Department of Electrical and Computer Engineering
North South University

Directed Research

“Bengali Text Summarizer:

An advanced summarizing of Bengali news articles”

Md. Tanjeelur Rahman Labib	**201 3677 642**
Tashin Mahmud Khan	**201 1819 042**
Md. Tasin Hossain Toha	**201 1664 042**
Md Saikot Hossain Sojib	**201 4055 642**

Faculty Advisor:

Rifat Ahmed Hassan

Lecturer

ECE Department

Summer, 2024

LETTER OF TRANSMITTAL

January, 2025

To

Dr. Rajesh Palit
Chairman,
Department of Electrical and Computer Engineering
North South University, Dhaka

Subject: **Submission of Directed Research Report on "Bengali Text Summarizer: An advanced summarizing of Bengali news articles"**

Dear Sir,
With due respect, we would like to submit our **Directed Research Report** on **"Bengali Text Summarizer: An Advanced Summarizing of Bengali News Articles"** as a part of our BSc program. The report deals with Bengali Text Summarizer which is an advanced summarizing of Bengali news articles. This project was beneficial to us in gaining experience in the practical field and applying it in real life. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report helpful and informative to have an apparent perspective.

Sincerely Yours,

....

Md. Tanjeelur Rahman Labib
ECE Department
North South University, Bangladesh

....

Tashin Mahmud Khan
ECE Department
North South University, Bangladesh

....

Md. Tasin Hossain Toha
ECE Department
North South University, Bangladesh

....

Md Saikot Hossain Sojib
ECE Department
North South University, Bangladesh

APPROVAL

Md. Tanjeelur Rahman Labib (201 3677 642), Tashin Mahmud Khan (201 1819 042), Md. Tasin Hossain Toha (202 1171 642), and Md Saikot Hossain (201 4055 642) from the Electrical and Computer Engineering Department of North South University have worked on the Directed Research Report titled "**Bengali Text Summarizer: An advanced summarizing of Bengali news articles**" under the supervision of **Rifat Ahmed Hassan** partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

……………………………………….

Rifat Ahmed Hassan
Lecturer
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh.

Chairman's Signature

……………………………………….

Dr. Rajesh Palit
Professor
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere, partially or entirely, for the award of any other degree or diploma. All project-related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been adequately acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students’ names & Signatures

1. Md Tanjeelur Rahman Labib

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Tashin Mahmud Khan

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

3. Md. Tasin Hossain Toha

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

4. Md Saikot Hossain Sojib

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, **Rifat Ahmed Hassan,** Lecturer, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice on the experiments, research, and theoretical studies carried out during the course of the current project and also in the preparation of the report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh, for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

“Bengali Text Summarizer:

An advanced summarizing of Bengali news articles”

Abstract— An efficient summary tool is critically needed given the explosive rise of digital Bangla content, mostly news items. In order to generate succinct and fluid summaries, this study suggests a Bangla text summarizer that makes use of machine learning and natural language processing. After being trained on over 80,000 articles, our sequence-to-sequence model with LSTM units and an attention mechanism performs admirably in handling the grammatical complexity of Bangla. It maintains excellent accuracy and relevance while drastically cutting down on the amount of time needed to process information. Our summarizer demonstrates promising results and provides the groundwork for future developments in text summarization, despite being computationally intensive and dealing with many linguistic nuances.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

Most digital content growth, especially in Bengali, seems to have happened in the genre of news articles. The fact is that so much information is coming that the readers do not know how to bear enough patience to go through the long, complete articles to grasp only the main factors. This problem can be eliminated by using automatic text summarization, whereby a concise yet coherent summary containing the essence of source documents is generated.

Although a lot of NLP tools are being used for a language like English, the same thing has not happened in the case of Bengali, with very few advanced language processing resources present. This work is motivated by a lack in effective tools designed specifically for Bengali text summarization, especially news articles. In this paper, we have tried to fill up this gap with some state-of-the-art transformer models like MT5 and provide a strong, scalable solution for Bengali news summarization.

1.2 Purpose and Goal of the Project

The primary purpose of this project is to develop an advanced Bengali Text Summarizer that can generate concise summaries of Bengali news articles in real time. The specific goals include:

- **Optimizing the MT5 Model for Bengali**: Training the transformer-based MT5 model for the Bengali language, which will eventually be fine-tuned with the current update to give better performance in the summarization of texts.

- **Scalability and Accessibility**: It should be made available to various groups of users: researchers, journalists, general readers.

1.3 Organization of the Report

This report is organized in such a way that an overview of the project is well-presented, logical, and sequential. The following table outlines the organization of the report, focusing on what is covered in each chapter:

Chapter	**Title**	**Description**
Chapter 2	Research Literature Review	Review existing research relevant to the project, identifying limitations and gaps that our work addresses.
Chapter 3	Methodologies	Details the methodologies used in the project, including system design, hardware and software components, and their implementation.
Chapter 4	Investigation/Experiment, Result, Analysis and Discussion	Presents the experiments conducted, results obtained, and provides an in-depth analysis and discussion of these results.
Chapter 5	Conclusion	Summarizes the project’s findings, discusses limitations, and suggests possible future improvements.

