SYLHET ENGINEERING COLLEGE

Department of Computer Science & Engineering

Affiliated with Shahjalal University of Science and Technology

Abstractive summarization of Bengali short story, using NLP

Submitted by

Nurul Islam Opu
Reg. No.: 2019331530

Mithun Chakraborty Tinni
Reg. No.: 2019331567

Department of Computer Science & Engineering

Submission Date : 22nd July,2025

Supervisor
Md Abu Naser Mojumder
Assistant Professor & Head
Department of Computer Science & Engineering
Sylhet Engineering College

Recommendation Letter from Thesis Supervisor

The thesis titled **"Abstractive summarization of Bengali short story, using NLP"** submitted by the group as mentioned below has been accepted as satisfactory in partial fulfilment of the requirements for the degree B. Sc. in Computer Science and Engineering in 22nd July, 2025.

Group Members:

Nurul Islam Opu (2019331530)

Mithun Chakraborti Tinni (2019331567)

Supervisor:

Md.Abu Naser Mojumder
Assistant Professor and Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Certificates of Acceptance

The thesis is titled **"Abstractive summarization of Bengali short story, using NLP"** submitted by **Nurul Islam Opu** and **Mithun Chakraborti Tinni**; Student ID. **2019331530** and **2019331567**; Session **2019-20**, to the Department of Computer Science and Engineering, Sylhet Engineering College, has been accepted as satisfactory in partial fulfilment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Internal
Nayan Kumar Nath
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Internal

Md. Lysuzzaman
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Internal

Md. Rasel Ahmed
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Internal

Md. Nojrul Islam
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Chairman

Md. Abu Naser Mojumder
Assistant Professor and Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet.

Member (External)

Mohammad Shahidur Rahman, Ph.D., SMIEEE
Professor
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Head of Dept. Chairman, Exam Committee

Acknowledgements

Above all, we express our heartfelt thanks to the Almighty, whose infinite mercy and unseen guidance have been our unwavering support throughout this journey. In both trials and triumphs, His blessings have guided our way and strengthened our determination.

We would like to express our heartfelt appreciation to our respected supervisor, **Md. Abu Naser Mojumder**, for his invaluable support, continuous encouragement, and insightful feedback. His guidance has been pivotal in shaping the direction and quality of our research, and we are truly grateful for his mentorship.

We extend our heartfelt gratitude to our esteemed teachers, **Md. Lysuzzaman** and **Nayan Kumar Nath**, whose unwavering commitment to education and profound expertise have greatly influenced our academic journey. Their continuous support and motivation have been truly inspiring throughout our academic journey.

Weareespeciallythankfulforthestrongcollaborationandmutualunderstandingwesharedas thesis partners. The joint effort in completing our thesis, "**Abstractive summarization of Bengali short story, using NLP**" was made possible through shared dedication, teamwork, and a commitment to learning.

Finally, we wholeheartedly acknowledge the unwavering support of our families, to whom we owe our deepest love and gratitude. Their selfless sacrifices, unconditional support, and constant prayers have been the foundation of our academic journey. This accomplishment would not have been possible without their steadfast faith in us.

Abstract

In this research work, we focused on developing an **abstractive summarization** system for **Bengali short stories**, a domain that has received limited attention and is considered low-resource compared to high-resource languages like English or other global languages. While automatic text summarization has seen advances in high-resource languages, Bengali remains underrepresented in this field due to its low-resource language status. This study presents abstractive summarization for Bengali short stories by constructing a new dataset of 430 stories paired with human-written summaries customized by us.

We then fine-tuned the multilingual **Bangla T5** model, adapted for Bengali as the **BanglaT5** model, on this dataset. The fine-tuning process enabled the model to generate semantically rich and grammatically coherent summaries that go beyond simple sentence extraction.

The performance of the model was evaluated using **BERTScore**, a semantic-based metric. The system achieved a **precision** of 0.5927, **recall** of 0.5398, and an **F1-score** of 0.5644, indicating a promising level of abstractive capability. The results highlight both the potential and challenges of building summarization systems in low-resource languages like Bengali.

These results presented additional challenges for us. Previously, we experimented with the **Bangla T5** model, but it yielded lower **ROUGE** scores compared to the current results.

This work contributes a novel dataset and baseline model for future research in Bengali text summarization and low-resource language Natural Language Processing (**NLP**). Future improvements may include incorporating larger pre-trained Bengali models, experimenting with **reinforcement learning** techniques, and expanding the dataset to include more diverse writing styles.

Keywords:

abstractive summarization, Bengali short stories, low-resource language, Bangla T5 model, BanglaT5, NLP, BERTScore, precision, recall, F1-score, pre-trained, reinforcement learning

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1

Background

In recent times, many Bengali short stories have been made available in digital formats. **Natural Language Processing (NLP)** enables computers to understand and communicate in human languages and efficiently handle various language-related tasks. By integrating computational linguistics with statistical modeling, machine learning, and deep learning, NLP allows machines to process and generate both text and speech. As a result, its applications have become increasingly popular and widely adopted across different domains.

One of the key applications of NLP is **text summarization**, which involves generating a shorter version of a document, article, or story while retaining its core meaning and essential information. The main goal is to extract the most important points from a lengthy text and present them in a concise format. This becomes especially valuable in an era of information overload, where reading through lengthy documents can be time-consuming.

There are two primary types of summarization techniques: **extractive** and **abstractive**. Extractive summarization selects and compiles important sentences directly from the original text. On the other hand, **abstractive summarization** attempts to comprehend the underlying meaning of the text and generate a new, shorter version that conveys the same information in original language. (Mim et al., 2021). While English and other languages have made significant advances in both types of summarization, Bengali remains underexplored, especially in the abstractive domain. Most of the Bengali summarization research relies heavily on extractive approaches due to the lack of high-quality annotated datasets and pretrained language models for Bengali. This gap highlights the urgent need for developing Bengali-specific resources and models that can support abstractive summarization tasks. Bangla T5 is pretrained on the mC4corpus, covering 101 languages, where the Bengali texts are also included.

Unlike T5, which was trained only on English data, Bangla T5 can handle a wide range of languages, making it a powerful tool for multilingual natural language processing tasks. Creating a custom dataset of Bengali short stories and applying in Bangla T5 (based on Bangla T5) can help address this gap and contribute to the growing parts of NLP research in low-resource languages.

1.2

Problem Statement:

Although there has been a lot of progress in summarization for English and other high-resource languages, Bengali still remains underdeveloped in this field. Most of the existing work on Bengali summarization depends on extractive methods, which often cannot reflect the natural flow and abstract style of human-written summaries. For abstractive summarization, Bengali lacks strong models and publicly available annotated datasets. To help solve this problem, we created our own dataset made up of Bengali short stories written by different authors.

