
A Hybrid Approach to Bangla Regional Text Classification Using BERT Ensemble and Region-Specific Lexical Oversampling

By

Babe Sultana
0122310021

Submitted in partial fulfillment of the requirements
of the degree of Master of Science in Computer Science and Engineering

January 6, 2026



Department of Computer Science and Engineering
United International University

Approval Certificate

This thesis, titled ‘**A Hybrid Approach to Bangla Regional Text Classification Using BERT Ensemble and Region-Specific Lexical Oversampling**’, submitted by **Babe Sultana**, Student ID: **0122310021**, has been accepted as **Satisfactory** in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 08.10.2025.

Board of Examiners

1.

Prof. Dr. Mohammad Nurul Huda
Professor
United International University

Supervisor

2.

Dr. Ohidujjaman
Assistant Professor
United International University

Co-Supervisor

3.

Prof. Dr. Dewan Md. Farid
Professor
United International University

Head Examiner

4.

Dr. Riasat Azim
Assistant Professor
United International University

Examiner-I

5.

Mr. Nahid Hossain
Assistant Professor
United International University

Examiner-II

6.

Professor Dr Khondaker Abdullah Al Mamun
Director IRIIC and MSCSE
United International University

Ex-Officio

Declaration

This is to certify that the work entitled ‘**A Hybrid Approach to Bangla Regional Text Classification Using BERT Ensemble and Region-Specific Lexical Oversampling**’ is the outcome of the research carried out by me under the supervision of ‘**Prof. Dr. Mohammad Nurul Huda, Professor and Ohidujjaman, Assistant Professor**’

Babe Sultana
MSCSE Program
Student ID: 0122310021
Dept. of Computer Science and Engineering
United International University
Dhaka, Bangladesh

In my capacity as supervisor of the candidate’s thesis, I certify that the above statements are true to the best of my knowledge.

Prof. Dr. Mohammad Nurul Huda
Professor
Dept. of Computer Science and Engineering
United International University

Abstract

Regional text analysis reflects the lived realities of diverse communities by capturing the linguistic richness and diversity present in various dialects. It bridges the gap between everyday regional usage and standardized language forms, thereby enhancing the inclusivity of language technologies. In this paper, we focus on five regional dialects in Bangladesh, namely Chittagong, Sylhet, Noakhali, Barishal, and Rangpur, using a dataset of 4,218 text samples. The dataset is validated by five regional experts and categorized into three tiers based on an assigned agreement criterion. Tier 1 represents a strictly filtered, high-confidence subset and is used primarily for evaluation. A set of region-specific special words, which belong exclusively to their respective regions and are validated by domain experts, is introduced. These words are used in a linguistically informed oversampling technique to balance the dataset in both experiments. In the first experiment, we demonstrate the effectiveness of the tiered dataset structure, where Tier 2 and Tier 3 (medium- and low-confidence subsets) are used for training, and Tier 1 (high-quality subset) is used for testing. In this setting, BanglaBERT achieves the best individual performance with 67.45% accuracy and a weighted F1-score of 67.62%. In the second experiment, we focus exclusively on the Tier 1 dataset, applying a wide range of machine learning and deep learning models to assess their effectiveness. The key contribution is a heterogeneous deep ensemble technique that combines three BERT models, BanglaBERT, BUETBERT, and DistilBERT, achieving an accuracy of 85.17% and a weighted F1-score of 84.84% on the Tier 1 dataset.

Acknowledgements

All the praises and thanks are to Allah, the most gracious and the ever merciful. HE has given me the strength, courage, and patience to carry out this work.

This thesis represents a great deal of time and effort, not only on my part but also on the part of my supervisor, Dr. Mohammad Nurul Huda, and my co-supervisor, Dr. Ohidujjaman. I am sincerely grateful for their continuous guidance, valuable feedback, and encouragement throughout the course of this research. Their expertise, patience, and motivation have been a constant source of inspiration to me.

I would also like to express my sincere appreciation to the respected members of my thesis committee for their insightful and constructive comments, which significantly contributed to improving the quality and depth of this thesis.

Furthermore, I am deeply indebted to my family for their unconditional love, patience, and continuous support. Their encouragement gave me the strength to overcome challenges and stay focused on my goals. My heartfelt gratitude also goes to my friends and well-wishers, whose moral support, motivation, and companionship made this journey more manageable and meaningful.

Finally, I wish to acknowledge with appreciation all those who, in one way or another, have supported, encouraged, and prayed for me during this endeavor. Without their contributions, this thesis would not have been possible.

Publication List

The main contributions of this research have been published, accepted, or are currently under review/preparation for journals and conferences. The related publications are listed below:

Journal Articles

1. BdRegionText: Resource Creation and Evaluation for Bangla Regional Text Classification with Machine Learning, Indonesian Journal of Electrical Engineering and Computer Science, 2025, Impact Factor: 0.641, published by the Institute of Advanced Engineering and Science (IAES) (Published)
2. A Hybrid Approach for Bangla Regional Text Classification Using Region-Specific Lexical Oversampling and BERT Ensemble Learning. Array, Impact Factor: 4.5, (Submitted)

Table of Contents

Table of Contents	vii
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Overview	1
1.2 Motivation	2
1.3 Problem Statement	3
1.4 Research Questions	4
1.5 Scope and Limitations	5
1.6 Research Contributions	5
1.7 Organization of the Thesis	6
2 Background and Literature Review	7
2.1 General Text Classification in Bangla and Other Languages	7
2.2 Ensemble Learning Approaches	8
2.3 Bangla Regional Speech and Text Research	9
2.4 Motivation and Research Gap	10
3 Annotated Corpus for Experimentation	12
3.1 Data collection	13
3.2 Annotation process	14
3.3 Filtering and tiered dataset construction	15
3.4 Regional lexicon for oversampling	17
3.5 Comparison analysis with relevant dataset	18
4 Methodology	21
4.1 Preprocessing	22
4.2 Proposed linguistically-informed oversampling technique	22
4.3 Conventional Method	23
4.3.1 BanglaBERT	24

4.3.2	BUETBERT	24
4.3.3	DistilBERT	24
4.4	Soft voting-based deep ensemble	25
5	Results and Discussion	26
5.1	Environmental setup	26
5.2	Experimental configuration for tiered datasets	27
5.3	Train and Test Data Split	27
5.4	Transformer-based model fine-tuning setup	28
5.5	Experiment 1: BERT-based model evaluation across all three tiered datasets	28
5.6	Experiment 2: ML, BERT, and ensemble model performance on tier 1 data	30
5.6.1	The impacts of applying Feature extraction techniques TF-IDF, TF-IDF + PCA, TF-IDF+SVD	31
5.6.2	The impacts of applying three embedding techniques named Word2Vec, FastText (Pre) and FastText	32
5.6.3	The Impact of BanglaBERT, BUETBERT, DistilBERT, and their Ensemble in the Dataset	32
5.6.4	Impacts of K-fold Cross-validation	35
5.7	Result Interpretation and Error Analysis	36
5.7.1	Confusion Matrix Analysis	36
5.7.2	ROC Curve Analysis	38
5.7.3	Error and Confusion Analysis	39
5.7.4	Overall discussion	40
6	Conclusion	42
6.1	Summary of Findings	42
6.2	Contributions of the Research	43
6.3	Limitations	43
6.4	Recommendations	43
6.5	Future Work	44
	References	47

List of Figures

3.1	Steps of datasets generation	12
3.2	Distribution of tiered and discarded dataset	15
4.1	The proposed system architecture	21
5.1	Confusion matrices for (a) BanglaBERT, (b) BUETBERT, and (c) DistilBERT	37
5.2	Confusion matrix for the ensemble of three BERT models.	37
5.3	ROC curves of (a) BanglaBERT, (b) BUETBERT, and (c) DistilBERT models.	38
5.4	ROC curve for soft voting ensemble of three BERT models	39

List of Tables

2.1	Summary of Existing Research Closely Related to Our Work	11
3.1	Five regional Bengali text samples with Bangla meaning and English translation	14
3.2	Criteria used to construct the tiered dataset.	15
3.3	Summary of the filtered dataset tiers based on annotator agreement.	16
3.4	Overview of the dataset by tier in numerical form (This section contains all numerical information, including the dataset’s maximum, minimum, average, and unique word counts and classwise instance counts.)	17
3.5	Sample region-specific special words used for oversampling	18
3.6	Comparison of existing Bangla regional text datasets.	19
5.1	Summary of the Experimental Configuration.	27
5.2	Transformer models and fine-tuning configurations for regional Bangla text classification	28
5.3	Performance of transformer-based models across all three tiered datasets before applying oversampling (summary metrics).	29
5.4	Precision, Recall, F1-score for regional classes along with overall Accuracy (%) and Weighted F1-score (%) for transformer-based Bangla embeddings (after oversampling) across all three tiered datasets.	29
5.5	Overall accuracy (%) comparison of feature extraction techniques: TF-IDF, TF-IDF + PCA, and TF-IDF + SVD, with and without oversampling	31
5.6	Overall accuracy (%) comparison of embedding techniques: Word2Vec, FastText (pre-trained), and FastText, with and without oversampling	32
5.7	Performance of transformer-based models before applying oversampling (summary metrics)	33
5.8	Precision, Recall, F1-score for regional classes along with overall Accuracy (%) and Weighted F1-score (%) for transformer-based Bangla embeddings and ensemble method (after oversampling)	33
5.9	Accuracy and weighted F1-score for training and testing sets, with and without oversampling	34

5.10 Five-fold cross-validation performance before and after oversampling. Values represent mean accuracy and weighted F1-score.	35
------------------------------------------------------------------------------------------------------------------------------------------	----

Chapter 1

Introduction

This chapter introduces the background of the research, highlights the motivation for the study, formulates the problem statement, and specifies the scope and limitations. The overall objective is to position the study within the broader research domain and demonstrate its novelty in the context of Bangla regional text classification.

1.1 Overview

This section provides an overview of the growing significance of regional text classification in the context of Bangladesh and beyond. With the rapid expansion of online communication, especially on social media, vast amounts of textual data are produced daily, reflecting regional linguistic diversity and cultural identity. Understanding and classifying such regional variations not only enriches natural language processing (NLP) research but also offers valuable insights into how people express emotions, traditions, and social belonging through their dialects. Although multi-modal research commonly considers both speech and text, regional variation creates additional complexity: two dialects may share similar spellings but differ in pronunciation, or they may differ in spelling while being phonetically close. Expanding research and reliable resources creation on regional dialects across both modalities, text and speech, will enable more robust AI systems, allowing speech-driven or text-based commands to better serve region-specific users with higher linguistic fidelity.

- Over the past twelve years, there has been a growing interest in text classification due to the massive amount of online data.
- Classifying and analyzing petabytes of data from social media and other platforms helps uncover trends, public interests, and opinions.
- In Asia, many countries consist of multiple regions where people often prefer to speak or write in regional dialects instead of the national language.
- In Bangladesh, Bengali (Bangla) is the official and most widely spoken language, yet people frequently communicate in their local dialects.

- Bengali is the fifth most spoken language in the world [1], but regional dialects remain deeply rooted in everyday communication.
- Children naturally acquire the language of their surroundings, often without following grammatical norms. Pronunciation and usage vary across regions.
- At home and in informal contexts, people prefer regional languages, while standard Bangla is used in education and official settings.
- On social media, many Bengalis express themselves in regional dialects. About 55 different regional languages are spoken in 64 districts of Bangladesh [2].
- Numerous Facebook pages and cultural events showcase regional identity, where people post, comment, and interact in their dialects.
- Dramas, songs, poems, and stories are also available in region-specific versions.
- Regional text classification, therefore, becomes an insightful way to understand the root of people’s emotions, culture, and identity.

1.2 Motivation

People strongly prefer platforms where they can express themselves in their regional dialect rather than in standard Bangla. For instance, the simple expression “আপনি ভালো আছেন?” (Are you fine?) can appear as “আম্নে ভালা আছেন্নি?” in Noakhali or “অনে গম আচন?” in Chittagong. Such differences highlight the linguistic richness across regions and the need to account for them in text analysis. Similarly, a common word like “পেঁপে” (papaya) may be pronounced and written as “ককিয়া” in Noakhali and “পোষা” in Barishal, showing how region-specific vocabulary can vary widely. Another fruit name “পেয়ারা” (guava) is written as “গইয়া” in Barishal, “গোব্বা” in Noakhali, and “গয়ম” in Chittagong, which indicates totally different pronunciation and spelling. These kinds of lexical and phonological shifts demonstrate that even very common and everyday words may change completely across dialects, making them unintelligible to people from other regions without prior exposure. For computational models, such differences pose a significant challenge because standard NLP systems are generally trained on formal or standard Bangla text and fail to capture these regional variations. Without region-specific datasets and lexicons, machine learning models misclassify or ignore these words, leading to poor performance in downstream tasks such as sentiment analysis, social media monitoring, or spam detection. Therefore, building expert-validated corpora and incorporating regional lexicons is essential to capture this diversity. By focusing on these dialectal features, our study ensures that classification models not only learn the general structure of Bangla but also recognize the subtle, localized patterns that truly reflect the way people communicate in different parts of Bangladesh. Now, the overall point is to describe the motivation of our research.

- Many young researchers are now interested in exploring regional text as a research domain.
- This study focuses on five major regional dialects of Bangladesh: Rangpur, Barisal, Noakhali, Sylhet, and Chittagong.
- Text data in regional languages can support important NLP applications such as:
 - Spam detection in regional dialects.
 - Social media monitoring.
 - Word categorization and sentiment analysis.
- While several works have explored Bangla text classification, only a few studies have targeted Bangla regional text specifically.

1.3 Problem Statement

This section discusses the dataset development and refinement process, emphasizing the challenges of resource scarcity in Bangla regional text research and the steps taken to overcome them. It reviews existing corpora and their limitations, highlighting the need for a more authentic and balanced dataset. The improved BdRegionText Version 2, used in this study, addresses these gaps through expert validation, inclusion of diverse dialects, and incorporation of region-specific lexical features to ensure linguistic richness and representational fairness across Bangla dialects.

