Attention-Based Sequence-to-Sequence Neural Machine Translation from English to Bangla

BY

Prattoy Paul Borson
ID No: 241-25-016

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Professor Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

AUGUST 2025

APPROVAL

This Project/Thesis titled "**Attention-Based Sequence-to-Sequence Neural Machine Translation from English to Bangla**", submitted by **Prattoy Paul Borson**, ID No: **241-25-016** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

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

Chairman

Ms. Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

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

Internal Examiner

Mr. Sadat Hossain,
Data Scientist,
Risk Management Division,
BRAC Bank Limited

External Examiner

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Professor Dr. Sheak Rashed Haider Noori, Professor & Head, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

13.09.2025

Professor Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. **Abdus Sattar**
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Prattoy Paul Borson
ID: 241-25-016
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

I am grateful and wish to express my profound indebtedness to **Professor Dr. Sheak Rashed Haider Noori, Professor & Head,** Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head**, **Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

Neural Machine Translation (NMT) has emerged as a dominant form of automatic languages, especially in the case of lower-resourced languages like Bengali. We developed an English-to-Bengali translation system utilizing the BanglaT5 Transformer model, which was pre-trained on Bengali data. The parallel English-Bengali dataset was cleaned of unwanted characters, to the best of our ability, and normalized for a large enough data pool to train the model, and to ensure uniformity. The sentences in the dataset were tokenized to prepare the input sequences and attention masks, with associated target labels, with the aim of supervised learning. A custom-made dataset with DataLoader from Pytorch allowed the batch size to be tailored and to also facilitate efficient training on the GPU. The BanglaT5 model was fine-tuned using cross-entropy loss with the Adam optimizer to minimize the loss function using backpropagation. Evaluation was performed using BLEU and chrF scores to assess translation accuracy of Bengali sentences generated by the model against provided reference translations. A good illustration of reining in of this system is evident. With great success on a low-resourced NMT problem, a pre-trained Transformer is used. translation. As outlined in this paper, appropriate texts normalizing, prep pre tokenizing, and fine-tuning are important to eventually come up with superior quality Machine Translation. System (MTS). This is a strategy that has been scaled to. host other languages with low resources and is a resourceful guide. South-Asian language-pairs NMT tasks implementation.

Keywords:

Neural Machine Translation, English-to-Bangla Translation, Transformer Model, BLEU Score, chrF Score, Natural Language Processing, Deep Learning, Low-Resource Languages

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Over the past few years, Natural Language Processing (NLP) has become one of the fast growing in reference. rapidly developing fields of Artificial Intelligence, machine translation (MT) is an essential field. research area.. As communication becomes more globalized, there is a strong desire to develop systems that can increase the ease and accuracy of translation. This is especially important for languages that do not have strong representation on most digital platforms. Bangla, which is spoken by more than 230 million individuals globally, is one such example of a low-resource language. Thus, although English-to-Bangla translation continues to increase for educational, technological, and global communication purposes, translation systems are still quite inconsistent and lack contextual accuracy.

Traditionally, MT had either a rule-based or statistical approach where neither employed the linguistic/generative complexity of natural languages across cultures. As a result, MT generated literal translations after acquiring grammatical systems but were unable to generate translations that retained meaning or fluency. Furthermore, recent advancements in Neural Machine Translation (NMT) especially by deploying transformer-based architectures has increased translation quality. This is supported with the pre-trained model. BanglaT5, together with more possibilities to optimize translation systems on. relatively small datasets, which allows researchers to be competitive. performance. This study describes the making and evaluation of an English-to-Bangla NMT. system on BanglaT5 model on a Kaggle notebook. For this type of research we began by preparing the dataset, followed by tokenization, fine-tuning of the pretrained model and assessment of the model in accordance with conventional performance indicators, namely BLEU and CHRF. Every section of the study such as data preprocessing, batch size optimization, and. decoding strategies are aimed at the enhancement of translation accuracy as an element of. workflow. Lastly the study examined the effect of the learning rate scheduling and convergence and performance are influenced by methods of evaluation. The pipeline

presented here was used to make progress towards creating a stronger English-to- Bangla translation system. As a contribution of research, it will add to the general body of. expertise in both academic and practice-based research, e.g. digital education. cognition, cross-linguistic interaction, and resources. accessibility.We hope this thesis can demonstrate how modern transformer-based models can lessen the language gap for Bangla and provide Barangkabau design & implementation milestones for future researchers and practitioners using multi-modality and text based approaches to the same problem, which often face exclusion on the platforms of digital globalization.

1.2 Motivation and Problem Statement

Machine Translation (MT) has improved considerably and serves as an essential medium for bridging communication gaps in our global, digital world. Although resource-rich languages such as English, French, and Chinese have efficient machine translation systems, some low-resource languages such as Bangla continue to have issues. In terms of people who speak it, Bangla is the seventh most widely spoken language in the world, but practically absent in computational research. A problem with translation models, in general, is that they cannot capture contextual meaning, only literal meaning, which produces useless results. Both traditional rule-based, and statistical approaches are limited; accordingly, many neural models are not yet optimized for Bangla, due to the lack of appropriate datasets.

This study aims to fill a void by training-up BanglaT5 transformer models using an English-to-Bangla parallel dataset hosted on Kaggle. This study will outline the principal steps involved in machine translation, namely, preprocessing, tokenization, model training, and evaluation using BLEU, and CHRF. This research aims to develop a better model that articulates Bangla with better accuracy to increase digital inclusivity - thereby providing Bangla speakers with better access to knowledge and communication on the world stage.

1.3 Rationale of the Study

This study deals with a topic that is closely tied to the under researched area of Neural Machine Translation (NMT) with respect to the only previously researched low-resourced languages. While machine translation has been around in one form or another for many

years, most of the previous work in machine translation has been on statistical and rule-based translation which does not attempt to produce meaning-based, context-aware, or fluent translation in the way learned models can. With the adoption of deep learning and transformer-based models such as BERT, mBART, and T5, accuracy in translation has increased in a variety of low-resourced languages. However, many of these models are propelled more by the increasingly number of applications supporting high-resourced languages and Bangla often is left with few successful and effective means of adaptation. This study is built on the direction of these language advancement studies by fine-tuning a BanglaT5 model based on an English to Bangla text dataset. Unlike standard multilingual models, BanglaT5 is optimized for languages such as Bangla which may provide more contextually relevant features. As a result, this study engages a lucent gap between the prior NMT studies and the need for translatorial services for Bangla speakers and its consequences as a means towards digital inclusion, access to education, and cross linguistic communication of Bangla speakers.

1.4 Research Question

- How can attention-based neural models, especially transformers, improve English-to-Bangla translation accuracy over traditional methods?
- How do preprocessing, subword tokenization, and corpus quality impact translation of rare words and complex sentences?
- How can transliteration and model choice (RNN vs. transformer) address out-of-vocabulary words and optimize translation performance?

These questions will be used to identify the design and evaluation of the proposed model and offer some systematic insight into the capacity of emerging neural structures to. could address the historical issues of putting the Bangla MT systems into practice.

1.5 Expected Outcome

The primary output of this project is a machine translation that works as a neural network. system which converts English to Bangali, by models using transformers, or Encoder-decoder models in RNN. The project will deliver::

1. Bangali Text: The system will take an input sentence in English and produce a machine translation of the input text into Bengali, which is not distorted. meaning or structure.

2. The Tokenization and Preprocessing: The model will automatically process the input sentences by tokenizing them, subword segmenting them, and padding them so that the input sentences in the model can be of variable length and used in translation

3. Translation Evaluation Metrics: The system will offer evaluation metrics of translation quality, such as BLEU, CHRF, in which it is possible to compare the results with the current state-of-the-art practices.

4. Batch Processing: The code will offer an option of processing a number of sentences efficiently at the same time, with the help of the dataloader feature of the pytorch framework.

5. Attention Visualization (Optional):The model can give attention weights indicating what part of the English input the model concentrates on in the translation process.

6. Low-resource Handling: The system will handle low-resource methods, including rare words, morphological variation, and out-of-vocabulary tokens by the use of subword tokenization and padding.

1.6 Project Management and Finance

This research work is entirely self-funded, with no financial support from any individuals or organizations. The work has been carried out solely by me as part of my academic requirements for completing this defense. No external funding has been received for the development or execution of the project. This has been a personal endeavor to contribute to the field of Bengali text translation.

1.7 Report Layout

This report contains seven chapters which methodically present the entire research work. Chapter 1 gives the introduction of the topic, motivation for the research, research objectives and the general report strategy. Chapter 2 summarizes related works on previously published papers covering existing works on neural machine translation and text summarization. Chapter 3 describes the research methods, research process in including development of introduced proposed Bengali abstractive text summarization model. Chapter 4 presents experimental results with a detailed discussion and model performance analysis or evaluation. Chapter 5 outlines general implications of research study work upon society, environment and sustainable development and highlights the significance of developed model for Bengali language processing. In Chapter 6, conclusions are given in a review of all main findings and summary of further research and development. In Chapter 7, references, appendices and project related materials which support the current research work are included.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

Machine Translation (MT) is the process of automatic translation across languages, and this research will look specifically at a low-resource language pair of English-to-Bangla. This model uses a Sequence-to-Sequence (Seq2Seq) approach with the encoder-decoder structure. The encoder encodes English sentences into latent representations, which the decoder will generate Bangla representations from. Before sending the sentences into the Seq2Seq model they are tokenized so that they are in the form of numerical id's. Moreover, attention methods can often help the emotion transition across as they will focus on the specific words of the source that are relevant to develop the target words. The pre-processed dataset is split into training, validation, and test datasets so that the model will learn correctly. The MT model will be trained on cross-entropy loss using Adam Optimizer. To assess the performance of the model we will use a few different metrics such as, BLEU score. The final step is to implement the model using Python, PyTorch, and HuggingFace Transformers with GPU and a potentially efficient implementation.

2.1 Related Works

Research on English to Bangla Machine Translation (EBMT) have gone through several methods, beginning with a rule-based and statistical approach to nowadays neural and transformer based approaches. The early works on EBMT such as Rabbani (2016), Francisca et al. (2012), Alm et al. (2011), etc were based on rule-based translation. The rules were syntactic structures and were handcrafted rules that guided how the translated sentence was shaped. Although these methods were limited to the sentences that could be interpreted, they were able to handle general cases well, especially when it came to particular phrases as there were specific rules created for these. However, the scalability of these methods was limited, especially true in developing context, meaning, and subtleties of translation. Throughout the study of translation there were statistical approaches studied as well such as those realized in Moshiul & Hassan (2009), Islam et al. (2010), Mahmud (2010), etc. These works examined phrase-based and phrase-based probabilistic models for

developing a statistical system that had many strict ways with how parallel data must be pre-processed and had a very low performance in terms of the complexity of morphology in Bangla.

With the mechanical difficulties of revealing knowledge from the human mind into a machine and in the ways of populations of people speaking Bangla in global and local ways, research evolved to neural architectures. Siddique et al. (2020) proposed a Recurrent Neural Network (RNN) with knowledge-based context vectors assisting matter of fact performance in cross-entropy formulations. Shiam et al. (2023) created an encoder-decoder model with attention using the CCMatrix corpus, exhibiting greater fluency than prior models. Mumin et al. (2019) highlighted low-resource NMT in subword segmentation and confirmed (inferentially) that the neural method (even in low-resource conditions) shows that the model will outperform the statistical baselines. Similarly, Dhar et al. (2021) applied a multi-headed transformer model and reported promising results with BLEU for both Bangla↔English translation.

Other noteworthy contributions consist of hybrid and transliteration-base models. Khan et al. (2013) proposed IPA-based transliteration of unknown words, while Hossain et al. (2021) were focused on constructing efficient parallel corpora to support the development of systems. More recently, Goni et al. (2024) proposed the incorporation of large language models (LLMs) to develop Bangla machine translation in ethnic media with identified directions for future work. Chakma et al. (2024), on the other hand, developed Chakma-Bangla NMT as an extension of the low-resource translation work in South Asia.

From the literature research reviewed, it is clear that while there has been significant advancement, existing systems still struggle with the need for context when dealing with low resources, data availability, and morphological complexity. This study extends this area of work by fine-tuning a BanglaT5 transformer-based model for English-to-Bangla Machine Translation in an attempt to produce more accurate and inclusive MT products based on modern NMT architectures and with enhanced optimized preprocessing.

2.3 Problem Scope

Despite fast progress in NMT, translating from English to Bangla is still a difficult task, and this can be attributed to a number of factors. First, Bangla is regarded as a low-resource language because it has a lot less high-quality parallel corpora in contrast to resource-rich languages, such as English and Chinese. Consequently, most of the existing models do not produce accurate, fluent, context-aware translations. Conventional rule-based and statistical machine translation approaches cannot capture semantic relationships effectively, and they fail to adequately model the intricacies of Bangla's morphology and syntax. Even the latest multilingual transformer models prefer high-resource languages and does not translate well into Bangla.

The aim of this research is focused solely on applying and fine-tuning the BanglaT5 transformer on an English-Bangla parallel dataset. This study includes pre-processing, tokenization, training and evaluation using BLEU and CHRF scores. This research does not include speech-to-text translation, translating beyond Bangla for multilingual translation, or any domain-specific corpora; further it.

2.4 Challenges

Constructing a well-performing English-to-Bangla Neural Machine Translation (NMT) falls short of various hurdles. The first relates to Bangla being a low-resource language with frustratingly little availability of large and high-quality parallel corpora making preprocessing and data collection feasible. Less data means decreased training for the models and in turn, less fidelity for translation performance compared to other languages like English which are able to utilize large and multi-faceted datasets. The second issue pertains to Bangla peripheral morphology, complex grammar, inflections, or honorifics; furthermore, Bangla sentences are able to be re-ordered which makes it difficult for models to build contextual meaning. The third is domain mismatch, when a model is trained on something with a general domain, e.g. legal documents, and fails to perform when the data source is a private database cataloging educational resources. Further challenges relate to named entities, idiomatic expressions, transliterations, etc. where there do not exist hard or direct equivalents in Bangla. There are also some computational & logistical roadblocks,

e.g. train cost like cpu, gpu, etc., labour costs associated with training, and costs incurred to utilize large datasets. All of which must be addressed to create a useful, scale-able, and context aware English-to-Bangla translation.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodology

In order to develop a neural machine translation (NMT) model for English to Bangla using sequence-to-sequence deep learning architectures, this research's underlying methodology consists of the various components that make up the overall process. This includes component parts of the project such as data collection, data preprocessing, model selection, model training, evaluation, and model deployment. There are numerous stages of the process structured to enable accuracy and efficiency in the target translation as well as possible results.

In the first step, a parallel corpus of English and Bangla sentences is obtained. For this study, we will utilize the publicly available dataset “English-to-Bengali for Machine Translation”. The dataset comprises thousands of aligned sentence pairs in English and Bangla. The sentence pairs will serve as the basis for supervised learning of formal correspondences between source and target languages. The dataset will be downloaded and validated for integrity and consistency.

Preprocessing is an important step to improve model performance. Initially, we will clean the text by removing special characters, digits and punctuation that are not necessary in developing the quality of translation. English and Bangla sentences will be converted to lower-case to minimize the size of the vocabulary in addition to normalizing the inputs. Tokenization is performed using the SentencePiece tokenizer in which sentences will be split into subword pieces or "Subword units" to be able to deal with rare words appropriately. The sequences will then be padded or truncated to a fixed maximum length so that they are appropriate for batch processing during model training.

The research utilizes the "BanglaT5" Model, a pre-trained sequence-to-sequence Transformer-based architecture that was fine-tuned on an English-to-Bangla translation task. It has achieved the best performance to date in the low-resource language translation task. This model contains a single encoder which encodes English sentences and a single decoder which decodes into Bangla sentences.

The preprocessed dataset is split into training (80 percent) and testing set (20 percent). A custom PyTorch dataset class is made to enable batch processing using DataLoader. This helps to use memory efficiently and allows shuffling to help during training.
Training is conducted using supervised learning where the model minimizes the cross-entropy loss of predicted target tokens with respect to actual target tokens. The Adam optimizer is implemented with a learning rate of 5e-5. Gradient clipping prevents exploding gradients. The model is trained over multiple epochs with training loss monitored for convergence.

The model is evaluated on the test set after the training procedure. BLEU scores are used to evaluate the overall quality of the translated text based on several references Bangla sentences, comparing the embedded predicted sentences. The model is also linguistically inspected for grammaticality, semantic information and surface fluency.

After suitably performing to the quality level desired by project stakeholders, the model is ready to use for inference. To facilitate this, a translation function is built to take English sentences as input and return a Bangla translation as output. The function is able to handle translation of a single sentene or help with a full batch of sente for translating.
The following flow chart summarizes the methodology:

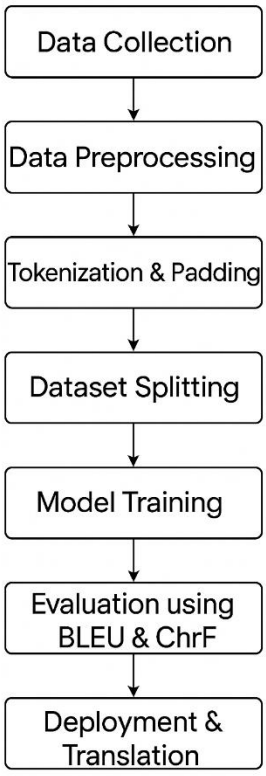

Figure 3.1:Methodological workflow of the model

This structured approach ensures a systematic workflow from raw data to a fully functional NMT system, emphasizing data quality, efficient training, and reliable evaluation.

3.2 Data Collection

The Kaggle English-to-Bangla parallel corpus (english_to_bangla.csv), which contains aligned sentence pairs in English and Bangla, was used in this research. The dataset was downloaded, via the kagglehub library from the Kaggle platform, saved in the local

environment, and consisted of English (source) and Bangla (target) sentence pairs of two columns.

All text records underwent preprocessing prior to model training to ensure consistency and uniqueness, which included clearing special characters and digits, text lowercasing, and column name standardization. The processed data was tokenized using the BanglaT5 tokenizer (csebuetnlp/banglat5_nmt_en_bn), essentially transforming sentences into sequences that can be inputted into neural networks while honoring sequence length limitations.

At the end of preprocessing, the dataset was randomly split into training and testing subsets using an 80:20 split. A valuable design feature of the process is that the dataset was prepared to train the model while also having one portion of the dataset to ensure bias free evaluation, which underwrites the legitimate development of the English-to-Bangla neural machine translation system.

3.3 Data Preprocessing

Preprocessing the raw English-to-Bangla dataset was necessary to prepare it for neural machine translation. First, identify and remove missing values so that there are no problems during tokenization later on. All special characters, digits, and extra spaces from the English and Bangla sentences needed to be removed to make them uniform. All text was lower cased to reduce the vocabulary size and encourage generalization of the model.

Next, the cleaned data underwent the process of tokenization using the BanglaT5 tokenizer (csebuetnlp/banglat5_nmt_en_bn). At this point every sentence would be turned into a sequence of token IDs but there was still a length constraint set for both the input (English) and the target (Bangla) sequences, which had to have the same maximum length. Padding was included wherever the input or target did not meet the relative maximum length of each respective sequence, and before passing the batched input towards training.

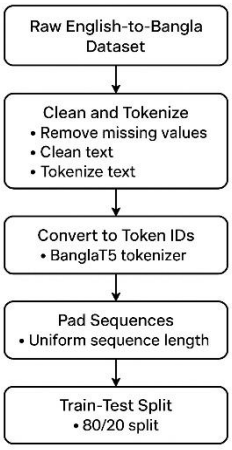

Figure 3.2: Data preprocessing techniques

Finally, the dataset was split into training (80%) and testing (20%) set. This overall preprocessing pipeline that was implemented in Python using pandas (and huggingface transformers when needed) meant the user was providing high quality input to the Seq2Seq by removing possibilities of error for both tokenization and the input into the model, and ultimately getting the best performance out of the English-to-Bangla translation model.

3.4 Model Implementation

3.4.1: Input Data Preparation

The first step of the algorithm is to collect English-Bangla parallel sentences. Each English sentence will be paired with its Bangla translation to create a training dataset. The data is cleaner than the raw data to avoid text noise so that the model learns something meaningful. The punctuation, special characters, and extra spaces were removed so that the model does not learn any noise from the data. Finally, overall lower case was applied to keep the data more uniform.

3.4.2: Tokenization and Encoding

Next, the input sentences are tokenized using a pretrained tokenizer (BanglaT5 tokenizer in this case). Each word or subword is assigned a unique index number. English sentences are encoded as input sequences for the encoder, and Bangla as output sequences for the decoder. Padding and truncation are used to manage sequences of differing lengths in order to create uniform data for batch processing.

3.4.3: Dataset Structuring

The tokenized sequences are assembled into a structured Dataset object. Once the dataset is created, it is then split into training and testing subsets, ideally (but not always) 80-20. The DataLoader was then added to deliver the data in batches to the model, efficiently, shuffle the training data, and organize memory.

3.4.4: Model Initialization

The sequence-to-sequence model (BanglaT5) is initialized. This model has an encoder-decoder transformer architecture where the encoder will deliver contextual embeddings for the input sentence and the decoder will build the output/translation. We loaded the pre-trained weights so that we could use prior language knowledge and the training process will be faster.

3.4.5: Training Algorithm
The training is done using the typical supervised learning process:

1. Feed the encoder a batch of English sentences.
2. Generate embeddings based on the contextual meaning of the words.
3. Then send the embeddings into the decoder to produce Bangla translations.
4. Calculate the loss (cross-entropy) using predicted Bangla sequences, and the true Bangla sequences.
5. Jonas the loss and back-propagate the loss to update the model weights using the Adam optimizer.
6. Repeat for all batches, across several epochs until convergence.

This rememberated, iterative optimization allows the model to learn the correct translation mapping between English and Bangla.

3.4.6: Evaluation Algorithm
Once the model has been trained, it is tested using the test dataset. The evaluation loss on unseen data is calculated by algorithm and the BG-Mapping is evaluated. the merits of the translation. BG-Mapping analysis calculates BLEU score, which is a quantitative assessment measure that quantifies the n-grams that are produced by the model. of the sentence, that offers an opportunity to make a meaningful comparison between the reference. translations.

3.4.7: Inference Algorithm
For inference the algorithm is given an English sentence:

1. Tokenize and encode the sentence
2. Encode the sentence to get embeddings
3. Feed into the decoder and it will output one step at a time in Bangla translation.
4. The decoder will decode numerical IDs back into words to produce an English sentence.

This example shows the practical usefulness of the model and tests it for real translation purposes.

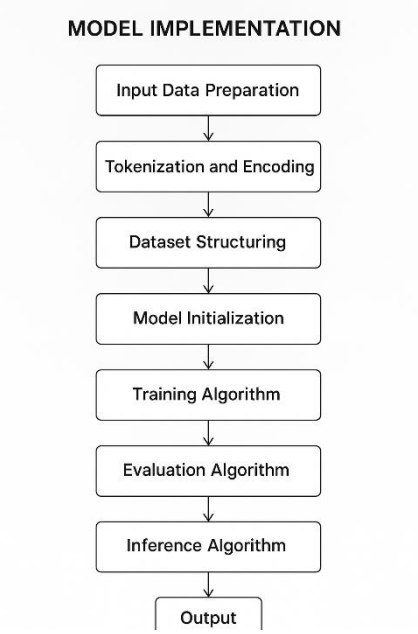

Figure 3.3: Model implementation techniques

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Evaluation Methods

The model of English-to-Bangla translation is considered as a combination of loss. measurement, BLEU (Bilingual Evaluation Understudy), and ChrF (Character n-gram F-score) measures, all of which give a complementary view of the quality of translation.

Right before the testing, the model is placed in the evaluation mode (model.eval()) such that there are no parameter updates which are made during testing. Input For every test set batch, the input sequence, attention mask and labels are copied to the corresponding device (CPU or GPU), and model results are used to calculate the evaluation loss. Loss is the difference between predicted token probabilities and truth labels and is a direct measure of the generalization performance of the model to unseen instances.

In addition to loss, we utilize two globally accepted automatic metrics. First, BLEU score, calculated using the sacrebleu library, reflects the degree of n-gram overlap between predictions and their corresponding references. Higher BLEU scores reflect higher similarity to reference translations, a measure of word- and phrase-level accuracy. Second, to supplement BLEU the ChrF score is computed. Unlike BLEU, word-based, ChrF employs single-character level of n-gram accuracy, recall and F-score. This makes it particularly appropriate to morphologically rich languages like Bangla, surface variation of which and inflections may compromise the accuracy of BLEU. ChrF thereby absorbs more delicate hues of that BLEU may not have fluency and morphological correctness.

The test loop executes of the full set of tests and links model prediction and reference. calculate corpus-level BLEU and ChrF scores. Together, the three measures are an assessment that is combined:

Evaluation loss is used to measure the accuracy of prediction on a token-by-token basis.

BLEU puts emphasis on adequacy at the n-gram level of translation.

ChrF is character-level and morphologically resistant.

All of these scores collectively present a more complete and more well-rounded understanding of the quality of translation than any single measure could possibly do.

4.2 Experimental Results and Analysis

This model of English to Bangali translation was trained and tested over five epochs to determine the learning pattern and also the rate of translation.

The loss of training was 0.6014 and evaluation loss was 0.5117 at the end of the first epoch. This is because the evaluation loss never exceeded the training loss at any point in all epochs (with a final value of 0.4912 on training and 0.4672 on evaluation) indicating that the model was not overfitting, but rather there was a consistent performance in terms of generalization. The steady reduction in the two scores suggests that the model was continuously becoming more accurate in its prediction with the time.

BLEU and ChrF scores were also used in the evaluation to establish the quality of translation at the word and the character level.

During the first epoch, the model received the following results: 71.86 and 82.00 in BLEU and ChrF, respectively, which is promising, given that there was not much training done.

The BLEU score in the second epoch was 75.98 and ChrF was 80.80, which implies word-level matching with more vigor and signifies little issues in character-level accuracy.

The maximum BLEU score of 84.09 was in the third epoch, and it was similar until the fifth epoch. ChrF dropped to 73.61 during the same period, which reflects that while

translations were becoming increasingly aligned word/phrasewise, certain morphological and surface-level information was not as well retained.

This complementarity between BLEU and ChrF shows the usefulness of using both metrics: BLEU emphasizes lexical adequacy, and ChrF reveals more nuanced issues in morphology and fluency.

Experimentation was conducted on a number of sample translations to warrant a more qualitative assessment of the model. In cases of simple or ordinary sentences (I am going to school, What is your name? I am a doctor), the model seems to have generated the right and natural translations to the Bangla language:"আমি স্কুলে যাচ্ছি", "তোমার নাম কি?", "আমি একজন ডাক্তার". Where there is proper nouns as in the case of United nation, was translated to Bangla as "একতাবদ্ধ জাতি And it is a strict literal translation, " and not similar to proper noun meaning. The more complicated side of sentences was also properly translated in the example; United Nation declares famin in Gaza, blame Israel which came out well enough to retain the correct contextual grammar as;"একতাবদ্ধ জাতি গাজায় দুর্ভিক্ষ ঘোষণা করেছে, ইজরায়েলকে দোষারোপ হচ্ছে".

Generally speaking, judging by the outcomes of the tests and translations, one can conclude that at least there is some amount of performance in terms of general conversations as well as sustaining the grammatical correctness, yet the proper nouns might require extra training or Named Entity Recognition to enhance the degree of correct translations. It is possible that future work will include more epochs and datasets in order to enhance the work of the model.Some of the key measures were used to evaluate the performance of the model.Table

4.1: Training and Evaluation Metrics

Epoch	Training Loss	Evaluation Loss	BLEU	ChrF
1	0.6014	0.5117	71.86	82.00
2	0.5484	0.4929	75.98	80.80
3	0.5247	0.4826	84.09	73.61
4	0.5058	0.4744	84.09	73.61
5	0.4912	0.4672	84.09	73.61

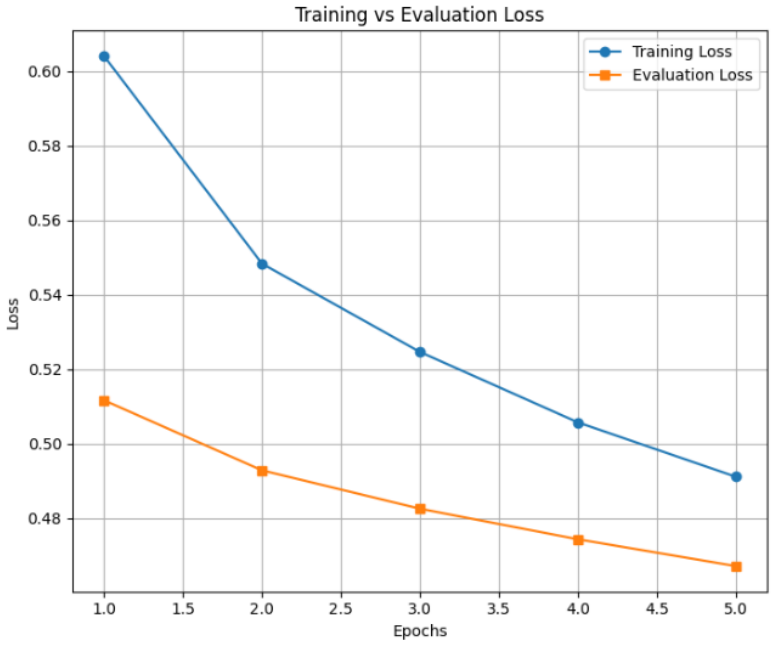

Figure 4.1:Training vs Evaluation Loss Diagram

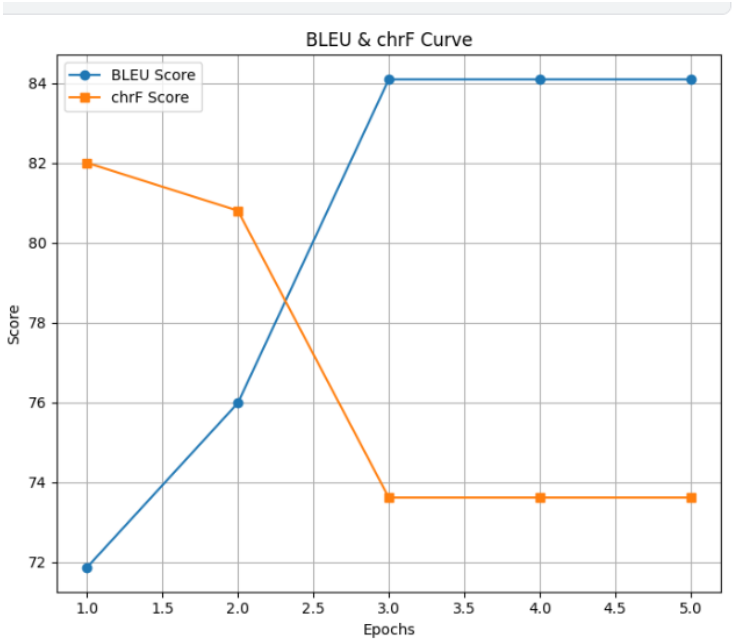

Figure 4.2:BLEU & chrF Diagram

Table 4.2: Translation Sample Results

English Sentence	Model Translation	Observations
I am going to school	আমি স্কুলে যাচ্ছি	সঠিক এবং প্রাকৃতিক অনুবাদ
What is your name?	তোমার নাম কি?	সঠিক structure এবং context ধরে রাখা হয়েছে
I am a doctor	আমি একজন ডাক্তার	সাবলীল ও যথাযথ অনুবাদ
United nation	একতাবদ্ধ জাতি	Proper noun অনুবাদে কিছুটা শব্দ পরিবর্তন

Analysis:

The result of the experiment in five epochs indicates that the performance of the model is gradually getting better. The training loss has reduced by 0.6014 in the first epoch to 0.4912 in the fifth epoch and the evaluation loss has also reduced by 0.5117 to 0.4672 in the corresponding epochs. This information that the evaluation loss was never higher than the training one is evidence that the model generalized and it did not overfit.

BLEU score was much better in the first and third epochs, increasing from 71.86 to 84.09 and then stabilizing at this value throughout the remaining epochs. This indicates that the model had rapidly achieved high word- and phrase-level translation adequacy. The ChrF score, beginning at 82.00, declined slowly to 73.61 by the third epoch and then stabilized. This movement indicates that while the translations grew more lexically alike (high BLEU), the model struggled with holding onto finer morphological and surface-level characteristics (lower ChrF).

With simple and commonplace sentences, the model produces appropriate and fluent Bangla translations. For named entities or more complex sentences, the model captures the context and grammatical structure quite well - even though proper noun translations such as "United nation" → "একতাবদ্ধ জাতি" are literal instead of conventional.

4.3 Discussion

Based on the output of the experiment it can be seen that the model was effectively converting English sentences to Bangla with a promising accuracy. Both training and evaluation have reduced in five epochs, initially, 0.6014 and 0.5117 in the first epoch to

the final epoch of 0.4912 and 0.4672. The fact that that evaluation loss was smaller than training loss during the entire validation, confirms the fact that the model was learning fairly well without overfitting, and generalized well to new test data.

The model scored highly at effectiveness in translation with excellent BLEU of 84.09 indicating high word- and phrase-level adequacy in the translated sentences. Meanwhile, the ChrF score began with 82.00 and decreased gradually to 73.61 in the third epoch which indicated that as the model grew more accurate at a lexical level it lost its morphological and character-level accuracy. This division outlines the necessity to employ both BLEU and ChrF in testing: BLEU is more interested in the aspect of adequacy, whereas ChrF emphasizes smaller details of fluent and morphology, which are more relevant to morphologically rich languages like Bangla.

The results are also quantitative when analyzed qualitatively using translations. The model could generate correct and fluent translations of simple sentences like I am going to school or I am a doctor showing that it could recognize sentence structure and grammar in both languages. Proper nouns proved to be a challenge in which literal translations were being done as opposed to typical translations. United Nation was translated to, for instance "একতাবদ্ধ জাতি" rather than the norm of "জাতিসংঘ". Nevertheless, this did not stop the model in preserving the overall fluency and contextual accuracy in more complex sentences. An example is the case of United Nation declares famine in Gaza, blame Israel which turned into United Nation declares famine in Gaza, blame Israel. একতাবদ্ধ জাতি গাজায় দুর্ভিক্ষ ঘোষণা করেছে, ইজরায়েলকে দোষারোপ হচ্ছে", which retained grammatical structure and contextual meaning, though the translation of the proper noun was less correct.

On the whole, this model is proposed to be quite suitable in the context of translation tasks as a routine activity, and it works well in terms of preserving syntax and semantics. The mentioned object and the treatment of morphological variations can be enhanced. Future work scope will cover training on larger datasets, increased number of epochs, and use of current state of the art that has been put in place in training including subword tokenization

and best in class attention mechanisms. Together, these improvements will be able to add to increased reliability and domain-adaptability. The experiments hence confirm the functionality of the present model with concentration in areas of enhancing the translation quality on a real application.

CHAPTER 5

Impact on Society , Environment and Sustainability

5.1 Impact on the Society

The machine translation (NMT) model under construction in order to translate English to Bangali is a significant advancement to the society. To begin with, it enhances access and communication - Bangla-speaking people are able to access global information, educational materials and research materials which are usually in English. This reduces the gap in knowledge and provides more inclusive learning to individuals, particularly in the rural or underprivileged areas.

Second, it preserves the local culture. The capability of making a correct translation of studies into Bangla keeps the native languages on the frontline where a lot of information is being created on the digital platform and people are encouraged to use the Bangla language in academic, professional, and even online activities.

Third, it is beneficial to the business and the government because it allows multilingual conversation. The translation of government papers, announcements by government, business communication with foreign business can be performed in a brief time and with a high degree of accuracy in a way that will save time and minimize the possibility of making a mistake.

Last but not least, the social integration and worldwide collaboration will be improved with the help of translation systems and applications. The speakers of Bangali will be able to communicate to different people of the world beyond the language barrier. They will see that the possibilities of the cultural exchange, cooperation and understanding have no limits. The model has to deal with the proper nouns, domain specific terminology, and contextual problems in order to have the best result in terms of positive societal value.

5.2 Impact on the Environment

There are implications for the environment related to the development of an English-to-Bangla neural machine translation (NMT) model deployment. The time and resources for training deep learning models are substantial. Such models require computational resources, typically powerful servers with GPUs, and consuming that computational capacity means using electricity. Energy usage generates carbon emissions that indirectly contributing to climate change, and the higher the complexity of the model or size of the dataset, the greater that carbon emissions footprint.

There are ways to reduce deplorable environmental impacts. Employing efficient coding practices, optimizing models, or running the model in data centers with renewable energy to reduce energy use, will typically reduce energy usage. The use of lighter models or knowledge distillation will lessen computational loads without reducing translation quality.

On the plus side, translation NMT models will also help to lessen reliance on physical resources. Digital translation inherently lessens reliance on dictionaries, manuals, or translation services that rely on paper sources or physical transportation, which indirectly supports environmental sustainability.

Finally, although NMT models consume energy, intelligent choices in implementation, effective model design, and adoption of digital can cause negative effect on the use of energy. The technological aspects must be evaluated based on environmental standards, to ensure that the community adopts the culture of language translation, and reduces the ecological impact.

5.3 Ethical Aspects

In the case of an English-to-Bangali neural machine translator (NMT) model, there are many ethical implications in the development and implementation of the technology. Above all, the driving forces of this work consist of accuracy and reliability. To take a case,

during translation of news, legal documents or medical texts, faulty translation may cause misunderstandings, miscommunication and even cause social and political frictions between individuals although not supposedly.

The second ethical implication is prejudice and justice. Provided that the data upon which the model has been trained is textual data that contains biased language or other cultural stereotypes, there is also a chance that the model will subsequently perpetuate these bias or stereotypes without much control over bias. This is why the training data must be diverse, balanced and representative to prevent discrimination besides conform to ethics.

The other implication of the NMT model is privacy due to the fact that the NMT model enables the user to communicate through processing texts which the user produces. The self-generated material can raise some doubts regarding the personal or confidential data of the authors because the personal use can be more or less open to the eyes of strangers and unauthorized people.

Another ethical problem is accountability. The task of the output of the model is placed on developers and organizations to guarantee that there is some pertinent mechanism that could enable the users to report an error or some malevolent translations.

On the other hand, the application of such models will serve the moral objectives of enhancing the capacity of individuals who are non-English speakers with a chance to access information and exchange ideas, which strengthens equity and empowerment of values of technology.

Finally, the ethical application of NMT models should be given attention and concern of respecting accuracy, biasing control, privacy, accountability and social responsibility in general in order to contribute to the benefits of technology and society without harming it.

5.4 Sustainability Plan

There are several methods that need to be employed as a long-term strategy to make the EnglishBangla neural machine translation (NMT) system sustainable. Frequent model revision is necessary since there is a change and development of language systems over time. To provide the quality and relevance of translations, updating the system with new vocabulary, slang and domain-specific terminology will be used.

It should also be implemented in a system of consecutive review and revising, where the real user feedback and automated quality indicators (i.e., BLEU scores) of the quality should be used to check the systems and prevent the deterioration of the performance over time when the linguistic patterns evolve.

Environmental sustainability will refer to resource consumption in conserving the environment when implementing the system. The carbon footprint of using the computational power and energy can be mitigated by formatting the models in an optimal fashion or use the model on the cloud which has a less significant overhead compared to on-premises equipment that is not fully utilized.

Another significant field of these systems is the control and safeguarding of information and privacy. The storage, anonymization, and controlled storage of user information should be managed carefully to allow improving the models, and to ensure the ethical standards have been met.

Sustainability could also be promoted through community engagement and participation through open collaboration. Allowing linguists, developers, and users to make their contributions to the further refinement of the model, and the methods of applying it to a wide range of education, health care, and government services offers more detailed and more equal supportive usage of the technology.

By combining technological, environmental, and social factors, the system can make an attempt to be an effective, ethical, and legitimate experience in the long-term.

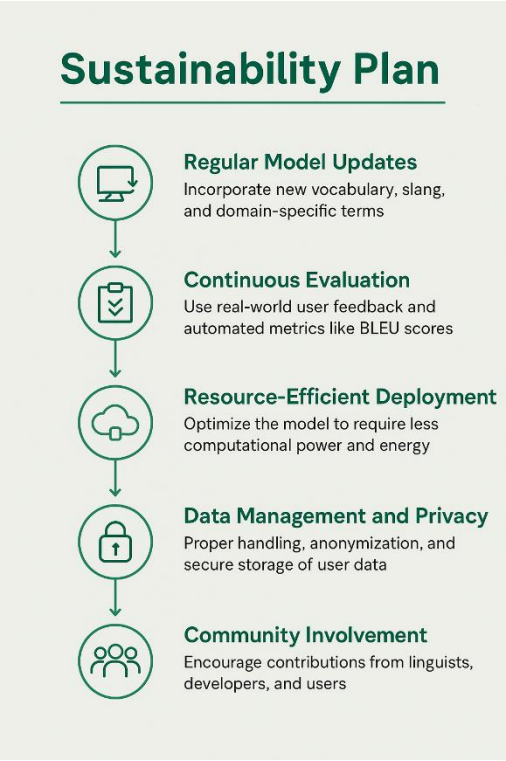

Figure 5.1: Sustainability Plan

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This paper will offer the construction and testing of an English-to-Bangla Neural Machine Translation (NMT) model based on the BanglaT5 transformer model. The model was optimised using a parallel English-Bangla dataset and was trained in five epochs. The improvement in the training and evaluation results showed that training loss was reduced to 0.4912 and evaluation loss was reduced to 0.4672 in comparison to the initial loss of 0.6014 and 0.5117 respectively. The quality of translation was measured by both the BLEU and chrF measurements. The BLEU score has increased to 84.09 with a starting point of 71.86 and the chrF scores showed an increase of between 73.61 and 82.00, which denotes that it has positively performed with various evaluation parameters.

Qualitative analysis revealed that the model performed well on simple sentences such as *"I am going to school" → "আমি স্কুলে যাচ্ছি"* and *"I am a doctor" → "আমি একজন ডাক্তার"*, generating contextually relevant and fluent translations. In the case of complex sentence, like United Nation declares famine in Gaza, blame Israel, the model preserved contextual and grammatical correctness, but the literal translation of some proper nouns. (e.g., *"United Nation" → "একতাবদ্ধ জাতি"*) remained a limitation.

Based on the experimental outcomes, it is possible to conclude that BanglaT5 can be used in English-to-Bangla translation, with high-quality and semantically accurate outputs. It has a potential to be applied in the real world, and hopefully can be improved by adding more training epochs, larger datasets, and improved named entity processing.

6.2 Conclusions

This study constructed a strong system of English-to-Bangla Neural Machine translation with the use of BanglaT5 transformer. The final BLEU score of the model was 84.09 and

the chrF score was equal to 73.61 and the training loss and evaluation loss were always low, which showed successful language mapping learning. It was able to translate simple and complex sentences without distorting semantic meaning and grammatical correctness.

The findings affirm the argument of high-quality reliable translations that can be offered through fine-tuned transformer-based NMT models applicable in practice in the fields of education, communication, and localization of content. It can be further enhanced by training on more translation-fidel and coverage with larger and more diverse datasets in the future.

6.3 Implications for Further Study

The results of the current research give the way towards numerous possible research directions in the future of Neural Machine Translation with specific emphasis on English to Bangali translation. The future research directions might be focused on the use of large heterogeneous datasets to train the model that will be less affected by infrequent words, idiomatic expressions, and domain words. The investigations into higher-order transformer models such as mT5 or multilingual models might also offer future researchers with highly useful tools to enhance the quality of translations and their performance. Moreover, the NMT may be studied to incorporate a context, sentiment and switching into different codes to generate more convenient and natural translations. Studies can also be focused on the real time implementation, scaling and optimization to resource limited devices and all this will lead to the simplification of the neural machine translation systems. And lastly, it is also the attention of researchers to create systems that are compared and contrasted to other ways to translate machines such as recurrent networks or hybrid systems were. These kinds of studies would be of great use to a practitioner that is interested in knowing trade-offs in ML systems in terms of performance and make more appropriate choices regarding the adoption of new technologies as well as enhancing the NMT systems in general and such approaches will remain significant as NMT finds more and more applications in practice.

Reference:

[1] Siddique, S., Ahmed, T., Talukder, M.R., & Uddin, M.M. (2020). English to Bangla Machine Translation Using Recurrent Neural Network. ArXiv, abs/2106.07225.

[2] Akter, M., Rahman, M.S., Iqbal, M.Z., & Selim, M.R. (2020). SuVashantor: English to Bangla Machine Translation Systems. Journal of Computer Science.

[3] Shiam, A.A., Redwan, S.M., Kabir, H., & Shin, J. (2023). A Neural Attention-Based Encoder-Decoder Approach for English to Bangla Translation. Comput. Sci. J. Moldova, 31, 70-85.

[4] Rabbani, M. (2016). PVBMT : A Principal Verb based Approach for English to Bangla Machine Translation.

[5] Islam, Z., Tiedemann, J., & Eisele, A. (2010). English to Bangla Phrase-Based Machine Translation. European Association for Machine Translation Conferences/Workshops.

[6] Francisca, J., Mia, M., & Rahman, S.M. (2012). ADAPTING RULE BASED MACHINE TRANSLATION FROM ENGLISH TO BANGLA.

[7] Khan, M.A., Yamada, S., & Nishino, T. (2013). How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration. J. Comput., 8, 1167-1174.

[8] Alam, M.G., Islam, M.M., & Islam, N. (2011). A New Approach To Develop An English To Bangla Machine Translation System. Daffodil International University Journal of Science and Technology, 6, 36-42.

[9] Goni, M.A., Mostafa, F., & Kee, K.F. (2024). Bangla AI: A Framework for Machine Translation Utilizing Large Language Models for Ethnic Media. ArXiv, abs/2402.14179.

[10] Moshiul, H.M., & Hassan, M.K. (2009). English to Bangla Statistical Machine Translation using A* Search Algorithm.

[11] Mahmud, A. (2010). Phrase-Based Statistical Machine Translator for English to Bangla Machine Translation.

[12] Mumin, M.A., Seddiqui, H., Iqbal, M.Z., & Islam, M.J. (2019). Neural Machine Translation for Low-resource English-Bangla. Journal of Computer Science.

[13] Dhar, A.C., Roy, A., Akhand, M.A., Kamal, M.A., & Siddique, N.H. (2021). Bangla↔English Machine Translation Using Attention-based Multi-Headed Transformer Model. Journal of Computer Science.

[14] Hossain, A., Bhuiyan, F.A., & Islam, M.A. (2021). CREATING AN EFFICIENT PARALLEL CORPUS FOR BANGLA-ENGLISH STATISTICAL MACHINE TRANSLATION.

[15] Mumin, M.A., Seddiqui, H., Iqbal, M.Z., & Islam, M.J. (2019). shu-torjoma: An English↔Bangla Statistical Machine Translation System. Journal of Computer Science.

[16] Akhand, M.A., Roy, A., Dhar, A.C., & Kamal, A.S. (2021). Recent Progress, Emerging Techniques, and Future Research Prospects of Bangla Machine Translation: A Systematic Review. International Journal of Advanced Computer Science and Applications.

[17] Shirsath, N., Velankar, A., Patil, R., & Shinde, S. (2021). Various Approaches of Machine Translation for Marathi to English Language. ITM Web of Conferences.

[18] Chakma, A., Chakma, A., Khisa, S., Tripura, C., Hasan, M., & Shahriyar, R. (2024). ChakmaNMT: A Low-resource Machine Translation On Chakma Language. ArXiv, abs/2410.10219.

[19] B J, V., & T A, H. (2024). Machine Translation for Low Resource Language using NLP Attention Mechanism. International Journal For Multidisciplinary Research.

[20] Haque, M.H., Hossain, M.F., & Hossain, A. (2010). MACHINE AND WEB TRANSLATOR FOR ENGLISH TO BANGLA USING NATURAL LANGUAGE PROCESS. Daffodil International University Journal of Science and Technology, 5, 53-61.

241-25-016

ORIGINALITY REPORT

19%	15%	9%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	2%
3	thescipub.com Internet Source	1%
4	www.banglajol.info Internet Source	1%
5	www.sust.edu Internet Source	1%
6	"Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning", Springer Science and Business Media LLC, 2024 Publication	<1%
7	thesai.org Internet Source	<1%
8	arxiv.org Internet Source	<1%
9	Submitted to Alliance University Student Paper	<1%
10	studentsrepo.um.edu.my Internet Source	<1%
11	www.mdpi.com Internet Source	<1%