TABLE 1.3 REPORT CONTENT ORGANIZATION

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

This chapter reviews literature and existing research related to text summarization in Bangla. The methodology and techniques followed and findings of previous studies are identified in this review along with their analysis to understand various contributions and weaknesses. It forms the background for the project, representing the current domain wherein other improvements are likely to be in demand. This article describes various approaches for text summarization, the differences between an extractive method and an abstractive technique, and discussions of how the methods have so far been implemented for the Bangla language. The following sections summarize some basic works that taught and influenced our approach to implementing an efficient Bangla Text Summarizer.

2.2 Automated Bangla Text Summarization by Sentence Scoring and Ranking

Published In	2013 International Conference on Informatics, Electronics and Vision (ICIEV)
Electronic ISBN	978-1-4799-0400-6
Authors	Md. Iftekharul Alam Efat, Mohammad Ibrahim, Humayun Kayesh
Publication Date	May, 2013
Paper Link	https://www.researchgate.net/publication/261212570_Aut%20omated_Bangla_text_summarization_by_sentence_scoring%20_and_ranking

Abstract:

Text summarization is one of the leading problems of natural language processing and deep learning in recent years. Text summarization contains a condensed short note on a large text document. Our purpose is to create an efficient and effective abstractive Bangla text summarizer what can generate an understandable and meaningful summary from a given Bangla text document. To do this we have collected various texts such as newspaper articles, Facebook posts etc. and to generate summary from those text we will be using our model. Our model works with bi-directional RNNs with LSTM in encoding layer and attention model at decoding layer. Our model works as sequence-to-sequence model to generate summary. There are some challenges we have faced while building this model such as text pre- processing, vocabulary counting, missing words counting, word embedding, unknown words find out and so on. In this model, our main goal was to make an abstractive summarizer and reduce the train loss of that. During our research experiment, we have successfully reduced the train loss to 0.008 and able to generate a fluent short summary note from a given text.

Bibliography:

Islam, M. A., Talukder, M. A. I., Abujar, S., Masum, A. K. M., & Faisal, F. et al. (July 2019). "Abstractive Bangla Text Summarization using Bi-directional RNNs with LSTM and Attention Mechanism." Dept. of CSE, Daffodil International University, Dhaka, Bangladesh.

2.3 Bangla Text Summarization Using Deep Learning

Published In	Not Published
Electronic ISBN	Not Available
Authors	Ahmed Sadman Muhib, Shakleen Ishfar, AKM Nahid Hasan

Publication Date	March, 2021
Paper Link	http://103.82.172.44:8080/xmlui/handle/123456789/1274

Abstract

In this thesis, we present our work regarding text summarization. Text summarization is the technique for generating concise and precise summaries of voluminous texts while focusing on the sections that convey useful information without losing the overall meaning. In this age of information, there are vast quantities of textual data available. Example sources include online documents, articles, news, and user reviews of various products and services. We can present the underlying information present in these texts concisely through summaries. However, generating summaries for such a large source of text documents by hand is troublesome. We can utilize neural machine summarization systems to generate summaries automatically. These systems leverage the power of deep learning models. Recently, with the invention of Transformer architecture, modern summarization systems have achieved revolutionary performance gains. Efficient transformer- based summarization systems exist for English and other popular languages, but not Bangla. In this research, we present an efficient transformer-based text summarization system for the Bangla language. We use subword encoding to eliminate the problem of rare and unknown words. We have created a large dataset, consisting of 600 thousand news articles, to train our model. We trained a 6 million parameter model that is capable of producing accurate summaries. We evaluated out summaries by observing it's generative performance.

2.4 An Approach for Bangla Text Summarization Using Word2Vector

Published In	2019 10th International Conference on Computing, Communication

	and Networking Technologies (ICCCNT)
Electronic ISBN	978-1-5386-5906-9
Authors	Sheikh Abujar, Abu Kaisar Mohammad Masum, Syed Akhter Hossain
Publication Date	July 2019
Paper Link	https://www.researchgate.net/publication/338358097_An%20_Approach_for_Bangla_Text_Summarization_using_Word2%20Vector