- A major challenge in this research area is the lack of reliable resources for Bangla regional text.
- Existing corpora are low-resource, with some regions having more data than others, making it hard to build balanced datasets.
- So far, only three datasets exist:
 - Vashantor [3] – a translation of standard Bangla text into five dialects.
 - Bhashamul [4] – a contest dataset recording IPA forms of dialects, often limited to single words or very short phrases.
 - BdRegionText (Version 1) [5] – our earlier dataset with four dialects, but not validated by experts.
- Limitations of prior datasets:
 - Vashantor relies on artificial translation, not natural usage.
 - Bhashamul contains very short sentences, lacking semantic richness for classification.

- BdRegionText V1 was not validated by regional experts.
- BdRegionText Version 2, the dataset used in this study, resolves these issues:
 - Includes five dialects.
 - Collected from authentic real-world sources.
 - Contains full phrases and sentences.
 - Validated by regional language experts to ensure accuracy.
- To contribute further, we also collected special words unique to each region, validated by experts.
 - These words enhance dataset richness and authenticity, as they carry cultural and contextual meanings.
 - Recently, a lexical dataset named BRWDS [6] was released, focusing on translating specific words across dialects.
 - Unlike BRWDS, our dataset emphasizes complete special regional words that are exclusive to a region.
 - Such words reflect regional history, lifestyle, and identity, making them essential for building accurate models.
 - Without them, classification systems may fail to capture subtle linguistic cues.
 - To strengthen their impact, oversampling techniques were applied to these words to:
 - * Address class imbalance.
 - * Amplify underrepresented regional features.
 - * Reinforce unique linguistic patterns of each region.
 - This strategy improves the classifier’s sensitivity, performance, and interpretability.

1.4 Research Questions

Based on the problem identified in Section 1.3, this study aims to investigate strategies for improving the classification of regional Bangla text. Specifically, the research focuses on enhancing dataset quality, leveraging linguistic knowledge in oversampling, and evaluating different machine learning and transformer-based approaches. The study addresses the following research questions:

- **RQ1:** Does a tiered and confidence-aware annotation strategy for preparing regional text resources improve the reliability of Bangla regional text classification?
- **RQ2:** Can a linguistically-informed oversampling method using a regional lexicon enhance classification performance across ML and BERT-based models?

- **RQ3:** Which approach—traditional ML, individual Bangla BERT models, or deep ensemble learning—achieves the most effective performance on the high-quality sample dataset?

These research questions are designed to guide the study systematically. RQ1 focuses on the quality and reliability of the dataset, ensuring that the regional texts used for training are accurately annotated. RQ2 investigates methods to address class imbalance, leveraging linguistic knowledge to improve model performance. RQ3 evaluates the effectiveness of different classification approaches, including traditional machine learning, individual Bangla BERT models, and ensemble techniques, to determine the most robust solution for regional Bangla text classification.

1.5 Scope and Limitations

The following section presents the scope and limitations of the study to define its boundaries and focus clearly. It highlights the specific regions, datasets, and objectives considered in this research, while also acknowledging the constraints and challenges that may influence the generalizability of the findings.

- Scope of the study:
 - Focuses on five Bangladeshi regional dialects: Rangpur, Barisal, Noakhali, Sylhet, and Chittagong.
 - Builds on BdRegionText Version 2, an authentic, expert-validated dataset.
 - Aims to explore regional text classification using modern NLP techniques.
- Limitations of the study:
 - Dataset size is small compared to high-resource languages.
 - Does not address all dialects of Bangladesh beyond the five selected ones.
 - Issues like code-switching (Bangla + English) and speech-based dialect recognition are not covered.
 - Findings are limited by data availability and may not generalize to all dialects.

1.6 Research Contributions

Now, to point out the overall contribution of this research article includes:

- Constructed a dataset comprising five regions of Bangladesh: Chittagong, Sylhet, Barishal, Noakhali, and Rangpur, by collecting 4,218 samples from diverse online sources. After validation by regional experts, we curated a cleaned and region-annotated Bangla dialect dataset containing 1,954 high-quality samples.

- Conducted comprehensive experimental analyses on the remaining samples by dividing them into two additional tiers: Medium-confidence and Low-confidence subsets to evaluate how model performance varies across differently reliable annotation groups.
- Developed a cleaned Regional Lexicon dataset comprising words that are exclusively used in specific regions. To address class imbalance, we applied an oversampling technique that leverages these region-specific keywords, enhancing the representation and learning of underrepresented classes.
- Fine-tuned three transformer-based models, BanglaBERT, BUETBERT, and DistilBERT, on the prepared dataset and built a heterogeneous ensemble framework that combines the strengths of all three models using soft voting strategies for our created dataset with high-quality samples.

1.7 Organization of the Thesis

The rest of the research article is organized as follows. Section 2 presents a discussion on the related works relevant to this research. Section 3 provides a comprehensive explanation of the corpus creation and annotation process. Section 4 provides a detailed description of the proposed ensemble framework’s architecture. Section 5 details the experimentation that evaluates the performance of the proposed model and presents a comparative analysis with other models. Finally, Section 6 concludes the study and discusses future research directions.

Chapter 2

Background and Literature Review

Text classification in Bangla has received considerable attention in different domains such as sentiment, spam, fake news, abusive language, and emotion detection. However, work on Bangla regional text classification is still limited due to the scarcity of region-wise resources. To better understand the research landscape, we divide the discussion into four parts: (i) general text classification in Bangla and other languages, (ii) ensemble learning approaches, (iii) Bangla regional speech and text research, and (iv) motivation and research gap.

2.1 General Text Classification in Bangla and Other Languages

A massive amount of research work has been done in various domains of text classification, like sentiment classification [7], spam SMS classification [8], fake news classification [9], identifying vulgarity in Bengali social media text [10], and emotion detection [11] in the Bangla language. However, few research studies have been done in this domain due to low resources and the limited availability of region-wise text in Bangla. This work mostly focuses on regional text classification in Bangla, and in this section, we mainly focus on regional text classification in the context of Bangladesh. It also includes the application of machine learning, deep learning, and ensemble model analysis in text classification problems. Many research works have been done on text classification in the Bangla language—some are lexicon-based, some use traditional machine learning and deep learning approaches, and some apply ensemble-based methods. First, considering the traditional machine learning (ML) techniques, where a classifier automatically extracts the most significant features from a given dataset. One research paper deals with Marathi document classification [12], where they applied SVM, Naïve Bayes, KNN, and Ontolog on a Marathi document classification dataset. Another study [13] highlighted the inter-connection between the Quran and Hadith by combining them in both the testing and

training phases. Various classification techniques, such as Naïve Bayes (NB), SVM, and KNN, were used along with term weighting strategies, particularly Term Frequency – Inverse Document Frequency (TF-IDF), and among them, SVM outperformed the others. This research [14] study deals with Indonesian local languages, focusing on five widely spoken ones. In their paper, they demonstrate an exploratory comparative study by applying some of the best machine learning algorithms, transformation methods, and feature extraction techniques to classify abusive language and hate speech. Their study also shows that the highest F1-score was achieved by applying the SVM algorithm combined with the CC transformation method and using unigram as the feature extraction technique. From neighboring India, this paper [15] presents sentiment analysis of regional languages on social media. The model analyzes customer tweets in Telugu, Tamil, Malayalam, Hindi, and Kannada, classifying them as positive, negative, or neutral using TextBlob and a text classification model, achieving 98% accuracy; however, it does not consider Bangla regional text from the Kolkata region. A recent study [16] on Indo-Aryan language identification demonstrated that integrating pre-trained language models with deep learning architectures and Keras embeddings yields strong performance for closely related languages. Their soft-voting ensemble of Bi-LSTM, GRU, Bi-GRU, and CNN-Bi-LSTM models, trained with augmented data, reported a leading 93% macro F1-score. However, most of the research works in text-based classification studies with various domains involve high-resource languages like English and Chinese, but many languages like Bangla and Iraqi dialect-based texts are considered low-resource languages. Here, this study [17] focuses on Iraqi dialect-based texts for sentiment analysis and proposes hybrid models called CNN-LSTM (Convolutional Neural Networks with Long Short-Term Memory), CNN-GRU (CNN with Gated Recurrent Unit), and AraBERT, which is a deep transformer model that enhances Iraqi sentiment analysis. They also show that among these, AraBERT notably performs better than all other models, achieving 90.18% accuracy.

2.2 Ensemble Learning Approaches

Ensemble learning is a machine learning technique that generates predictions from multiple independent models to create a more accurate and powerful combined forecast. Each machine learning model has its own advantages and disadvantages of its own. We can combine the strengths of these many models with the aid of ensemble models. For Bangla domain research, previous studies have applied traditional machine learning (ML) and deep learning to experiment with their work. However, ensemble techniques are now mostly used for improved performance, as ensemble methods enhance performance by combining the strengths of individual models. In this paper, the authors [18] demonstrate that they created a stacked ensemble model of 4 machine learning models—Naïve Bayes, Random Forest, Support Vector Machine, and XGBoost—to predict whether a comment is a religious hate comment or not. And they achieved 96.6% accuracy for the stacked ensemble model architecture, whereas individual machine learning models only secured

80.7% accuracy, which is a drastic increase in performance. This research [add after] also works with ensemble techniques, where they used an ensemble technique that employs a voting procedure, which is an ensemble of 4 machine learning algorithms named LR, SVM, RF, and MLPC to classify Bangla-English code-mixed spam or ham SMS. They got 96.92% accuracy, which is greater than the individual model accuracy. In this study [19], they also proposed an ensemble of deep learning models with CNN, LSTM, and BiLSTM, where they achieved a 0.81 weighted F1-score for the detection of abusive content in Bengali-English code-mixed texts. This research study [20] works with multi-class textual emotion classification, where they also used an ensemble of three models named LR, RF, and SVM, and they achieved a 62.39% weighted F1-score, which is greater compared to other individual models. Since transformer-based models have become increasingly successful in NLP tasks, recent studies have explored ensemble approaches that leverage various BERT variants to further enhance performance. BERT-based ensemble techniques utilize the complementary strengths of multiple pre-trained models to produce more reliable and accurate predictions, especially in challenging text classification tasks. The research study [21] exploring the transformer Ensemble approach for classifying Cyberbullying Bangla text proposes an ensemble method combining multiple transformer models: multilingual BERT, XLM-RoBERTa, DistilBERT, BanglaBERT, and Bangla-BERT-Base. Their ensemble approach achieved an accuracy of 87.61% and a weighted F1-score of 87.59%. A new ensemble method called MaxOfAvgProb is proposed by this research paper [22], where they present an ensemble approach combining DistilBERT, XLM-RoBERTa, Bangla-BERT-Base, and BanglaBERT. They applied this method to two publicly available datasets: the Bengali Depression Corpus [23] and a depression severity dataset from [24]. The model achieved an F1-score of 63.47% and an accuracy of 62.90% on [23], and an F1-score of 86.35% and an accuracy of 86.45% on [24].

2.3 Bangla Regional Speech and Text Research

Before starting the discussion on regional text classification in Bangla, we found some research works related to Bangla regional dialects in speech. Authors [2] demonstrated their research on Bangla regional speech recognition, focusing on 7 regional dialects and creating a dataset containing 30 hours of Bangla speech from these 7 regions. Another study [25] focused on five districts in Bangladesh: Rangpur, Kishoreganj, Narail, Chittagong, and Narsingdi, resulting in 178 speech recordings totaling the first 39 hours of a regional speech corpus. To enable applications such as virtual voice assistants and regional language processing tools, their research aims to facilitate the development of Automatic Speech Recognition (ASR) systems customized to regional Bangla dialects. Now, moving to the discussion on regional text classification, it is important to note that in the context of Bangla regional text classification, most research so far has primarily focused on dataset creation, as we already discussed earlier in the introduction section. One such dataset is the Vashantor Benchmark Dataset for Automated Translation of Bangla Re-

gional Dialects to Bangla Language. In their study, the authors applied mBERT and Bangla-BERT-base models, achieving an accuracy of 85.86% for Bangla-BERT-base and 84.36% for mBERT. But the Vashantor dataset mainly focuses on translating one Bangla sentence into multiple regional forms. It lacks diversity in terms of different forms of regional text and does not include distinct lexicons from various regions. Another dataset is Bhashamul [4], where they created a corpus using real-world language formats. However, the minimum sentence length is only one, making it more like lexicon data samples rather than actual texts. Such short sentences make it difficult to determine the region they belong to and do not provide sufficient information to identify the context. They applied multiple transformer-based models, and ByT5 performed the best with a 1.995 Word Error Rate (public score). Another group of authors [26] presented ONUBAD, a large and freely available dataset for the automatic translation of Chittagong, Sylhet, and Barisal dialects into Standard Bangla using a Neural Machine Translation (NMT) system, which includes 980 sentences per regional dialect. From their study, we came to know that the minimum sentence length is 1, which may impact model performance, as models may face difficulties in learning meaningful patterns and may also become biased. In their study, no performance evaluation section is found where they discuss various evaluation metrics results to help readers understand how their dataset performs when applying different machine learning or deep learning models. Also, for translation research, widely recognized metrics like WER, METEOR, BLEU, and ROUGE scores are important to understand performance, but no such data has been provided in their paper.

2.4 Motivation and Research Gap

So, after the above discussion, we may consider that Bangla regional dialect text-based research is still comparatively low, though interest is now growing. To increase the reflection of real-world usage, improve language technology, or support social and policy research, regional dialect text-based research is important. These dialects carry deep cultural, emotional, and social meanings that are often lost when translated into the standard language. Thus, classifying texts based on regional dialects is not merely a technical undertaking; rather, it is a culturally and linguistically significant endeavor that supports inclusive research, technology development, and communication. And lastly, we can say that various research works have been done with Bangla regional dialects, but we are the first who work not only with Bangla regional text, but also parallelly develop a cleaned Regional Lexicon dataset comprising words that are exclusively used in specific regions, as we have already mentioned in the introduction section. As far as we know, no researchers have yet taken the initiative to work on this type of task. Because region-specific terms have distinct linguistic, cultural, and contextual connotations that general-purpose lexicons are unable to capture, it is crucial to create a cleaned Regional Lexicon dataset that includes only words used in certain places. These words are essential markers for determining the regional origin of a text since they are ingrained in local communication and are frequently

Table 2.1: Summary of Existing Research Closely Related to Our Work

Prior Work	Key Findings	Limitations
Paper [3]	Created a dataset for automated translation of Bangla into regional dialects; applied mBERT and Bangla-BERT-base, achieving $\sim 85\%$ accuracy.	Focuses on translation rather than classification; lacks lexicon diversity; sentences are very short (length = 1), reducing contextual information.
Paper [4]	Built a corpus in real-world formats; ByT5 achieved the best performance with a WER of 1.995.	Samples are mostly single words or minimal text; insufficient context for identifying dialect; difficult for classification tasks.
Paper [26]	Introduced a parallel translation dataset for Chittagong, Sylhet, and Barisal dialects (980 sentences each).	Very short sentences; only three dialects covered; no performance evaluation with standard metrics (BLEU, ROUGE, WER); no classification focus.
Paper [2, 25]	Developed regional speech corpora covering 5 - 7 dialects; contributed to ASR system development.	Focused on speech, not text; datasets relatively small (30–39 hours); no exploration of text-based regional classification.
Paper [18, 19, 20, 21]	Demonstrate significant improvements over single models for tasks such as hate speech, abusive language, and spam detection.	Applied mostly to standard Bangla or code-mixed text; no targeted use in Bangla regional dialect classification.

absent from standard Bangla. By adding these unique words, we can increase the dataset’s semantic richness, capture dialectal subtleties, and improve classification accuracy.

To better position our work within the existing body of research, we reviewed prior efforts on Bangla regional text and related NLP tasks. Table 2.1 summarizes the most relevant studies, highlighting their key findings as well as their limitations. As can be seen, existing works have largely focused on translation or speech-based resources, with limited attention to the classification of regional Bangla text. Furthermore, many datasets consist of short or artificially constructed sentences, and very few have been validated by linguistic experts. These limitations underscore the need for a comprehensive, expert-validated corpus and robust classification approaches, which we aim to address in this study.

Chapter 3

Annotated Corpus for Experimentation

This section describes the complete pipeline used to build the BdRegionText v2 corpus, beginning with data collection and continuing through annotator validation, filtering, and tiered dataset construction. Its purpose is to clearly present how raw regional texts were gathered, pre-processed, and annotated, and how a region-specific lexical resource was integrated to enrich the linguistic quality of the corpus. Finally, this section outlines a comparison with existing related datasets to highlight the contribution and impact of BdRegionText v2 as a new resource for Bangla regional NLP research.

Social media and online platforms such as YouTube, Facebook, and Instagram heavily

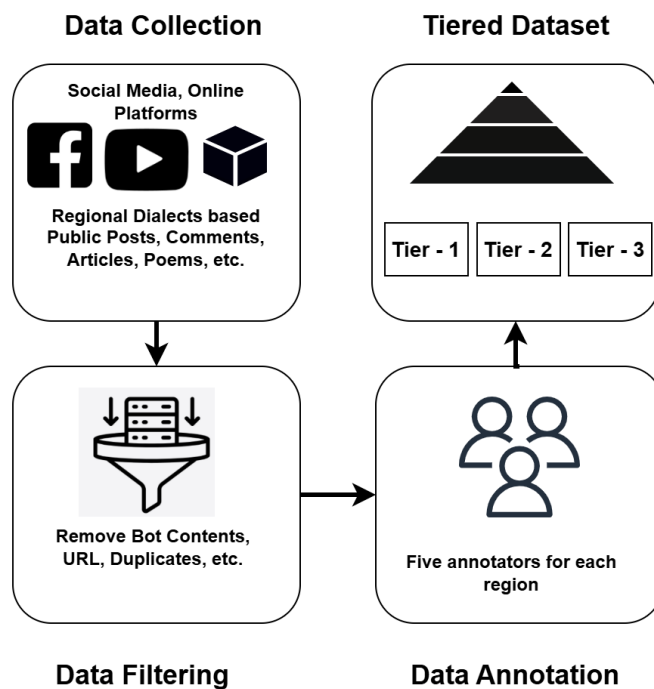


Figure 3.1: Steps of datasets generation

influence today’s world. These platforms have become an integral part of daily life, especially for the younger generation, who spend a significant amount of time on them. As a result, social media has emerged as a top-tier platform for tracking human interaction, trending topics, public opinions, and user interests through comments, posts, and other online activities. Nowadays, people tend to express their thoughts, emotions, and advice on online social media using their regional texts rather than pure English or their first language. As Bengali people, we also tend to express our thoughts through regional texts on social media platforms. Additionally, people are delighted and proud to text in their regional dialects, according to an analysis of internet platforms that highlights this rapidly expanding trend. This motivation influenced us to work with the Bangla regional text. We have constructed a dataset by collecting texts from various sources, such as Facebook group posts, Facebook and YouTube comments, poem writings, regional dialect-based article writings, and more. The figure 3.1 represents the whole procedure of the dataset construction process, and the details of each segment will be discussed below.

3.1 Data collection

Public posts and comments on Bangladeshi social media sites, such as Facebook and YouTube, were used to gather data from the period 2023 to 2024 and ensuring public posts only, no private messages, compliance with platform guidelines, and no personally identifiable information. The dialect data is used strictly for academic research, not for profiling. The model is not intended for discrimination, surveillance, or ranking dialects; its intended use is language preservation, dialect research, and NLP improvement. We focused on users who are likely to employ regional dialects in casual writing. Bot content, duplicates, and messages that were too brief or generic were filtered out. During the data collection process, human verification was systematically incorporated to ensure the authenticity and linguistic correctness of the regional texts. Each collected text was manually reviewed to confirm that it contained meaningful content and accurately represented the intended regional dialect. Since regional Bangla communities on social media often comment on or correct linguistic usage within their own groups or pages, these user-generated interactions served as an additional verification layer. When community members indicated in the comment section that a particular word, phrase, or sentence did not belong to their region, or suggested that a different regional variant was typically used, such texts were excluded from the dataset. These steps reflect a careful and responsible data curation process aimed at minimizing noise, preventing mislabeling, and maximizing the reliability of the collected regional dialect data. Since many users often use code-mixed sentences in various regional texts, we selected only pure Bangla-written (used only Bangla characters) texts from specific regions. The corpus excludes code-mixed sentences where English words appear as part of English phrases or expressions. However, commonly used English-origin loanwords transliterated into Bangla script like ‘কুইজ’ (Quiz), ‘ভিডিও’ (Video), ‘কমেন্ট’ (Comment), ‘ফোন’ (Phone), ‘মার্কেট’ (Market) .etc were retained, as

they are fully naturalized in contemporary Bangla and do not disrupt dialectal patterns. A total of 4,218 text samples were gathered from comments and unofficial public posts, written content on several social media sites in Bangladesh. We picked these examples to represent five different regional dialect zones: Chittagong, Barishal, Rangpur, Noakhali, and Sylhet, even though Bangladesh has more than 55 recognized dialects. These dialect zones were chosen because they reflect popular speech communities with unique lexical and phonetic patterns that frequently show up on social media. Due to uneven digital participation, uneven regional representation online, and a lack of native speakers available for annotation, it was not possible to gather trustworthy user-generated text for all dialects. Therefore, rather than being a comprehensive representation of all Bangla dialects, the current dataset should be viewed as a targeted resource created for five high-usage regions. Future work will include adding more dialects to the dataset. The table 3.1 represents the sample text of each region, along with its actual Bangla and English meanings for proper understanding.

Table 3.1: Five regional Bengali text samples with Bangla meaning and English translation

Region	Regional Text	Bangla Meaning	English Meaning
Chittagong	আবু ছালিকের বউ ইবে বেশি মিছা হতা হয়!	আবু ছালিকের বউটাই সবচেয়ে বেশি মিথ্যা কথা বলে	Abu Salek's wife lies the most.
Barishal	এত্তের একটা জেলায় বোলে মোর লইগ্লা পোলা নাই।	এত বড় একটা জেলায় বলে আমার জন্য কোনো ছেলে নেই।	In such a big district, there's not a single boy for me!
Rangpur	তুমরা জানিয়া করবেন কি?	তোমরা জেনে করবে কি?	Even if you know, what can you do about it?
Noakhali	আম্নেরা জানেন ল্যাংলা আম অনেক মিষ্টি অয়।	আপনারা জানেন লাংড়া আম অনেক মিষ্টি হয়।	You all know Langra mangoes are very sweet.
Sylhet	সারা বাংলাদেশো এরা রইদ পড়ছে খুব	সারা দেশে আজ খুব রোদ পড়ছে	It's very sunny all over Bangladesh.

3.2 Annotation process

For the annotation process, we engaged five native Bangla speakers from Chittagong, Barishal, Rangpur, Noakhali, and Sylhet, each representing a different region, to annotate the collected data. As stated earlier, the data collection process included human verification to ensure the authenticity of regional texts. Accordingly, one native annotator was assigned per region to complete the final annotation stage. To reduce bias and maintain fairness:

- Five native Bangla speakers from five different regions were selected as annotators.
- The same Excel file containing 4,218 unlabeled texts was sent to each annotator.
- The original regional designations were hidden from the annotators, who simply marked "Yes" when a sentence fit their dialect and "No" otherwise.

- This produced a Yes/No binary matrix for every sentence in each of the five zones.

3.3 Filtering and tiered dataset construction

The original regional designations were hidden from the annotators, and the collected Excel files were compiled together to construct the tiered dataset.

Table 3.2: Criteria used to construct the tiered dataset.

Original Region	Regional Text	Chittagong (Yes/No)	Barishal (Yes/No)	Rangpur (Yes/No)	Noakhali (Yes/No)	Sylhet (Yes/No)	Selected Tier
Chittagong	বন্দা পাঞ্জাবি ইবে হতো?	Yes	No	No	No	No	Tier 1
Rangpur	তুমরা জানিয়া করবেন কি	No	No	Yes	No	No	Tier 1
Barishal	আসল বরিশাইল্ল্যা চেনতে তুমি চাও?	No	Yes	No	Yes	No	Tier 2
Barishal	গাজী ট্যাংকের লাহান মজবুত, শক্ত, জং পড়বেনা, মরচাও	No	No	No	No	No	Tier 3
Sylhet	আইছেরে ভাই কলি কাল, ছাগিয়ে চাটে বাঘর গাল।	No	No	No	Yes	No	Discarded

This table presents 3.2 examples illustrating how the three-tiered dataset was created. The following steps were followed:

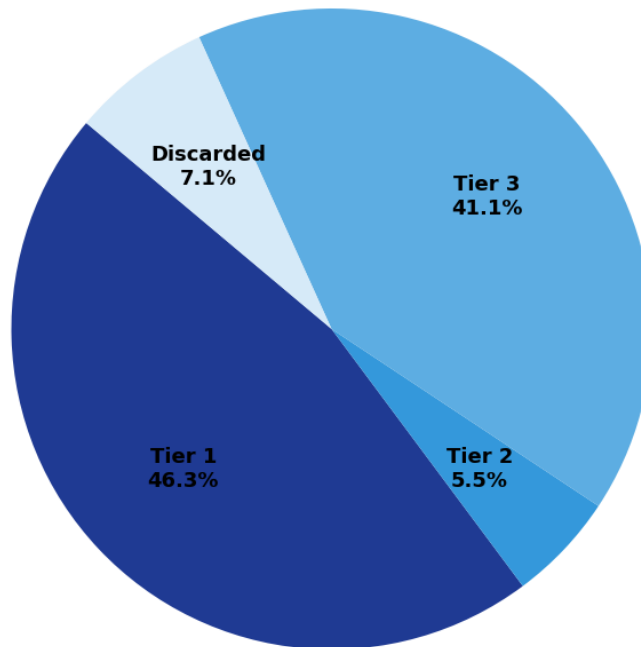


Figure 3.2: Distribution of tiered and discarded dataset

- **Tier 1 (High-quality subset):** Exactly one annotator marked “Yes”, and that annotator belonged to the sentence’s true region. All four annotators from the other

regions marked “No”. This represents the strongest form of positive evidence: only the native speaker of the correct region recognized the sentence as their dialect.

- **Tier 2 (Medium-confidence subset):** Two annotators marked “Yes”, and among them, one was from the true region. This indicates that the sentence contains regional characteristics, but with weaker exclusivity or overlapping lexical cues.
- **Tier 3 (Low-confidence subset):** All five annotators marked “No”. These sentences lacked clear region-specific markers or were too general for any annotator to confidently claim as their own dialect.

Following the above rules, we structured the dataset into three tiers and discarded a small portion that did not meet the assigned agreement criteria. Figure 3.2 illustrates how the full dataset was proportionally distributed after validation by the annotators. Specifically, 46.3% of the total data was classified as Tier 1, while 5.5% and 41.1% were categorized as Tier 2 and Tier 3, respectively. Lastly, 7.1% of the samples were discarded because they failed to meet any condition and were found to be misleading or confusing to the annotators. Only the Strict Dataset was used for training and evaluating the proposed classification models, ensuring the highest possible label reliability. Out of 4,218

Table 3.3: Summary of the filtered dataset tiers based on annotator agreement.

Dataset	Size	Agreement Type	Usage
Tier 1	1,954	1 Yes (true class), 4 No	Experiment 1 & 2
Tier 2	232	2 Yes (1 true class), 3 No	Experiment 1
Tier 3	1,732	All No	Experiment 1

collected texts, 1,954 were categorized as Tier 1, 1,732 as Tier 3, and 232 as Tier 2. The remaining texts were discarded. Table 3.3 summarizes the size, agreement type, and experimental usage of each tier. We employed a deterministic filtering approach rather than using statistical agreement metrics, such as Cohen’s or Fleiss’ Kappa. Cohen’s Kappa is typically used in scenarios where all annotators evaluate the same items from a shared perspective—for example, medical diagnoses of a single disease or English teachers grading the same grammatical errors. In contrast, our annotation setting is fundamentally different. Each text sample was manually collected from region-specific public posts or comments by a researcher familiar with that region (e.g., a sample originating from Region A was first validated by a human curator knowledgeable about Region A). After this initial verification, at least one native speaker from the same region reviewed the sample to confirm its dialectal authenticity. Thus, every sample effectively received two levels of validation—one from the collector and one from a native annotator. Because annotators did not share a uniform evaluation perspective across all regions, statistical inter-rater agreement metrics such as Cohen’s Kappa were not directly applicable. We acknowledge that, despite the multi-stage verification process, some degree of subjective bias may remain. To mitigate this, we adopted a deterministic positive-evidence rule: a sample was

included in the high-confidence Tier-1 subset only if it received an unambiguous “Yes” vote from its corresponding region’s native annotator. This ensured that Tier-1 contains only strong, confidently identified region-specific texts. In the future, we plan to involve another five annotators again so that each sample is evaluated by multiple raters, enabling statistically robust agreement measures and further strengthening the dataset.

Now, after finalizing all the Tiered datasets, in Table 3.4, we have demonstrated the minimum and maximum word counts in each text, the average word count for each tier-wise dataset, the number of unique words present in each class, and the number of class-wise instances for each tier.

Table 3.4: Overview of the dataset by tier in numerical form (This section contains all numerical information, including the dataset’s maximum, minimum, average, and unique word counts and classwise instance counts.)

Tier	Region	Minimum Word Count	Maximum Word Count	Average Word Count	Unique Word Count	Instance Count
Tier 1	Chittagong	2	48	8	2084	727
	Barishal	3	24	7	1239	394
	Rangpur	2	19	7	437	130
	Noakhali	2	26	8	999	261
	Sylhet	2	42	11	1878	442
Tier 2	Chittagong	2	10	5	235	63
	Barishal	4	13	7	267	55
	Rangpur	4	13	7	51	10
	Noakhali	3	22	7	208	38
	Sylhet	2	33	11	453	66
Tier 3	Chittagong	2	16	6	624	172
	Barishal	2	23	7	1410	422
	Rangpur	2	24	7	961	308
	Noakhali	2	24	8	1582	411
	Sylhet	2	28	7	1393	419

3.4 Regional lexicon for oversampling

To address class imbalance, we also took additional steps to make a significant contribution to this research on dialects. We collected special words from each region that are unique to that specific region. The same procedure is also followed here to collect the region-specific unique words, similar to the collection of regional text from the five regions. Here, the collector also utilized human interaction to fully verify the region-specific words in the first

step of collection. To validate these region-specific special words, we involved the same five annotators, each representing one of the five regions. The following steps were carried out to develop the Regional Lexicon for Oversampling:

- The terms were manually selected and validated by all five annotators
- A special word was selected when one annotator marked “Yes”, and that annotator belonged to the sentences true region indicating that the word is used in their region and genuinely belongs to that region. In contrast, the other annotators marked “NO”, meaning the word is not used in their regions for that particular meaning. Only in such cases was the word considered region-specific and selected for that region.
- In cases where two or more annotators marked “Yes”, or when all annotators marked “NO”, those words were discarded from the Regional Lexicon for Oversampling. Only words with a single “Yes” from the corresponding region and “NO” from all others were retained.
- These region-specific words were then used to oversample the underrepresented classes during training, enhancing the model’s robustness and overall performance.

Now, table 3.5 illustrates sample special words from five regions, along with their English meanings and typical usage contexts. And we believe that such words help preserve semantic and cultural nuances that are crucial for accurate regional classification.

Table 3.5: Sample region-specific special words used for oversampling

Region	Special Word	Meaning (English)	Usage Context
Chittagong	দইজ্জা	Sea beach	Commonly used to refer to sea beach.
Barishal	বিলোই	Cat	Regional term for ‘cat’.
Rangpur	চ্যাংড়া/চ্যাংড়ি	Boys/Girls	Used to refer to young boys and girls.
Noakhali	হাকনা কেলা	Ripe banana	Dialectal phrase for ripe banana.
Sylhet	ফুরি	Girl	Refers to a young girl in casual speech.

3.5 Comparison analysis with relevant dataset

As we have already discussed previously, the various existing datasets that were developed in Bangla regional dialects vary significantly in terms of size, scope, and linguistic validation. Since the linguistic characteristics and dialectal variations of Bangla spoken in Bangladesh differ from those of Bangla used in certain regions of India, particularly in phonetics and regional usage, this comparison focuses exclusively on Bangladeshi works. This

Table 3.6: Comparison of existing Bangla regional text datasets.

Dataset	Year	Region	Key Findings	Limitations
Vashantor [3]	2023	Chittagong, Noakhali, Sylhet, Barishal, Mymensingh	Vashantor: A collection of 32,500 sentences representing five regional Bangla dialects, combining Bangla, Banglish, and English.	Content is largely translated into regional variants (e.g., standard Bangla → Sylhet), which may not reflect everyday usage or full linguistic diversity.
Bhashamul [4]	2024	Rangpur, Tangail, Kishoreganj, Narail, Narsingdi	Bhashamul: Six-region dataset with 8,941 test and 30,311 training samples.	A minimum sentence length of one word provides too little context for reliable regional identification; it is unclear whether regional experts validated the data.
ONUBAD [26]	2025	Chittagong, Barishal, Sylhet	ONUBAD: Comparable corpus with English translations for 1,540 words, 130 clauses, and 980 sentences per dialect.	One-word minimum sentences hinder semantic understanding, especially for tasks like sentiment, hate speech, or emotion classification.
BdRegionText V1 [5]	2025	Chittagong, Noakhali, Rangpur, Barishal	BdRegionText V1: 2,573 texts across four regional dialects.	Not validated by regional experts, raising concerns about linguistic authenticity.
This work	2025	Sylhet, Rangpur, Barishal, Chittagong, Noakhali	BdRegionText V2: 4,218 texts across five dialects, validated by regional experts; introduces a 3-tier structure with Tier 1 (1,954 samples) as the highest-confidence subset.	Tier 3 (lower-confidence subset) may be re-validated by additional experts to expand Tier 1 and improve linguistic richness.

ensures a more coherent, contextually aligned, and linguistically meaningful evaluation of prior research. In this table, we have highlighted the key findings of their works along with the limitations that we have addressed in our study. Although there is still more to be done to further enrich this research, we believe that the creation of additional resources for Bangla regional dialects will have a significant impact on Bangla NLP tasks. Now, Table 3.6 presents a comparative overview of the most prominent works. For instance, the Vashantor dataset primarily focused on generating translated variants of standard Bangla into regional forms, which raises questions about the authenticity of naturally occurring dialectal features. Similarly, the Bhashamul dataset introduced a large-scale resource covering six regions, yet the absence of clear expert validation limits its reliability for fine-grained classification tasks. The ONUBAD corpus emphasized parallel English translations, but its short-length samples restrict deeper semantic analysis. BdRegionText V1, though covering four dialects, also lacked regional expert verification, reducing linguistic authenticity. In contrast, our proposed BdRegionText V2 dataset expands coverage to five

regions, ensures expert validation for quality assurance, and introduces a three-tier confidence structure that balances size with reliability. This structured approach makes the dataset more suitable not only for regional text classification but also for broader downstream tasks such as sentiment analysis, code-mixed text processing, and cross-dialectal linguistic research.

Chapter 4

Methodology

This chapter illustrates the system’s general design, which is then examined in detail to reveal the specific elements of the proposed strategy. The sequential sequence of procedures,

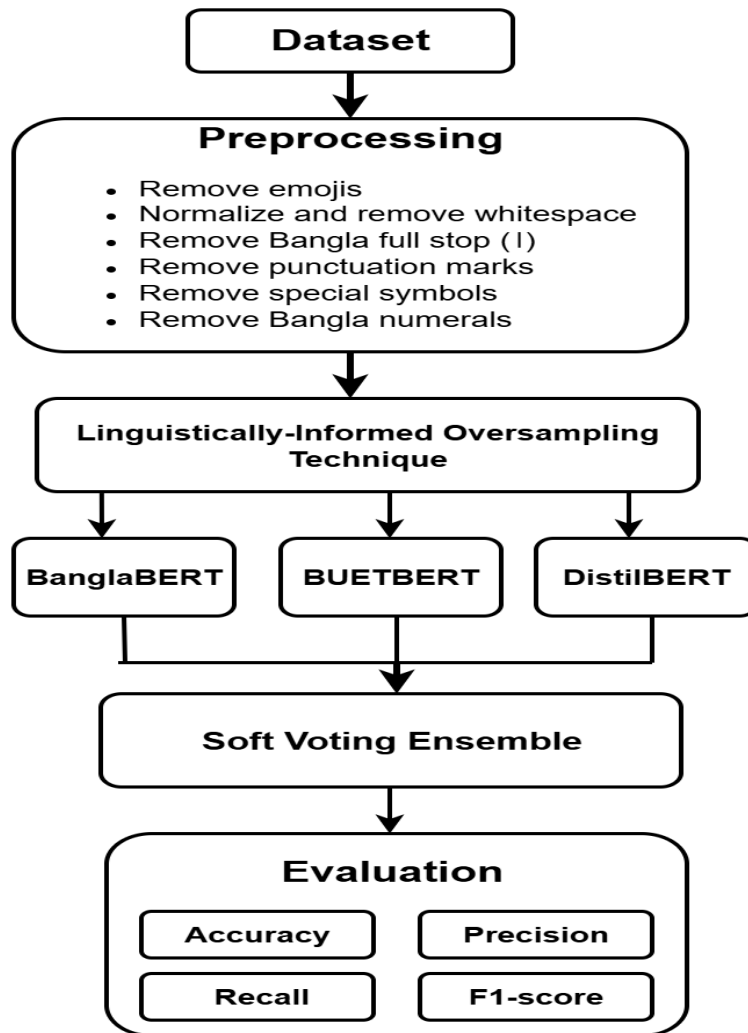


Figure 4.1: The proposed system architecture

including preprocessing, oversampling, model training with multiple BERT variations, and the final soft voting ensemble used for prediction, is highlighted by this design. A high-level overview of the complete pipeline is provided in Figure 4.1, clearly illustrating how each part contributes to the final categorization output.

4.1 Preprocessing

The original Bangla text data underwent a thorough preparation procedure to guarantee high-quality input for the ensuing oversampling and classification steps. Normalizing the input and eliminating noise, which could otherwise impair model performance, was the main goal of this step. First, to remove non-textual features that do not contribute to semantic understanding, all emojis were eliminated using the demoji library. Then, a single space character was used to standardize all Unicode space characters and various types of whitespace. This includes not only regular spaces but also non-breaking spaces and other uncommon Unicode whitespace forms. To prevent tokenization issues, the Bangla full stop (।), character—frequently used in Bangla scripts to indicate sentence boundaries—was replaced with a space. Following this, all punctuation marks, including both Western and Bangla-specific punctuation, were removed from the text. Special symbols such as rare characters, typographic signs, and currency symbols were also eliminated. In addition, all Bangla numerals (০-৯) were removed to reduce sparsity and exclude numeric information that does not support text classification tasks. Finally, to ensure clean token boundaries, the text was further processed by stripping leading and trailing spaces and collapsing multiple internal spaces into a single space. It is important to note that stop words were not removed in this step, as each region has its own unique vocabulary and spelling variations—removing such words might discard significant features. By standardizing the textual input and reducing noise, this preprocessing step helped enhance the performance of the oversampling technique and the subsequent BERT-based classification models.

4.2 Proposed linguistically-informed oversampling technique

In this section, we discuss how we addressed the class imbalance problem in this research. To handle class imbalances in this regional Bangla dataset, we developed a linguistically informed oversampling method that differs from traditional approaches. Specifically, we first identified region-specific words—dialectal terms unique to each region—as described in Section 3.4, where we detailed the preparation of these special regional dialect words.

Here, the algorithm 1 demonstrates the proposed Linguistically-Informed Oversampling Technique. Here, RSW = Region-Specific Word dictionary, and the table 3.5 demonstrates some examples of Region-specific special words. Given a dataset D containing N samples and a set of class labels $C = \{c_1, c_2, \dots, c_k\}$, the goal is to balance the dataset such that all classes reach the size of the majority class, denoted as N_{\max} . For each class c_i , the subset D_i includes all instances with that label. If the number of samples $n_i = |D_i|$

Algorithm 1. Linguistically-informed oversampling based on region-specific words

Input: Dataset D with N samples; class labels $C = \{c_1, c_2, \dots, c_k\}$; region-specific word map RSW ; majority class size N_{\max}

Output: Balanced dataset D_{balanced}

```

1  $D_{\text{balanced}} \leftarrow D$ 
2 for each class  $c_i \in C$  do
3    $D_i \leftarrow \{x \in D \mid \text{label}(x) = c_i\}$ 
4   if  $|D_i| < N_{\max}$  then
5      $\text{special\_words} \leftarrow RSW[c_i]$ 
6      $D_i^{\text{special}} \leftarrow \{x \in D_i \mid x \text{ contains any word in } \text{special\_words}\}$ 
7     while  $|D_i| < N_{\max}$  do
8       select  $s \in D_i^{\text{special}}$  uniformly at random
9       append a copy of  $s$  to  $D_{\text{balanced}}$ 
10       $D_i \leftarrow D_i \cup \{s\}$ 
11 return  $D_{\text{balanced}}$ 

```

is less than N_{\max} , the algorithm references a predefined region-specific word map RSW to extract class-specific linguistic cues. These cues are used to filter D_i into D_i^{special} , a subset of samples that contain any of the region-specific words. While D_i remains smaller than N_{\max} , the algorithm randomly samples from D_i^{special} and appends duplicate instances to the dataset. The final output, D_{balanced} , ensures uniform class distribution while preserving regional lexical diversity. For example, if the majority class had 500 samples and the 'Nowoakhali' class had 300, we replicated texts with unique 'Nowoakhali' words until we reached 500. During oversampling, this technique preserves semantic richness while increasing the presence of class-representative features. Instead of mindlessly copying every sample, this helps the model learn linguistic patterns that distinguish each region by selectively using duplicate examples that have significant dialectal features.

4.3 Conventional Method

This research paper evaluated the performance improvement associated with the suggested Linguistically-Informed Oversampling Technique by comparing it with the most popular text categorization methods. These models mostly belong to the fields of Deep Learning (DL) and Machine Learning (ML). Logistic Regression (LR), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Random Forest, Stochastic Gradient Descent (SGD), Gradient Boosting, and XGBoost are some of the conventional machine learning classifiers that we tested with. These models' classification performance was compared after they were trained with the proper hyperparameters. For regional text classification, we used three transformer-based deep learning models that have already been trained: DistilBERT, Bangla-BERT, and BUET-BERT. These models can comprehend contextual representations of Bangla text and are based on the BERT architecture.

4.3.1 BanglaBERT

The transformer model BanglaBERT is a monolingual BERT-based model that has been pre-trained solely on a sizable Bangla corpus [27]. It is tailored for Bangla syntax and semantics and adheres to the standard BERT architecture. In order to predict the original tokens based on contextual information, the model employs the Masked Language Modeling (MLM) pretraining objective, where 15% of input tokens are randomly masked and predicted based on context. In terms of mathematics, the MLM goal reduces the cross-entropy loss:

$$\mathcal{L}_{\text{MLM}} = - \sum_{i \in \mathcal{M}} \log P(x_i | x_{\setminus i}) \quad (4.1)$$

where $x_{\setminus i}$ denotes the sequence with the masked token and \mathcal{M} is the set of masked positions, where x_i is the original token at position i . Due to its monolingual specialization, BanglaBERT achieves notable performance gains over multilingual baselines like mBERT in syntactic and semantic evaluations and its broad corpus coverage. BanglaBERT is particularly suitable for handling rare regional lexical markers and morphologically rich dialect words, making it highly relevant for our regional text classification task.

4.3.2 BUETBERT

Another Bangla monolingual transformer model, BUETBERT, was created by BUET’s CSE department. Compared to its multilingual counterparts, BUETBERT performs better at comprehending contextual references after being trained on a rich and varied corpus of Bangla text. Similar to BanglaBERT, it pretrains using MLM. Using the same setup as BanglaBERT, we refine the csebuetnlp/banglabert model for downstream classification. A fully connected softmax layer processes the input’s final representation, $h_{[\text{CLS}]}$:

$$P(y | X) = \text{softmax}(Wh_{[\text{CLS}]} + b) \quad (4.2)$$

The cross-entropy loss over the class labels is minimized to fine-tune the model, where $W \in \mathbb{R}^{k \times d}$, $b \in \mathbb{R}^k$, d is the hidden size, and k is the number of output classes. BUETBERT consistently surpasses both multilingual and certain monolingual models on classification benchmarks, attributed to its more extensive and domain-rich training corpus. BUETBERT’s specialization on certain Bangla corpora allows it to capture regionally relevant linguistic patterns, though it has a smaller training corpus than BanglaBERT. This makes BUETBERT a strong complement for regional text classification, especially in combination with lexical oversampling.

4.3.3 DistilBERT

DistilBERT is a distilled version of BERT that operates 60% faster and uses 40% fewer parameters while maintaining 97% of BERT’s functionality. Through the use of a knowledge distillation process, the student model (DistilBERT) is trained to minimize the Kullback-

Leibler (KL) divergence to mimic the behavior of a larger teacher model (BERT):

$$\mathcal{L}_{\text{KD}} = \text{KL}(P_T \parallel P_S) = \sum_i P_T(i) \log \frac{P_T(i)}{P_S(i)} \quad (4.3)$$

Where the output probability distributions of the student and teacher models are denoted by P_T and P_S , respectively. DistilBERT provides a lightweight and efficient alternative, making it especially suitable for resource-limited environments, while still delivering competitive performance. The efficiency of DistilBERT enables faster experimentation, and when combined with an oversampling strategy, it exhibits improved learning of rare region-specific tokens.

4.4 Soft voting-based deep ensemble

A collection of heterogeneous deep learning models, based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, was trained using the balanced dataset following the preprocessing and oversampling stages. Specifically, three distinct BERT variants were employed, each initialized with different model configurations or pre-trained checkpoints, and were trained in parallel on the same input data. By incorporating BanglaBERT, BUETBERT, and DistilBERT, the ensemble was designed to capture diverse linguistic patterns and modeling behaviors. Each model was independently fine-tuned and saved for subsequent inference. To aggregate the predictions, a soft voting ensemble technique was applied. Unlike hard voting, which relies solely on predicted class labels, soft voting combines the probability distributions output by each model across the target classes. These probability vectors were merged, typically by averaging, and the final prediction was assigned to the class with the highest averaged probability. This approach allows the ensemble to account for the confidence levels of individual models, resulting in predictions that are more stable and accurate. The use of soft voting is particularly beneficial when the constituent models exhibit complementary strengths, as it leverages their collective knowledge and mitigates individual prediction errors. Ultimately, this ensemble strategy enhances the robustness and generalization ability of the overall classification framework.

Chapter 5

Results and Discussion

This chapter outlines the practical aspects and analytical outcomes of the study, focusing on both the implementation and evaluation phases. It begins with a detailed description of the experimental setup, including the computational environment, dataset preparation with tiered configurations, the details of the train and test data split, and the fine-tuning of transformer-based models. The primary goal is to demonstrate how the proposed methodology was systematically implemented to ensure reliable and reproducible results. Following the implementation details, the chapter presents and analyzes the results obtained from the experiments. The discussion is organized into two distinct parts, corresponding to the two experimental setups. Experiment 1 provides a comparative evaluation across all three tiers, where Tier 1 is exclusively used as a testing set to assess the robustness and generalization capabilities of the top-performing BERT models in a diverse and challenging environment. Experiment 2, on the other hand, focuses solely on Tier 1, showcasing results derived from various combinations of machine learning models, feature extraction techniques, and BERT-based approaches, including the proposed ensemble architecture.

5.1 Environmental setup

The experimental configuration was selected to ensure the effective operation and reproducibility of the suggested system. The approach’s viability for application in resource-constrained contexts is demonstrated by the low computational resources required. The main details of the setting and system used for the experiment are listed below:

- Operating System: Windows 10
- Processor: Intel Core i5-4300M CPU @ 2.60GHz
- Memory (RAM): 8 GB
- Development Platform: Google Colab (Cloud IDE)
- Programming Language: Python

5.2 Experimental configuration for tiered datasets

As already mentioned in the corpus creation section, we developed a tiered dataset, where Tier 1 represents the strictly cleaned data and serves as the primary focus of this research study. However, an additional experiment was conducted using the other two datasets, Tier 2 and Tier 3. Although Tier 2 and Tier 3 were also collected from regional sources during the initial data collection phase, they failed to meet the annotation agreement strategy and thus were not included in the main pipeline. Therefore, the core experimental setup, including the proposed system architecture diagram, focuses solely on Tier 1. Within this setting, we applied a variety of machine learning (ML) and BERT-based models, exploring multiple feature extraction and embedding techniques, and concluded with a soft voting ensemble strategy as a major contribution. To assess the broader applicability of our models, we also conducted another Experiment, where we utilized all three tiers (Tier 1, Tier 2, and Tier 3) and applied the three best-performing BERT models individually. This helped demonstrate how well these models generalize across noisier and less strictly annotated data. Now, Table 5.1 represents the overall experimental configuration for tiered datasets.

Table 5.1: Summary of the Experimental Configuration.

Experiment	Dataset
Experiment 1	Tier 1, Tier 2, Tier 3
Experiment 2	Tier 1

5.3 Train and Test Data Split

For all experiments on Tier 1, the dataset was divided into a training set and a testing set a ratio of 80:20 was maintained. 20% of the data was used for model testing, and the remaining 80% was used for training. To ensure compatibility across all experiments, a 20% stratified fixed test set was selected from the 1954 Tier 1 samples and kept aside. The remaining 80% was used for training and for applying oversampling techniques. Although we also present k -fold cross-validation results to demonstrate the strengths of individual BERT models under both the hold-out and the k -fold settings, stratified k -fold cross-validation was used in these experiments as well, and the proposed oversampling technique was applied only to each training fold. Stratification was essential because the tiered datasets are imbalanced by construction, and preserving proportional class representation prevents sampling bias during model evaluation. No external class weights or loss re-weighting methods were used. For BERT-based models, prediction thresholds were kept at the default softmax decision rule, as the multiclass setting inherently selects the highest-probability label. No threshold optimization was applied because the task does not involve

binary discrimination. These choices align to maintain comparable conditions across the ML pipeline and the BERT experiments.

5.4 Transformer-based model fine-tuning setup

In this research work, the three transformer-based models named BanglaBERT, BUETBERT, and DistilBERT are fine-tuned for the regional Bangla text classification using HuggingFace’s `AutoModelForSequenceClassification` and the `Trainer` framework. The texts were tokenized using the corresponding pre-trained tokenizer with a maximum sequence length of 128, including truncation and padding. All models were trained for 3 epochs with a batch size of 8. The fine-tuning configuration employed a learning rate of 2×10^{-5} , linear learning-rate scheduling, a warmup ratio of 0.1, and a weight decay of 0.01. A dropout probability of 0.2 was applied to both the hidden states and attention layers. Evaluation was performed at the end of each epoch, and accuracy and weighted F1-score were computed using a custom evaluation function. The same configuration was applied for non-oversampled and oversampled training sets. Table 5.2 lists the transformer models used in this study along with their corresponding HuggingFace identifiers and pre-training types.

Table 5.2: Transformer models and fine-tuning configurations for regional Bangla text classification

Model	Huggingface Identifier	Pre-training Type
BanglaBERT	sagorsarker/bangla-bert-base	Bangla-only, monolingual
BUETBERT	csebuetnlp/banglabert	Bangla-only, monolingual
DistilBERT	distilbert-base-multilingual-cased	Distilled multilingual

5.5 Experiment 1: BERT-based model evaluation across all three tiered datasets

For Experiment 1, we used the three-tier structure of this dataset and applied BERT-based models to observe their performance. In this experiment, we followed the same procedure and methodology, including the same preprocessing steps. After preprocessing, we applied the Linguistically-Informed Oversampling Technique and then used the three BERT models: BanglaBERT, BUETBERT, and DistilBERT. Here, the Tier 2 and Tier 3 datasets were used for training, while the Tier 1 dataset, our high-quality subset with strong annotator confidence, was used for testing the models. In Table 5.3, we have presented

Table 5.3: Performance of transformer-based models across all three tiered datasets before applying oversampling (summary metrics).

Model	Accuracy (%)	Weighted F1 (%)
BanglaBERT	62.03	61.74
BUETBERT	42.48	33.38
DistilBERT	54.50	51.18

the performance of the three BERT models on the three-tiered dataset before applying oversampling, and in Table 5.4, we have shown the results after applying the proposed Linguistically-Informed Oversampling technique. The comparison clearly demonstrates

Table 5.4: Precision, Recall, F1-score for regional classes along with overall Accuracy (%) and Weighted F1-score (%) for transformer-based Bangla embeddings (after oversampling) across all three tiered datasets.

Model	Region	Precision	Recall	F1-Score	Accuracy	F1-Score (Weighted)
BanglaBERT	Rangpur	0.58	0.75	0.65	67.45	67.62
	Barishal	0.74	0.75	0.74		
	Noakhali	0.53	0.72	0.61		
	Sylhet	0.59	0.79	0.67		
	Chittagong	0.90	0.54	0.67		
BUETBERT	Rangpur	0.53	0.78	0.63	59.26	59.65
	Barishal	0.63	0.83	0.72		
	Noakhali	0.38	0.68	0.49		
	Sylhet	0.47	0.39	0.43		
	Chittagong	0.92	0.52	0.67		
DistilBERT	Rangpur	0.52	0.75	0.62	65.04	65.58
	Barishal	0.67	0.79	0.72		
	Noakhali	0.43	0.70	0.53		
	Sylhet	0.65	0.67	0.66		
	Chittagong	0.91	0.53	0.67		

that oversampling improves model performance. Balancing the regional classes allowed the models to learn region-specific patterns more effectively. Since the training set for this experiment is smaller than the test set, the models struggled on the imbalanced data, but their performance improved once the oversampling technique was applied. An interesting observation in this experiment is that all BERT-based models achieved slightly higher weighted F1-scores than their corresponding accuracy scores. In Table 5.4, BanglaBERT performs the best with an accuracy of 67.45%, followed by DistilBERT and BUETBERT with accuracies of 65.04% and 59.26%, respectively. BanglaBERT also achieves the highest weighted F1-score of 67.62%. Across all models, the weighted F1-score remains slightly higher than the accuracy, indicating that the models are not biased toward the dominant

class (e.g., Chittagong) and are successfully capturing meaningful patterns from the less frequent classes. This trend suggests that the models generalize well across regional dialects rather than overfitting to any single region. Tier 2 and Tier 3 also play an important conceptual role in the overall learning framework. Tier 2 contains mixed or partially ambiguous dialectal cues, while Tier 3 represents very weak or noisy signals that annotators could not confidently label. Exposure to this range of variation helps the models learn broader region-specific characteristics that commonly appear in real-world social media text. These tiers also offer promising opportunities for future work. For example, Tier 1 can be used as a high-confidence seed set, with Tier 2 and Tier 3 integrated into semi-supervised learning or domain-adaptation pipelines, such as pseudo-labeling or consistency regularization. These approaches can further improve model robustness by allowing the model to learn from both clean and noisy dialectal distributions. Although such methods were beyond the scope of the present study, the tiered dataset structure provides a strong foundation for future extensions.

5.6 Experiment 2: ML, BERT, and ensemble model performance on tier 1 data

In this section, we have explored several feature extraction and embedding techniques to analyze the performance of our dataset, Tier 1 and observe how it responds to different approaches. Additionally, different BERT models were applied to evaluate and compare their performance on this dataset. Normally, regional text classification behaves differently from standard text classification. Many regions share a large portion of vocabulary, while the distinct regional cues often appear in a few rare or uniquely used words. These differences are not captured well by simple word-frequency statistics like TF-IDF. In this setting, identifying subtle contextual and subword-level patterns becomes more important than counting how often a word appears. BERT-based models are effective here because they learn contextual meaning, token interactions, and subword representations, enabling them to capture regional expressions that traditional feature extraction techniques fail to represent. As a result, transformer models outperform TF-IDF and similar statistical methods in regional Bangla text classification, where context plays a central role. The comparative analysis across traditional feature extraction methods (TF-IDF, TF-IDF + PCA, TF-IDF + SVD), static and pretrained embeddings (Word2Vec, FastText, pretrained FastText) provides valuable insights into their impact on regional, imbalanced Bangla text classification. And also in this section, we have also discussed the BERT-based models and their soft voting ensemble after applying them to this dataset, along with a detailed analysis of their performance.

5.6.1 The impacts of applying Feature extraction techniques TF-IDF, TF-IDF + PCA, TF-IDF+SVD

This table 5.5 represents an overall comparative analysis for 7 traditional machine learning algorithms named Logistic Regression (LR), GradientBoost, Multi-layer Perceptron Classifier (MLPC), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and XGBoost, with three feature extraction techniques: TF-IDF, TF-IDF + PCA, and TF-IDF + SVD, for understanding how and which feature extraction method performs on this strictly cleaned dataset without and with applying proposed oversampling techniques. We just demonstrated the accuracy of the percentage value for each model in this table.

Table 5.5: Overall accuracy (%) comparison of feature extraction techniques: TF-IDF, TF-IDF + PCA, and TF-IDF + SVD, with and without oversampling

Model	Without Oversampling			With Oversampling		
	TF-IDF	TF-IDF+PCA	TF-IDF+SVD	TF-IDF	TF-IDF+PCA	TF-IDF+SVD
LR	63	60	60	65	59	59
GradientBoost	59	59	62	57	59	59
MLPC	63	65	66	62	66	63
Random Forest	60	56	58	57	57	56
SGD	67	68	67	65	60	61
SVM	65	59	59	64	58	56
XGBoost	58	64	61	55	60	60

Across the three feature extraction settings, TF-IDF remains the most reliable representation for traditional ML models without oversampling, except for LR. After oversampling techniques, LR performs better than without oversampling (63 to 65). PCA and SVD offer only slight improvements in accuracy because these models already perform well on sparse TF-IDF features, and dimensionality reduction eliminates some discriminative, high-frequency regional terms. The slight gains seen in MLPC with PCA suggest that lower-dimensional inputs help neural and boosting models avoid overfitting, but the effect is modest. After oversampling, performance does not improve consistently because ML models treat TF-IDF features as independent word counts; duplicating samples does not introduce new lexical variation and may even amplify noise. Overall, the results indicate that classical ML methods remain stable across feature settings, but none of the combinations extract region-specific linguistic patterns as effectively as transformer models.

5.6.2 The impacts of applying three embedding techniques named Word2Vec, FastText (Pre) and FastText

On this unbalanced regional dataset, Table 5.6 shows that FastText (pretrained) embeddings deliver the strongest and most consistent performance across all models. Since these vectors are trained on large Bangla corpora, they can represent uncommon or morphologically rich regional words through subword n-grams, making them particularly effective when spelling variations or district-specific terms appear in the data. In contrast, Word2Vec does not model subword information and struggles with rare regional terms. Still, it performs moderately across most classifiers and shows notable improvement after oversampling, especially for GradientBoost (58 to 64), Random Forest(55 to 65), SVM(38 to 46), and XGBoost (60 to 66), where the accuracy increases enough to approach some FastText (pretrained) results.

Table 5.6: Overall accuracy (%) comparison of embedding techniques: Word2Vec, FastText (pre-trained), and FastText, with and without oversampling

Model	Without Oversampling			With Oversampling		
	Word2Vec	FastText (Pre)	FastText	Word2Vec	FastText (Pre)	FastText
LR	46	64	37	44	63	27
GradientBoost	58	62	59	64	62	59
MLPC	61	68	37	58	68	36
Random Forest	55	63	57	65	65	57
SGD	51	67	37	41	61	39
SVM	38	65	37	46	64	25
XGBoost	60	65	59	66	64	62

The weakest performance is observed for custom-trained FastText. Because it is trained only on a limited and imbalanced dataset, the embeddings become underfitted, and oversampling further amplifies noise by repeatedly exposing the model to duplicated low-frequency regional words. As a result, MLPC, SGD, and SVM degrade noticeably under this setting. However, for GradientBoost, Random Forest, and XGBoost, custom FastText still performs slightly better than Word2Vec in the oversampled setting (for example, XGBoost rises from 59% to 62%). Overall, pretrained FastText remains the most reliable embedding choice for regional and low-resource Bangla text, particularly when class imbalance is present.

5.6.3 The Impact of BanglaBERT, BUETBERT, DistilBERT, and their Ensemble in the Dataset

The performance of BanglaBERT, BUETBERT, DistilBERT, and their ensemble is evaluated both before and after applying oversampling. The summary metrics before over-

sampling (Table 5.7) show that all three individual transformer models achieve moderate accuracy, with BanglaBERT leading slightly (74.68% accuracy, 73.97% weighted F1). Dis-

Table 5.7: Performance of transformer-based models before applying oversampling (summary metrics)

Model	Accuracy (%)	Weighted F1 (%)
BanglaBERT	74.68	73.97
BUETBERT	72.12	67.84
DistilBERT	74.17	71.42
Ensemble (BanglaBERT + BUETBERT + DistilBERT)	80.31	77.91

tilBERT performs comparably with 74.17% accuracy, while BUETBERT lags at 72.12% accuracy and 67.84% weighted F1. The ensemble model improves performance to 80.31% accuracy and 77.91% weighted F1, demonstrating that combining the strengths of multiple models helps overcome individual weaknesses even without any data-level augmentation.

Table 5.8: Precision, Recall, F1-score for regional classes along with overall Accuracy (%) and Weighted F1-score (%) for transformer-based Bangla embeddings and ensemble method (after oversampling)

Model	Region	Precision	Recall	F1-Score	Accuracy	F1-Score (Weighted)
BanglaBERT	Rangpur	0.54	0.58	0.56	77.74	77.43
	Barishal	0.78	0.75	0.76		
	Nowoakhali	0.71	0.56	0.62		
	Sylhet	0.77	0.77	0.77		
	Chittagong	0.84	0.91	0.88		
BUETBERT	Rangpur	0.66	0.73	0.69	78.26	77.77
	Barishal	0.79	0.82	0.81		
	Nowoakhali	0.59	0.42	0.49		
	Sylhet	0.72	0.81	0.76		
	Chittagong	0.89	0.88	0.89		
DistilBERT	Rangpur	0.75	0.69	0.72	81.84	81.76
	Barishal	0.75	0.81	0.78		
	Nowoakhali	0.77	0.65	0.71		
	Sylhet	0.82	0.83	0.82		
	Chittagong	0.88	0.90	0.89		
Ensemble (BanglaBERT + BUETBERT + DistilBERT)	Rangpur	0.70	0.62	0.65	85.17	84.84
	Barishal	0.86	0.87	0.87		
	Nowoakhali	0.81	0.65	0.72		
	Sylhet	0.85	0.86	0.86		
	Chittagong	0.88	0.95	0.91		

After applying oversampling (Table 5.8), all models show consistent improvement

across class-wise precision, recall, and F1-scores. DistilBERT becomes the strongest single model, achieving 81.84% accuracy and 81.76% weighted F1. It performs particularly well for Sylhet (F1 = 0.82) and Chittagong (F1 = 0.89). BanglaBERT also improves, reaching 77.74% accuracy and 77.43% weighted F1, with strong performance in Sylhet (F1 = 0.77) and Chittagong (F1 = 0.88). BUETBERT, while lighter and faster, remains the weakest among the three with 78.26% accuracy and 77.77% weighted F1, showing lower performance for Nowoakhali (F1 = 0.49). However, it maintains stable precision and recall across the remaining regions. The ensemble model once again provides the strongest overall performance after oversampling. With 85.17% accuracy and 84.84% weighted F1, it outperforms all individual models. Class-wise, it delivers consistently high scores for Barishal (F1 = 0.87), Sylhet (F1 = 0.86), and Chittagong (F1 = 0.91). Even in more challenging regions such as Nowoakhali and Rangpur, it achieves a better balance compared to individual transformers. These results confirm that ensemble learning is highly effective for regional Bangla text classification. By combining contextual strengths from different Bangla transformers, the ensemble reduces model-specific bias and improves robustness across regional variations. Oversampling further stabilizes minority-class learning, contributing to the improved performance observed across all metrics.

As shown in Table 5.9, we evaluated overfitting by comparing the training and test performance on the Tier-1 dataset both before and after applying the proposed oversampling technique. Tier-1 is a small dataset (1954 samples) that contains strong dialectal cues, and models such as BanglaBERT, BUETBERT, and DistilBERT are high-capacity transformers. These models can easily memorize training examples, so high training accuracy is expected, especially when the dataset is small. This naturally creates a noticeable train-test gap without implying harmful overfitting.

Table 5.9: Accuracy and weighted F1-score for training and testing sets, with and without oversampling

Dataset	Accuracy (%)	Weighted F1 (%)
Without Oversampling		
Train	92.58	91.81
Test	80.31	77.91
With Oversampling		
Train	99.59	99.59
Test	85.17	84.84

For the non-oversampled setting, the model achieved 92.58% training accuracy with 91.81% weighted F1, while the test set reached 80.31% accuracy and 77.91% weighted F1. After applying the proposed oversampling, training accuracy increased to 99.59% with a 99.59% weighted F1 score. More importantly, the test accuracy improved to 85.17%, with a corresponding 84.84% weighted F1 score. Because F1 summarizes both precision and recall, the substantial F1 improvement confirms that predictions became more balanced

across dialect classes. The rise in test accuracy from 80.31% to 85.17%, along with the increase in weighted F1, indicates that the oversampling strategy enhances generalization rather than causing harmful overfitting. The ensemble performs reliably on unseen Tier-1 samples, demonstrating that the proposed oversampling technique is effective for handling dialect imbalance.

5.6.4 Impacts of K-fold Cross-validation

In this section, we describe how each BERT model behaves on the Tier-1 dataset when evaluated using stratified 5-fold cross-validation. Because Tier-1 contains only 1954 samples with significant class imbalance, we use Stratified K-Fold to maintain proportional class distribution in every fold. Oversampling is applied only to the training portion of each fold, ensuring that the test split remains untouched and unbiased. Table 5.10 shows the 5-fold cross-validation results before and after oversampling. All models exhibit a clear performance gain, demonstrating that the proposed lexical-aware oversampling strategy improves model stability across multiple train–test partitions. BUETBERT benefits the most, with +12.23% accuracy and +18.56 F1, indicating that it is highly sensitive to class imbalance. DistilBERT also shows great improvement, while BanglaBERT, already the strongest model under imbalance, still gains consistently with +2.87% accuracy and +3.04 F1. The positive ΔAcc and ΔF1 across all models confirm that the oversampling method improves generalization rather than introducing randomness or overfitting.

Table 5.10: Five-fold cross-validation performance before and after oversampling. Values represent mean accuracy and weighted F1-score.

Model	Without Oversampling		With Oversampling		ΔAcc	ΔF1
	Acc (%)	F1 (%)	Acc (%)	F1 (%)		
BanglaBERT	75.89	75.37	78.76	78.41	+2.87	+3.04
BUETBERT	66.99	60.95	79.22	79.51	+12.23	+18.56
DistilBERT	71.60	68.76	80.30	80.07	+8.70	+11.31

Although BUETBERT and DistilBERT show slightly lower performance in the fixed holdout evaluation compared to the averaged 5-fold results, this is expected. All Tier-1 experiments use the same test set to maintain consistent comparative analysis, so results naturally differ from cross-validation partitions. BanglaBERT, however, achieves higher performance in 5-fold evaluation (78.76% accuracy and 78.41% F1), compared to the fixed test set (77.74% accuracy and 77.43% F1), indicating that oversampling reinforces lexical and contextual cues more effectively when the training data is reshuffled. And we already know that BanglaBERT is trained on a much larger and linguistically diverse Bangla corpus, enabling stronger contextual understanding, uses richer subword representations, retains more semantic depth than BUETBERT, and has more model capacity than DistilBERT, which is a compressed and optimized variant. Regional Bangla cues often appear in low-frequency or morphologically rich tokens, and BanglaBERT’s tokenizer handles

these more effectively, especially when oversampled data reinforces rare lexical patterns. Together, these factors explain why BanglaBERT maintains strong performance across both imbalanced and oversampled settings, while BUETBERT and DistilBERT benefit more dramatically from oversampling due to their lower baseline capacity.

5.7 Result Interpretation and Error Analysis

In this section, we discuss the confusion matrix and ROC curve results for individual BERT models and the deep ensemble model, followed by the error and confusion analysis for misclassified samples. Since our dataset preparation involved strict cleaning and the proposed oversampling technique, all visual analyses (confusion matrices and ROC curves) are reported for the oversampled Tier-1 experiments, while numerical tables provide comparisons before and after oversampling. For the evaluated BERT-based models, we include confusion matrix and ROC curve analyses to offer a comprehensive understanding of model performance beyond accuracy and F1-score on the Tier-1 dataset. These visualizations illustrate how effectively the models distinguish regional dialects, reveal systematic misclassification patterns, and highlight the trade-offs between true positive and false positive rates. By examining these findings, we aim to better understand each model’s strengths and weaknesses in handling linguistic overlap and class imbalance. Furthermore, the error and confusion analysis identifies the causes behind incorrect predictions, explains why misclassifications occur for specific regional dialects, and suggests directions to improve model performance.

5.7.1 Confusion Matrix Analysis

In Figure 5.1, we present the confusion matrices of the three individual BERT models. BanglaBERT shows strong performance on the high-resource classes, with a large number of correctly classified samples for Barishal (59), Sylhet (68), and Chittagong (133). However, it still confuses some Rangpur and Nowoakhali instances with Barishal or Sylhet, indicating that it struggles slightly with the minority regions. BUETBERT improves the recognition of low-resource classes: it correctly classifies 19 Rangpur samples and reduces misclassification for Sylhet (71), where BanglaBERT is lower(68). DistilBERT further strengthens predictions for Nowoakhali (34) and Sylhet (73), while maintaining correctness for Chittagong (131) than BUETBERT, but it remains somewhat noisy for Rangpur(18) and Barishal(64), where for these two classes BUETBERT slightly performs better. And the interesting fact is that more minority class Rangpur, where oversampled text is most considered, for this class, DistilBERT performs better than BanglaBERT because DistilBERT handles oversampled data more smoothly. Where BanglaBERT overfits the duplicates, memorizes noise instead of patterns, BUETBERT and DistilBERT benefit from oversampling.

The confusion matrix of the soft-voting ensemble, shown in Figure 5.2, highlights

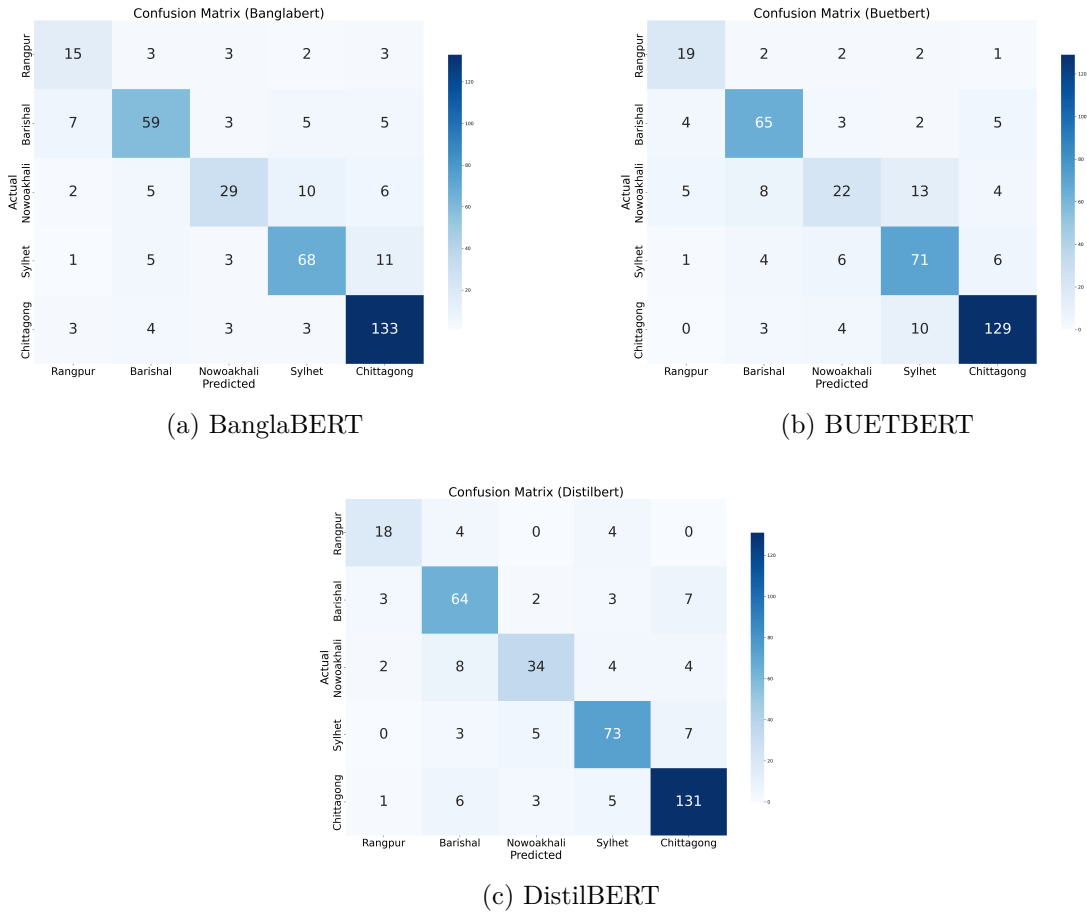


Figure 5.1: Confusion matrices for (a) BanglaBERT, (b) BUETBERT, and (c) DistilBERT

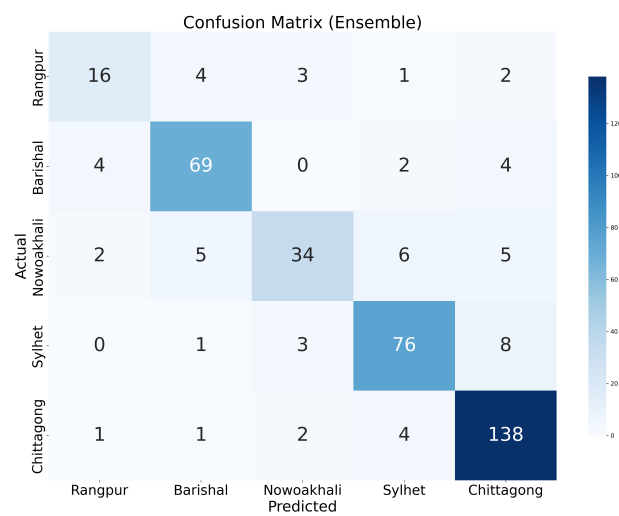


Figure 5.2: Confusion matrix for the ensemble of three BERT models.

the benefit of combining these three models. Correct predictions increase across all regions, Barishal (69), Nowoakhali (34), Sylhet (76), and Chittagong (138), except Rangpur (16). Because the Rangpur dialect is not heavily confused with other regions, it does not appear among the major confusion pairs in the individual confusion matrices. All are showing improved or at least stabilized counts compared to the individual models. Off-diagonal entries shrink noticeably, especially for the previously problematic pairs such as Chittagong→Sylhet and Nowoakhali→Sylhet. This confirms that the ensemble effectively aggregates complementary strengths from BanglaBERT, BUETBERT, and DistilBERT, leading to more reliable and consistent regional predictions.

5.7.2 ROC Curve Analysis

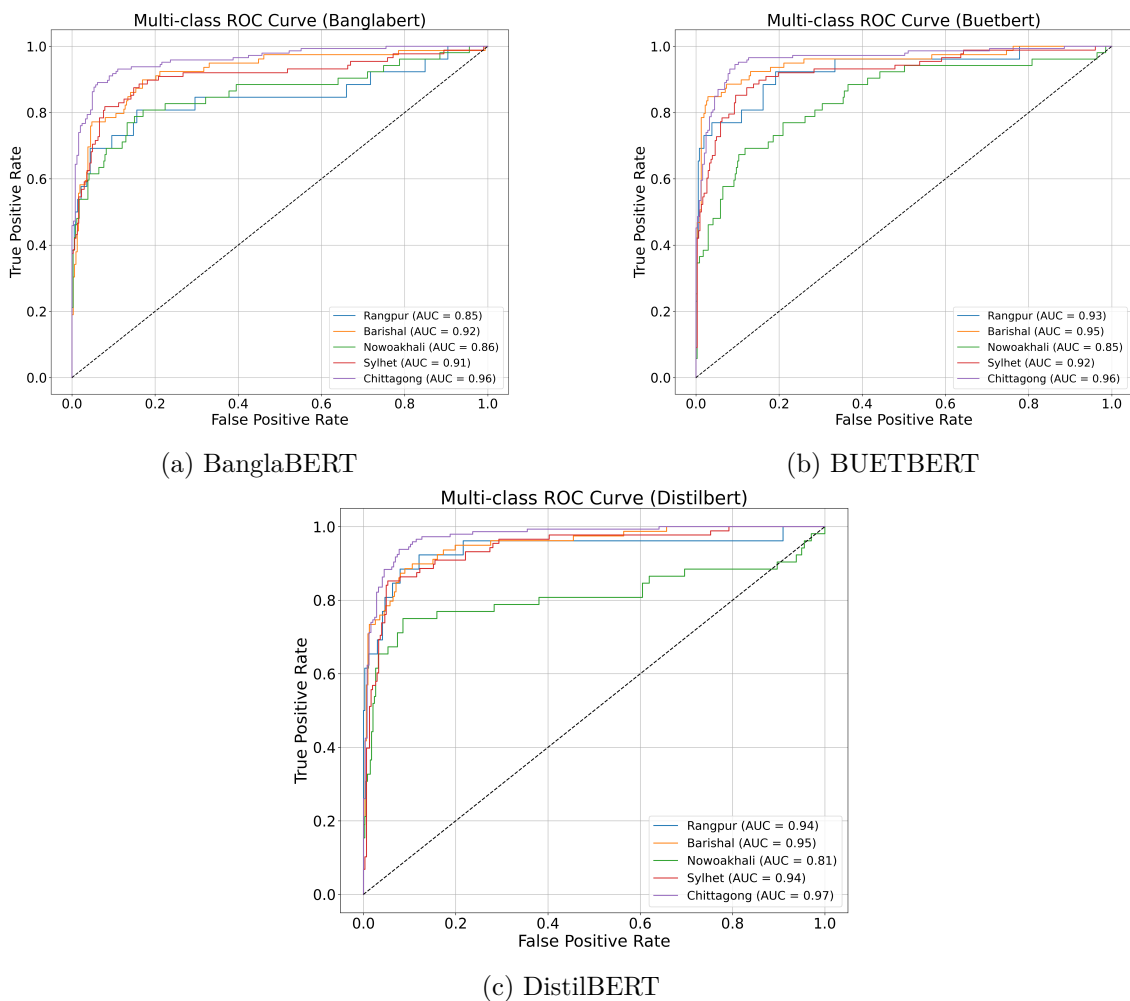


Figure 5.3: ROC curves of (a) BanglaBERT, (b) BUETBERT, and (c) DistilBERT models.

In Figure 5.3, we represent the multiclass ROC curves for BanglaBERT, BUETBERT, and DistilBERT. BanglaBERT attains AUC values in the range of 0.85–0.97, performing particularly well for Barishal, Sylhet, and Chittagong, while Rangpur and Nowoakhali re-

main comparatively weaker. BUETBERT yields more balanced discrimination, with AUC scores around 0.85–0.96, and improves the curve for Rangpur (about 0.93), indicating better sensitivity to that low-resource region. DistilBERT also achieves high AUCs (roughly 0.81–0.97); it provides strong separation for Barishal, Sylhet, and Chittagong but exhibits a noticeably lower AUC for Nowoakhali, reflecting the difficulty of this class and mostly gets confused with other dialect pairs. The ROC curve of the ensemble model in Figure 5.4

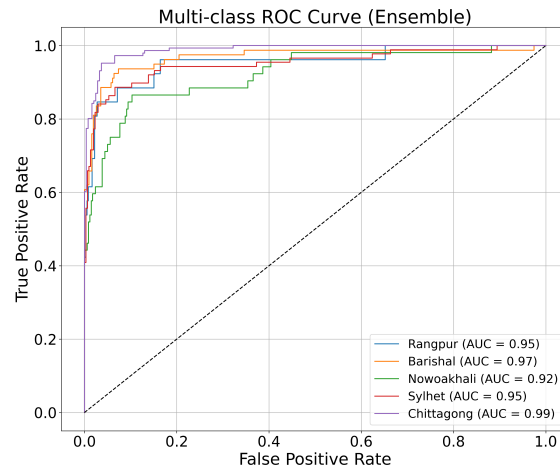


Figure 5.4: ROC curve for soft voting ensemble of three BERT models

shows a consistent improvement over the individual BERT variants. The ensemble reaches AUC values of approximately 0.95 for Rangpur, 0.97 for Barishal, 0.92 for Nowoakhali, 0.95 for Sylhet, and 0.99 for Chittagong, indicating very strong class separability across all five regions. Compared to the single models, the ensemble raises the AUC for the weaker classes (Rangpur and Nowoakhali) while preserving or slightly improving performance for the stronger ones. Altogether, the confusion matrices and ROC curves demonstrate that the soft-voting ensemble not only boosts overall accuracy and F1-score but also delivers more stable and discriminative behaviour across all dialect classes, which is crucial for reliable Bangla regional text classification in this low-resource setting.

5.7.3 Error and Confusion Analysis

In this section, an error analysis was conducted on the misclassified test samples. According to the confusion matrices, for the BanglaBERT model, the top two confusion pairs are Sylhet \rightarrow Chittagong (11 samples) and Nowoakhali \rightarrow Sylhet (10 samples). Similarly, for the BUETBERT model, these pairs are also the top confusion pairs: Nowoakhali \rightarrow Sylhet (13 samples) and Chittagong \rightarrow Sylhet (10 samples). For the DistilBERT model, Sylhet \rightarrow Chittagong (7 samples) is considered as the second-highest confusion pair, followed by Sylhet \rightarrow Nowoakhali (5 samples). These regions share several phonological and morphological similarities. For example, the word "ভালা" (*vala*, English meaning "good") is used with the same spelling and meaning across Nowoakhali, Sylhet, and Chittagong, along with several other words with identical meanings. Another word "ইস্কুল" (*eskul*, En-

glish meaning “School”) is used for the same spelling and meaning for Nowoakhali and Barishal. And if we see for the DistilBERT model top confusion pair is Nowoakhali → Barishal (8 samples). The models, relying primarily on word-level signals, often failed to capture these subtle structural nuances.

Regarding the Rangpur class, it appears less frequently in confusion pairs like Rangpur → Chittagong, 0 samples for DistilBERT, and only 1 sample for BUETBERT. The Rangpur dialect belongs to the North-Central (Varendri) or Northern Bengali group, making it geographically and linguistically distinct from the Eastern and Southern groups, such as Barishal, Noakhali, Chittagong, and Sylhet [28]. Its texts are clearly different from the others. The fact that Rangpur belongs to a fundamentally different primary cluster (Northern/Varendri) than Noakhali and Barishal (Eastern/Vanga) reflects their distinct historical development and linguistic features, making Rangpur texts relatively easier to distinguish when region-specific vocabulary is used. For the ensemble BERT model, the most frequent confusion pair observed in individual models, Noakhali → Sylhet (10 and 13 samples), is reduced to 6 samples, demonstrating that combining model strengths reduces ambiguity. Overall, the error analysis indicates that dialect pairs sharing a larger portion of common region-specific vocabulary exhibit higher confusion, while dialects with more distinctive lexical patterns (such as Rangpur) are classified more reliably. This further supports the motivation behind our proposed linguistically informed oversampling strategy rather than random sampling, which emphasizes region-specific lexical cues to strengthen the model’s ability to distinguish closely related dialects.

5.7.4 Overall discussion

In this section, we present an overall discussion of how regional linguistic characteristics influence model behavior and why the proposed lexical-based oversampling technique performs well for BERT models. Regional Bangla dialects exhibit a mixture of shared vocabulary and unique region-specific expressions. These subtle lexical cues carry important semantic and phonological differences that cannot be captured by simple word-frequency statistics.

BERT models learn contextual meaning, token interactions, sentence structure, subword patterns, and deep semantic relationships. When minority-class samples are duplicated through oversampling, these models receive additional exposure to underrepresented regional cues. Instead of memorizing duplicates, BERT strengthens the underlying contextual representation of these dialectal patterns. This reinforcement explains the consistent improvement observed in all three transformer variants after applying the proposed oversampling strategy. In contrast, TF-IDF is a purely lexical representation that counts word occurrences without modeling context or semantics. Duplicating the same minority documents produces identical sparse vectors, which do not introduce any new information. In some cases, oversampling even shifts TF or IDF distributions in undesirable ways, which explains why TF-IDF based classifiers do not consistently benefit from oversampling. A

similar trend is observed in other static embedding techniques (Word2Vec, FastText pre-trained, FastText self-trained, TF-IDF+PCA, TF-IDF+SVD): although some models improve slightly, the gains remain limited because these methods do not fully capture the morphological or contextual nuances present in regional dialects.

Across all experiments, a clear pattern emerges: models that rely on contextual or subword-level information benefit most from our linguistically informed oversampling strategy. This is especially relevant for regional text classification, where distinctions often depend on fine-grained lexical variations, dialect-specific morphemes, and subtle semantic cues. Our oversampling method intentionally amplifies lexical signals of the minority class, rather than performing blind duplication. As a result, it enhances the representation space for low-resource dialects, allowing BERT models to generalize more effectively. The substantial performance improvements across all BERT variants, both in stratified 5-fold cross-validation and in the Tier-1 holdout evaluation, confirm that the proposed oversampling strategy reinforces minority linguistic patterns in a meaningful and task-specific way. The method does not cause overprediction or bias; instead, it strengthens dialect-specific cues and stabilizes performance across different data splits. This provides strong evidence that the proposed technique is well-suited for imbalanced, low-resource Bangla regional text classification and contributes a valuable methodological advancement to the field.

Chapter 6

Conclusion

This chapter summarizes the study’s major findings, highlights its contributions, acknowledges limitations, and provides recommendations and directions for future work. The results presented in the earlier chapters demonstrate the effectiveness of the proposed methods and their applicability in addressing regional text classification challenges in Bangla. We also reflect on the broader implications of this research in terms of linguistic diversity and NLP development for low-resource languages. Finally, the chapter outlines potential avenues for extending this work, ensuring its relevance and adaptability to future advancements in the field.

6.1 Summary of Findings

The main findings of this research highlight how curated datasets, enriched lexicons, and diverse modeling approaches impact Bangla regional text classification performance:

- This study addressed key issues in Bangla regional text classification, including data scarcity, dialectal variation, and class imbalance.
- We created BdRegionText v2, a cleaned and expert-validated regional dataset covering five main dialects: Rangpur, Sylhet, Barishal, Noakhali, and Chittagong.
- To improve dataset quality, we developed a specialized regional lexicon containing distinct, region-specific words.
- To handle class imbalance, we applied a region-specific keyword oversampling strategy.
- Several conventional machine learning models were assessed using feature representations such as Word2Vec, pretrained FastText, TF-IDF, TF-IDF + PCA, and TF-IDF + SVD.
- Additionally, we refined three Bangla transformer models (BanglaBERT, BUET-BERT, and DistilBERT) and proposed a heterogeneous ensemble using soft voting.

- Overall, the study demonstrated the effectiveness of combining expert-validated datasets, enriched lexicons, and ensemble approaches to advance Bangla regional text classification.

6.2 Contributions of the Research

The research contributions illustrate how this work advances Bangla NLP and regional dialect classification:

- Development of BdRegionText v2, an expert-validated dataset covering five Bangla dialects.
- Creation of a regional lexicon with culturally significant, region-specific words.
- Proposal of a keyword-based oversampling strategy to mitigate class imbalance.
- Comprehensive evaluation of machine learning and transformer-based approaches for regional text classification.
- Introduction of a heterogeneous ensemble framework to improve performance and generalization.
- Contribution of open research resources (BdRegionText v2 and Regional Lexicon) to support future Bangla NLP studies.

6.3 Limitations

Despite the study’s achievements, certain limitations must be acknowledged:

- The dataset, although validated, remains smaller compared to high-resource language corpora.
- The study is restricted to five dialects, excluding many other Bangla regional variations.
- Code-switching (Bangla–English mixing) and spoken dialect recognition were not addressed.
- Computational limitations restricted large-scale pre-training on regional corpora.

6.4 Recommendations

Based on the outcomes of this research, several practical recommendations can be made:

- Researchers and practitioners can use the BdRegionText v2 dataset and regional lexicon to build more robust NLP systems for Bangla dialects.

- The proposed ensemble classification framework can be adapted for other low-resource languages facing dialectal diversity.
- Policymakers and educational institutions can leverage these resources to promote regional inclusivity in digital platforms and applications.
- Developers of social media monitoring tools and sentiment analysis systems can benefit from integrating region-specific models for more accurate insights.

6.5 Future Work

Looking ahead, future research can extend this work in several directions to further improve Bangla regional NLP:

- Expanding BdRegionText v2 to cover additional dialects, enlarging the Tier-1 subset through iterative expert validation, and leveraging Tier-2 and Tier-3 samples to refine the dataset further.
- Conduct large-scale pre-training on region-specific corpora to improve transformer-based models.
- Explore code-mixed Bangla–English regional texts to reflect real-world language usage.
- Investigate speech-based dialect classification alongside textual data.
- Enhance classification methods by integrating context-aware embeddings, meta-learning approaches, or multimodal techniques.

References

- [1] Firoj Alam, SM Habib, Dil Afroza Sultana, and Mumit Khan. Development of annotated bangla speech corpora. *BRAC University*, 2010. uri: <http://hdl.handle.net/10361/633>.
- [2] Prommy Sultana Hossain, Amitabha Chakrabarty, Kyuheon Kim, and Md Jalil Piran. Multi-label extreme learning machine (mlelms) for bangla regional speech recognition. *Applied Sciences*, 12(11):5463, 2022. doi: <https://doi.org/10.3390/app12115463>.
- [3] Fatema Tuj Johora Faria, Mukaffi Bin Moin, Ahmed Al Wase, Mehidi Ahmmed, Md Rabiuis Sani, and Tashreef Muhammad. Vashantor: a large-scale multilingual benchmark dataset for automated translation of bangla regional dialects to bangla language. *arXiv preprint arXiv:2311.11142*, 2023. doi: <https://doi.org/10.48550/arXiv.2311.11142>.
- [4] SM Islam, Sadia Ahmmed, and Sahid Hossain Mustakim. Transcribing bengali text with regional dialects to ipa using district guided tokens. *arXiv preprint arXiv:2403.17407*, 2024. doi: <https://doi.org/10.48550/arXiv.2403.17407>.
- [5] Babe Sultana, S. M. Mirajul Hoque, Md Gulzar Hussain, and Mohammad Nurul Huda. BdRegionText: Bangladeshi Regional Dataset, 2024.
- [6] Umme Aiman, MD Nakibul Islam, Hana Sultan Chowdhury, Md. Sadekur Rahman, Md. Tarek Habib, and Mahady Hasan. BRWDS: A Multipurpose Dataset For Bangla Regional Word Detection, 2025.
- [7] Rezaul Haque, Naimul Islam, Mayisha Tasneem, and Amit Kumar Das. Multi-class sentiment classification on bengali social media comments using machine learning. *International journal of cognitive computing in engineering*, 4:21–35, 2023.
- [8] Babe Sultana, Zakia Afrin, Farhana Ryhan Kabir, and Dewan Md Farid. Bilingual spam sms detection using machine learning. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2023.
- [9] Rimon Barua, MM Rahman, and Usman Gani Joy. Comparative analysis of bangla news classification: a study of fake news detection and multiclass classification using bert and fasttext. *International Journal of Computers and Applications*, 47(5):475–485, 2025.

-
- [10] Salim Sazzed. Identifying vulgarity in bengali social media textual content. *PeerJ Computer Science*, 7:e665, 2021.
- [11] Sara Azmin and Kingshuk Dhar. Emotion detection from bangla text corpus using naive bayes classifier. In *2019 4th international conference on electrical information and communication technology (EICT)*, pages 1–5. IEEE, 2019.
- [12] Pooja Bolaj and Sharvari Govilkar. Text classification for marathi documents using supervised learning methods. *Int. J. Comput. Appl*, 155(8):6–10, 2016. doi: <https://doi.org/10.5120/ijca2016912374>.
- [13] Sankirti Sandeep Shiravale, Sanjeev S Sannakki, and R Jayadevan. Text region identification in indian street scene images using stroke width transform and support vector machine. *SN Computer Science*, 2(5):357, 2021. doi: <https://doi.org/10.1007/s42979-021-00745-y>.
- [14] Ajeng Dwi Asti, Indra Budi, and Muhammad Okky Ibrohim. Multi-label classification for hate speech and abusive language in indonesian-local languages. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–6. IEEE, 2021.
- [15] Kakuthota Rakshitha, H M Ramalingam, M Pavithra, H D Advi, and Maithri Hegde. Sentimental analysis of indian regional languages on social media. *Global Transitions Proceedings*, 2(2):414–420, 2021.
- [16] Paliwal Mohan Subhash, CR Kavitha, Deepa Gupta, et al. Indo-aryan dialect identification using deep learning ensemble model. *Procedia Computer Science*, 235: 2886–2896, 2024.
- [17] Hafedh Hameed Hussein and Amir Lakizadeh. A systematic assessment of sentiment analysis models on iraqi dialect-based texts. *Systems and Soft Computing*, page 200203, 2025.
- [18] Md Nahid Hasan, Raiyan Azim, Mahmudul Hasan, and Md Monarul Islam. A stacked ensemble model to identify bangla religious hate comments. In *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, pages 1–6. IEEE, 2024.
- [19] Iftekhar Fahim, Shawly Ahsan, and Mohammed Moshiul Hoque. Abusive comment detection from bengali-english code-mixed social media texts using ensemble of deep learning. In *International Conference on Artificial Intelligence and Knowledge Processing*, pages 252–267. Springer, 2024.
- [20] Tanzia Parvin and Mohammed Moshiul Hoque. An ensemble technique to classify multi-class textual emotion. *Procedia Computer Science*, 193:72–81, 2021.

-
- [21] Md Nesarul Hoque and Md Hanif Seddiqui. Exploring transformer ensemble approach to classify cyberbullying text for the low-resource bengali language. In *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*, pages 1–6. IEEE, 2024.
- [22] Md Nesarul Hoque, Umme Salma, Md Jamal Uddin, and Sadia Afrin Shampa. Depression intensity identification using transformer ensemble technique for the resource-constrained bengali language. *Journal of Engineering Advancements*, 5(02):27–34, 2024.
- [23] Md Nesarul Hoque and Umme Salma. Detecting level of depression from social media posts for the low-resource bengali language. *Journal of Engineering Advancements*, 4(02):49–56, 2023.
- [24] Muhammad Khubayyeb Kabir, Maisha Islam, Anika Nahian Binte Kabir, Adiba Haque, and Md Khalilur Rhaman. Detection of depression severity using bengali social media posts on mental health: study using natural language processing techniques. *JMIR Formative Research*, 6(9):e36118, 2022.
- [25] Md Rezuwan Hassan. *A character gram modeling approach towards Bengali Speech to Text with Regional Dialects*. PhD thesis, Brac University, 2023.
- [26] Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*, 58:111276, 2025.
- [27] Md Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10: 91855–91870, 2022.
- [28] Wikipedia Contributors: Bengali dialects. https://en.wikipedia.org/wiki/Bengali_dialects. Last Edited: 19 September 2025.