Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language

Preprint · November 2023

DOI: 10.48550/arXiv.2311.11142

6 authors, including:

Fatema Tuj Johora Faria
Ahsanullah University of Science & Tech
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Mukaffi Bin Moin
Ahsanullah University of Science & Tech
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Ahmed Al Wase
Ahsanullah University of Science & Tech
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Mehidi Ahmmed
Khulna University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language

Fatema Tuj Johora Faria[1*], Mukaffi Bin Moin[1], Ahmed Al Wase[1], Mehidi Ahmmed[2], Md. Rabius Sani[1], Tashreef Muhammad[3]

[1]Ahsanullah University of Science and Technology, Dhaka, Bangladesh.
[2] Khulna University, Dhaka, Bangladesh.
[3]Southeast University, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): fatematujjohorafaria142@gmail.com

Contributing authors: mukaffi28@gmail.com; ahmed.alwasi34@gmail.com; mehidi0308@gmail.com; rshridoy010113@gmail.com; tashreef.muhammad@seu.edu.bd

Abstract

The Bangla linguistic variety is a fascinating mix of regional dialects that adds to the cultural diversity of the Bangla-speaking community. Despite extensive study into translating Bangla to English, English to Bangla, and Banglish to Bangla in the past, there has been a noticeable gap in translating Bangla regional dialects into standard Bangla. In this study, we set out to fill this gap by creating a collection of **32,500** sentences, encompassing Bangla, Banglish, and English, representing five regional Bangla dialects. Our aim is to translate these regional dialects into standard Bangla and detect regions accurately. To achieve this, we proposed models known as mT5 and BanglaT5 for translating regional dialects into standard Bangla. Additionally, we employed mBERT and Bangla-bert-base to determine the specific regions from where these dialects originated. Our experimental results showed the highest BLEU score of 69.06 for Mymensingh regional dialects and the lowest BLEU score of 36.75 for Chittagong regional dialects. We also observed the lowest average word error rate of 0.1548 for Mymensingh regional dialects and the highest of 0.3385 for Chittagong regional dialects. For region detection, we achieved an accuracy of 85.86% for Bangla-bert-base and 84.36% for mBERT. This is the first large-scale investigation of Bangla regional dialects to Bangla machine translation. We believe our findings will not only pave the way for future work on Bangla regional dialects to Bangla machine translation, but will also be useful in solving similar language-related challenges in low-resource language conditions.

Keywords: Bangla Regional Dialects Translation, Neural Machine Translation, Region Detection, Regional Dialects Corpus

1 Introduction

Neural Machine Translation (NMT) [1] represents an innovative technology in Natural Language Processing (NLP), bringing about a significant transformation in the way we approach automated translation tasks. Unlike traditional rule based or statistical machine translation systems, NMT relies on deep neural networks to directly translate text from one language to another. Translation, the core function of NMT [2], can be categorized into two primary types: sentence-level translation and word-level translation. Sentence translation involves translating entire sentences or phrases from one language to another, preserving the meaning and context. Word translation, on the other hand, focuses on individual words or short phrases and their corresponding translations. These two types of translation serve distinct purposes [3], with sentence translation allowing comprehensive document translation, while word translation supports finer-grained language analysis and understanding. While a few years ago, the focus was on achieving high-quality translations for widely spoken and well-resourced languages, the current improvements in translation quality have highlighted the importance of addressing low-resource languages and dealing with more diverse and remarkable translation challenges [4].

The Bangla language is spoken by about 228 million people as their first language and an additional 37 million people speak it as a second language. Bangla is the fifth most spoken first language and the seventh most spoken language overall in the world [5]. Bangladesh has 55 regional languages spoken in its 64 districts, while the majority of the population speaks two different varieties of Bengali. Some people also speak the language of the region they live in. The variations in the Bengali language extend beyond vocabulary to differences in pronunciation, intonation, and even grammar. A regional language, which is also called a dialect, is a language that children naturally learn without formal grammar lessons, and it can differ from one place to another. These regional languages can cause changes in the way the main language sounds or is written. Even though there are these regional differences, the Bangla language in Bangladesh can be categorized into six main classes: Bangla, Manbhumi, Varendri, Rachi, Rangpuri, and Sundarbani [6].

Our research addresses a significant gap in the field of machine translation, specifically about the Bengali language. While previous research has mostly concentrated on translating between Bengali to English [2], English to Bengali [5], English to Banglish [7], Manipuri to Bengali [8], Bangla to Banglish [9], and even Hindi to Bangla [10], it has largely skipped over translating various Bengali regional dialects to and from Bengali. Furthermore, previous research concentrated mostly on distinguishing regions from audio speech [6] [11] rather than giving complete translations of regional speeches into Bengali. The overall lack of prior work or dedicated datasets for regional dialect translation emphasizes the importance of our research in filling these significant knowledge gaps. While a variety of dialects is a source of cultural pride, it also poses an enormous communication challenge. Multiple dialects existing can lead to misconceptions, miscommunications, and a communication gap between individuals from different regions. In such a situation, the need for effective machine translation capable of overcoming dialectal differences becomes critical. Our research tackles this critical problem by presenting the "Vashantor" dataset, a useful resource for automated translation of Bangla regional dialects to Bangla Language, enabling enhanced communication and understanding across all of these different geographic regions.

Recognizing Bangladesh's linguistic diversity, we have chosen to concentrate on translating the Bangla language from five distinct regional dialects found in the Chittagong, Noakhali, Sylhet, Barishal, and Mymensingh regions. For example, when we translate the Chittagong regional dialect, "বউতদিন ফর তোঁয়ারে দেইলাম" into Bengali, it becomes "অনেক দিন পর তোমাকে দেখলাম". Similarly, when we translate the Noakhali regional dialect, "মেলা দিন হর তোমার দেয়া হাইলাম" into Bengali, it also turns into "অনেক দিন পর তোমাকে দেখলাম". Likewise, the Sylhet regional dialect, "বাক্কাদিন বাদে তুমারে দেখলাম" is translated as "অনেক দিন পর তোমাকে দেখলাম" in Bengali. These translations show how different regional dialects can be expressed in the Bengali language. These regions have been identified as the most significant contributors to the research due to the pronounced linguistic variations and communication challenges presented by their unique dialects. To the best of our knowledge, this is the first attempt to translate Bangla regional dialects into Bangla.

In our research, we used state-of-the-art translation models to make it easier to translate different regional dialects into Bangla. Additionally, we developed a system to identify the specific regions associated with the text in our corpus. In essence, our work has made several noteworthy contributions, which can be summarized as follows:

- We created a comprehensive dataset, including:
  - 2,500 samples for Bangla, Banglish, and English each.
  - 12,500 samples for regional Bangla dialects and regional Banglish dialects each.
  - 12,500 samples for region detection.
- We validated the dataset using Cohen's Kappa and Fleiss's Kappa.
- We applied cosine similarity to quantitatively assess variations and similarities among Bangla regional dialects and standard Bangla language.
- We employed machine translation evaluation metrics for assessing the quality of machine translation output, including: Character Error Rate (CER), Word Error Rate (WER), Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Metric for Evaluation of Translation with Explicit ORdering (METEOR)
- We utilized performance metrics, including Accuracy, Precision, Recall, F1 score, and Log loss for the region detection task.

The remaining part of the paper is structured as follows: Section 2 provides a thorough review of related works that serve as the foundation for our study. Moving forward, Section 3 delves into the details of our dataset creation process. The Section 4 explores the complexities of the models used for dialect-to-Bangla translation and region detection. In Section 5, we thoroughly investigate the evaluation metrics used to evaluate both regional dialect translation and region detection. The Section 6 presents and goes further on our proposed methodology, demonstrating the analytical process and planning that explains our research goals. The Section 7 diligently describes the results of the experiment and gives a comprehensive understanding. The Section 8 focuses further on identifying future research opportunities and providing a road map for the current study's continuation and advancement. The Section 9 brings the research to a conclusion by bringing together the many components of our study and offering a final summary that describes the key results, contributions, and significance of our research.

2 Related Works

In this section, we have given a brief summary of previous research which is relevant to our research. The summary is broken down into four primary subsections: Section 2.1 Unsupervised Neural Machine Translation (UNMT), focusing on translation learning without paired examples, using monolingual data from both languages involved; Section 2.2 Sequence-to-Sequence (Seq2Seq) Neural Machine Translation employing an encoder-decoder setup to convert text from one language to another; Section 2.3 Transformer Based Neural Machine Translation utilizing self-attention to capture word dependencies for improved translation performance and efficiency; and Section 2.4 Adversarial Neural Machine Translation, integrating adversarial learning techniques to differentiate between human and machine translations, aiming to enhance translation quality. Table 1 presents a comprehensive summary of the existing works on machine translation that have been discussed within this study.

2.1 Unsupervised Neural Machine Translation

Lenin et al. [12] introduced unsupervised Machine Translation models for low-resource languages, emphasizing their ability to work without parallel sentences. It focused on Manipuri-English, highlighting linguistic disparities and challenges. It concluded that unsupervised Machine Translation for such language pairs is feasible, based on experiments. They compared Unsupervised Machine Translation (USMT) and UNMT models, focusing on models like Monoses, MASS, and XLM. Initial findings showed that USMT was more effective for Manipuri-English. It highlighted challenges in adapting unsupervised MT methods to this pair and evaluated the strengths and weaknesses of USMT and UNMT models. It used a Manipuri-English corpus from newspapers and evaluated using BLEU scores. Monoses obtained the best BLEU score of 3.13 for English to Manipuri and a score of 6.37 for Manipuri-English outperforming the UNMT systems. They established a baseline and encouraged further research for the low-resource Manipuri-English language pair. In another research, Guillaume et al. [13] explored the possibility of machine translation in the absence of parallel data, which is a significant challenge in the field. They offered a model for mapping monolingual corpus from two distinct languages into a common latent space. They demonstrated their approach's strong performance in unsupervised machine translation. While it might not outperform supervised approaches with abundant resources, it produced outstanding results. It matched the quality of a supervised system trained on 100,000 sentence pairs from the WMT dataset, for example. It achieved strong BLEU scores in

the Multi30K-Task1 dataset, particularly 32.76 in the English-French pair. The research also examined the model's performance in various settings, demonstrating its versatility. The primary outcome was establishing the practicality of unsupervised machine translation using shared latent representations, with outstanding results across a wide range of language pairs.

2.2 Sequence-to-Sequence (Seq2Seq) Neural Machine Translation

The work of Rafiqul et al. [2] focused on translating Bengali to English, which overcomes the difficulties of Bangla's complicated grammatical rules and large vocabulary. To improve performance, they employed a Seq2Seq learning model with attention-based recurrent neural networks (RNN) and cross-entropy loss metrics. They built a model with less than 2% loss by carefully building a dataset with over 6,000 Bangla-English Seq2Seq sentence pairs and precisely analyzing training parameters. Shaykh et al. [5] on the contrary, focused on the task of English to Bangla translation using RNN, especially an encoder-decoder RNN architecture. Their method included a knowledge-based context vector to aid in exact translation between English and Bangla. The study highlighted the importance of data quality, with 4,000 parallel sentences serving as the foundation. Particularly, they overcame the issue of varying sentence lengths by using a combination of linear activation in the encoder and tanh activation in the decoder to achieve optimal results. In addition, their findings highlighted the superiority of GRU over LSTM as well as the significance of attention processes implemented via softmax and sigmoid activation functions.

2.3 Transformer-Based Neural Machine Translation

Laith et al. [4] proposed an innovative Transformer-Based NMT model tailored for Arabic dialects, addressing challenges in low-resource languages, particularly their unique word order and scarce vocabulary. Their approach employed subword units and a common vocabulary, as well as the WordPiece Model (WPM) for exact word segmentation, sparsity reduction, and translation quality enhancement, particularly for unknown (UNK) terms. Key contributions included a shared vocabulary approach between the encoder and decoder, as well as the usage of wordpieces, which resulted in higher BLEU scores. The research indicated a considerable improvement in translation quality through comprehensive testing including diverse Arabic dialects and translation jobs to Modern Standard Arabic (MSA). Furthermore, the study examined the impact of characteristics such as the number of heads in self-attention sublayers and the layers in encoding and decoding subnetworks on the model's performance. On the contrary, Soran et al. [14], provided a novel transformer-based NMT model for low-resource languages, with an emphasis on the Kurdish Sorani Dialect. This model employed attention approaches and data from several sources to get a BLEU score of 0.45, suggesting high-quality translations. The addition of four parallel datasets, Tanzil, TED Talks, Kurdish WordNet, and Auta, expanded the system's domain adaptability. Because of its six-layer encoder and decoder architecture, which was improved by multi-head attention, the model offered excellent translation capabilities.

2.4 Adversarial Neural Machine Translation

Lijun et al. [15] introduced a novel approach called Adversarial-NMT for NMT. Adversarial-NMT reduces the distinction between human and NMT translations, in contrast to standard methods that attempt to maximize human translation resemblance. It uses an adversarial training architecture with a CNN as the adversary. The NMT model aims to produce high-quality translations to deceive the adversary, and they were co-trained using a policy gradient method. Adversarial-NMT greatly increases translation quality compared to strong baseline models, according to experimental results on English to French and German to English translation tasks. For English to French, they employed the top 30,000 most frequent English and French words, and for German to English, they utilized the top 32,009 most frequent words. Comparing their Adversarial-NMT to the baseline models, it performed a better translation on about 59.4% of the sentences. Furthermore, Wenting et al. [16] discussed the issue of short sequence machine translation from Chinese to English by introducing a generative adversarial network (GAN). The GAN consists of a generator and a discriminator, with the generator producing sentences that are indistinguishable from human translations and the discriminator separating these from human-translated sentences. To evaluate and direct the generator, both dynamic discriminators and static BLEU score targets are used during the training phase. When compared to typical recurrent neural network (RNN) models, experimental results on an English-Chinese translation dataset showed a more than 8% improvement in translation quality. The proposed approaches' average BLEU scores were 28.2.

Table 1: A summary of various existing unsupervised, sequence-to-sequence (Seq2Seq), transformer-based, adversarial network-based neural machine translation works.

Types	Authors	Year	Contribution
Unsupervised Neural Machine Translation	Lenin et al. [12]	2021	Used unsupervised Machine Translation models for the low-resource Manipuri-English language pair.
	Guillaume et al. [13]	2017	Explored machine translation in the absence of parallel data and put out a strategy for aligning monolingual corpora.
	Zhen et al. [17]	2018	Utilized an extension to extract high-level representations of the input phrases, which consists of two independent encoders sharing partial weights.
Sequence-to-Sequence (Seq2Seq) Neural Machine Translation	Rafiqul et al. [2]	2023	Focused on Bengali to English translation, addressing Bengali's complex syntax and rich vocabulary.
	Shaykh et al. [5]	2021	Concentrated on English to Bangla translation employing an encoder-decoder RNN structure with a knowledge-based context vector for precise translation.
	Arid et al. [1]	2019	Investigated several neural machine translation techniques for Bangla-English, and with average improvements of 14.63% and 32.18%.
Transformer-Based Neural Machine Translation	Laith [4] et al.	2021	Presented a Transformer-Based NMT model intended for Arabic dialects, solving issues in low-resource languages through the use of subword units.
	Soran et al. [14]	2023	Suggested a unique transformer-based NMT model adapted for the low-resource Kurdish Sorani Dialect, earning an impressive BLEU score of 0.45 for high-quality translations.
	Dongxing et al. [18]	2022	Presented the interacting-head attention method, which improves multihead attention by allowing more extensive and deeper interactions among tokens in different subspaces.
Adversarial Neural Machine Translation	Lijun et al. [15]	2017	Developed Adversarial-NMT technique with a CNN to minimize differences between human and NMT translations.
	Wenting et al. [16]	2022	Utilized a GAN for Chinese-English translation which surpassed RNNs by 8% in translation quality on an English-Chinese dataset, and achieved an average BLEU score of 28.2.
	Wei et al. [19]	2020	Explored a reinforcement learning paradigm which includes a discriminator as the terminal signal in order to limit semantics.

3 Corpus Creation

3.1 Data Collection

In the process of curating the "Vashantor" dataset, we diligently selected speech text data from a wide range of sources to ensure its authenticity and quality. The name of our dataset was intentionally chosen to be "Vashantor" or "ভাষান্তর" in Bangla, which means "Translation" in English. The choice of "Vashantor" shows a deeper cultural connection, especially in regard to the Bangla language itself. It indicates the dataset's focus on Bangla or translations involving Bangla, highlighting the language's significance within the context of the dataset. We gathered the dataset in Bangla, Banglish (mix of Bangla and English, using the English alphabet

Table 2: Translator Information For Bangla Regional Dialects

Region	Translator	Educational Status	Language Expertise	Age	Gender
Chittagong	Translator 1	Undergraduate	Dialect Expert	25	Female
	Translator 2	Undergraduate	Dialect Expert	24	Male
	Translator 3	Graduate	Dialect Expert	27	Male
Noakhali	Translator 1	Undergraduate	Dialect Expert	24	Male
	Translator 2	Undergraduate	Dialect Expert	23	Female
	Translator 3	Graduate	Dialect Proficient	26	Male
Sylhet	Translator 1	Undergraduate	Dialect Proficient	25	Female
	Translator 2	Graduate	Dialect Expert	27	Male
Barishal	Translator 1	Undergraduate	Dialect Expert	24	Male
	Translator 2	Undergraduate	Dialect Proficient	24	Male
	Translator 3	Undergraduate	Dialect Expert	25	Male
Mymensingh	Translator 1	Undergraduate	Dialect Expert	23	Male
	Translator 2	Undergraduate	Dialect Expert	25	Male

to write Bangla), and English, which include five regional Bangla dialects. The primary sources for our data collection included websites, social media platforms, and discussion boards. By extracting text data from these sources, we aimed to capture natural language as it is used in regular dialogues among individuals. We prioritized the selection of text data that closely resembled typical conversations, discussions, and interactions between people. This approach allowed us to assemble a dataset that accurately reflects the language used in real-world communication. By focusing on regular dialogue, we ensured that the "Vashantor" dataset is not only comprehensive but also representative of everyday language usage.

3.2 Translation Process

In the translation process, we engaged individuals with expertise in each of the five regions, ensuring that the translations were both accurate and consistent. For the Chittagong, Noakhali, and Barishal dialects, three individuals were involved in the translation process. Two translators worked on the Sylhet and Mymensingh regions. Each person played a vital role in understanding the variations of their respective dialects, using their linguistic expertise to translate the text effectively. The translation process was conducted cooperatively, with regular consultations to maintain accuracy and consistency across the dataset. This approach allowed us to capture the distinct features of each dialect while ensuring the dataset's overall quality and reliability.

3.3 Translation Guideline

We provided our translators with guidelines that emphasized authenticity while allowing for regional variability to maintain uniformity in translations. We recognized the value of linguistic variety and incorporated it into our dataset. For example, in the Chittagong region, the word "তার" (English translation: "His") can be translated as "ইবার" or "ইতার". Similarly, "সাথে" (English translation: "With") can be expressed as "লই" or "ফোয়ারে", and "গল্প" (English translation: "Story") can be written as "গল্ফ" or "কিচ্চা". In Barishal region, multiple words like "বড় ভাই" (English translation: "Elder Brother") conceivably translated as "ন্যাভাই" or "মেয়াভাই". Furthermore, the term "বোকা" (English translation: "fool") is also spoken as "গোজ্জা" or "বোগদা". Our translation guidelines allowed for different word choices with equivalent meanings, embracing the various writing styles and linguistic diversity found across different regions.

3.4 Translator Identity

We engaged with a team of competent and qualified translators, each with specialized expertise in their respective regions, to create the "Vashantor" dataset. Their qualifications and linguistic competence were essential in assuring the dataset's accuracy and validity. The translators' identities, allocated regions, and linguistic skills are highlighted in the Table 2.

3.5 Regional Dialect Variations

Our dataset covers a wide range of dialects and regional differences, showcasing the linguistic diversity across five distinct regions. We used cosine similarity [20], a measure of linguistic similarity between two spoken languages, to assess the relationship between Bangla and these regional dialects. This allowed us to quantify linguistic differences and similarities between dialects. For instance, the cosine similarity between the standard Bangla sentence "আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?" (English translation: "Do you not like to study at all?") and The presentation of equivalents in several dialects in Table 3 shows how these linguistic variances relate to the Bangla language.

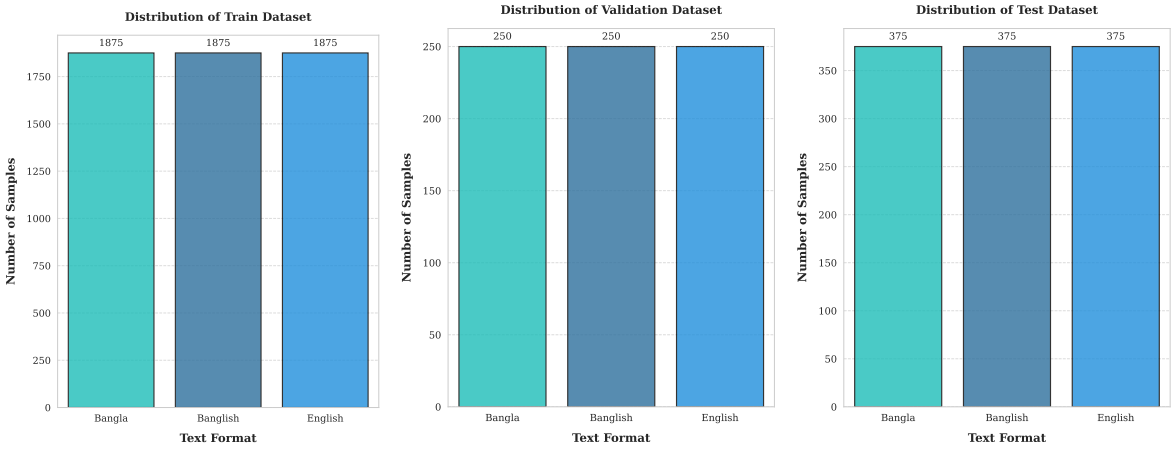
Table 3: Cosine Similarity Between Bangla Text and Five Bangla Regional Dialects

Bangla Text: আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?
Chittagong Dialect Text: অনর কি ফরালেহা গইরতো এবেরেও গম ন লাগে?
Cosine Similarity Between These Two Texts: 0.00
Bangla Text: আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?
Noakhali Dialect Text: আন্নের কি হরালেয়া কইত্তে একছের ভালা লাগে না?
Cosine Similarity Between These Two Texts: 0.00
Bangla Text: আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?
Sylhet Dialect Text: আফনার কিতা পড়ালেখা করতে এখদম ভালা লাগে নানি?
Cosine Similarity Between These Two Texts: 0.38
Bangla Text: আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?
Barishal Dialect Text: আমনের কি পড়াল্যাহা করতে এক্কালেই ভালো লাগে না?
Cosine Similarity Between These Two Texts: 0.40
Bangla Text: আপনার কি পড়ালেখা করতে একদমই ভাল লাগে না?
Mymensingh Dialect Text: আপনের কি পড়ালেহা করতে এক্কেরেই ভালা লাগে না?
Cosine Similarity Between These Two Texts: 0.86

The analysis of several Bangla speech corpora, through the use of the Term Frequency - Inverse Document Frequency (TF-IDF) [21] algorithm for getting average cosine similarity scores, provides valuable insights into linguistic relationships and reveals different levels of similarity. The Bangla and Mymensingh Speech Corpus has the most similarity, with a significant average cosine similarity score of 0.0288. Following closely after, the Bangla and Sylhet Speech Corpus had the most similarity, with a score of 0.0216. In contrast, comparisons with the Bangla and Chittagong Speech Corpus showed a lower average cosine similarity score of 0.0099. Similarly, the Bangla and Noakhali Speech Corpus shared a similarity score of 0.0139, while the Bangla and Barishal Speech Corpus had an average cosine similarity of 0.0124. These results show a decreasing linguistic similarity slope as one moves from Mymensingh and Sylhet to Noakhali, Barishal, and Chittagong.

3.6 Translation Quality Control Process

In our Translation Quality Control process, we implemented two essential metrics, Cohen's Kappa [22] and Fleiss' Kappa [23], to carefully assess translation quality and inter-annotator agreement, assuring the highest level of dependability for our dataset. Cohen's Kappa (K_1) was applied specifically to evaluate the translations for the Sylhet and Mymensingh regions. This metric involved the assessment of translations by Translators 1 and 2. Cohen's Kappa (K_1) for Sylhet and Mymensingh regions:



(a) Train Dataset (b) Validation Dataset (c) Test Dataset

Figure 1: Core Data Information

$$K_1 = 1 - \frac{1 - \kappa}{1 - \kappa_{\max}} \tag{1}$$

Where:

K_1 = Cohen's Kappa (K_1) for the Sylhet and Mymensingh regions.
κ = Cohen's Kappa coefficient for the agreement between Translators 1 and 2.
$\kappa_{\max}$ = Maximum possible agreement (usually equals 1).

Fleiss' Kappa (K_2) was employed for assessing the quality of translations in the Chittagong, Noakhali, and Barishal regions. Unlike Cohen's Kappa, Fleiss' Kappa extends the assessment to involve three Translators 1, 2, and 3. Fleiss' Kappa K_2 for Chittagong, Noakhali, and Barishal regions:

$$K_2 = \frac{1}{N(N-1)} \left[\sum_{j=1}^{k} \left(\frac{1}{N} \sum_{i=1}^{N} n_{ij}(n_{ij} - 1) \right) - \frac{1}{N(N-1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{N} n_{ij} \right)^2 \right] \tag{2}$$

Where:

K_2 = Fleiss' Kappa (K_2) for the Chittagong, Noakhali, and Barishal regions.
N = The total number of raters or translators (in this case, 3: Translators 1, 2, and 3).
k = The number of categories or ratings (commonly used when assessing quality).
n_{ij} = The number of raters who rated the i_th subject in the j_th category.

In terms of translation quality, the Chittagong region demonstrated a Fleiss' Kappa rating of 0.83, while the Sylhet region showed notable agreement with a Cohen's Kappa value of 0.87. On the other hand, the Noakhali region exhibited a Fleiss' Kappa rating of 0.91, while the Mymensingh region displayed a Cohen's Kappa value of 0.92, and the Barishal region demonstrated strong unity among independent translators with a Fleiss' Kappa rating of 0.93.

3.7 Dataset Statistics

We have carefully organized the "Vashantor" dataset to ensure comprehensive coverage for each region. The dataset statistics in the table below showcase the distribution of training, testing, and validation data for the five regions. Initially, we manually split the texts into 75% for training, 15% for testing, and 10% for the validation set presented in Figure 1 and Table 4. In the Table 5, Table 6, and Table 7, we provide an

Table 4: Regional Dialects Data Information

Region	Text Format	Number of Training Samples	Number of Testing Samples	Number of Validation Samples	Total Sample
Chittagong	Chittagong Bangla	1875	375	250	2500
	Chittagong Banglish	1875	375	250	2500
Noakhali	Noakhali Bangla	1875	375	250	2500
	Noakhali Banglish	1875	375	250	2500
Sylhet	Sylhet Bangla	1875	375	250	2500
	Sylhet Banglish	1875	375	250	2500
Barishal	Barishal Bangla	1875	375	250	2500
	Barishal Banglish	1875	375	250	2500
Mymensingh	Mymensingh Bangla	1875	375	250	2500
	Mymensingh Banglish	1875	375	250	2500

overview of speech corpus size, maximum text length, and minimum text length. This breakdown offers valuable insights into the variation in text lengths.

Table 5: Dataset Length for Core Data Collection

Text Format	Speech Corpus Size (in words)	Highest Text Length (in words)	Lowest Text Length (in words)
Bangla	72,439	19	2
Banglish	81,514	19	2
English	76,615	26	2

Table 6: Dataset Length for Different Regional Bangla Dialects

Region	Speech Corpus Size (in words)	Highest Text Length (in words)	Lowest Text Length (in words)
Chittagong	72,483	19	2
Noakhali	72,181	22	2
Sylhet	73,999	20	2
Barishal	77,494	19	2
Mymensingh	74,503	19	2

3.8 Benchmarking against Existing Datasets

To establish a comprehensive regional dialect dataset, we conducted comparisons with existing datasets, including those for English to Bangla and Bangla to English translations. Notably, our "Vashantor" dataset stands out for its distinctive incorporation of regional dialect translations. The comparative analysis in Table 8 presents a comprehensive overview of our dataset in relation to existing datasets.

Table 7: Dataset Length for Different Regional Banglish Dialects

Region	Speech Corpus Size (in words)	Highest Text Length (in words)	Lowest Text Length (in words)
Chittagong	77,599	19	2
Noakhali	78,045	22	2
Sylhet	82,424	20	2
Barishal	82,587	19	2
Mymensingh	81,751	19	2

Table 8: Benchmarking against Existing Datasets

Paper	Translation Directions	Regional Dialects to Bangla	Regional Dialects to Banglish	Region Detection	Dataset	Availability
This paper	All mentioned directions	Yes	Yes	Yes	32,500	will be publicly available
Rafiqul et al. [2]	Bangla to English	No	No	No	9,482	Not available
Shaykh et al. [5]	English to Bangla	No	No	No	4000	Not available
Prommy et al. [6]	No	No	No	Yes	30 hour	Publicly available
Kishorjit et al. [8]	Bangla to Manipuri	No	No	No	20,687 words	Not available
Niladri et al. [10]	Hindi to Banglish	No	No	No	80000	Not available
Mohammad et al. [24]	English to Bangla	No	No	No	70,614	Publicly available
Nafisa et al. [25]	Bangla to English	No	No	No	2,660	Not available
S.M. et al. [11]	No	No	No	Yes	9,303 voices	Not available

3.9 Challenges Faced

While creating the "Vashantor" dataset, we faced various challenges that made the process more complicated. These challenges included:

- Difficulty finding language experts
- Intra-regional language variations
- Diverse typing styles
- Spelling mistakes

People spoke in surprisingly diverse ways even within the same regions, which added another layer of complexity. The translators need to have a thorough understanding of the languages they were translating into was a big obstacle. They had to use caution when translating words from Bangla. For instance, when translating "সবাই" (English translation: "We"), they had to make a precise decision between "বেঞ্জুনে" and "বেকে". Translators had their unique typing styles, making consistency a challenge. An example of this is the

different spellings of "খাউয়া" (English translation: "Eat") and "খাওয়া", which mean the same thing but are spelled differently. Dealing with these various challenges was crucial to make sure the "Vashantor" dataset is known for its quality, accuracy, and language diversity.

3.10 Availability and Usage

We have structured the "Vashantor" dataset in easily accessible formats, primarily available in JSON and CSV, catering to the convenience of researchers and practitioners. These formats enable easy integration into a wide range of natural language processing applications and machine learning models. Once published, scholars and practitioners who want to use the dataset can do so through our dedicated online repository, which will make it available for academic and research purposes. This publicly available policy will promote the use of the dataset in language studies, dialect analysis, machine translation, and other domains. By providing straightforward access and a well-organized structure, we aim to facilitate the broadest possible usage of the "Vashantor" dataset within the research community.

4 Dialect-to-Bangla Translation and Region Detection Models

4.1 Regional Dialects to Bangla Language Translation Models

4.1.1 mT5

mT5, or Massively Multilingual Pre-trained Text-to-Text Transformer [26], is a multilingual version of the T5 text-to-text transformer model. It is a state-of-the-art language model with a robust encoder-decoder architecture. It has been pre-trained on a vast and diverse dataset comprising 101 languages sourced from the web. mT5 comes in various model sizes, ranging from 300 million to 13 billion parameters, allowing for high-capacity and powerful language models. One of its standout features is its exceptional competence in multilingual translation tasks, making it an ideal choice for projects involving the translation of text between different languages.

4.1.2 BanglaT5

BanglaT5 [27] is a state-of-the-art sequence-to-sequence Transformer model designed for the Bengali language. It is based on the original Transformer architecture and has been pretrained on the extensive "Bangla2B+" dataset, which contains 5.25 million documents gathered from a carefully selected list of web sources, totaling 27.5 GB of text data. The model architecture is the base variant of the T5 model, featuring 12 layers, 12 attention heads, a hidden size of 768, and a feed-forward size of 2048. Authors of another research [28] suggest two unique methods: aligner ensembling, which combines multiple sentence aligners to improve alignment accuracy, and batch filtering, which improves corpus quality by filtering out low-quality sentence pairings.

4.2 Region Detection Models

4.2.1 mBERT

The pretrained mBERT [29] model is designed for use with the top 104 languages and employs masked language modeling for self-supervised pretraining. It learns bidirectional sentence representations and sentence relationships. The publicly available model is consistent with BERT-base-cased in terms of its architectural specifications. It features 12 layers, 768 hidden units, 12 attention heads, and a total of 110 million parameters, mirroring the configuration of BERT-base-cased. mBERT can be fine-tuned on various downstream tasks and is particularly useful for tasks where the input text may be in multiple languages. It's versatile for multilingual tasks.

4.2.2 Bangla-bert-base

Bangla-bert-base [30] is a monolingual pretrained language model that follows the BERT architecture and makes use of mask language modeling for the Bengali language. The Bengali commoncrawl corpus and the Bengali Wikipedia Dump Dataset were transformed into the BERT format, with each sentence on a separate line and an extra line to indicate document separation. The BNLP package is used to generate the model's vocabulary, which consists of 102025 tokens and is made available on GitHub and the Hugging Face model hub. The publicly available model has 12 layers, 768 hidden units, 12 attention heads, and 110 million parameters; it is consistent with the bert-base-uncased architecture. One Google Cloud GPU was used for a

total of one million steps of training. An improved BERT variation titled BanglaBERT performs remarkably well through a range of Bengali NLP tasks.

5 Evaluation Metrics

We explore the evaluation metrics used in this section to assess our translation models' and region detection models' performance. We categorize the evaluation into two primary components: Translation Metrics and Region Detection Metrics.

5.1 Dialect-to-Bengali Translation Metrics

5.1.1 Character Error Rate

Character Error Rate (CER) [31] is a metric used to evaluate the quality of character-level text generation. It assesses the accuracy of generated text by measuring character-level errors, including substitutions, insertions, and deletions when compared to ground truth text. The CER score is typically expressed as a percentage or fraction, with lower values indicating higher accuracy in the generated text.

5.1.2 Word Error Rate

Word Error Rate (WER) [32] is a significant metric used to evaluate the accuracy and quality of generated text. WER quantifies the difference between the generated text and ground truth text in terms of words. It measures the accuracy of text output by considering word-level errors such as substitutions, insertions, deletions, and word order changes. WER is crucial for assessing the performance of translation systems. A lower WER score indicates higher accuracy, with scores closer to zero signifying that the generated translation is more faithful to the reference translation.

5.1.3 BLEU Score

BLEU (Bilingual Evaluation Understudy) [33] a tool for checking how good machine-generated translations are. It looks at how accurate and smooth the machine-generated text is. BLEU helps us see if the machine's translation matches human-made translations. To calculate the BLEU score, it considers factors like precision, recall, and a brevity penalty to give a complete assessment of the generated text. It works by comparing the similarity of n-grams, which are groups of n words, in the machine-generated text to the ground truth text. BLEU scores range from 0 to 1, with higher scores meaning better quality translations, especially for languages like Bengali.

5.1.4 ROUGE Score

We use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [34], which include ROUGE-1, ROUGE-2, and ROUGE-L, to evaluate machine-generated translations. These scores help us assess the quality and fluency of the translations. ROUGE-1 looks at how many single words in the machine's text match the ground truth text. ROUGE-2 checks for pairs of words (bigrams) that match, giving us a more detailed analysis of language accuracy. ROUGE-L examines the longest common sequence of words in both the machine's text and the ground truth text, providing insights into content coherence and flow.

5.1.5 METEOR Score

Meteor (Metric for Evaluation of Translation with Explicit ORdering) [35] is a powerful evaluation metric designed for assessing translation quality in the context of regional dialects to Bengali language translation. This metric is designed to assess the quality of translations by comparing them to human-crafted references.

5.2 Region Detection Metrics

5.2.1 Accuracy

Accuracy [36] in the context of region detection measures the proportion of correctly classified regions to the total number of regions in our dataset. It quantifies how well our model can accurately assign a text to its actual region. The score can be calculated as:

$$Accuracy = \frac{Number\ of\ Correctly\ Detected\ Regions}{Total\ Number\ of\ Regions} \tag{3}$$

5.2.2 Precision

Precision [37] is the ratio of true positives (correctly predicted instances of a specific region) to the sum of true positives and false positives (instances where the model incorrectly predicted the region). Precision is calculated as follows:

$$\text{Precision} = \frac{Correctly\ Detected\ Regions}{Total\ Detected\ Regions} \tag{4}$$

5.2.3 Recall

Recall [37], in the context of region detection for regional dialects, is a metric that measures the ability of a model to correctly identify and retrieve text samples belonging to a specific region from the "Vashantor" dataset. To put it another way, recall evaluates how well the model recognizes and categorizes text into its appropriate regional categories, ensuring that only a small number of samples are ignored or misclassified. The formula for recall is as follows:

$$\text{Recall} = \frac{Number\ of\ Correctly\ Detected\ Texts\ for\ a\ Specific\ Region}{Total\ Number\ of\ Texts\ Belonging\ to\ that\ Region} \tag{5}$$

5.2.4 F1 Score

The F1 Score [37] for region detection is the harmonic mean of precision and recall. It is particularly useful when we want to balance the trade-off between false positives and false negatives in the context of region detection. The score can be calculated as:

$$\text{F1 Score} = \frac{2 \cdot Precision\ for\ Region\ Detection \cdot Recall\ for\ Region\ Detection}{Precision\ for\ Region\ Detection + Recall\ for\ Region\ Detection} \tag{6}$$

5.2.5 Logarithmic Loss

Logarithmic Loss [38], commonly known as Log Loss, is a metric used to evaluate the performance of classification models in the context of multiclass classification, where we can classify texts into one of several regions (Chittagong, Noakhali, Sylhet, Barishal, Mymensingh). It quantifies the accuracy of a model's predicted class probabilities, rewarding accurate and confident predictions while penalizing uncertain or inaccurate ones. The Logarithmic Loss (Log Loss) for region detection is calculated as:

$$\text{Log Loss} = -\frac{1}{N} \sum [y \cdot \log(p) + (1 - y) \cdot \log(1 - p)] \tag{7}$$

Where:

N : Total number of texts.
y : True label (1 if the text belongs to a specific region, 0 otherwise).
p : Predicted probability that the text belongs to the specific region.

6 Proposed Methodology

This section describes the proposed methodology for translating different regional dialects to their corresponding standard Bengali language and region detection task, which is broken down into nine primary phases. Each input text gets two separate translation alternatives under the proposed method. In addition, we examine the quality of all translation options against the reference translation and the effectiveness of region detection models using a variety of performance metrics. The main phases of the proposed method for translations, taking into account five regional dialect sentences "অনে কি চন্না আই এডেত্তুন চলি জাই?" in Chittagong region, "আন্নে কি চান আই এডেত্তুন চলি যাই?" in Noakhali region, "আফনে চাইন নি আমি ইন তাকি যাই গি?" in Sylhet region, "আমনে কি চান মুই এইহানে গোনে চইল্লা যাই?" in Barishal region, "আফনে কিতা চান আমি এইহান থাইক্কা চইল্লা যাই?"

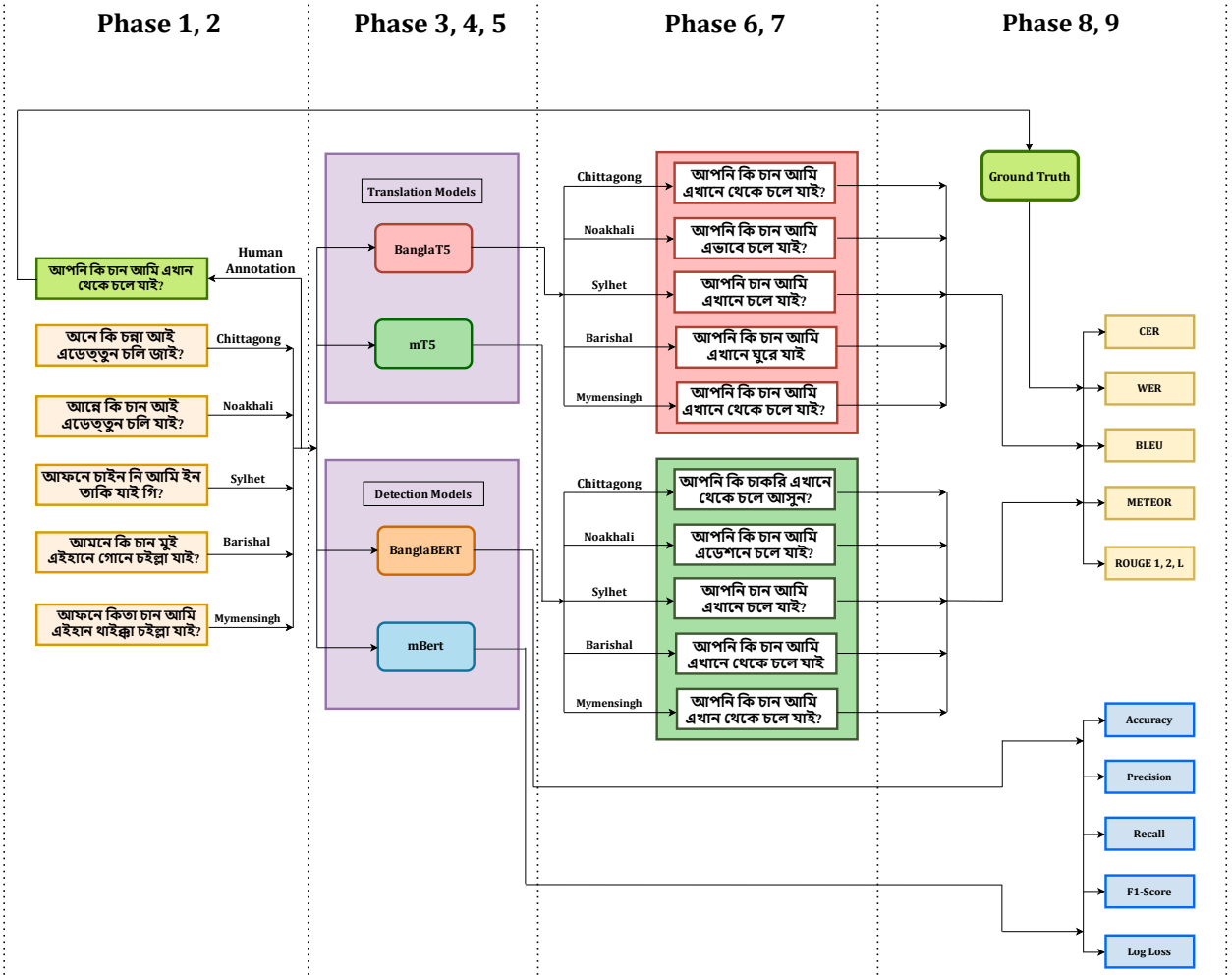


Figure 2: Methodology for two separate models based on sample data as input text and translated texts as output, as well as region detection to determine which input text corresponds to to which region.

in Mymensingh region, (English translation: "Do you want me to leave here?") as an example, are presented in Figure 2 and are discussed more below.

Phase 1) Input Text: In this phase, the input text is selected from our "Vashantor" dataset, a vast collection of text that includes a wide range of regional dialects.

Phase 2) Obtain Human Translation: During this phase, we collect human-generated translations from regional dialects into standard Bangla. These human translations act as the foundation for accurate translation by serving as the ground truth. While regional dialects vary greatly across Chittagong, Noakhali, Sylhet, Barishal, and Mymensingh, the human-generated Bangla translations are stable throughout all of them. These translations serve as a universal bridge, ensuring that the final output is in standard Bangla regardless of the original text's regional dialect. This phase ensures that the translation process follows a common, recognized Bangla language to provide clear and precise communication.

Phase 3) Regional Dialect Translation Models: During this phase, we train translation models that are specific to each regional dialect's linguistic characteristics. To translate from dialects to the standard Bangla, we employ models such as "mt5-small" and "BanglaT5". These models are particularly developed to comprehend the variety and complexity of five regional dialects, ensuring accurate and contextually appropriate translations.

Phase 4) Region Detection Models: In this Phase, we concentrate on region detection, which is an important part of our process. We employ powerful models, namely "mBERT" and "Bangla-bert-base," which

have been fine-tuned to effectively recognize the specific region from which the input text originates. The use of these models is motivated by their exceptional capacity to capture small variations in language associated with each regional dialect.

Phase 5) Hyperparameter Tuning: At this phase, we focus on improving our models' performance by fine-tuning their hyperparameters. Hyperparameters are external configuration variables that control how our translation and region detection models perform. We want to increase the accuracy, efficiency, and overall quality of our models by improving these hyperparameters. In Section 7.3, we present the detailed hyperparameter tuning procedures for both translation and region detection models.

Phase 6) Generation of Translation Options: Using the two different models, we generate two alternative translations for each input text. This offers us ten possible translations for the five regional dialects. This method allows us to analyze various translation options and select the one that most closely represents the standard Bangla language.

Phase 7) Post-processing Enhancement: We intend to improve our translations and region detection in this phase. We accomplish this by carefully enhancing the translated text using specialized methods. These methods examine grammar, punctuation, and the text's overall soundness. These complex methods entail looking at the entire text to ensure consistency and that the translation sounds authentic. In addition, we edit any grammar errors and change the style and tone to match the situation at hand.

Phase 8) Translation Quality Assessment: We apply five types of metrics to determine the quality of our translations from regional dialects to Bangla. These metrics assist us in measuring various aspects of translation accuracy and fluency. These metrics include: CER, WER, BLEU, METEOR, ROUGE(ROUGE-1, ROUGE-2, and ROUGE-L). In Section 7.2, we dig into an in-depth analysis of the scores obtained from these metrics. This analysis compares the performance of individual translation models and gives qualitative insights into translation quality.

Phase 9) Evaluation of Region Detection: During this phase, we want to ensure that our models can correctly identify the region from where the input text originates. We utilize many metrics to assess how effectively it operates. Accuracy, Precision, Recall, F1 Score, and Log Loss are examples of these metrics. In Section 7.2, We perform a complete metric score analysis to evaluate and compare the performance of different region detection models, providing significant insights into their accuracy and effectiveness.

7 Experimental Results and Analysis

7.1 Experimental Setup

The experiments were conducted on two different setups. The first setup used Google Colaboratory, with Python 3.10.12, PyTorch 2.0.1, a Tesla T4 GPU (15 GB), 12.5 GB of RAM, and 64 GB of disk space. The second setup used the Jupyter Notebook environment, with Python 3.10.12, PyTorch 2.0.1, an NVIDIA GeForce RTX 3050 GPU (8 GB), 16 GB of RAM, and a 512 GB NVMe SSD.

7.2 Experiments

The Table 9 provides an overview of the performance of two machine translation models, mT5 and BanglaT5, in translating Bangla regional text from five distinct regions (Chittagong, Noakhali, Sylhet, Barishal, and Mymensingh) to standard Bangla. The performance is measured in terms of four metrics: character error rate (CER), word error rate (WER), BLEU score, and METEOR score. BanglaT5 outperforms mT5 across all four metrics in four regions, except for Sylhet. This indicates that BanglaT5 excels in translating Bangla regional dialects to standard Bangla. In Chittagong, for instance, BanglaT5 exhibits a CER of 0.2040 and a WER of 0.3385, while mT5 records a CER of 0.2308 and a WER of 0.3959. Additionally, BanglaT5 achieves higher BLEU and METEOR scores, with a BLEU score of 44.03 and a METEOR score of 0.6589, whereas mT5 scores 36.75 in BLEU and 0.6008 in METEOR. Similarly, in Noakhali, BanglaT5's performance surpasses that of mT5 with a CER of 0.1863 and a WER of 0.3214, whereas mT5 has a CER of 0.2035 and a WER of 0.3870. Moreover, BanglaT5 achieves an impressive BLEU score of 47.38 and a METEOR score of 0.6802, surpassing the corresponding scores for mT5, which are 37.43 in BLEU and 0.6073 in METEOR. In Sylhet, mT5 has a CER of 0.1472 and a WER of 0.2695, while BanglaT5 has a CER of 0.1715 and a WER of 0.2802. Additionally, mT5 secures a higher BLEU score of 51.32, compared to BanglaT5's BLEU score of 51.08. Moving on to Barishal, mT5, and BanglaT5 exhibit competitive performance, with mT5 having a CER of 0.1480 and a WER of 0.2644, while BanglaT5 records a CER of 0.1497 and a WER of 0.2459. Furthermore,

Table 9: CER, WER, BLEU, METEOR scores of all the Bangla regional dialect translation models

Region	Models	CER	WER	BLEU	METEOR
Chittagong	mT5	0.2308	0.3959	36.75	0.6008
	BanglaT5	0.2040	0.3385	**44.03**	0.6589
Noakhali	mT5	0.2035	0.3870	37.43	0.6073
	BanglaT5	0.1863	0.3214	**47.38**	0.6802
Sylhet	**mT5**	0.1472	0.2695	**51.32**	0.7089
	BanglaT5	0.1715	0.2802	51.08	0.7073
Barishal	mT5	0.1480	0.2644	48.56	0.7175
	BanglaT5	0.1497	0.2459	**53.50**	0.7334
Mymensingh	mT5	0.0796	0.1674	64.74	0.8201
	BanglaT5	0.0823	0.1548	**69.06**	0.8312

in terms of translation quality, BanglaT5 surpasses mT5 with a BLEU score of 53.50 and a METEOR score of 0.7334, while mT5 scores 48.56 in BLEU and 0.7175 in METEOR. In Mymensingh, both models demonstrate comparable CER and WER, with BanglaT5 maintaining a slight advantage in BLEU score of 69.06 and METEOR score of 0.8312 over mT5, which scores 64.74 in BLEU and 0.8201 in METEOR. Overall, the study found that Chittagong has the lowest performance metrics scores for both BanglaT5 and mT5, while Mymensingh has the highest performance metrics scores.

The presented Table 10 outlines the ROUGE scores for Bangla regional dialect translation models, mT5 and BanglaT5, across various regions. BanglaT5 consistently outperforms mT5 in terms of recall, precision, and f1-score across all regions and versions rogue-1, rogue-2, and rogue-L. Notably, BanglaT5 exhibits superior performance, showcasing its effectiveness in capturing the nuances of Bangla regional dialects. Specifically, BanglaT5 consistently outshines mT5, with the Mymensingh region consistently exhibiting the highest performance for both models. Mymensingh scores surpass 0.70 and reach a peak of 0.84 in recall, precision, and f1-score. In contrast, the Chittagong region tends to display comparatively lower scores, ranging from 0.4662 to 0.7321 for both models. This suggests that BanglaT5 is particularly effective in translating Mymensingh dialect, while Chittagong dialect poses a greater challenge for both models.

The Table 11 presents a comparative analysis of two region detection models: mBERT and Bangla-bert-base. In terms of overall performance metrics, Bangla-bert-base exhibits a slightly higher accuracy of 85.86% compared to the accuracy of mBERT, which is 84.36%, indicating a marginally better ability to correctly classify instances. Additionally, Bangla-bert-base also demonstrates a lower log loss of 0.8804, suggesting more accurate probabilistic predictions compared to mBERT's log loss of 0.9549. The confusion matrices for these two models are displayed in Figure 3a and Figure 3b. Moving on to the region-specific metrics, both models are evaluated on their precision, recall, and f1-score for five distinct regions: Chittagong, Noakhali, Sylhet, Barishal, and Mymensingh. In the Chittagong region, Bangla-bert-base shows a precision of 0.8840, recall of 0.9147, and an f1-score of 0.8991, indicating a balanced performance in correctly identifying instances of Chittagong. On the other hand, mBERT demonstrates lower a precision of 0.8779 and a lower recall of 0.9013 in the same region. For the Barishal region, mBERT exhibits a precision of 0.9437 and a recall of 0.9412, resulting in an f1-score of 0.9424. Bangla-bert-base, however, shows a slightly lower precision of 0.9301 and higher recall of 0.9599, yielding a marginally higher f1-Score of 0.9447. Similar variations in precision, recall, and f1-score are observed across the other regions. A comparison of two BERT models is presented in Figure 4.

7.3 Hyperparameter Settings

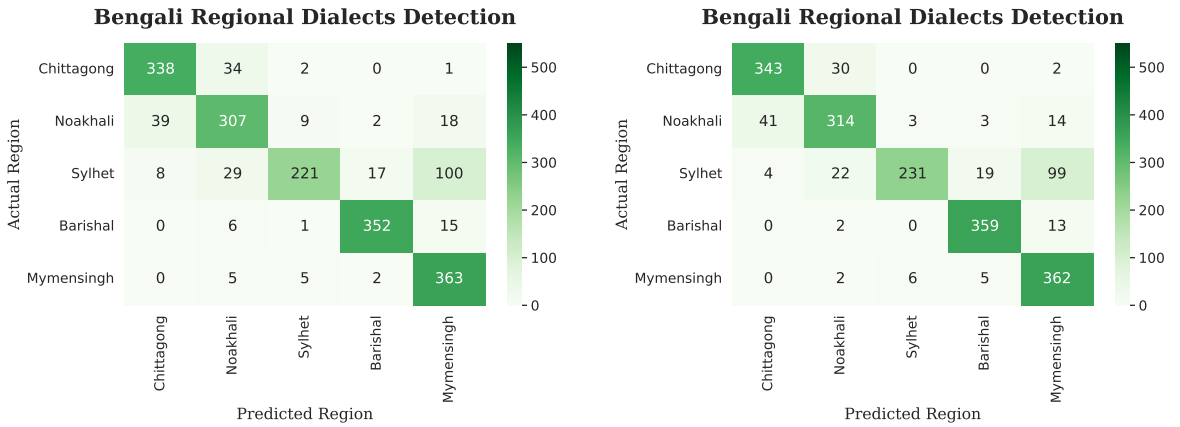
The Table 12 shows the hyperparameters used to train a regional dialects translation model for five different regions in Bangladesh: Chittagong, Noakhali, Sylhet, Barishal, and Mymensingh. In the hyperparameter tuning for Bangla regional dialects to Bangla translation using two models, mT5 and BanglaT5. Key hyperparameters include a learning rate of 0.001, a fixed batch size of 16, and varying numbers of epochs for each region and model, such as the highest number of epochs is observed in the Chittagong region for the BanglaT5 model, reaching 53, while the lowest is found in the Mymensingh region for the same model, with 28 epochs. The optimization algorithm employed is AdamW. Additionally, a sequence length of 128 is set,

Table 10: ROUGE scores of all the Bangla regional dialect translation models

Region	Translation Model	Version	Recall	Precision	F1-Score
Chittagong	mT5	r-1	0.6563	0.6820	0.6659
		r-2	0.4662	0.4854	0.4733
		r-L	0.6526	0.6784	0.6623
Chittagong	BanglaT5	r-1	0.7082	0.7321	0.7172
		r-2	0.5217	0.5413	0.5290
		r-L	0.7032	0.7272	0.7123
Noakhali	mT5	r-1	0.6670	0.6753	0.6642
		r-2	0.4765	0.4745	0.4712
		r-L	0.6615	0.6723	0.6642
Noakhali	BanglaT5	r-1	0.7282	0.7312	0.7245
		r-2	0.5517	0.5632	0.5590
		r-L	0.7221	0.7321	0.7232
Sylhet	mT5	r-1	0.7487	0.7721	0.7584
		r-2	0.5851	0.6028	0.5923
		r-L	0.7472	0.7703	0.7568
Sylhet	BanglaT5	r-1	0.7493	0.7721	0.7578
		r-2	0.5881	0.6054	0.5944
		r-L	0.7477	0.7705	0.7562
Barishal	mT5	r-1	0.7545	0.7635	0.7628
		r-2	0.5877	0.5968	0.5885
		r-L	0.7524	0.7623	0.7585
Barishal	BanglaT5	r-1	0.7735	0.7759	0.7729
		r-2	0.6084	0.6107	0.6082
		r-L	0.7732	0.7755	0.7726
Mymensingh	mT5	r-1	0.8418	0.8458	0.8431
		r-2	0.7176	0.7214	0.7189
		r-L	0.8418	0.8458	0.8431
Mymensingh	BanglaT5	r-1	0.8407	0.8355	0.8362
		r-2	0.7128	0.7088	0.7091
		r-L	0.8407	0.8355	0.8362

Table 11: Performance Overview of all region detection models

Models	Accuracy	Log Loss	Region	Precision	Recall	F1-Score
mBERT	84.36%	0.9549	Chittagong	0.8779	0.9013	0.8895
			Noakhali	0.8058	0.8187	0.8122
			Sylhet	0.9286	0.5893	0.7210
			Barishal	0.9437	0.9412	0.9424
			Mymensingh	0.7304	0.9680	0.8326
Bangla-bert-base	85.86%	0.8804	Chittagong	0.8840	0.9147	0.8991
			Noakhali	0.8486	0.8373	0.8430
			Sylhet	0.9625	0.6160	0.7512
			Barishal	0.9301	0.9599	0.9447
			Mymensingh	0.7388	0.9653	0.8370



(a) mBERT (b) Bangla-bert-base

Figure 3: Heat map representation of the region detection confusion matrix

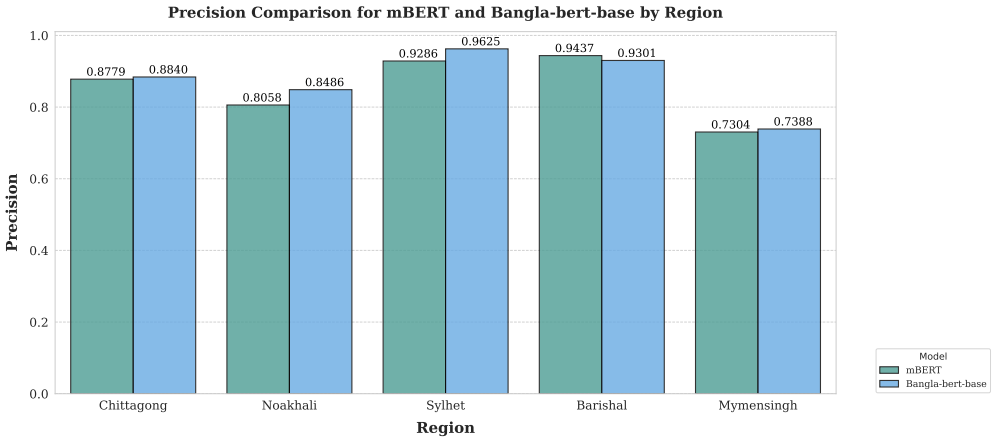
Table 12: Hyperparameter Tuning for Bangla Regional Dialects to Bangla Translation

Region	**Models**	**Learning Rate**	**Batch Size**	**Number of Epochs**	**Optimizer**	**Sequence Length**
Chittagong	mT5 BanglaT5	0.001 0.001	16 16	50 53	AdamW AdamW	128 128
Noakhali	mT5 BanglaT5	0.001 0.001	16 16	45 40	AdamW AdamW	128 128
Sylhet	mT5 BanglaT5	0.001 0.001	16 16	43 45	AdamW AdamW	128 128
Barishal	mT5 BanglaT5	0.001 0.001	16 16	35 35	AdamW AdamW	128 128
Mymensingh	mT5 BanglaT5	0.001 0.001	16 16	30 28	AdamW AdamW	128 128

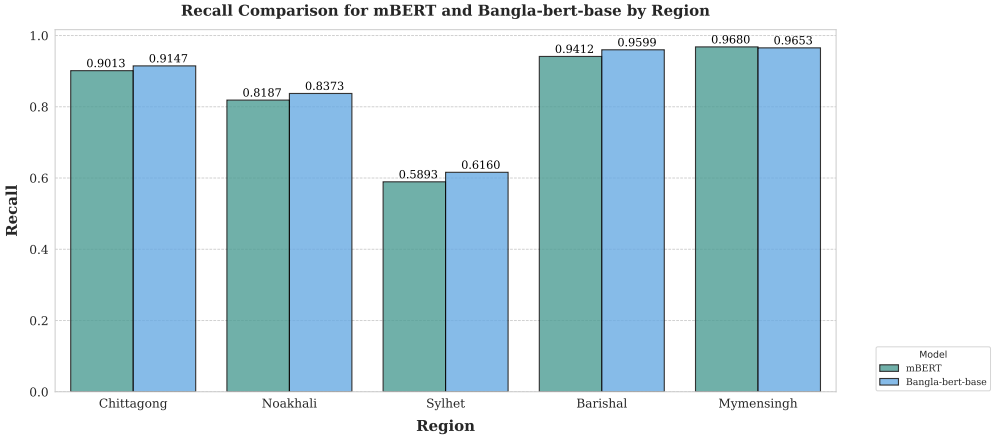
a critical parameter for tasks dealing with sequential data like natural language processing. Moving on to Table 13, the hyperparameter tuning results for region detection are presented, focusing on two pre-trained BERT models: mBERT and Bangla-Bert-Base. For all regions, both models are trained with consistent hyperparameter values, including a learning rate of 0.00002, a batch size of 16, 10 epochs using the AdamW optimizer, and a sequence length of 128.

7.4 Performance Comparison

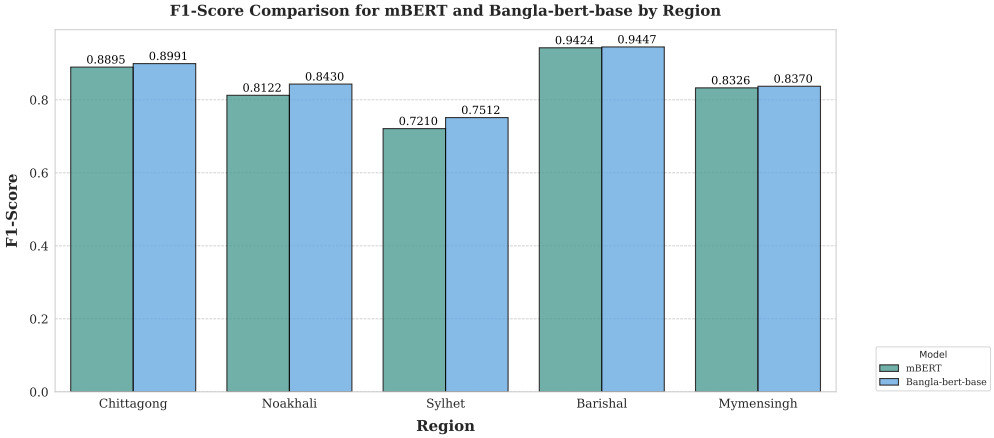
We've explained a translation error analysis in Table 14, outlining the source text (ST), its reference translation (RT), the model-generated translation (MT), and its English translation (ET). This comprehensive assessment includes an in-depth examination of error types, including lexical, grammatical, and contextual. Lexical errors appear as mistranslations, improper word choices, or semantic inequalities, and include flaws at the word or vocabulary level. Grammatical errors are faults with sentence structure, syntax, and grammatical rules, such as incorrect conjugation of verbs or inappropriate sentence formation. Contextual errors are differences in expressing the intended meaning within a larger context, which frequently result from a lack of understanding of contextual details in the source text. Each detected error is evaluated for severity and labeled as either major or minor. Major errors, which have a significant influence on overall coherence and accuracy, have a significant impact on translation accuracy. Minor errors, on the other hand, despite having a lower impact, nonetheless contribute to the overall assessment of translation quality by addressing finer concerns that may hinder readability. The Correction/Notes encompasses precise correction ideas marked by 'INS' (Insert) for adding content, 'DEL' (Delete) for eliminating core issues, and 'SUB' (Substitute) for



(a) Precision Comparison



(b) Recall Comparison



(c) F1score Comparison

Figure 4: Comparing Precision, Recall, and F1 Score Performance of mBERT and Bangla-bert-base

Table 13: Hyperparameter Tuning for Region Detection

Region	Models	Learning Rate	Batch Size	Number of Epochs	Optimizer	Sequence Length
Chittagong	mBERT	0.00002	16	10	AdamW	128
	Bangla-bert-base	0.00002	16	10	AdamW	128
Noakhali	mBERT	0.00002	16	10	AdamW	128
	Bangla-bert-base	0.00002	16	10	AdamW	128
Sylhet	mBERT	0.00002	16	10	AdamW	128
	Bangla-bert-base	0.00002	16	10	AdamW	128
Barishal	mBERT	0.00002	16	10	AdamW	128
	Bangla-bert-base	0.00002	16	10	AdamW	128
Mymensingh	mBERT	0.00002	16	10	AdamW	128
	Bangla-bert-base	0.00002	16	10	AdamW	128

modifying words or phrases. Additional notes provide further insights into translation details, allowing for a more complete understanding of issues faced as well as suggestions for improvement.

8 Future Research Directions

The detection of slang will be an important part of our future work in the field of regional dialects to Bengali language translation. Slang, which refers to informal words and phrases used within distinct dialects, is common in numerous regions. In the Barishal region, for example, "সামার ফো" (English translation: "son of an asshole") is a common informal slang word, and in the Mymensingh region, "গোলামের পুত" (English translation: "son of a slave") is a typical informal slang word. The ability to recognize and handle slang is crucial for showing accurate and culturally suitable translations. In addition, our following research will include the extension of sentiment analysis to include emotion recognition within regional dialects. This enhancement intends to give a deeper awareness of emotions such as joy, anger, sadness, and surprise, as well as to add depth to our translation abilities. In addition to these works, we will concentrate on cross-regional translation, which aims to overcome language differences across dialects and encourage simpler communication. We may improve translation accuracy and relevancy by recognizing the unique characteristics of each regional dialect. For example, in Chittagong region, a phrase like "তোঁয়ার কি মন হারাফ নে?" (English translation: "Are you upset?") may be translated as "তুমার কিতা মন খারাফ নি?" in Sylhet region, emphasizing the need for cross-regional linguistic comprehension. In addition, we will evaluate dynamic writing styles used in Bengali. For instance, various forms like ("চলো" , "চোল" , "চোলো" , and "চল") are all acceptable for the word "come on".

9 Conclusion

Our research has made considerable advances in the translation of Bangla regional dialects into the standard Bangla language. We not only recognized substantial linguistic variances across Bangla's several regional dialects, but we also created translation models and datasets that successfully bridge these dialectal gaps. We have shown the usefulness of our models in generating accurate and socially acceptable translations by thoroughly evaluating them using a range of performance metrics. The models consistently provided notable BLEU scores across different regions, with the Chittagong region reaching the highest at 44.03 using BanglaT5, and the Mymensingh region achieving an amazing peak of 69.06 using the same model. Furthermore, the Noakhali region has the highest BLEU score of 47.38 when utilizing BanglaT5. Additionally, our model achieved a stunning BLEU score of 51.32 in the Sylhet region with mT5, while our models performed excellently in the Barishal region, obtaining the highest BLEU score of 53.50 using BanglaT5. In terms of region detection, our Bangla-bert-based model obtained an accuracy of 85.86%, slightly exceeding mBERT, which achieved an accuracy of 84.36%. Our findings will help people from various regional dialect-speaking communities communicate and understand each other better.

Table 14: An in-depth analysis of translation errors across every region.

Region	**Examples**	**Error Type**	**Severity**	**Correction /Notes**
Chittagong	**ST:** আর ডোর বাই ইবা তার বন্ধুর ফয়ারিত এহন বিরি ন হায়দে **RT:** আমার বড় ভাই এখন তার বন্ধুদের সাথে সিগারেট খায় না **MT:** আমার বড় বোন এখন তার বন্ধুদের সাথে বিয়ে করে নাহ **ET:** My elder brother does not smoke cigarettes with his friends now	Lexical & Contextual	Major	“বোন” SUB “ভাই”, “বিয়ে” SUB “সিগারেট”, “করে” SUB “খায়”, DEL “হ” from নাহ”
Noakhali	**ST:** আইজগা সিলেটে বৃষ্টি অইব **RT:** আজ সিলেটে বৃষ্টি হবে **MT:** আজকে সিলেট বৃষ্টি হবে **ET:** Today it will rain in Sylhet	Grammatical	Minor	DEL “কে” from “আজকে”, “সিলেট” SUB “সিলেটে”
Sylhet	**ST:** আমার বান্ধবী কিচ্ছু ভাবতে ফারে না আফনাকে ছাড়া **RT:** আমার বান্ধুবি কিছু ভাবতে পারেনা আপনাকে ছাড়া **MT:** আমার গার্লফ্রেন্ড কিছু চিন্তা করতে পারে না আপনাকে ছাড়া **ET:** My girlfriend can’t think of anything without you	Lexical	Minor	“গার্লফ্রেন্ড” SUB “বান্ধুবি”, “চিন্তা” SUB “ভাবতে”, DEL “করতে” before “পারে”
Barishal	**ST:** হ্যায় বিয়াইত্যা হেইতে মামলা করতে চাইতেয়াছে না **RT:** সে বিবাহিত তাই মামলা করতে চাচ্ছে না **MT:** সে বিয়ে করে তাই মামলা করতে চাইছে না **ET:** She is married so does not want to sue	Contextual	Major	“বিয়ে” SUB “বিবাহিত”, DEL “করে” before “তাই”, “চাইছে” SUB “চাচ্ছে”
Mymensingh	**ST:** তোমার হাশির কাসে আমার বেগ কিসু হার মাইন্না যায় **RT:** তোমার হাসির কাছে আমার সব কিছু হার মেনে যায় **MT:** তোমার বোনের কাছে আমার সব হার মেনে যায় **ET:** I lose everything to your smile	Lexical & Contextual	Major	“বোনের” SUB “হাসির”, INS “কিছু” after “সব”

10 Declarations

10.1 Ethical Approval and Consent to participate

Not Applicable.

10.2 Human and Animal Ethics

Not Applicable.

10.3 Conflicts of Interest

The authors declare that they have no conflicts of interest.

10.4 Funding

This research was carried out with no external funding.

10.5 Authors' contributions

Faria and Mukaffi defined the research scope, conducted the study, collected data, performed coding, executed the majority of experiments, and drafted the manuscript. Wase contributed to data collection, conducted several experiments, analyzed the writing quality, and addressed grammatical errors in the paper. Mehidi and Rabius participated in data collection and analysis. Tashreef executed some experiments and provided critical editing for the manuscript.

References

[1] Md. Arid Hasan, Firoj Alam, Shammur Chowdhury, and Naira Khan. Neural machine translation for the bangla-english language pair, 12 2019.

[2] Rafiqul Islam, Mehedi Hasan, Mamunur Rashid, and Rabea Khatun. Bangla to english translation using sequence to sequence learning model based recurrent neural networks. In Md. Shahriare Satu, Mohammad Ali Moni, M. Shamim Kaiser, and Mohammad Shamsul Arefin, editors, *Machine Intelligence and Emerging Technologies*, pages 458–467, Cham, 2023. Springer Nature Switzerland.

[3] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level machine translation: Methods and evaluation, 12 2019.

[4] Laith H. Baniata, Isaac. K. E. Ampomah, and Seyoung Park. A transformer-based neural machine translation model for arabic dialects that utilizes subword units. *Sensors*, 21(19), 2021.

[5] Shaykh Siddique, Tahmid Ahmed, Md Talukder, and Md. Mohsin Uddin. English to bangla machine translation using recurrent neural network. *International Journal of Future Computer and Communication*, pages 46–51, 06 2020.

[6] 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), 2022.

[7] Redwan Rizvee, Asif Mahmood, Shakur Mullick, Sajjadul Hakim, and Seth Darren. A robust three-stage hybrid framework for english to bangla transliteration. *International Journal on Natural Language Computing*, 11:15, 02 2022.

[8] Kishorjit Nongmeikapam, Ningombam Herojit Singh, Sonia Thoudam, and Sivaji Bandyopadhyay. Manipuri transliteration from bengali script to meitei mayek: A rule based approach. In Chandan Singh, Gurpreet Singh Lehal, Jyotsna Sengupta, Dharam Veer Sharma, and Vishal Goyal, editors, *Information Systems for Indian Languages*, pages 195–198, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[9] Naushad UzZaman. *Phonetic encoding for Bangla and its application to spelling checker , transliteration , cross language information retrieval and name searching.* N/A, 2005.

[10] Author Name. Problems and challenges in hindi to bangla translation: Some empirical observation and workable solutions. *Translation Today*, 13(1), 2004.

[11] S. M. Saiful Islam Badhon, Habibur Rahaman, Farea Rehnuma Rupon, and Sheikh Abujar. Bengali accent classification from speech using different machine learning and deep learning techniques. In Samarjeet Borah, Ratika Pradhan, Nilanjan Dey, and Phalguni Gupta, editors, *Soft Computing Techniques and Applications*, pages 503–513, Singapore, 2021. Springer Singapore.

[12] Lenin Laitonjam and Sanasam Ranbir Singh. Manipuri-English machine translation using comparable corpus. In John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu, editors, *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual, August 2021. Association for Machine Translation in the Americas.

[13] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

[14] Soran Badawi. A transformer-based neural network machine translation model for the kurdish sorani dialect. *UHD Journal of Science and Technology*, 7(1):15–21, Jan. 2023.

[15] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation, 2018.

[16] Wenting Ma, Bing Yan, and Lianyue Sun. Generative adversarial network-based short sequence machine translation from chinese to english. *Scientific Programming*, 2022:7700467, Jan 2022.

[17] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing, 2018.

[18] Dongxing Li and Zuying Luo. An improved transformer-based neural machine translation strategy: Interacting-head attention. *Computational Intelligence and Neuroscience*, 2022:2998242, Jun 2022.

[19] Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. A reinforced generation of adversarial examples for neural machine translation, 11 2019.

[20] Dani Gunawan, C Sembiring, and Mohammad Budiman. The implementation of cosine similarity to calculate text relevance between two documents. *Journal of Physics: Conference Series*, 978:012120, 03 2018.

[21] Shahzad Qaiser and Ramsha Ali. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 07 2018.

[22] Nicole Blackman and John Koval. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine*, 19:723–41, 04 2000.

[23] Rosa Falotico and Piero Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470, 03 2014.

[24] Md Abdullah Al Mumin, Abu Awal Md Shoeb, Md Reza Selim, and M Zafar Iqbal. Supara: A balanced english-bengali parallel corpus. *SUST Journal of Science and Technology*, 16(2):46–51, 2012.

[25] Nafisa Nowshin, Zakia Sultana Ritu, and Sabir Ismail. A crowd-source based corpus on bangla to english translation. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–5, 2018.

[26] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[27] Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[28] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online, November 2020. Association for Computational Linguistics.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[30] Sagor Sarker. Banglabert: Bengali mask language model for bengali language understanding, 2020.

[31] Jacob Wobbrock and Brad Myers. Analyzing the input stream for character- level errors in unconstrained text entry evaluations. *ACM Trans. Comput.-Hum. Interact.*, 13:458–489, 12 2006.

[32] Klaus Zechner and Alex Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, page 186–193, USA, 2000. Association for Computational Linguistics.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[34] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[35] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[36] Vincent Labatut and Hocine Cherifi. Accuracy measures for the comparison of classifiers, 2012.

[37] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[38] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5511–5518, Apr. 2020.