Abstract

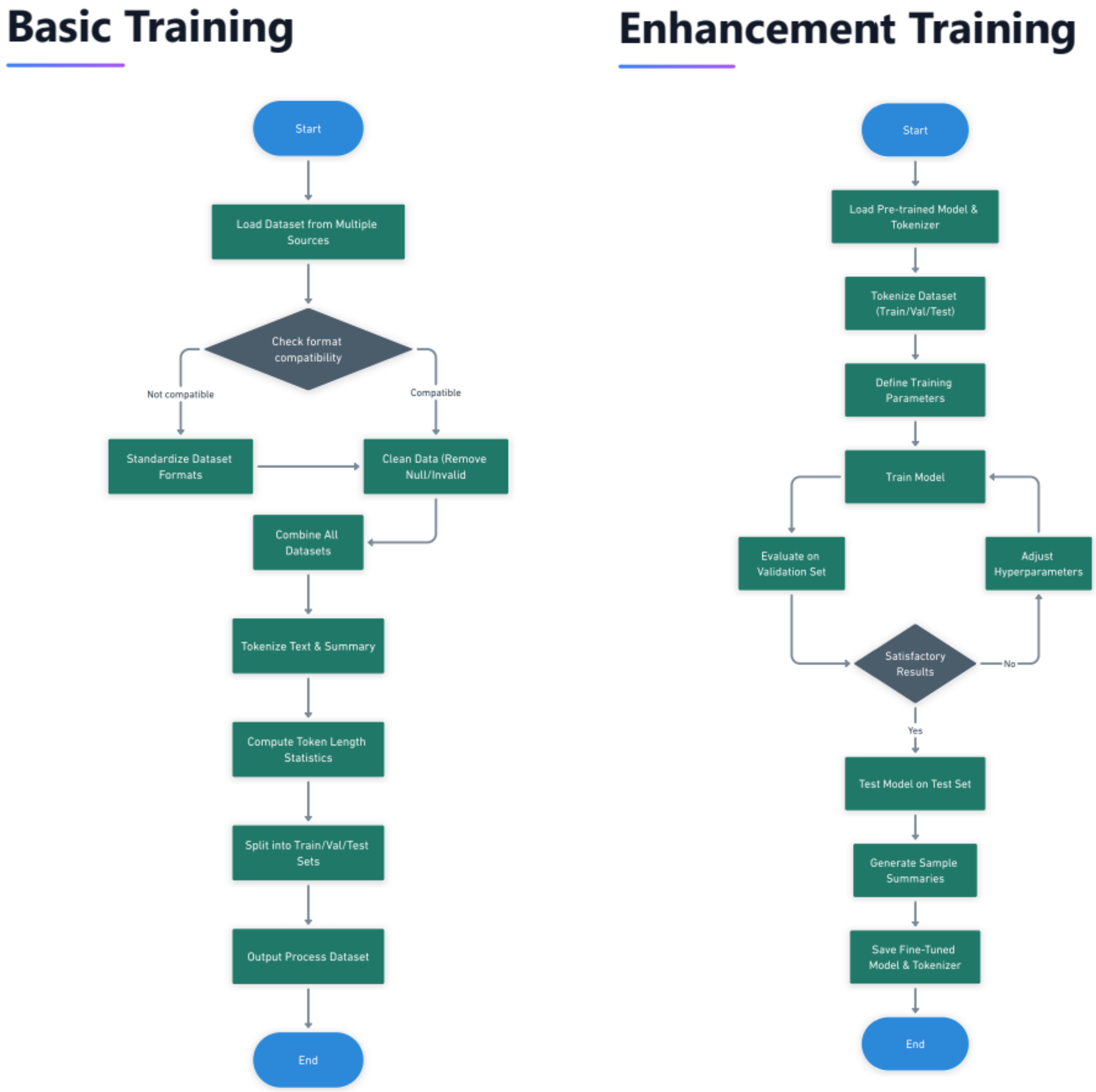
Text Summarization is one of the mentionable research areas of Natural language processing. Several approaches have already been developed in this concern. Such as -Abstractive approach and extractive approach. Most recent recurrent neural network methods are producing much better results. Several mentionable research has already been discussed for English language summarizer, but a few have already done for the Bangla language. There are so many prerequisites for data analysis purpose- word2vector is one of them. Understanding the vector representation of any text leads the way to identify the key main points of that specific text and helps to measure the relationship of that text with other texts in similarity/dissimilarity. Generated matrix using word2vector can easily applicable for identifying top-ranked sentence/words, either domain specific or in general form. In this paper, a word2vector approach has been discussed in the context of text summarization for the Bangla language.

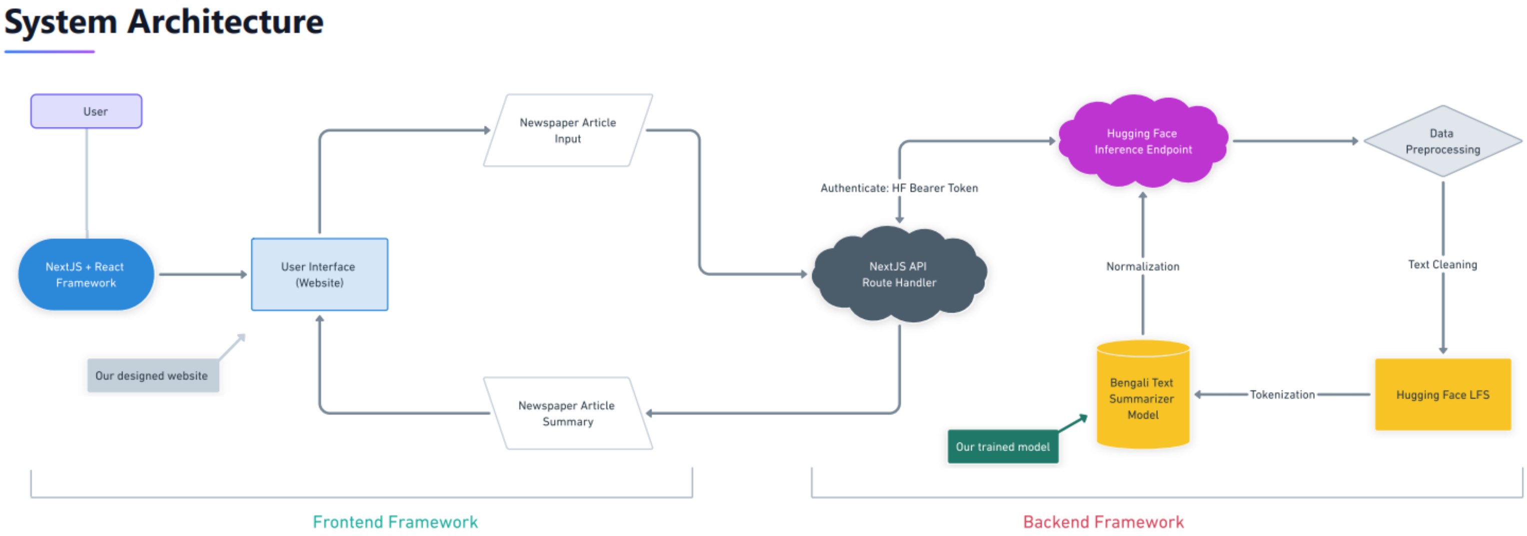
Chapter 3 Chapter 3 Methodology

3.1 System Design

The design of the **Bengali Text Summarizer** project involves multiple stages, starting from data preparation to model deployment in a web application. The following sections describe the design through diagrams and conceptual flow:

Flowchart of the System-


Basic Training
Start
Load Dataset from Multiple Sources
Check format compatibility
Not compatible
Compatible
Standardize Dataset Formats
Clean Data (Remove Null/Invalid
Combine All Datasets
Tokenize Text & Summary
Compute Token Length Statistics
Split into Train/Val/Test Sets
Output Process Dataset
End
Enhancement Training
Start
Load Pre-trained Model & Tokenizer
Tokenize Dataset (Train/Val/Test)
Define Training Parameters
Train Model
Evaluate on Validation Set
Adjust Hyperparameters
Satisfactory Results
No
Yes
Test Model on Test Set
Generate Sample Summaries
Save Fine-Tuned Model & Tokenizer
End



3.2 Hardware and/or Software Components

Dataset Preparation

- **Dataset Source:** Over **20,000 Bengali article-summary pairs** collected from various news websites.
- **Exploratory Data Analysis (EDA):** Performed to ensure the dataset quality, check distribution, and remove outliers or noisy data.
- **Preprocessing Techniques:** Included text cleaning, normalization of Bengali characters, and removing redundant or irrelevant content.

Applied AI/DL Models

- **Transformer Model:** Fine-tuned **MT5** (Multilingual T5) model trained on 15,000 Bengali text-summary pairs using the Seq2SeqTrainer.

- **Tokenization:** Used **MT5Tokenizer** for breaking input text into subword units compatible with the MT5 model.
- **Hyperparameter Optimization:** Optimized parameters such as learning rate, maximum sequence length, number of beams for beam search, and length penalty.
- **Evaluation Metrics:** Used **ROUGE scores** to assess the quality of summaries.

Table 3.1: List of Software/Hardware Tools

Tool	**Functions**	**Other similar Tools (if any)**	**Why selected this tool**
Python 3.8+	Programming language for implementing data preprocessing, training, and backend.	R, Julia	Python offers extensive libraries and community support for NLP and ML.
PyTorch	Deep learning framework for training and fine-tuning the MT5 model.	TensorFlow, Keras	PyTorch provides flexibility and seamless integration with transformers.
Transformers Library	Provides pre-trained MT5 models, tokenizers, and utilities for fine-tuning.	Fairseq	Hugging Face Transformers is widely used for state-of-the-art NLP tasks.
Pandas	Data manipulation and preprocessing during dataset preparation.	Dask, Modin	Pandas is efficient for handling large datasets and has excellent support.
NumPy	Provides support for numerical computations and array manipulation.	SciPy, CuPy	Essential for preprocessing tasks and widely compatible with Python libraries.
NVIDIA GPU	Accelerates training of the MT5 model by parallelizing computations.	AMD GPUs, TPUs	NVIDIA GPUs are widely supported by PyTorch and ensure faster model training.
Local Machine/Cloud Server	Hosts the backend and handles requests between the user interface and the model API.	AWS, Google Cloud, Azure	Offers scalability and flexibility for deploying web applications and APIs.

3.2 Hardware and/or Software Implementation

This project primarily involves **software implementation** for building a Bengali Text Summarizer. The system is designed and implemented using a modular approach, comprising the following steps:

1. Data Collection and Preprocessing

- **Preprocessing Techniques:**
    - Text normalization to standardize Bengali characters.
    - Removing special characters, stop words, and unnecessary whitespace.
    - Tokenizing the Bengali text using the **MT5Tokenizer**, which converts the input text into a sequence of subwords suitable for the model.

2. Model Development

- **Base Model:** The **MT5 (Multilingual T5)** transformer model was chosen as the base architecture.
- **Fine-Tuning:** The model was fine-tuned on **25,000 Bengali article-summary pairs** using the Seq2SeqTrainer for supervised learning. Key aspects of model fine-tuning included:

    - **Hyperparameters:**
        - Maximum sequence length: 512 tokens.
        - Batch size: 16.
        - Learning rate: 5e-5.
        - Number of epochs: 5.

- **Training Techniques:**
    - **Beam Search:** Used during generation to improve the quality of summaries.
    - **Length Penalty:** Adjusted to ensure summaries are concise without losing key details.
- **Evaluation:** Used **ROUGE metrics** (ROUGE-1, ROUGE-2, and ROUGE-L) to evaluate the quality of generated summaries.

3. Real-time summarization with API Integration

- The fine-tuned MT5 model was deployed.
- Real-time processing was enabled by sending the user's input text to the API, generating the summary.
- The cloud-based deployment ensures scalability and reliability for multiple concurrent users.

4. Testing and Debugging

- **Functional Testing:** Ensured that all components, including text preprocessing, summarization, and UI, work seamlessly together.
- **Performance Testing:** Evaluated the summarizer's accuracy using held-out test data.
- **Error Handling:** Implemented to manage issues such as invalid user input, API failures, or timeouts.

Key Outcomes

The implementation resulted in a fully functional **Bengali Text Summarizer** that combines a state-of-the-art transformer model. It provides users with real-time summarization capabilities, delivering high-quality, human-like summaries of Bengali news articles.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

4.1 Experiment Design

Training Loss Evolution:

- Tracked the training loss across multiple epochs to monitor the model's convergence.
- Loss values were recorded for each model version, starting from initial rounds of training with 10,000 data points to subsequent fine-tuning on 20,000 data points.

Evaluation Metrics:

- Used **BLEU**, **METEOR**, and **BERTScore** for quantitative evaluation of the summaries, focusing on semantic similarity and grammatical accuracy.

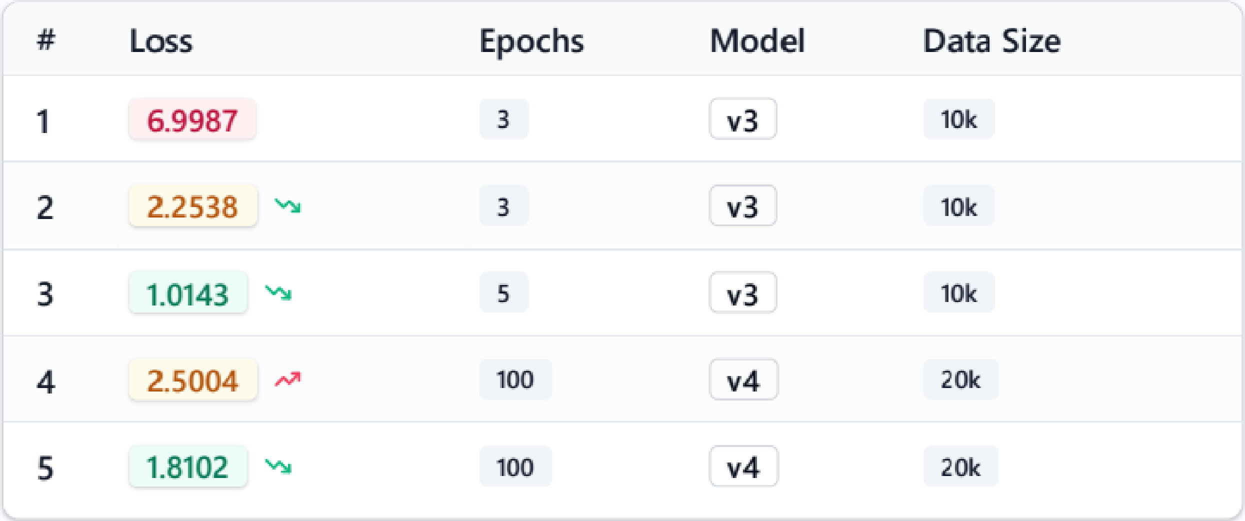
User Interface and Real-Time Processing:

- Tested the responsiveness and real-time performance of the web interface to ensure instant summarization of Bengali text input.

4.2 Results

Training Loss Evolution:

Table 4.1 shows the evolution of the training loss across different training setups and dataset sizes:



#	Loss	Epochs	Model	Data Size
1	6.9987	3	v3	10k
2	2.2538 ↘	3	v3	10k
3	1.0143 ↘	5	v3	10k
4	2.5004 ↗	100	v4	20k
5	1.8102 ↘	100	v4	20k

Table 4.1: Training loss evolution across different rounds and model versions.

Analysis:

- The gradual decrease in loss from 6.9987 to 1.0143 for v3 (10k dataset) demonstrates effective learning during initial training.
- Fine-tuning on the larger 20k dataset resulted in further performance improvements in model version v4.

Evaluation Metrics:

Figure 4.1 compares the BLEU and METEOR scores for training, validation, and test sets.

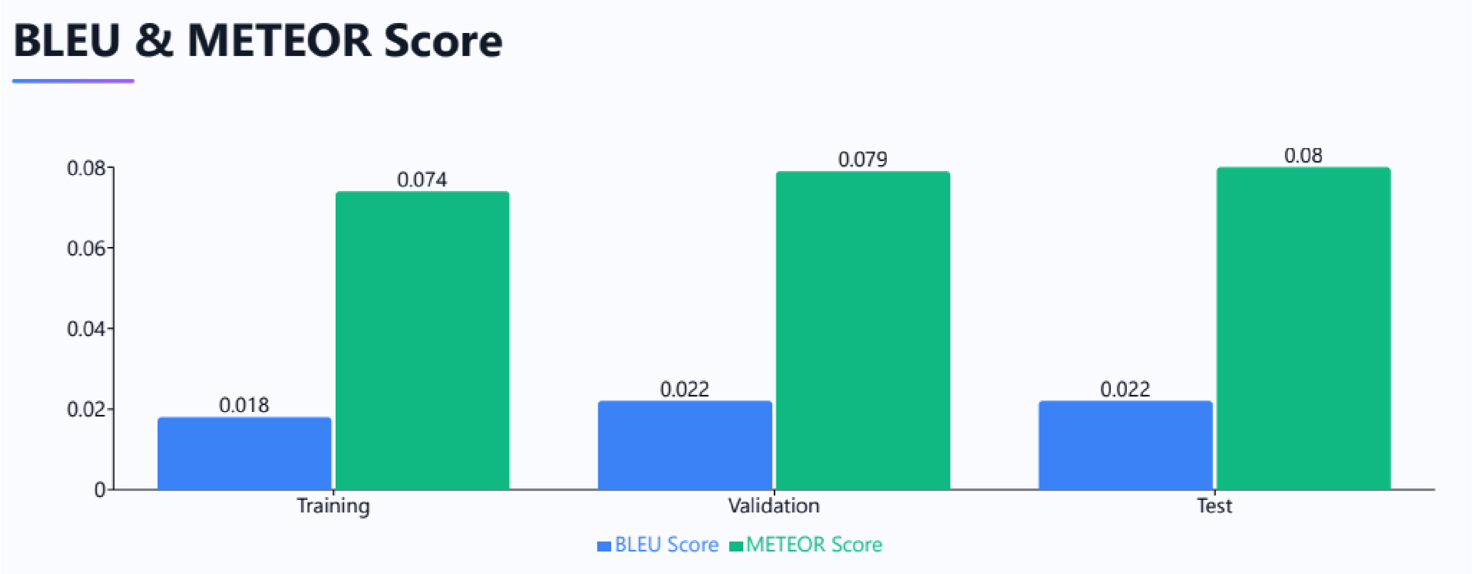


Figure 4.1: BLEU and METEOR scores across datasets (Training, Validation, Test).

Analysis:

- Higher BLEU and METEOR scores on the test set indicate a strong generalization of the model.
- BLEU scores demonstrate syntactic correctness, while METEOR measures semantic overlap between summaries and reference texts.

Content Coverage

The summarizer was evaluated on how much content it retained from the original article. Figure 4.2 presents the percentage of content covered by the generated summaries.

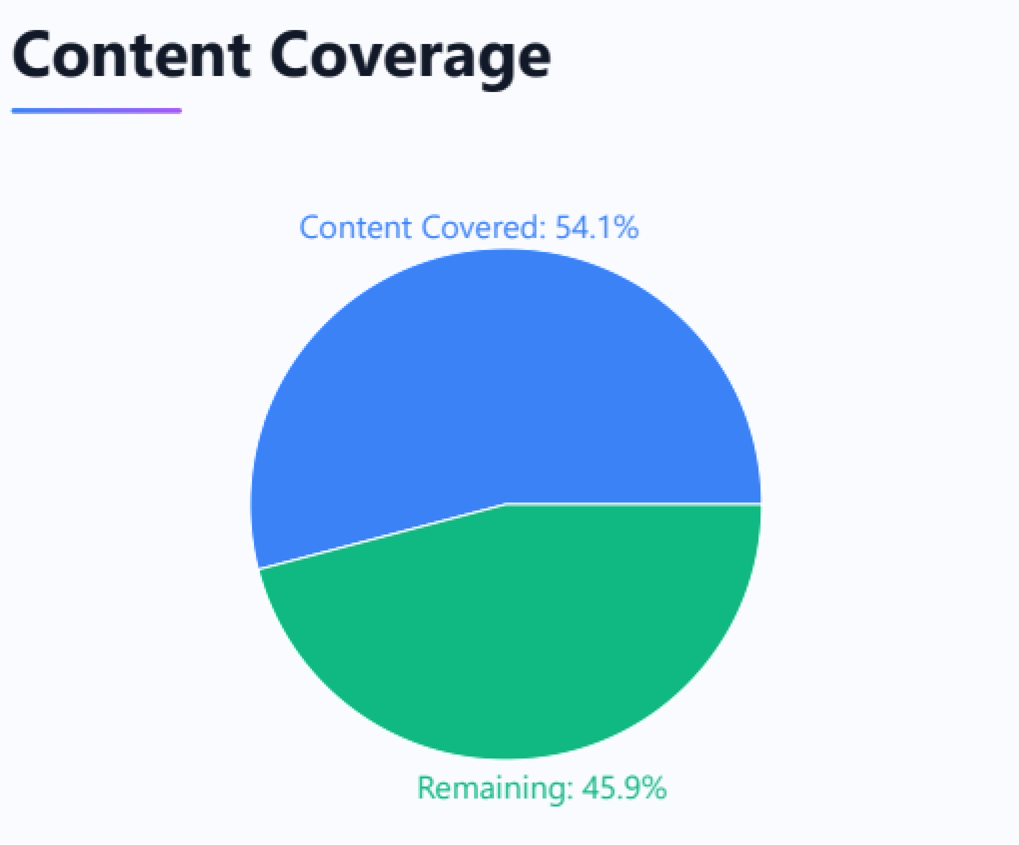


Figure 4.2: Content coverage by the summarizer.

Analysis:

- While 54.1% of key content is captured by the summaries, the remaining 45.9% reflects areas where the model may have truncated or omitted certain details. This highlights potential improvement opportunities for handling longer texts.

BERTScore Evaluation

To evaluate semantic similarity, **BERTScore** was used to measure precision, recall, and F1 score for the generated summaries (Figure 4.3).

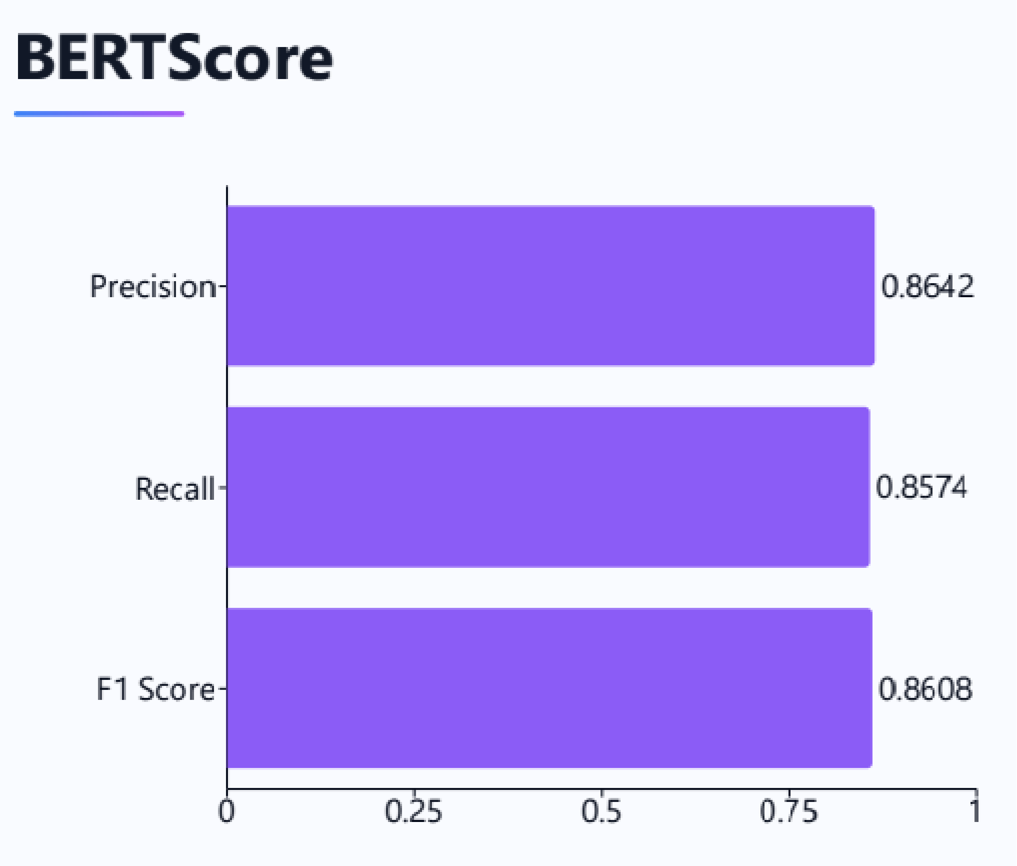


Figure 4.3: BERTScore results for generated summaries.

Analysis:

- High precision and recall scores indicate that the model accurately captures important semantic relationships while avoiding excessive irrelevant details.

4.3 Discussion

1. **Model Performance:**
   - The **fine-tuned MT5 model** achieved significant improvement in loss reduction, BLEU, and METEOR scores, showcasing its suitability for Bengali text summarization.
   - The training with progressively larger datasets (from 10k to 20k articles) enhanced the model's generalization capabilities.
2. **Challenges:**
   - The model struggled to cover all relevant content for very long articles, as seen in the content coverage results (54.1%). Techniques such as hierarchical attention or chunking could improve this in the future.
   - BLEU and METEOR scores, though satisfactory, indicate room for improvement in generating more contextually rich summaries.

Chapter 5 Conclusions

5.1 Summary

The Bangla Text Summarization System is one such effort to bridge the important gap in the lack of accessible tools for summarizing Bangla-language content. This project applied a combination of the pre-trained model Google/mt5-small and robust NLP techniques to successfully create a scalable and efficient method of generating concise summaries of Bangla news articles. This is made by integrating using Colab Pro for model training; seamless coordination among hot cutting-edge AI. This prototype, not with standing other challenges with datasets and computational capacities, did all that was put forth to upscale Bangla digitally and enhance Bangla content more accessible to speakers.

5.2 Limitations

While the project achieved significant milestones, certain limitations were identified:

Dataset Scarcity:

A major challenge was the lack of high-quality, annotated datasets specific to Bangla, which impacted the overall accuracy and generalizability of the model.

Model Constraints:

The model’s performance was constrained by computational resources, particularly the inability to train longer summaries beyond 200 tokens without crashes in Google Colab.

Fine-tuning a larger model with more parameters was not feasible within the given infrastructure.

Quality of Generated Summaries:

The generated summaries were functional but not always of high quality, with occasional issues in grammatical coherence and semantic accuracy due to dataset limitations.

Limited Evaluation Metrics:

Due to low-quality reference summaries, ROGUE scores could not be computed, restricting the evaluation to BLEU, METEOR, and BERTScore metrics.

5.3 Future Improvement

Several improvements can be implemented to enhance the project further:

Improved Dataset Collection:

Focus on building a larger, high-quality, annotated Bangla dataset by collaborating with local news organizations, universities, or crowdsourcing platforms.

Enhanced Model Training:

Upgrade computational resources to fine-tune larger models or explore more sophisticated architectures like T5-large or GPT-based models for better performance.

Advanced Preprocessing Techniques:

Implement advanced preprocessing techniques to handle the nuances of Bangla text, such as better tokenization, semantic labeling, and handling colloquial or regional variations.

Quality Optimization:

Employ reinforcement learning or feedback-based methods to improve the grammatical accuracy and coherence of the generated summaries.

Broader Evaluation Metrics:

Explore alternative evaluation methods to assess the quality of summaries without relying heavily on reference datasets, such as human evaluation or newer NLP benchmarks.

Scalability Enhancements:

Migrate to more robust cloud platforms, such as AWS or Azure, for scalable deployment to handle more significant traffic and ensure lower latency.

Inclusion of Multilingual Support:

Extend the project to include multilingual summarization capabilities, leveraging the pre-trained model’s inherent support for multiple languages.

User Experience Improvements:

Enhance the frontend interface for better accessibility and usability, incorporating user feedback to improve the interaction design and features.

Chapter 6 References

1. Md. Iftekharul Alam Efat, Mohammad Ibrahim, Humayun Kayesh, “Automated Bangla Text Summarization by Sentence Scoring and Ranking,” 2013 International Conference on Informatics, Electronics, and Vision (ICIEV), May 2013. https://www.researchgate.net/publication/261212570_Automated_Bangla_text_summariz ation_by_sentence_scoring_and_ranking
2. A. S. Muhib, S. Ishfar, A. N. Hasan, & A. R. M. Kamal, “Bangla Text Summarization Using Deep Learning,” Islamic University of Technology (IUT), Department of Computer Science and Engineering (CSE), March 2021. http://103.82.172.44:8080/xmlui/handle/123456789/1274
3. Sheikh Abujar, Abu Kaisar Mohammad Masum, Syed Akhter Hossain, “An Approach for Bangla Text Summarization Using Word2Vector,” 2019 10th International Conference on Computing, Communication, and Networking Technologies (ICCCNT), July 2019. https://www.researchgate.net/publication/338358097_An_Approach_for_Bangla_Text_S ummarization_using_Word2Vector
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., “Attention Is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), 2017. https://arxiv.org/abs/1706.03762
5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Journal of Machine Learning Research, 2020. https://arxiv.org/abs/1910.10683
6. Das, A., Sarkar, K., & Chakrabarti, A., “A Comprehensive Study on Low-Resource NLP Techniques for Indian Languages,” International Journal of Computational Linguistics, vol. 14, pp. 125–137, 2022.
7. Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C., “A Neural Probabilistic Language Model,” Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf