://doi.org/10.1016/j.sasc.2025.200244.

[15] Islam, S. Ahmmed, and M. S. Hossain, "Transcribing Bengali Text with Regional Dialects to IPA using District
Guided Tokens," *arXiv (Cornell University)*, Mar. 2024, doi: https://doi.org/10.48550/arxiv.2403.17407.

[16] Tabia Tanzin Prama and M. M. Anwar, "Sylheti to Standard Bangla Neural Machine Translation: A Deep Learning-Based Dialect Conversion Approach," *Lecture notes in networks and systems*, pp. 208–217, Jan. 2025, doi: https://doi.org/10.1007/978-3-031-78925-0_21.

[17] Faria, M. B. Moin, A. A. Wase, M. Ahmmed, M. R. Sani, and T. Muhammad, "Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.11142.

[18] S. Chowdhury, D. R. A. Remal, S. T. Pasha, A. Islam, and S. R. H. Noori, "ChatgaiyyaAlap: A dataset for conversion from Chittagonian dialect to standard Bangla," *Data in Brief*, vol. 59, p. 111413, Feb. 2025, doi: https://doi.org/10.1016/j.dib.2025.111413.

[19] T. Mahmud, Michał Ptaszyński, and F. Masui, "Automatic Vulgar Word Extraction Method with Application to Vulgar Remark Detection in Chittagonian Dialect of Bangla," *Applied sciences*, vol. 13, no. 21, pp. 11875–11875, Oct. 2023, doi: https://doi.org/10.3390/app132111875.

[20] J. Sikder, P. Chakraborty, U. K. Das, and K. Dhar, "A hybrid approach for Bengali sentence validation," *Artificial Intelligence Review*, vol. 57, no. 11, Oct. 2024, doi: https://doi.org/10.1007/s10462-024-10795-2.

[21] J. Alam, P. Roy, Mahfuzur Rahman Shuvo, N. Hasan, and M. M. Rahman, "Fine-Tuning Large Language Models for Regional Dialect Comprehended Question answering in Bangla," *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–6, Jan. 2025, doi: https://doi.org/10.1109/SCEECS64059.2025.10940303.

[22] M. R. Islam, A. Ahmad, and M. S. Rahman, "Bangla text normalization for text-to-speech synthesizer using machine learning algorithms," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, pp. 101807–101807, Oct. 2023, doi: https://doi.org/10.1016/j.jksuci.2023.101807.

[23] A. J. M. Jakaria, R. R. Chowdhury, J. J. Konia, D. Roy, and N. T. A. Meem, "A Comparative Study on different Machine Learning Approaches for Categorizing Bangla Documents," *International Journal of Computer Applications*, vol. 186, no. 61, pp. 32–39, Jan. 2025, doi: https://doi.org/10.5120/ijca2025924391.

[24] Umme Ayman, C. Saha, Azmain Mahtab Rahat, and Sharun Akter Khushbu, "BanglaBlend: A Large-Scale Nobel Dataset of Bangla Sentences Categorized by Saint and Common Form of Bangla Language," *Data in Brief*, vol. 58, pp. 111240–111240, Dec. 2024, doi: https://doi.org/10.1016/j.dib.2024.111240.

[25] Azmine Toushik Wasi, R. Islam, M. R. Islam, T. H. Rafi, and D.-K. Chae, "Exploring Bengali Religious Dialect Biases in Large Language Models with Evaluation Perspectives," Jul. 25, 2024.

https://www.researchgate.net/publication/382638382_Exploring_Bengali_Religious_Dialect_Biases_in_Large_Language_Models_with_Evaluation_Perspectives

[26] M. Arafat, Z. S. Raha, B. Paul, and T. Muhammad, "Bridging Dialects: Translating Standard Bangla to Regional Variants Using Neural Models," pp. 885–890, Dec. 2024, doi: https://doi.org/10.1109/iccit64611.2024.11022371.

[27] R. K. Das, M. Islam, and S. A. Khushbu, "BTSD: A curated transformation of sentence dataset for text classification in Bangla language," *Data in Brief*, vol. 50, p. 109445, Oct. 2023, doi: https://doi.org/10.1016/j.dib.2023.109445.

[28] N. Sultana, R. Yasmin, B. Mallik, and M. S. Uddin, "ONUBAD: A comprehensive dataset for automated conversion of Bangla regional dialects into standard Bengali dialect," *Data in Brief*, vol. 58, p. 111276, Feb. 2025, doi: https://doi.org/10.1016/j.dib.2025.111276.

[29] Istiak Tanvir, A. Akter, and T. Islam, "An Investigation of Bias in Bangla Text Classi cation Models," Mar. 13, 2025. https://www.researchgate.net/publication/389788664_An_Investigation_of_Bias_in_Bangla_Text_Classi_cation_Models?enrichId=rgreq-1701b0d99307ed1fb6cd9c6f96002b89-XXX&enrichSource=Y292ZXJQYWdlOzM4OTc4ODY2NDtBUzoxMTQzMTI4MTMxNTUwMzM3NEAxNzQxODQxNDgyMzMx&el=1_x_3&_esc=publicationCoverPdf

213-15-4447

ORIGINALITY REPORT

13%	10%	9%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	arxiv.org Internet Source	1%
4	aclanthology.org Internet Source	<1%
5	ltu.diva-portal.org Internet Source	<1%
6	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
7	Submitted to United International University Student Paper	<1%
8	www.sci.muni.cz Internet Source	<1%
9	"Proceeding of the 2nd International Conference on Machine Intelligence and Emerging Technologies", Springer Science and Business Media LLC, 2025 Publication	<1%
10	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent	<1%