1.3

Objectives

- To create a custom dataset of Bengali short stories along with their abstractive summaries.
- To fine-tune a pre-trained multilingual Text-to-Text Transfer Transformer model (Bangla T5) using the developed dataset.
- To evaluate the quality of the generated summaries using which measures the overlap between machine-generated and human-written summaries.
- To analyze the performance and accuracy of the model on the training data to further improve the quality of abstractive summaries.
- To establish a foundation for future research in abstractive summarization of Bengali short stories.
- Provide the system with a dedicated **16 GB** GPU, **32 GB** shared GPU memory, and **64** GB RAM

1.4

Limitations

- Computation resource limits when fine-tuning large models like BanglaT5 Less Evaluation metrics such as: numbers of BERT Score precession recall and F1 score in percentage .
- Limited ability to handle poetic texts, which are common in Bengali literature.
- **Small dataset size** (only 430 stories) as compare to other languages.

Chapter 2

Literature Review

2.1

Related works

Our research paper is about **"Abstractive summarization of Bengali short stories using NLP."**

Since 2007, Bangla Text Summarization (BTS) has improved a lot, with different models and methods being used to create short summaries from long texts.There are two main types of text summarization: extractive and abstractive. Extractive summarization identifies and combines existing sentences from the original text to create a summary. Abstractive summarization, on the other hand, generates a new, condensed summary by paraphrasing and interpreting the original text.Abstractive model used here Bangla T5-multilingual-XLSum BART-large-CNN. Extractive Approaches include sentence ranking and extraction step, including: (i) statistical ones,(ii) graph-based ones,(iii) semantic-based ones.Limitations are Abstractive models need retraining or fine-tuning whenever documents from different languages or domains are used.Abstractive models require large datasets, long training times, and specialized hardware.Abstractive summarization performs best when models are fine-tuned on domain-matched data.Extractive methods are more stable across datasets but lack abstraction and naturalness.BLEU and ROUGE score show similar ranking trends for both summarization ."(Giarelis et al., 2023)" .

The task of chart-to-text summarization belongs to the broader domain of Natural Language Generation ,specifically focused on generating descriptive text from structured, non-linguistic data such as charts and tables.Chart-to-text summarization is the task of creating descriptive summaries from structured data like tables or charts.This dataset facilitates the analysis of LLMs (Bangla T5, BanglaT5, Gemma) in Bengali chart-totext summarization.Metrics used to evaluate these tasks include ROUGE, BLEU, WER, and CER, which help assess structure, fluency, and correctness.All tested models (Bangla T5, BanglaT5, Gemma) perform poorly on the Bengali chart-to-text summarization task.Although a benchmark dataset (Bengali ChartSumm) was proposed, its size and diversity are limited."(nahidaakter et al.,2025)".

"(Chowdhury et al.,2021)" They use NCTB Dataset (Abstractive), that have 139 samples and BNLPC Dataset (Extractive),200 samples.Best performance (extractive) reaches ~61% for ROUGH-1 , Best performance (abstractive) reaches ~12% for ROUGE-1.introduced BenSumm, the first graph-based unsupervised model designed to generate abstractive summaries from Bengali text documents. The model requires only a Part-of-Speech (POS) tagger and a pre-trained language model trained on Bengali data. To evaluate the performance of BenSumm, the authors also developed a human-annotated dataset comprising document-summary pairs, facilitating both assessment of their model and future benchmarking in Bengali abstractive summarization research. Experimental results demonstrated that BenSumm outperformed several well-established unsupervised extractive summarization systems, despite not relying on human-annotated reference summaries.However, a key limitation of BenSumm lies in its inability to generate novel words, highlighting the need for future research through the integration of multi-sentence compression and paraphrasing to enhance its abstractive capabilities.

"(Vahid Nejad Mahmood Abadi & Fahimeh Ghasemian,2024) ",this paper about to improve Persian abstractive text summarization using a hybrid method combining: Bangla T5 transformer model ,Three-phase fine-tuning,Reinforcement Learning (RL).This work is designed to reduce exposure bias and increase the quality and fluency of generated summaries. Their experiments show that the final model significantly outperforms existing baseline models like mBART and PEGASUS, achieving state-of-the-art ROUGE scores (ROUGE-1: 53.89, ROUGE-2: 37.25, ROUGE-L: 50.23).However,the proposed method is evaluated only on the language ;Persian, so its effectiveness on other low-resource languages (like Bengali or Tamil) remains untested.

"(Joglekar et al.,2025)",Text Summarization in local language, focused on focuses on abstractive text summarization techniques for regional languages,Hindi.They used state-of-the-art Transformer based models to generate coherent, contextually accurate summaries from local language content. In this research ,the dataset utilized news articles and corresponding human-generated summaries.Where the maximum text lengths were 500 words and the maximum summaries were 50 words.The model was trained by 20 epochs.The results obtained from the implementation of the Transformer based model through ROUGE scores .The model achieved a ROUGE-1 score with a F1 measure of approximately 0.67, a precision of 0.86, and a recall of 0.55.Where the scores of ROUGE-2 was very low.The results give a foundation for further exploration and improvement in automated text summarization

methodologies.

"(Efat et al.,2013)", introduced an extraction-based summarization technique tailored for Bangla text documents,addressing the deficiency of such works in comparison to English.It focused on the necessity of automated systems,given the increasing volume of Bangla documents.The authors proposed leverages word frequency technique,sentence positional value,cue words and document skeleton,with respect to precision.Here the results give the 83.57% agreement between machine created and human making summarizations.

"(Abujar et al.,2017)",the authors developed a heuristic-based model that processes Bengali texts using linguistic and statistical techniques to generate concise summaries. From the research, it explained, how the Bengali sentences should be scored after removing the stop words. They use a three-phase process combining linguistic preprocessing, sentence scoring, and final refinement. They used scoring formula, based on Python and NLTK. There proposed method gave a close summary to human generated ones. They didn't show any numerical accuracy but a figure shown high correlation between system and human output scores.

2.2

Research Gap

This research focuses on abstractive summarization of Bengali short stories, an area that has seen limited progress compared to high-resource languages. Bengali text summarization research began around 2007, primarily using extractive techniques based on statistical, graph-based, or semantic methods. Recent studies have explored **abstractive models** like Bangla T5, BART, and BanglaT5, but these require large annotated datasets, domain-specific fine-tuning, and powerful hardware, making them challenging for low-resource languages. Studies such as "Giarelis et al. (2023) "emphasized the limitations of abstractive methods across languages and domains. "Chowdhury et al. (2021)" introduced BenSumm, an unsupervised graph-based model for Bengali abstractive summarization using POS tags and pre-trained Bengali models. Although it outperformed extractive baselines, it lacked true abstraction due to the absence of novel word generation. In a related domain, Nahida Akter et al. (2025) worked on Bengali chart-to-text summarization using Bangla T5, BanglaT5, and Gemma but found poor performance due to dataset limitations. Other works, such as Mahmood Abadi & Ghasemian (2024), proposed a hybrid Bangla T5 approach using reinforcement learning for Persian summarization, showing state-of-the-art performance but lacking multilingual validation. Similarly, Joglekar et al. (2025) explored abstractive summarization for Hindi using Transformer models,

achieving good ROUGE-1 scores but struggling with ROUGE-2. Earlier Bengali works by Efat et al. (2013) and Abujar et al. (2017) relied on heuristic and statistical extractive methods, showing high similarity with human-generated summaries but lacking abstraction. These studies highlight the growing interest in Bengali summarization but also underline the need for better datasets, fine-tuned models, and further research in abstractive approaches for low-resource languages.

Chapter 3

Background study

This section describes the fundamental theory of the study.This includes an overview of machine learning and discuss the basic concept of machine learning algorithm,that used in our study.

3.1

Machine Learning:

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on developing systems that can learn from data and improve their performance over time without explicit programming. It enables computers to identify patterns, make predictions, and make decisions by analyzing large datasets.It uses various algorithms, which are sets of instructions, to analyze data and build models.Machine learning learns from data.It's algorithms are classified into 2 categories ,"Supervised" and "Unsupervised". Supervised Learning uses labeled data,providing with input data and expected output.On the other hand, Unsupervised Learning uses unlabled data only given input data and finding relationshipwithen the data without any guidance.

The models created by ML algorithms can be used to make predictions or decisions on new, unseen data by using mathematical models based on training datasets.

3.2

Machine Learning Techniques in NLP:

NLP (Natural Language Processing) is not a learning algorithm ,but it is a field of study focused on enabling machines to understand and process human language.NLP uses a variety of ML techniques to enable computers to generate and understand human like languages. In this section ,there are some great machine learning techniques applied in the domain of NLP.

Supervised Learning

Text classification : This model is used to categorize text into organized groups with respect

to training data.

Named Entity Recognition (NER): It involves training a model to identify and classify named entities (such as:name date , locations) within text.

UnSupervised Learning

Word Embeddings:

Techniques that represent words as dense vectors based on their meanings and context in large text corpora.

Clustering Documents:

Grouping similar documents together based on their content without predefined categories.

Topic Modeling:

An unsupervised method to discover hidden topics in a collection of documents based on word patterns without labeled data.

Transfer Learning in Machine Learning

Transfer learning uses pre-trained models from one machine learning task or dataset to improve performance and generalization ability on a related task or dataset. In NLP, a pretrained model like BanglaT5 can be fine-tuned on Bengali summarization tasks using a small set of dataset , this is transfer learning.

3.3

Used of Supervised Learning in our works

Supervised learning is a type of machine learning where the algorithm learns from a labeled training dataset for training, where the correct output is known for each input ,to learn the relationship between inputs and outputs and then predicts outcomes for new, unseen data.

3.3.1

Loss Function

A loss function is a type of *objective function***,** which in the context of data science refers to any function whose minimization or maximization represents the objective of model training.It finds the difference between the predictive value and the actual target value of ML algorithm. A loss value of zero in the result means perfect performance. It gives us back a non-negative number indicating the disagreement or error between the targeted and predicted outputs. There are 2 types of Loss Function:

- **Mean Square Error (MSE) / L2 Loss :** Used primarily for regression problems and its numerical equation is : **$MSE = (1/n) * \Sigma(y_i - \bar{y})^2$**

Where n = number of total datasets

y_i= predicted value

$\bar{y}$=target value

- **Mean Absolute Error (MAE)/L1 Loss :** The MAE is another regression loss function that measures the average absolute difference between the predicted and actual target values.Its neumerical formula is: $MAE = (1/n) * \Sigma| y_i - \bar{y}|$

3.3.2

Gradient Descent

Gradient descent is a type of algorithms used to search for parameters that minimize the loss function.This is declared by the loss function's gradient with respect to the various weightsIt . is an iterative process that finds the best weights and bias that minimize the loss of the model. It calculates the slope of the cost function with respect to each parameter. A hyperparameter that controls the step size during parameter updates. The gradient indicates the direction of the steepest increase in the cost function. Stochastic Gradient Descent (SGD**)** uses a single training example to calculate the gradient at each iteration. It allows models to learn from data and improve their accuracy over time.

3.3.3

Optimization

These algorithms help models learn from data by finding the best possible solution through iterative updates. Different parameters are available for optimization of network. Besides model parameters, optimization involves tuning hyperparameters, which are settings that control the learning process itself, such as the learning rate(alpha) or regularization strength.

3.3.4

Learning Rate

The learning rate α is the 1st hyperparameters to be modified after choosing an architecture.

Here the α is called the learning rate.It is a hyper-parameter used to govern the pace at which an algorithm updates or learns the values of a parameter estimate.The hyper-parameter controls the rate of learning or speed of a model. The higher the alpha the larger portion of the current gradient is consider and smaller the alpha the smaller is the considered gradient.

3.4

Model Overview

BanglaT5 is a transformer-based language model developed to address the lack of high-performing NLP tools for the Bengali language. It is based on the T5 (Text-to-Text Transfer Transformer) architecture, which reframes all language tasks—such as summarization, translation, classification, and question answering—into a unified text-to-text format. Unlike multilingual models like Bangla T5 that are trained across numerous languages, BanglaT5 is pre-trained exclusively on a diverse and extensive Bengali corpus, making it better suited for capturing the syntactic, semantic, and grammatical nuances of the Bengali language.

The pretraining dataset for BanglaT5 includes a large collection of Bengali text from sources like Bengali Wikipedia, news articles, books, and web content. This focused training allows the model to generate more fluent, contextually accurate, and culturally relevant outputs in Bengali. Because of its monolingual design, BanglaT5 has shown improved performance in downstream NLP tasks like summarization and translation when compared to general-purpose multilingual models.

In this research, BanglaT5 is fine-tuned specifically for the task of abstractive summarization of Bengali short stories. The model leverages its strong language understanding capabilities to generate concise and coherent summaries that are not simple extractions, but instead are reworded representations that retain the core meaning of the original story. Its architecture allows it to generalize well in low-resource settings, though performance is still highly dependent on the quality and size of the training data.

Overall, BanglaT5 plays a crucial role in advancing NLP research for Bengali, a low-resource but widely spoken language. It demonstrates the potential of language-specific pretraining in improving the quality of text generation tasks and offers a strong foundation for future developments in Bengali natural language applications.

3.5

A Comprehensive Exploration of Bengali Short Stories and there Urgencies

Bengali short stories lead to a unique and important place in South Asian literature, that reflects the diverse socio-political and cultural landscape of Bengal. The stories related of human emotions, to displacement and the search for meaning in a changing society. Given the cultural richness and narrative complexity of Bengali short stories, they serve as an excellent domain for natural language processing tasks such as summarization, translation, and sentiment analysis. However, despite their literary importance, the computational study of Bengali short stories remains under-explored, particularly in the area of abstractive summarization, due to limited sources and tools.

Bengali is the 7th most popular language in the world,spoken by over 265 million individuals globally. This language has its unique characteristics, not directly translatable to the other language. This is necessary to develop the summarizing process of Bengali short story. For this, the reader can get interest about what story he/she should study fully.

3.6

Text Summarization Method

Text summarization in NLP is the process of shortening a longer text while keeping the main information and meaning unchanged. The goal is to create a short and accurate summary that captures the main idea of the original document. There are two primary types of text summarization:

i. Extractive summarization: This technique automatically creates summaries by picking and joining important parts from the original text. It tries to keep the original meaning while making the text shorter.

ii. Abstractive summarization: This method creates new sentences by understanding the meaning of the original text.

Both methods have been used in different languages for research. However, the focus on Bengali short story summarization is still very limited. This gap shows a clear opportunity for more work and research in this area.

3.7

Challenges in Abstractive Summarization of Bengali Short Story:

Abstractive summarization of Bengali short stories presents different challenges for the complexity of the language and lack of the resources.

- limited availability of sufficient datasets**:** One major issue is the limited availability of insufficient datasets for training models. Unlike English, Bengali has small and low quality of story summary,which makes it challengeable to create a fine-tuned model.

- Linguistic Intricacies**:** The inherent complexities of the Bengali language itself represent a significant hurdle . Bengali short stories often contain large numbers of **linguistic intricacies**, such as idiomatic expressions ,complex sentence structures and poetic language, which make abstractive summarization more challenging for machine learning models.

3.8

Broader context

As Natural Language Processing (NLP) continues to advance, automatic text summarization has become a valuable tool for handling large volumes of text. While languages like English and other high-resource languages have made significant progress in both extractive and abstractive summarization, Bengali still lags behind in this area. In this broader context, it is important to recognize that this research is valuable not only for the Bengali language community but also as a contribution to the wider field of NLP. Moreover, developing effective summarization techniques for Bengali short stories is important not just for technological development, but also for preserving and promoting regional literature in the digital age.

Chapter 4

Methodology

This section describes the methodological frame work followed in our study. It covers the data acquisition and pre-processing strategies, the deep learning models implemented, the embedding techniques used, the experimental setup, and the evaluation metrics applied to measure model performance.

4.1

Data Collection and Annotation

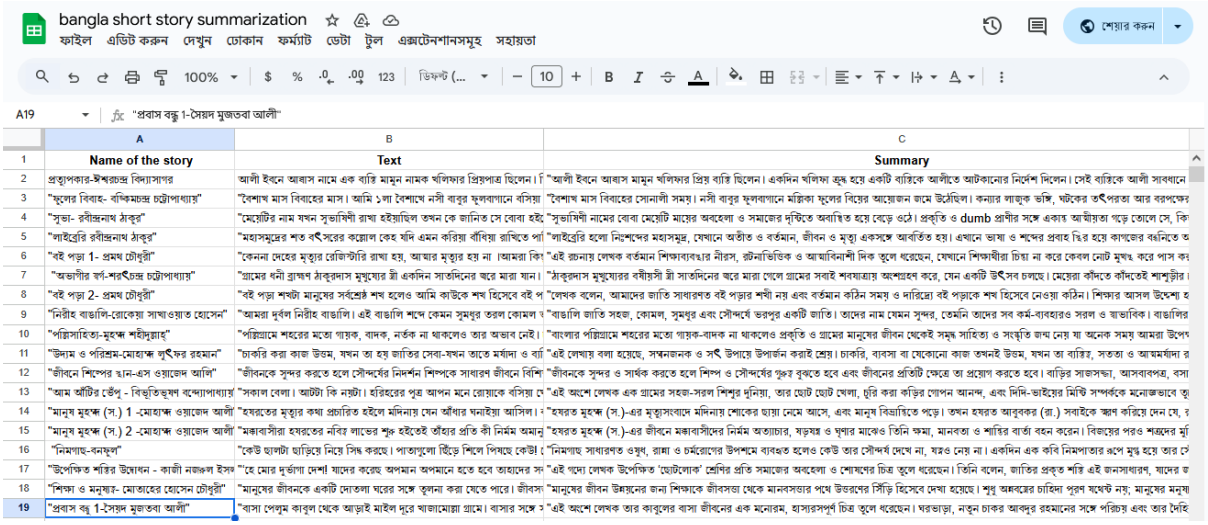
For this research, a diverse corpus of Bangla short stories was compiled from various sources. Primary data was extracted from NCTB (National Curriculum and Textbook Board) books, which include educational short stories. Additional data was collected from online articles, newspapers, and literary websites, ensuring a mix of narrative styles and domains. Each story was manually reviewed to ensure completeness and relevance. For supervised learning, human-annotated abstractive summaries were prepared for each story, focusing on capturing the main idea, key events, and emotional tone within 3–5 lines.

We have created our data of Bengali short story books, written by various popular writer.

Table 4.1: Total Dataset

Total number of story (rows)	**Columns**
430	3,(name of the story, text, summary)

Here are some story collection for our dataset:



bangla short story summarization

	A	B	C
1	Name of the story	Text	Summary
2	প্রত্যুপকার-ঈশ্বরচন্দ্র বিদ্যাসাগর	আলী ইবনে আবাস নামে এক ব্যক্তি মামুন নামক খলিফার প্রিয়পাত্র ছিলেন। ি	"আলী ইবনে আবাস মামুন খলিফার প্রিয় ব্যক্তি ছিলেন। একদিন খলিফা ক্রুদ্ধ হয়ে একটি ব্যক্তিকে আলীতে আটকানোর নির্দেশ দিলেন। সেই ব্যক্তিকে আলী সাবধানে
3	"ফুলের বিবাহ- বঙ্কিমচন্দ্র চট্টোপাধ্যায়"	"বৈশাখ মাস বিবাহের মাস। আমি ১লা বৈশাখে নসী বাবুর ফুলবাগানে বসিয়া	"বৈশাখ মাস বিবাহের সোনালী সময়। নসী বাবুর ফুলবাগানে মল্লিকা ফুলের বিয়ের আয়োজন জমে উঠেছিল। কন্যার লাজুক ভঙ্গি, ঘটকের তৎপরতা আর বরপক্ষ
4	"সুভা- রবীন্দ্রনাথ ঠাকুর"	"মেয়েটির নাম যখন সুভাষিণী রাখা হইয়াছিল তখন কে জানিত সে বোবা হইবে	"সুভাষিণী নামের বোবা মেয়েটি মায়ের অবহেলা ও সমাজের দৃষ্টিতে অবাঞ্ছিত হয়ে বেড়ে ওঠে। প্রকৃতি ও dumb প্রাণীর সঙ্গে একায় আত্মীয়তা গড়ে তোলে সে, কি
5	"লাইব্রেরি রবীন্দ্রনাথ ঠাকুর"	"মহাসমুদ্রের শত বৎসরের কল্লোল কেহ যদি এমন করিয়া বাঁধিয়া রাখিতে পা	"লাইব্রেরি হলো নিঃশব্দের মহাসমুদ্র, যেখানে অতীত ও বর্তমান, জীবন ও মৃত্যু একসঙ্গে আবর্তিত হয়। এখানে ভাষা ও শব্দের প্রবাহ স্থির হয়ে কাগজের বন্ধনিতে অ
6	"বই পড়া 1- প্রমথ চৌধুরী"	"কেননা দেহের মৃত্যুর রেজিস্টারি রাখা হয়, আমার মৃত্যুর হয় না ।আমরা কি	"এই রচনায় লেখক বর্তমান শিক্ষাব্যবস্থার নীরস, রটনাভিত্তিক ও আয়াবিনাশী দিক তুলে ধরেছেন, যেখানে শিক্ষার্থীরা চিন্তা না করে কেবল নোট মুখস্থ করে পাস কর
7	"অভাগীর স্বর্গ-শরৎচন্দ্র চট্টোপাধ্যায়"	"গ্রামের ধনী ব্রাহ্মণ ঠাকুরদাস মুখুয্যের স্ত্রী একদিন সাতদিনের জ্বরে মারা যান।	"ঠাকুরদাস মুখুয্যেরর বর্ষীয়সী স্ত্রী সাতদিনের জ্বরে মারা গেলে গ্রামের সবাই শবযাত্রায় অংশগ্রহণ করে, যেন একটি উৎসব চলছে। মেয়েরা কাঁদতে কাঁদতেই শ্মশুরীর
8	"বই পড়া 2- প্রমথ চৌধুরী"	"বই পড়া শখটা মানুষের সর্বশ্রেষ্ঠ শখ হলেও আমি কাউকে শখ হিসেবে বই প	"লেখক বলেন, আমাদের জাতি সাধারণত বই পড়ার শখী নয় এবং বর্তমান কঠিন সময় ও দারিদ্র্যে বই পড়াকে শখ হিসেবে নেওয়া কঠিন। শিক্ষার আসল উদ্দেশ্য হ
9	"নিরীহ বাঙালি-রোকেয়া সাখাওয়াত হোসেন"	"আমরা দুর্বল নিরীহ বাঙালি। এই বাঙালি শব্দে কেমন সুমধুর তরল কোমল ব	"বাঙালি জাতি সহজ, কোমল, সুমধুর এবং সৌন্দর্যে ভরপুর একটি জাতি। তাদের নাম যেমন সুন্দর, তেমনি তাদের সব কর্ম-ব্যবহারও সরল ও স্বাভাবিক। বাঙালির
10	"পল্লিসাহিত্য-মুহম্মদ শহীদুল্লাহ্"	"পল্লিগ্রামে শহরের মতো গায়ক, বাদক, নর্তক না থাকলেও তার অভাব নেই।	"বাংলার পল্লিগ্রামে শহরের মতো গায়ক-বাদক না থাকলেও প্রকৃতি ও গ্রামের মানুষের জীবন থেকেই সমৃদ্ধ সাহিত্য ও সংস্কৃতি জন্ম নেয় যা অনেক সময় আমরা উপেক্ষ
11	"উদ্যম ও পরিশ্রম-মোহাম্মদ লুৎফর রহমান"	"চাকরি করা কাজ উত্তম, যখন তা হয় জাতির সেবা-যখন তাতে মর্যাদা ও ব্যা	"এই লেখায় বলা হয়েছে, সম্মানজনক ও সৎ উপায়ে উপার্জন করাই শ্রেয়। চাকরি, ব্যবসা বা যেকোনো কাজ তখনই উত্তম, যখন তা ব্যক্তি, সততা ও আত্মমর্যাদা র
12	"জীবনে শিল্পের স্থান-এস ওয়াজেদ আলি"	"জীবনকে সুন্দর করতে হলে সৌন্দর্যের নিদর্শন শিল্পকে সাধারণ জীবনে বিশি	"জীবনকে সুন্দর ও সার্থক করতে হলে শিল্প ও সৌন্দর্যের গুরুত্ব বুঝতে হবে এবং জীবনের প্রতিটি ক্ষেত্রে তা প্রয়োগ করতে হবে। বাড়ির সাজসজ্জা, আসবাবপত্র, বসা
13	"আম আঁটির ভেঁপু - বিভূতিভূষণ বন্দ্যোপাধ্যায়"	"সকাল বেলা। আটটা কি নয়টা। হরিহরের পুত্র আপন মনে রোয়াকে বসিয়া ম	"এই অংশে লেখক এক গ্রামের সহজ-সরল শিশুর দুনিয়া, তার ছোট ছোট খেলা, চুরি করা কড়ির গোপন আনন্দ, এবং দিদি-ভাইয়ের মিষ্টি সম্পর্ককে মনোজ্ঞভাবে তু
14	"মানুষ মুহম্মদ (স.) 1 -মোহাম্মদ ওয়াজেদ আলী"	"হযরতের মৃত্যুর কথা প্রচারিত হইলে মদিনায় যেন আঁধার ঘনাইয়া আসিল।	"হযরত মুহম্মদ (স.)-এর মৃত্যুসংবাদে মদিনায় শোকের ছায়া নেমে আসে, এবং মানুষ বিভ্রান্তিতে পড়ে। তখন হযরত আবুবকর (রা.) সবাইকে ঝাণ করিয়ে দেন যে, হ
15	"মানুষ মুহম্মদ (স.) 2 -মোহাম্মদ ওয়াজেদ আলী"	"মক্কাবাসীরা হযরতের নবির লাভের শুরু হইতেই তাঁহার প্রতি কী নির্মম অমানু	"হযরত মুহম্মদ (স.)-এর জীবনে মক্কাবাসীদের নির্মম অত্যাচার, ষড়যন্ত্র ও ঘৃণার মাঝেও তিনি ক্ষমা, মানবতা ও শান্তির বার্তা বহন করেন। বিজয়ের পরও শত্রুদের মু
16	"নিমগাছ-বনফুল"	"কেউ থালটা ছাড়িয়ে নিয়ে সিদ্ধ করছে। পাতাগুলো ছিঁড়ে শিলে পিষছে কেউ (	"নিমগাছ সাধারণত ওষুধ, রান্না ও চর্মরোগের উপশমে ব্যবহৃত হলেও কেউ তার সৌন্দর্য দেখে না, যত্নও নেয় না। একদিন এক কবি নিমপাতার রূপে মুগ্ধ হয়ে তার সৌ
17	"উপেক্ষিত শক্তির উন্মেষন - কাজী নজরুল ইসল	"হে মোর দুর্ভাগা দেশ যাদের করেছ অপমান অপমানে হতে হবে তাহাদের স	"এই গদ্যে লেখক উপেক্ষিত 'হোটলোক' শ্রেণির প্রতি সমাজের অবহেলা ও শোষণের চিত্র তুলে ধরেছেন। তিনি বলেন, জাতির প্রকৃত শক্তি এই জনসাধারণ, যাদের জ
18	"শিক্ষা ও মনুষ্যত্ব- মোতাহের হোসেন চৌধুরী"	"মানুষের জীবনকে একটি দোতলা ঘরের সঙ্গে তুলনা করা যেতে পারে। জীবন	"মানুষের জীবন উন্নয়নের জন্য শিক্ষাকে জীবনজ থেকে মানবসত্তার পথে উত্তরণের সিঁড়ি হিসেবে দেখা হয়েছে। শুধু অক্ষরজ্ঞান চাহিদা পূরণ যথেষ্ট নয়; মানুষের মনুষ্য
19	"প্রবাস বন্ধু 1-সৈয়দ মুজতবা আলী"	"বাসা পেলুম কাবুল থেকে আড়াই মাইল দূরে খাজামোয়া গ্রামে। বাসার সঙ্গে ১	"এই অংশে লেখক তার কাবুলের বাসা জীবনের এক মনোরম, হাস্যরসপূর্ণ চিত্র তুলে ধরেছেন। ঘরভাড়া, নতুন চাকর আবদুর রহমানের সঙ্গে পরিচয় এবং তার দৈহি

Fig 4.1: Dataset Sample

4.1.1

DataPre-processing

The text data was tokenized using a pretrained tokenizer corresponding to the model being used (BanglaT5 or Bangla T5-small). Both the stories (as input) and their summaries (as target) were tokenized with a maximum input length of 512 tokens and maximum target length of 128 tokens, with truncation and padding applied as needed. The dataset was wrapped into the Hugging Face Dataset object to allow seamless integration with the training pipeline.

a. Data labelling

Each record in the dataset was structured into two fields:

- Text: the full body of the short story
- Summary: the corresponding human-written abstractive summary

b. Data Cleaning

Text normalization was performed to remove unnecessary characters, punctuation noise, redundant newlines, and formatting artifacts. Unicode issues were resolved, and all texts were standardized in UTF-8 format. Additionally, a filtering process excluded overly short or irrelevant samples.

4.2

Model Architecture and Training

This study employed a Bangla-adapted version of the T5 model for abstractive summarization. T5 (Text-To-Text Transfer Transformer) treats every NLP task as a text-to-text problem. We used a multilingual variant — Bangla T5-small, which supports Bangla and has been pre-trained on multilingual data.

Bangla T5 (such as `csebuetnlp/banglat5`) is based on the Bangla T5-small architecture. **Bangla** T5 model focuses on specially on Bengali language.And it has deep understanding of syntax, style.

Model Used: Bangla T5 (based on Bangla T5-small)

- Transformer-based encoder-decoder architecture
- Fine-tuned on the collected Bangla short story dataset
- Input format: summarize: <Bangla story>
- Output: generated abstractive summary in Bangla

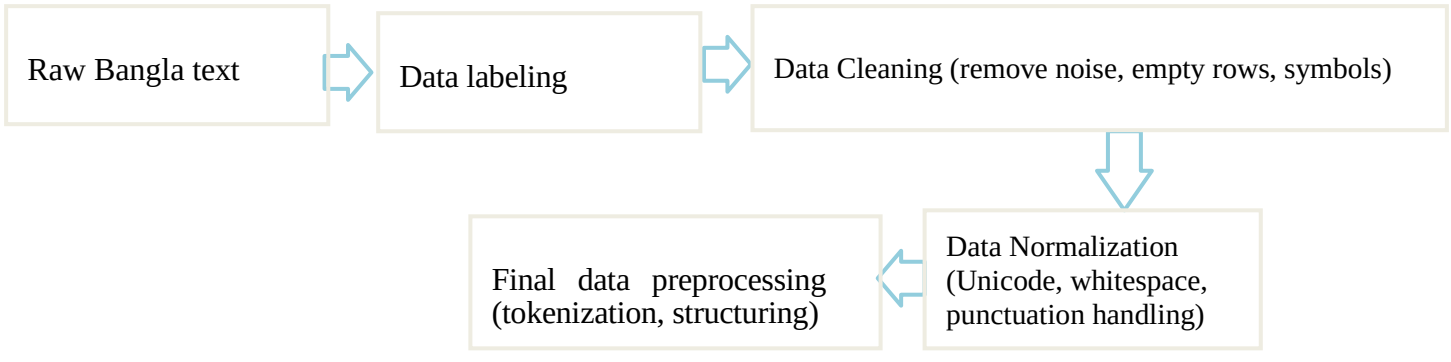


Fig 4.2: Data preprocessing

4.3

Bangla T5 Model Architecture

The Bangla T5 model is based on Google's T5 (Text-to-Text Transfer Transformer) architecture, which is an encoder-decoder transformer that handles all NLP tasks in a text-to-text format. In our thesis, a multilingual variant (Bangla T5-small) is used and fine-tuned specifically on Bangla short story summarization tasks. It converts all tasks into text data and perform its operation.

Key Components:

1. **Input Format**
   Every task is framed as:
   summarize: <Bangla story text>

This prompt guides the model to treat the task as summarization.

2. **Tokenize**r
   The input text is tokenized using the Sentence Piece tokenizer trained on multilingual text. This breaks Bangla text into subword tokens.

3. **Encoder**
   The encoder consists of multiple transformer layers (12 in Bangla T5-small). It processes the input tokens and produces contextualized hidden representations.

4. **Decoder**
   The decoder takes the encoder output and generates the target tokens (summary), one by one, attending to both previously generated tokens and the encoder's outputs.
5. **Linear + Softmax Layer**
   The decoder's output is passed through a linear layer and a softmax to generateprobabilities for each token in the vocabulary.

6. **Output (Detokenizer)**

The output token IDs are detokenized back into Bangla text the abstractive summar

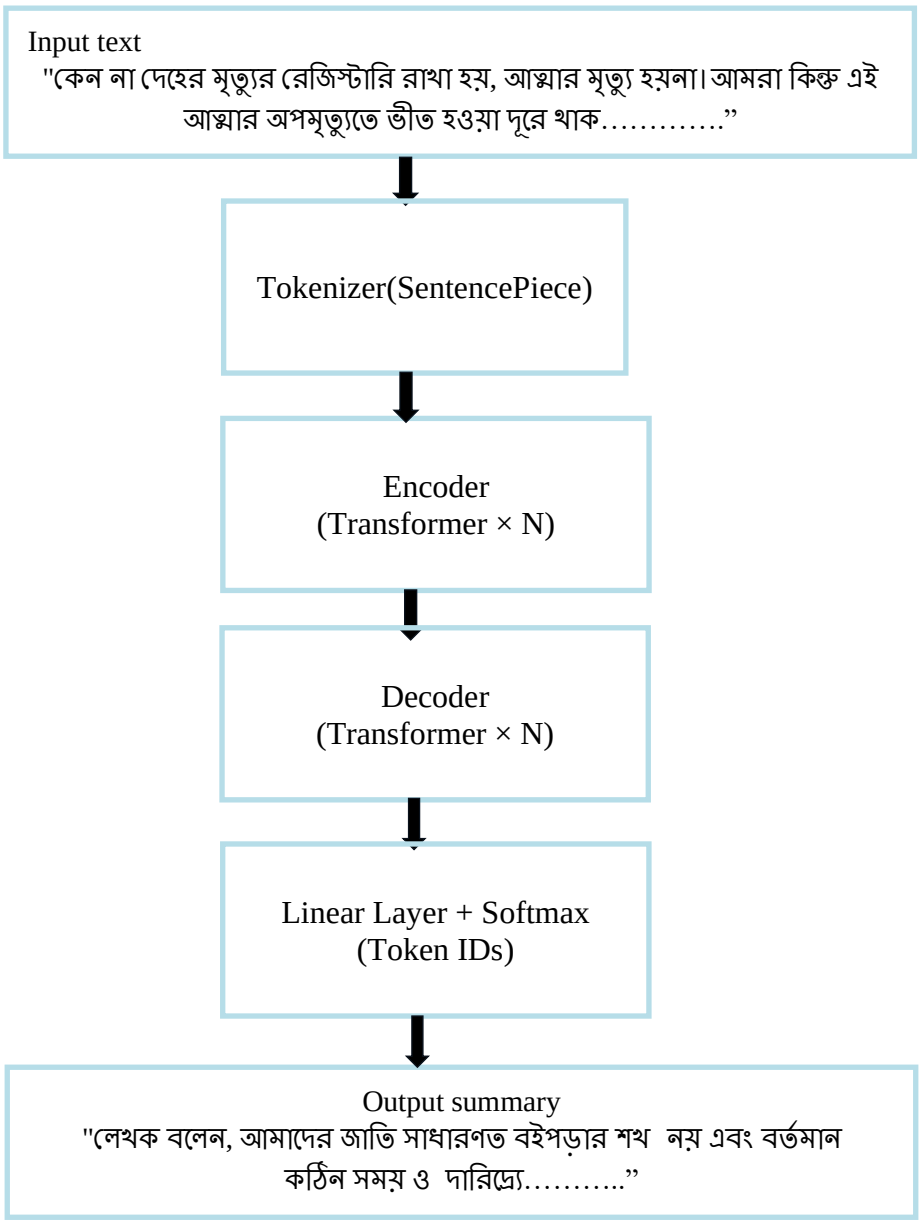


Fig 4.3: T5 Architecture Diagram

4.4

Model training

Dataset split:

Table 4.2: Dataset spliting

Total number of rows	430
Number of training data (70% of total data)	301
Number of test data (30% data)	84
Validation data	45

4.5

Experimental Setup

- ➔ **ProgrammingLanguage**: Python
- ➔ **DeepLearningFramework**: PyTorch, Tokenizer
- ➔ **Tools**: Hugging face transformer
- ➔ **Hardware**: NVIDIAGPU- enabled environment (RTX-9800)
- ➔ **Environment**: visual studio code
- ➔ **Training Parameters**:
    - ◆ Epochs:30
    - ◆ Optimizer: Adam
    - ◆ Batchsize:2
    - ◆ Learning rate: 0.00005
    - ◆ Gradient accumulation step: 2

:

4.6

Performance Evaluation

The model was evaluated using Bert score to compare machine-generated summaries with

reference summaries.

To evaluate model effectiveness, the following metrics were used:

- Precision(P)
- Recall(R)
- F1-Score(F1)

4.6.1

Mathematical Formulas:

Let TP,TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

$$\text{Precision} = \frac{TruePositive(TP)}{TruePositive(TP)+FalsePositive(FP)}$$

$$\text{Recall} = \frac{TruePositive(TP)}{TruePositive(TP)+FalseNegative(FN)}$$

$$\text{F1score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Chapter 5

Result and Discussion

In this research, a transformer-based Bangla-T5 model was fine-tuned for the task of Bangla short story summarization. The model was trained for a total of 30 epochs. During training, both training and validation losses showed a consistent downward trend, indicating effective learning by the model. Although some fluctuations were observed in the early epochs, from around the 20th epoch onward, the validation loss stabilized, reaching its lowest point of 1.3126 at epoch 28, suggesting that the model was generalizing well without significant overfitting.

To evaluate the quality of the generated summaries, BERT Score was used. The model achieved a Precision of 0.5927, Recall of 0.5398, and F1-score of 0.5644. While these values are moderate, they are considered reasonably good given the limitations of available Bangla language resources and the relatively small dataset size used in this work.

The generated summaries were found to be relevant and context-aware, capturing the core message of the stories. However, in some instances, the summaries exhibited repetition or lacked clarity. These issues could potentially be addressed through improved subword tokenization, more refined dataset cleaning, and advanced decoding techniques such as beam search,

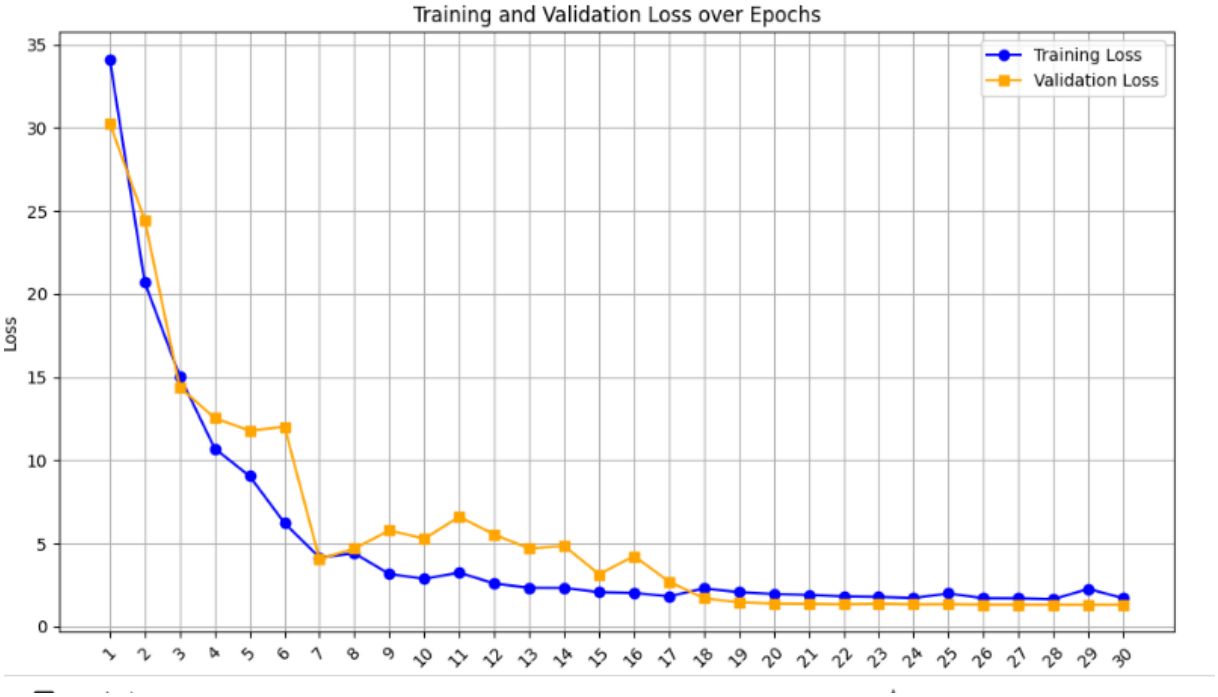


Fig 5.1: Training loss and validation loss over epochs

Table 5.1 : Result of the work

Evaluation	**Value**
Precision	0.5927
Recall	0.5398
F1 score	0.5644

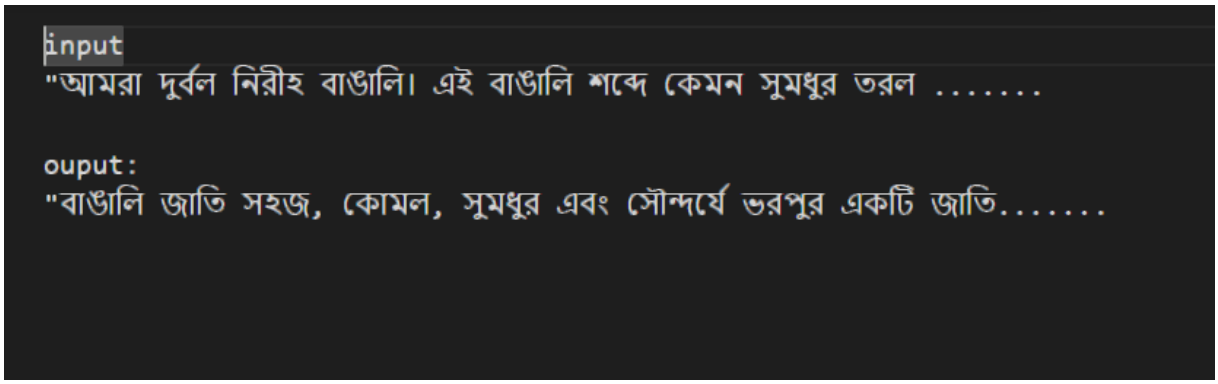


Fig. 5.2: Sample input and output

Chapter 6

Conclusion and Future Work

Limitations & Future works:

The study faced several challenges, including a relatively small dataset of only 430 stories, limited computational resources for fine-tuning large models like Bangla T5, low bert score percentages, and difficulty in accurately summarizing poetic or stylistic expressions common in Bengali literature.

To overcome those problems following points are necessary for the future works we think:

 Collecting more Bengali short stories written by various authors to improve the model's learning capacity.

 Include summaries written by experts from native speakers to ensure quality.

 While our study utilized a system with a dedicated 16 GB **GPU**, 32 GB shared GPU memory, and 64 GB RAM, fine-tuning large models like Bangla T5 still posed computational challenges. Utilizing more powerful hardware or adopting parameter-efficient training techniques could reduce resource requirements.

 Compare BERTScor with human evaluations for better accuracy assessment.

 Research in Low-Resource NLP Tools such as, POS taggers, tokenizers, and syntactic parsers tailored for Bengali language.

 Use reinforcement learning techniques or human-in-the-loop systems to refine the output and improve summary naturalness.

 Finally, to better catch the poetic and complex texts common in Bengali literature, future models could be fine-tuned specifically on such content or supported with stylistic data augmentation techniques.

Conclusion

In this research, we aimed to develop an abstractive summarization system for Bengali short stories, a domain that has received limited attention compared to high-resource languages. To address the scarcity of datasets and pre-trained models for Bengali, we created a custom dataset comprising 430 Bengali short stories along with corresponding human-written summaries. The Bangla T5 model was fine-tuned on this dataset, and its performance was evaluated using BERT scores. While the model demonstrated some ability to extract key information and generate coherent summaries, the overall quality of the abstractive summaries remained limited—largely due to the small dataset size. In future work, we plan to expand the dataset by including more stories and high-quality human-generated summaries. We believe that with more data and extended research efforts, the quality and effectiveness of the generated summaries will significantly improve. Overall, this study lays the groundwork for continued research in NLP for Bengali and other low-resource languages, particularly in the field of literary text summarization.

Reference

1. Radia Rayan Chowdhury et al., "Unsupervised Abstractive Summarization of Bengali Text Documents,"*arXiv preprint arXiv:2102.04490*, 2021. [Online]. Available: https://arxiv.org/abs/2102.04490

2. Sheikh Abujar et al., "A heuristic approach of text summarization for Bengali documentation,",in Proc. 8th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), IIT Delhi, India, Jul. 2017

3. Vahid Nejad Mahmood Abadi & Fahimeh Ghasemian, "Enhancing Persian text summarization through a three-phase fine-tuning and reinforcement learning approach with the Bangla T5 transformer model, "Scientific Reports, vol. 14, no. 1, pp. 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-78235-3

4. Md. Iftekharul Alam Efat et al. "Automated Bangla Text Summarization by Sentence Scoring and Ranking," in *Proc. 2013 Int. Conf. Informatics, Electronics & Vision (ICIEV)*, Dhaka, Bangladesh, 2013. doi: 10.1109/ICIEV.2013.6572687.

5. Pushkar Joglekar et al.," Text Summarization in local language," in *Proc. 2025 IEEE Int. Students' Conf. on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, 2025. doi : 10.1109/SCEECS64059.2025.10940704.

6. Fahmida Afroja Hoque Barsha and Mohammed Nazim Uddin ," Comparative Analysis of BanglaT5 and Pointer Generator Network for Bengali abstractive story summarization," in *Proc. 2023 Int. Conf. on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Dhaka, Bangladesh, 2023. doi: 10.1109/ICICT4SD59951.2023.10303633.

7. DIftekharul Mobin et al.," A Review of the State-of-the-Art Techniques and Analysis of Transformers for Bengali Text Summarization," *Big Data and Cognitive Computing*, vol. 9, no. 5, p. 117, Apr. 2025. doi: 10.3390/bdcc9050117.

8. Kanta Prasad Sharma et al.," A Systematic Review on Text Summarization: Techniques, Challenges, Opportunities," *Expert Systems*, vol. 42, 2025, Art. no. e13833, doi: 10.1111/exsy.13833.

9. Tahmid Hasan et al., “Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation,” in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.207

10. Linting Xue et al.," Bangla T5: A Massively Multilingual Pre-trained Text-to-Text Transformer," *arXiv preprint* arXiv:2010.11934, Mar. 2021. [Online]. Available: