BENGALI TEXT SUMMARIZATION: A HYBRID METHODOLOGY USING SEQUENCE TO SEQUENCE RNNS

BY

TASKIN KHALEQUE
ID: 221-25-085

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

Supervised By

Sheak Rashed Haider Noori, PhD
Professor & Associate Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2022

APPROVAL

This Project/Thesis titled **"BENGALI TEXT SUMMARIZATION: A HYBRID METHODOLOGY USING SEQUENCE TO SEQUENCE RNNS"**, submitted by Taskin Khaleque, ID No: 221-25-085 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 17-01-2023.

BOARD OF EXAMINERS

17/01/23

Chairman

Dr. Touhid Bhuiyan, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

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

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Safaet

Md. Safaet Hossain
Associate Professor & Head
Department of Computer Science and Engineering
City University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Associate Head, Department of Computer Science and Engineering,** 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:

15.01.2023

Dr. Sheak Rashed Haider Noori
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted by

Taskin Khaleque
ID: 221-25-085
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

I have given my efforts to this thesis. However, it would not have been possible without the kind support and help of many individuals. I would like to express my deepest appreciation to all those who provided me the possibility to complete this report.

At first, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessings which allowed me to complete this thesis successfully.

I would like to express my grateful appreciation for **Dr. Sheak Rashed Haider Noori, Professor & Associate Head**, Department of Computer Science & Engineering, Daffodil International University for being my adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped me to do my thesis work in proper way.

I also give my deepest thanks to all the faculty members and staff of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In modern world, technology has transformed our lives for the better. However, human attention spans are shortening, and the amount of time people want to spend reading is dwindling at an alarming rate. As a result, it's critical to provide a quick overview of key news or article by creating a brief summary of the most important news piece, as well as the most intuitive summary in accordance with the synopsis. There are enormous amounts of textual data available in this era of information. Online documents, articles, news, and customer evaluations of various goods and services are a few examples of sources. The purpose of document summarizing is to find the core meaning of the original material. However, it is impossible to create custom summaries for such a vast supply of text documents. Humans have the ability to make abstraction by reading a article. However, summarizing using computer is always a hard problem. Abstractive text summarization is used to improve the topic coverage of automatic summaries by paying more attention to the semantics of the words and experimenting with rephrasing the input sentences in a human-like manner improve soundness and readability. Although there has been a lot of prominent study on abstractive summary in the English language, there have only been a few publications on Bengali abstractive news summarization (BANS). In this thesis, we proposed a hybrid model for extracting summary from long articles that combines both extractive and abstractive approaches. In the extractive part, BERT (BERTSUM) is used to find the most relevant sentences from the document then using sequence to sequence (seq2seq) based bidirectional Long Short-Term Memory (LSTM) network model with attention at encoder-decoder to generate the summary. Experiments were carried out using publicly available Kaggle datasets (Bengali newspaper dataset). The results verify our method and show that the suggested hybrid model produces a compact and engaging summary. We evaluated our summaries by observing its generative performance. 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 0.018 and able to generate a fluent short summary note from a given text.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

SHORT FORM	**ABBREVIATION**
NLP	Natural Language Processing
RNN	Recurrent Neural Network
NLTK	Natural Language Tools Kit
LSTM	Long Short-Term Memory
BNS	Bengali News Summarization
Sequence to Sequence Model	Seq2seq
Bi-directional LSTM	Bi-directional Long Short-Term Memory
DL	Deep Learning
ML	Machine Learning

CHAPTER 1
INTRODUCTION

1.1 Introduction

Text summarization is the task of creating a concise and fluent summary of a larger document, while preserving the most important information and overall meaning. It is a useful tool for a variety of applications, including information retrieval, content analysis, and automated essay grading. There are various approaches to text summarization, including extractive and abstractive methods. Extractive summarization involves selecting key phrases and sentences from the original text, while abstractive summarization involves generating new phrases and sentences that capture the meaning of the original text.

Nowadays, technology is rapidly changing, and everything can be accomplished in the blink of an eye. The majority of individuals retrieve information using technology such as smartphones, computers, and others. In addition, studying newspapers online is one of them, so before studying any kind of information article, a meaningful summary is crucial to pique the appetite of any user to read this specific information. The demand for automatic text summarization systems is spiking these days. The extractive approach aims to extract the most important phrases or sentences from the entire manuscript, whereas the abstractive techniques try to paraphrase new words and phrases which is already available in the content. It generates a brief summary of the original text that focuses on the important points. The generated summaries can also additionally encompass extra terms and sentences now no longer discovered in the authentic text. Abstractive precis makes use of deep learning-primarily based totally strategies for era of summaries. Here, the sentences generated from it is able to or might not be accurate. The major challenge for the seq2seq model is to tackle a document. Moreover, a large document will hamper the performance of the existing seq-2seq model.

In this research, we recommend a pipeline technique that use both extractive and abstractive approaches. The suggested hybrid model consists of an extractive mechanism that extracts important sentences and phrases from the document and apply a reinforced abstractive mechanism that uses the extractive mechanism's key sentences/phrases to construct a succinct summary.

1.2 Motivation

Text summarization is a part of Natural Language Processing. It is also considered the most challenging as well as interesting job which process a concise and meaningful summary of the text. It is time consuming to read the whole document in order to get the proper meaning of the document. As a result, to find the exact meaning is quite challenging. To overcome this kind of problem, automatic text summarization is a must which summarizes the text while also counting the number of documents, words, and frequently used words. The world we live in is becoming increasingly intertwined with science and technology and will continue to be so for the foreseeable future. Now, we spend a lot of time in the internet reading article, books, web pages, newspaper and so on. Moreover, sometime we get bored to find the necessary information among those article due to unstructured data and dizzy meaning.

In this modern era, many tools related to NLP has been developed for languages such as English, Spanish. However, we have left behind due to not having advanced research in this field for Bengali language. That is why we should come forward to contribute in this field. In NLP, there are many core problems and summarization in one of them. An advanced text summarization can help us in understanding the meaning of the long text in a short time. For language like, English, Spanish etc. they have developed advanced NLP tools and design model for summarization purposes but we have limited tools that can't support the summarization problem. Without proper research, it is impossible to find the solution of this problem. Therefore, in our research work, we recommend a technique by integrating both extractive and abstractive mechanisms using a pipelined approach to summarize Bangla documents.

When it comes to share our emotion and expressing ourselves, Mother language is always preferred. In our Bengali language, NLP resources are not adequate so we need to focus on this area and new technologies must be developed.

1.3 Rationale of the Study

The use of natural language processing (NLP) techniques for text summarization has been an active area of research in recent years, with a growing interest in applying these methods to under-resourced languages like Bengali. However, there are still many challenges to overcome in order to develop effective text summarization systems for Bengali.

Language is a structured system of communication by which people can share their thoughts and understanding of issues. Language consists of grammar and vocabulary. It is considered the primary means of communication of humans, and can be conveyed through spoken, sign, or written language. Due to the continuous evolution of humankind and the geographical distribution of human civilization, numerous languages are created [11]. Bangla is a widely used

Language with native speakers from Bangladesh. However, western part of India also use Bangla as their regional language. Considering the number of speakers and demographical perspective, Bangla is an essential language in the world [12].

Bangla ranked the sixth most widely used language globally, with 268 million native speakers around the globe. It is the mother tongue of Bangladesh and also a regional language of India.

This vast amount of digital content in Bangla is an excellent source of qualitative and quantitative analysis [4]. Natural Language Processing (NLP), a branch of computer science that deals with computational analysis and of text and speech. Some of the example of NLP's are text summarization, question answering, and fact identification.

Furthermore, this study aims to propose a novel seq2seq based approach using the state-of-the-art models such as transformer based models, which have been proven to be effective in various NLP tasks such as machine translation, text summarization and have a good performance on other languages, but the effectiveness of these models on Bengali is yet to be proven.

In summary, the rationale of this study is to develop an effective text summarization system for Bengali, addressing the challenges associated with the complexity of the Bengali language and lack of labeled data. Additionally, this study will propose a novel approach based on the state-of-the-art models.

1.4 Research Questions

- What is Bangla Text Summarization?
- How does summarization work?
- What are the advantages of summarization?
- How to preprocess unstructured Bengali data?
- How extractive and abstractive work?
- How does Bengali text summarization Model works?

1.5 Expected Output

Our main interest is to use a hybrid approach to summarize Bangla text. Data compressing for Bangla language is a new study. In general, there are two kinds of approaches for extracting information from raw text input and using it for a summarization model: extractive and abstractive. Extractive approaches only choose the most significant sentences from a text (without necessarily understanding their meaning), the result summary only includes a portion of the original content and abstractive models make use of more sophisticated NLP to comprehend the semantics of the text and produce a useful summary. Our purpose for our research is to amalgamate both the approaches to maintain a remarkable efficiency of this method. In this research paper, we try to explain our ideas on the execution phase in order to improve precision and reduce total loss when developing the model.

1.6 Report Layout

This report is divided into five chapters.

The first chapter provides an overview of the entire project. There are numerous sections such as Introduction, Motivation, Rationale of the Study, Research Questions, Expected Output, and Report Outline Research.

In chapter 2, we will discuss about the background of our research topic.

In chapter 3, we will discuss about the methodologies employed in our study.

In chapter 4, we will discuss about the obtained results and discussion.

In chapter 5, we will discuss about the conclusion.

CHAPTER 2
BACKGROUND

2.1 Introduction

In general, there are two kinds of approaches for extracting information from raw text input and using it for a summarization model: extractive and abstractive.

Extractive summarization: The process of identifying important sections of the text and concatenating them to form a summary. This is done by selecting sentences or phrases from the original text that are most representative of its meaning. Extractive summarization is typically seen as a simpler task than abstractive summarization, as it doesn't require the generation of new phrases or sentences. However, this approach can lead to a summary that is not as coherent or fluent as one produced by an abstractive method.

Abstractive summarization: Abstractive models make use of more sophisticated NLP (i.e., word embeddings) to comprehend the semantics of the text and produce a useful summary. Because they require several parameters and data, abstractive approaches are consequently significantly more difficult to train from scratch.

In this approach, summary is generated from scratch based on the input information. Abstractive machine learning algorithms, which can construct new sentences or phrases which represent the most significant information from the source text, can help overcome the grammatical flaws of extraction techniques.

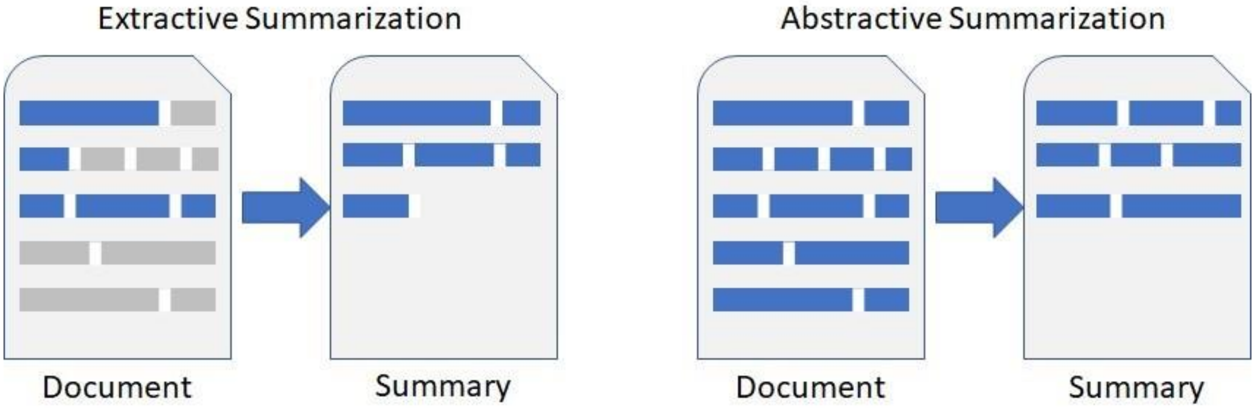


Figure 2.1: Extractive & Abstractive Summarization

2.2 Literature Review

The text summary is a short condensed note on a large text document [6]. There exist various kinds of abstractive summarization approaches stated in Yeasmin et al. [1]. In 2017, Abujar et al. [2] demonstrated a heuristic approach summarizing for Bengali language that deals with extractive method, suggest a new type of sentence scoring procedure, and define a set of text analysis rules based on the heuristics for the first time. Following that, Talukder et al. [3] introduced a model that used bi-directional RNNs with LSTM in the encoding layer and an attention model in the decoding layer to minimize train loss to 0.008. However, due to a shortage of Bengali dataset, the model was only trained using a little amount of data. Aside from that, the model generates summaries with a finite number of words and hence cannot manage a text with an endless number of words. P. Bhattacharjee et al. [4] compiled a dataset of over 19k articles and their human-written summaries from bangla.bdnews24.com in 2020, and also proposed a LSTM-RNN based attention model, in which, it applied attention to both the encoder and the decoder layer. And, (Abujar et al., 2020) [5] propose the extension of the text generation approach for the Bengali language. J. Tan et al. [6], investigation towards neural abstractive summary generation significantly improves the performance of neural sentence summarization models for English language but it has not been definitively proven because of the challenge that remains end-to-end method based on understanding the whole document.

Again, Abujar et al [9] demonstrated that Word2vec is essential for text summarization. It saves all related words as a numeric number, which is useful when working with significant or non-important values in an LSTM cell. They use a medium dataset to build word embedding in Bengali language, and create a nice Bengali Word2vec file. The use of a word2vector technique in the context of text summarization for the Bengali language is explored in this work.

In study Dhar et al. [10] proposed a hybrid pointer generating network that addresses the problems of inaccurately replicating factual facts and phrase repetition. They use a hybrid pointer generator network that can produce words out of vocabulary and improve accuracy in recreating real information, as well as a coverage mechanism that inhibits repetition, to supplement the attention-based sequence. The provided approach outperforms the Bengali state-of-the-art in both qualitative and quantitative evaluations of the model.

2.3 Research Summary

During our research, we have studied both extractive and abstractive mechanism and find out the benefits and efficiency of both the approaches. At first we are more interested in the deep leaning model. However, it cannot handle a long document. So, we designed a hybrid model. Our dataset is a publicly available dataset which contains more than 100000 newspaper article and their corresponding summary. There are a few others column in our dataset that we have cleaned during the pre-processing stage. Pre-processing text is necessary for any summarization work. This stage includes splitting the text and remove stop words. Then we have inserted the Bengali contractions and counted the vocabulary of the dataset. W2V provides a numeric value in the vocabulary file. We have used a pre-trained w2v file. We have used both extractive and abstractive mechanisms using a pipelined approach to produce a concise summary. In the extractive part, we have used pre-trained model BERT and in the abstractive part a chain model based on attention model. This model includes an encoder and decoder that employ bi-directional LSTM cells. At first, we provide the input vector as an input of the encoder layer and receive another related word vector as output. When passing the sequence special token such as PAD's, UNK's used.

2.4 Challenges

There are several challenges associated with using a seq2seq model for text summarization in the Bengali language:

Lack of large labeled datasets: Bengali is a relatively less-resourced language and obtaining large labeled datasets for training and evaluating text summarization models can be a significant challenge. This can make it difficult to train high-performing models and may limit the ability to compare the performance of different approaches.

Complex morphological and syntactic structure: Bengali language has complex morphological and syntactic structure that can make it challenging to accurately understand the meaning of the text. This complexity can make it difficult for seq2seq models to accurately capture the meaning of the text and generate fluent and coherent summaries. Handling of non-standard words and expressions: Bengali has many non-standard words and expressions that are used in colloquial language, these are often found in texts from social media or informal conversation, it can be a challenge for seq2seq models to understand and accurately summarize these texts.

Handling of code-mixed text: Bengali is often used as a code-mixed language with English, this can make it challenging for seq2seq models to accurately understand the meaning of the text and generate fluent and coherent summaries as it might require additional preprocessing steps to handle code-mixing.

Also, in the pre-processing stage, some coding is necessary to clean the text data. For instance, Unicode is required when deleting punctuation from text, and there is only one way to handle it properly that is raw coding. The second is to remove stop words. A large structured dataset is another challenge. In last, a huge data collection can yield a vast vocabulary, and a broad vocabulary facilitates the creation of an ideal summary.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this section, we'll outline the entire research process. As each research work is different in terms of the methods or techniques used to solve it. The methodology includes all the methods that were used in the research work and, providing a brief description of each component.

Figure-3.1 indicates the workflow of our version.

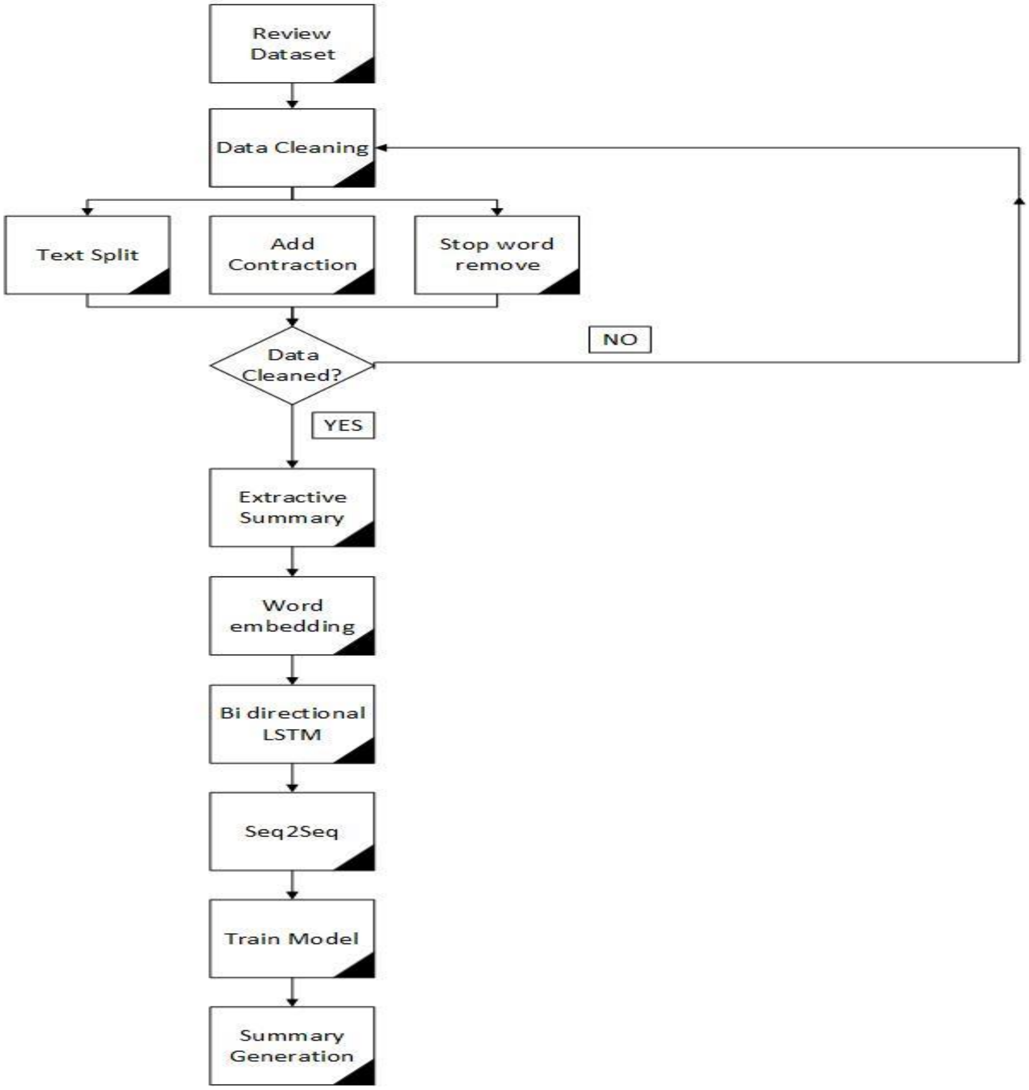


Figure 3.1: Process workflow

3.2 Research Subject and Instrumentation

The title of our research topic is "Bengali Text Summarization: A Hybrid Methodology Using Sequence To Sequence RNNs". In Bengali, NLP it is a major research area. A high configuration PC with GPU and other instruments are required in a deep learning model. This model's needed instruments are listed below.

Table 3.1Software and Tools

Hardware and Software	Development Tools
Core i7	Python 3.7
1 TB HDD	TensorFlow 1.15
Google Colab with 12 GB GPU and 4000 GB RAM	NLTK
	Pandas
	Numpy
	Windows OS

3.3 Data Collection and Preprocessing

Larger dataset can make the model more accurate and, for our model we need a handsome amount of data so we have used a dataset of 100k Bengali news articles from bdnews24. In this dataset there are eight columns. However, we have kept only two columns, the article and corresponding summary column.

For data preprocessing, we have followed the following steps

1) Contractions in Bengali are formed when two or more words are combined to form a single word, with one or more letters omitted. Replace contractions with their longer bureaucracy additionally there are a few different instances wherein we’ve eliminated the contractions such as “বি. দ্র.” ,”ড.”, “মোঃ” etc. In this manner we've

got eliminated all of the pointless characters from the dataset. Some example of Bengali contraction are showing in the 3.2 table.

2) By using NLTK(Natural Language Toolkit) in python for Bengali language, we have eliminated the stop words.
3) We have also created a unwanted word list by analyzing different datasets and use it to remove those from the datasets
4) Raw coding to remove punctuation from Bengali text.
5) Handle missing values.

Table 3.2 Contraction List

Short Form	Long Form
"বি.দ্র "	"বিশেষ দ্রষ্টব্য"
"ড."	"ডক্টর"
"ডা."	"ডাক্তার"
"ইঞ্জি:"	"ইঞ্জিনিয়ার"
"রেজি:"	"রেজিস্ট্রেশন"
"মি."	"মিস্টার"
"মু."	"মুহাম্মদ"
"মো."	"মোহাম্মদ"

Figure 3.2 shows the text cleaning steps

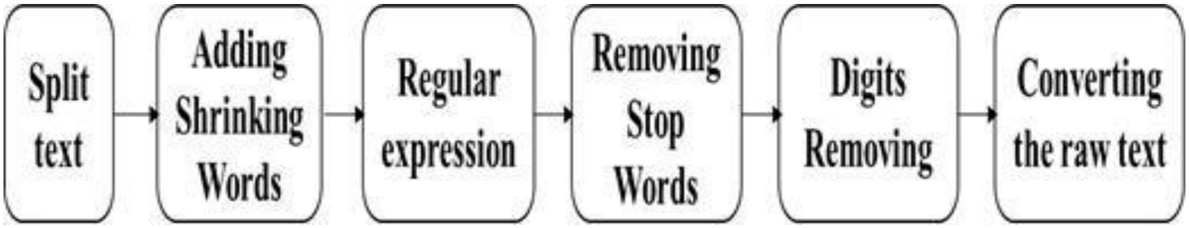


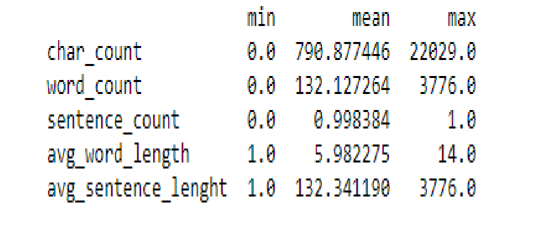
Figure 3.2: Data Preprocessing

Figure 3.3 shows the dataset before and after the cleanup process.

Table 3.3 Clean text & summary

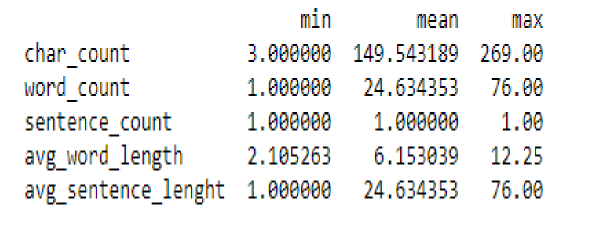
Original Article	Clean Article	Original Summary	Clean Summary
একুশে পদকপ্রাপ্ত বরেণ্য নজরুল সংগীতশিল্পী ও স্বাধীনতা সংগ্রামের কণ্ঠযোদ্ধা শাহীন সামাদের উপস্থাপনায় আসছে বিজয় দিবস উপলক্ষে বাংলাদেশ টেলিভিশনে প্রচারের জন্য নির্মিত হয়েছে বিজয়ের গানের বিশেষ সংগীতানুষ্ঠান বিজয় নিশান উড়ছে ঐ মাহবুবা ফেরদৌসের প্রযোজনায় অনুষ্ঠানে সংগীত পরিবেশন করেছেন মার্লিন, রাশেদ, অপু, সাব্বির, রন্টি দাশ, লুইপা, প্রিয়াংকা, সুস্মিতা, নন্দিতা ও বাঁধন মার্লিন গান গেয়েছেন জন্ম আমার ধন্য হলো মাগো, সাব্বির গেয়েছেন যে মাটির বুকে ঘুমিয়ে আছে, রাশেদ গেয়েছেন মাগো ভাবনা কেন ও রন্টি গেয়েছেন ও মাঝি নাও ছাইড়া দে মাঝি পাল উড়াইয়া দে অনুষ্ঠানে উপস্থিত ছিলেন সংগীত পরিচালক সুজেয় শ্যাম	একুশে পদকপ্রাপ্ত বরেণ্য নজরুল সংগীতশিল্পী স্বাধীনতা সংগ্রামের কণ্ঠযোদ্ধা শাহীন সামাদের উপস্থাপনায় আসছে বিজয় দিবস উপলক্ষে বাংলাদেশ টেলিভিশনে প্রচারের নির্মিত হয়েছে বিজয়ের গানের সংগীতানুষ্ঠান বিজয় নিশান উড়ছে মাহবুবা ফেরদৌসের প্রযোজনায় অনুষ্ঠানে সংগীত পরিবেশন মার্লিন রাশেদ অপু সাব্বির রন্টি দাশ লুইপা প্রিয়াংকা সুস্মিতা নন্দিতা বাঁধন মার্লিন গান গেয়েছেন জন্ম ধন্য মাগো সাব্বির গেয়েছেন মাটির বুকে ঘুমিয়ে রাশেদ গেয়েছেন মাগো ভাবনা রন্টি গেয়েছেন মাঝি নাও ছাইড়া দে মাঝি পাল উড়াইয়া দে অনুষ্ঠানে উপস্থিত সংগীত পরিচালক সুজেয় শ্যাম	একুশে পদকপ্রাপ্ত বরেণ্য নজরুল সংগীতশিল্পী ও স্বাধীনতা সংগ্রামের কণ্ঠযোদ্ধা শাহীন সামাদের উপস্থাপনায় আসছে বিজয় দিবস উপলক্ষে বাংলাদেশ টেলিভিশনে প্রচারের জন্য নির্মিত হয়েছে বিজয়ের গানের বিশেষ সংগীতানুষ্ঠান বিজয় নিশান উড়ছে ঐ মাহবুবা ফেরদৌসের প্রযোজনায় অনুষ্ঠানে সংগীত পরিবেশন করেছেন মার্লিন, রাশেদ, অপু,	একুশে পদকপ্রাপ্ত বরেণ্য নজরুল সংগীতশিল্পী স্বাধীনতা সংগ্রামের কণ্ঠযোদ্ধা শাহীন সামাদের উপস্থাপনায় আসছে বিজয় দিবস উপলক্ষে বাংলাদেশ টেলিভিশনে প্রচারের নির্মিত হয়েছে বিজয়ের গানের সংগীতানুষ্ঠান বিজয় নিশান উড়ছে মাহবুবা ফেরদৌসের প্রযোজনায় অনুষ্ঠানে সংগীত পরিবেশন মার্লিন রাশেদ অপু

Table 3.4 Text Length Summary



```
                    min        mean      max
char_count          0.0  790.877446  22029.0
word_count          0.0  132.127264   3776.0
sentence_count      0.0    0.998384      1.0
avg_word_length     1.0    5.982275     14.0
avg_sentence_lenght 1.0  132.341190   3776.0
```

Article Overview



```
                         min        mean    max
char_count          3.000000  149.543189  269.00
word_count          1.000000   24.634353   76.00
sentence_count      1.000000    1.000000    1.00
avg_word_length     2.105263    6.153039   12.25
avg_sentence_lenght 1.000000   24.634353   76.00
```

Summary Overview

3.4 Executional Requirement

Firstly we have used pre-trained model BERT to find the most important sentences from our dataset and then feed the extractive data to our propose model to get an abstractive Bengali text summarizer.

BERT, or Bidirectional Encoder Representations from Transformers, a method of pre-training language representations which obtains state-of-the-art results on a large array of tongue Processing (NLP) tasks and considers the context from either side (left and right) of a word. RNN, Attention mechanisms, and Transformers are utilized in BERT to interpret human languages. BERT extractive summarization allows for control over the number of sentences and characters used in the summary. BERT was trained using only a noticeable text corpus.

Sequence-to-sequence (seq2seq) models are a type of neural network architecture that are commonly used for a variety of natural language processing tasks, including text summarization.

The basic idea behind a seq2seq model is to use two recurrent neural networks (RNNs) -- one as the encoder and the other as the decoder. The encoder reads the input sequence, such as a paragraph of text, and generates a fixed-length context vector that is then passed to the decoder. The decoder uses this context vector to generate the output sequence, such as a summary of the input text.

Here's an outline of the steps that seq2seq use to train a seq2seq model for text summarization:

1. Collect and preprocess a dataset of input-output pairs, where the input is the text to be summarized and the output is the summary.

2. Tokenize the input and output text into word or sub-word units, and build a vocabulary of tokens.

3. Create a seq2seq model with an encoder and a decoder, both of which are RNNs. The encoder should be a bidirectional RNN that reads the input sequence in both forward and backward directions. The decoder should be a unidirectional RNN that generates the output sequence.

4. Train the model by feeding it the input-output pairs, and using a suitable loss function to calculate the difference between the predicted output and the true output.

5. Once the training is done, we use the trained model to generate summaries of new input text by feeding it through the encoder and using the decoder to generate the summary.

6. Finally, fine tune the model using Hyperparameter tuning and Beam Search technique to get best results.

Apart from the above steps, we employed bi-directional RNN with LSTMs to construct our encoding layer. The outputs must be concatenated because we are employing a bi-direction RNN, and Luong for my attention style has been employed. By using it, the model may learn more quickly and give better outcomes.

In order to create bidirectional recurrent neural networks (RNNs), the first recurrent layer of the network is duplicated, resulting in two layers lying side by side. Afterward, giving the input sequence to the first layer in its original form as input and giving the second layer a reversed version of the input sequence. Long STM networks or LSTMs are Neural Networks that are utilized in an exceedingly spread of tasks. With LSTM, this method has been shown to be highly effective. Because there is evidence that the context of the entire utterance is used to interpret what is being said rather than a linear interpretation, the usage of giving the sequence bi-directionally was initially validated within the area of speech recognition

We have used our word embedding matrix and both the encoding and decoding sequences will use these embedding. Previously there are only a few works were executed for Bengali textual content summarization and that's why we attempted to make a textual content summarizer which could generate a right precis from a given textual content. We have used TensorFlow CPU version-1.13.1.

3.5 Applied Mechanism

After cleaning the data, we have used neural networks that takes text vocabulary as the input and output a new sequence in another domain. We have followed the following steps:

Step-1: The texts are padded into sequences with the same length to get a feature matrix.

Step-2: After getting the feature matrix, we use word embedding mechanism where words from the vocabulary are mapped to vectors of real numbers. We have used a Bengali pre-trained word vectors called "bn_glove.39M.300d" which contains 39M (39055685) tokens, 0.18M(178152) vocab size.

Step-3: Encoder-Decoder, the encoder processes the input sequence and returns its own internal states that serve as the context for the decoder, which predicts the next word of the target sequence, given the previous ones.

Step-4: Model utilized for training and prediction, we created two neural networks, one for training and the other, the "Inference Model," to make predictions by using some of the layers from the trained model. Both neural networks have an encoder-decoder structure.

For the feature engineering task, we have completed some data analysis to find the right sequence size, as our data has different lengths. Apart from this, we have also calculated the value of how many numbers model must remember. Fig 3.4 shows the length distribution and Fig 3.3 shows the word frequency for both the article and summary.

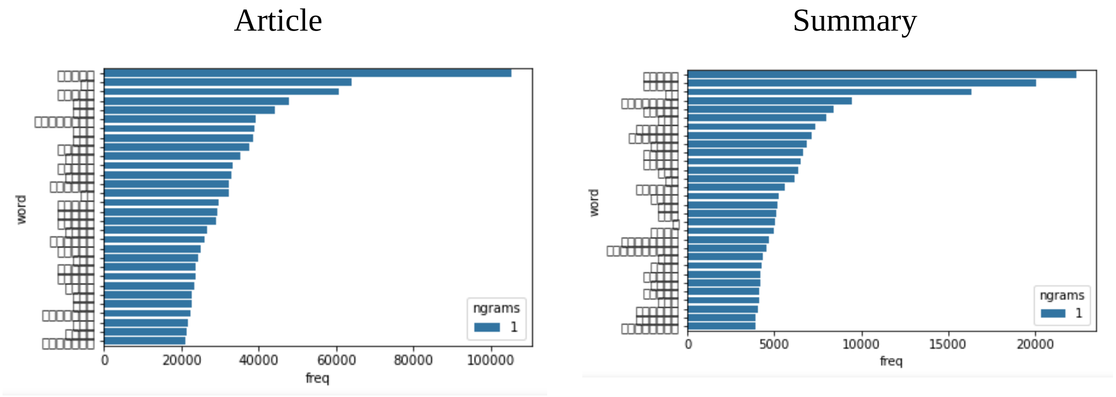


Figure 3.3: Word frequency

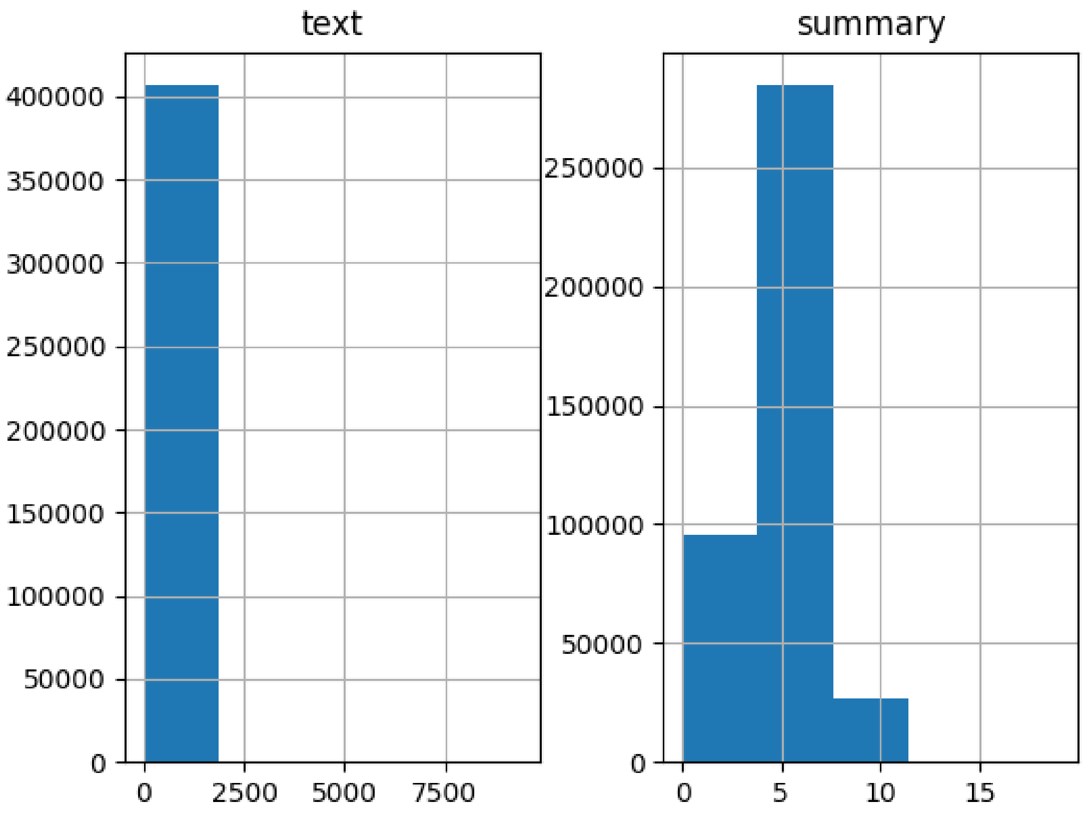


Figure 3.4: Word count for article and summary

A sequence-to-sequence (seq2seq) model for text summarization can be mathematically represented as a function that maps an input sequence, x, to an output sequence, y. The function is parameterized by a set of weights, θ, which are learned during the training process.
The main building blocks of a seq2seq model are the encoder and the decoder. The encoder is a recurrent neural network (RNN) that reads the input sequence, x, one token at a time and updates a hidden state, h. At each time step, t, the encoder computes the hidden state, h_t, as a function of the previous hidden state, h_{t-1}, and the current input token, x_t.

The decoder is also an RNN that generates the output sequence, y, one token at a time. The decoder's hidden state, s, is initialized with the encoder's final hidden state, h_T. At each time step, t, the decoder computes the current output token, y_t, as a function of the previous hidden state, s_{t-1}, and the current input token, y_{t-1}.

The most widely used form of encoder-decoder is RNN based such as LSTM and GRU, let's consider encoder RNN as f_enc and decoder RNN as f_dec

The mathematical representation of the encoder can be written as:
h_t = f_enc(h_{t-1}, x_t)

And the mathematical representation of the decoder can be written as:

y_t = f_dec(s_t, y_{t-1})
where s_t = h_T (initial hidden state of decoder)

The overall seq2seq model can be written as:
y = f(x, θ) = f_dec(f_enc(x, θ_enc), θ_dec)

The goal of training the model is to learn the best set of weights, θ, that minimize the difference between the predicted output, y, and the true output, y_true. This difference is usually measured by a loss function, such as cross-entropy loss.

At each time step during the training phase, we input the encoder a sentence's words one by one in order. If there is a sentence, for instance “একুশে পদকপ্রাপ্ত বরেণ্য নজরুল সংগীতশিল্পী”, then at time step t=1, the word “একুশে” is fed, then at time step t=2, the word “পদকপ্রাপ্ত” is fed, and so on. If, for instance, the sequence is composed of the words x1, x2, x3, and x4, the encoder in training would look as follows:

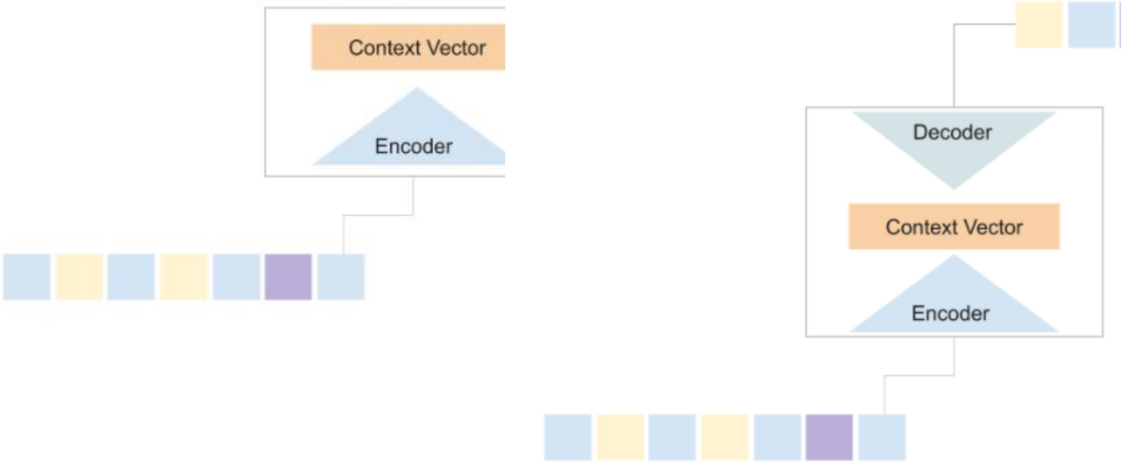


Figure 3.5 Encoder & Decoder Layer

Depending on the data encoded by the encoder, the decoder is trained to begin producing the output sequence. Before feeding the target sequence—in this case, the summary we wish to predict—to the decoder, special tokens called "<start>" and "<end>" are inserted. While decoding the test sequence, the target sequence remains unclear. As a result, we begin predicting the target sequence by feeding the decoder with the first word, which is invariably the <start> token. The sentence is concluded by the <end> token.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Results & Analysis

We trained our transformer model for 60 epochs. The loss and accuracy metrics are shown in figure 4.1

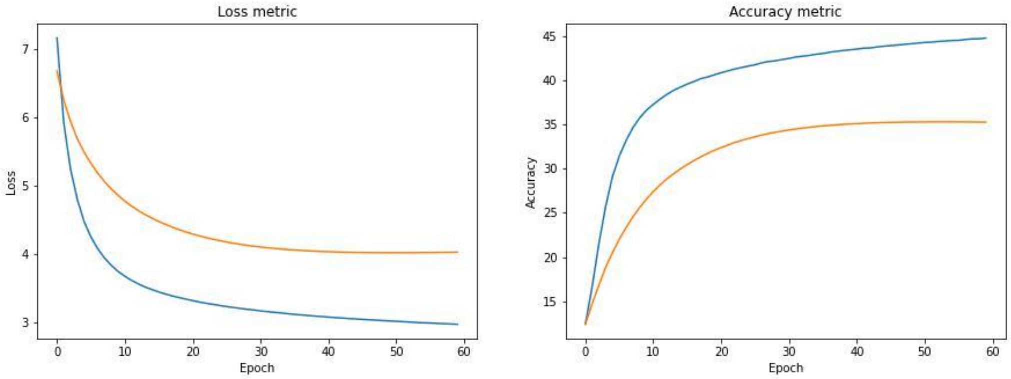


Figure 4.1: Loss and accuracy graph for summarizer model

From metrics figure we can see that our peak accuracy was around 45%. This might seem low, but there are several limitations of the automatic evaluation metric we used. First of all, even if a sentence is written differently but conveys the same meaning, it will be assigned a high loss and low accuracy

These are the hyper parameters that we used to train this model. The 60 epochs ensures that our model becomes trained, and only stops training as a result of early stopping (when the loss stops decreasing).

Table 4.1 Value of Parameter

Parameter	Value
epochs	60
batch	128
num_layers	2
rnn_size	256
learningrate	0.01
probability	0.75

To create a summary, we will select a sentence from the dataset as the input, and we will choose the summary length at random. Here is a favorable answer from the computer following a short training session using our model and dataset.

Table 4.2 Sample output example from the model

Original Text:	রবীন্দ্রসংগীতশিল্পী অণিমা রায় এবার ভারতের বাংলা ছবিতে গান করছেন। ছবির নাম ‘রিইউনিয়ন’। গত সোমবার কলকাতার এক হোটেলে ছবির মহরত অনুষ্ঠানে উপস্থিত ছিলেন তিনি। অণিমা রায় জানান , দিনটি তার জন্য সত্যি বিশেষ ছিল। এবারই প্রথম তিনি ভারতের ছবিতে প্লেব্যাক করছেন। সেটা অবশ্যই রবীন্দ্রসংগীত। ‘রিইউনিয়ন’ ছবির গানের সংগীত আয়োজন করছেন জয় সরকার। ছবিতে মুখ্য চরিত্রে অভিনয় করছেন পরমব্রত চট্টোপাধ্যায় ও প্রিয়াঙ্কা সরকার। যৌথভাবে পরিচালনা করছেন নবারুণ সেন ও মুরারি রক্ষিত। ‘রিইউনিয়ন’ ছবির মহরত অনুষ্ঠানে অণিমা রায় ও জয় সরকার সামাজিক যোগাযোগমাধ্যম ফেসবুকে অণিমা লিখেছেন, ‘প্রার্থনা করবেন বন্ধুরা, সিনেমাটি যেন সবার ভালোবাসা পায়, আর আমার গাওয়া গানটি যেন অন্য মাত্রা যোগ করে আমার দেশের সম্মান রাখতে পারে।’ আরও জানালেন, কলকাতা সফরে গিয়ে এবার দ্বৈত কণ্ঠের আরেকটি গান গেয়েছেন তিনি।
Predicted Summary	ভারতের বাংলা ছবিতে গান করছেন অণিমা রায় গত সোমবার কলকাতার এক হোটেলে ছবির মহরত অনুষ্ঠানে উপস্থিত ছিলেন তিনি

Original Text:	১৯৭১ সালের ২৪ মার্চ কালরাতে গণহত্যার পর সশস্ত্র মুক্তিযুদ্ধে ঝাঁপিয়ে পড়েছিলেন নিরস্ত্র বাঙালি। মুক্তিযুদ্ধ শুরুর এক মাস পর ২৫ এপ্রিল জল-স্থল ও আকাশ পথে বরিশালে হামলা চালায় পাকিস্তানি হানাদার বাহিনী। তারা সার্কিট হাউস, বিএম স্কুল, নতুন বাজার পুলিশ ফাঁড়িসহ বিভিন্ন স্থানে অবস্থান নিয়ে থাকে। ২ মে তারা স্থায়ী ঘাঁটি স্থাপন করে নগরীর বান্দ রোডের পানি উন্নয়ন বোর্ড (ওয়াপদা) অফিস এবং কীর্তনখোলা তীরবর্তী খাদ্য বিভাগের সিএসডি (ত্রিশ গোডাউন) গোডাউন কম্পাউন্ডে। সেখানেই তারা চালাতে থাকে পৈশাচিকতা। সামরিক ট্রাকে করে তারা ঘাঁটিতে ধরে আনত মুক্তিকামী সাধারণ মানুষ এবং নারীদের। তারপর তাদের ওপর চালাত ইতিহাসের বর্বরতম নিষ্ঠুরতা। প্রত্যক্ষদর্শী তৎকালীন খাদ্য বিভাগের কর্মচারী আবদুল হাকিম আলী সরদার জানিয়েছেন, দাফতরিক কাজে প্রতিদিন হানাদারদের ক্যাম্পে গিয়ে দেখেছেন, ওই ক্যাম্পে মুক্তিযোদ্ধা এবং মুক্তিকামী সাধারণ মানুষকে দলে দলে ধরে এনে নির্মম নির্যাতন চালানো হতো। হানাদাররা বেয়নেট দিয়ে কুপিয়ে এবং রাইফেল থেকে গুলি চালিয়ে হাজার হাজার মানুষকে হত্যা করে লাশ ফেলে দিত গোডাউন লাগোয়া খালে। আবার নদীর তীরে নিয়েও হত্যা করে লাশ নদীতে ভাসিয়ে দিত। শুধু তাই নয়, নারীদের ধরে এনে চালানো হতো যৌন নির্যাতন।
Predicted Summary	১৯৭১ সালের ২৪ মার্চ কালরাতে গণহত্যার পর সশস্ত্র মুক্তিযুদ্ধে ঝাঁপিয়ে পড়েছিলেন নিরস্ত্র বাঙালি পাকিস্তানি হানাদার বাহিনী তারপর তাদের ওপর চালাত বর্বরতম নিষ্ঠুরতা

CHAPTER 5
CONCLUSION

5.1 Summary of the study

This project is based on Bengali NLP. In this project, we aim to develop a hybrid model for summarizing abstract texts in Bangla. This approach can be used to automatically summarize Bengali text. The entire job was finished in just under five months. There are various parts to the project work and research. Below is a step-by-step breakdown of the project's overall synopsis.

Sl. No	**Steps**
1	Collection of Bengali dataset
2	Data pre-processing
3	Implement BERT
4	Organize the summary from step 3
5	Collect Word2vec
6	Vocabulary count
7	Load pre-trained word2vec
8	Add token
9	Define Encoder and Decoder with LSTM
10	Build seq2seq model
11	Train model
12	Output result analysis

5.2 Conclusion

We present a hybrid model for summarizing news stories and generating precise summary in this research. The suggested model uses an extractive method to choose the most important phrases and sentences, after which it employs a seq2seq model with bidirectional LSTM encoders and decoders. Using the bdnews datasets, this model was trained, tested, and verified. However, Additionally, there aren't enough effective words to vector, and there isn't even a lemmatizer for the Bengali language. Future efforts will be made to address these issues in the hopes of creating a better text summarizing model for Bengali. We are certain that the methodology provided here may be used to other summarizing tasks in a variety of fields.

In conclusion, RNNs are a promising approach for abstractive text summarization, using either encoder-decoder architectures or transformer architectures. While there is still room for improvement, RNN-based models have achieved good performance on a variety of text summarization tasks. Further research is needed to better understand the strengths and limitations of RNNs for text summarization, and to develop new methods for improving their performance.

5.3 Recommendations

In order to improve the performance of the model, we are attempting to expand the dataset and the summary of that dataset for the remaining stages of our research. We will be building many models for the summarization for the Bengali language that can be useful to identify the better performance. We are now only working with short sequences; a stronger summarizer is required for Bengali literature, which has extensive sequences.
The following list of recommendations for text summarization includes some of them.

- Create a large dataset and its corresponding summary for better result.
- Reduce the size of the text to keep the original idea
- Automatically summarizes text retrieval system

5.4 Indication for further study

The model has some limitations. As every research work changes every moment therefore, we have to adopt those changes and update our model in future.

- Large dataset
- Addition of more sequences/chains
- Expansion of research.

- No restrictions on text length.

We have a plan to develop some web api so that it can be accessed anywhere. After completing the full research, we have also some other plan such as mobile app development and website platform to access it more easily.

REFERENCES

[1] Yeasmin, S., Tumpa, P.B., Nitu, A.M., Uddin, M.P., Ali, E., Afjal, M.I, “Study of abstractive text summarization techniques,” American Journal of Engineering 6(8), 253–260 (2017)

[2] S. Abujar, M. Hasan, M. S. I. Shahin, and S. A. Hossain, “A heuristic approach of text summarization for Bengali documentation,” *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2017, doi: 10.1109/icccnt.2017.8204166.

[3] M. A. I. Talukder, S. Abujar, A. K. M. Masum, F. Faisal, and S. A. Hossain, “Bengali abstractive text summarization using sequence to sequence RNNs,” *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2019, doi: 10.1109/icccnt45670.2019.8944839.

[4] P. Bhattacharjee, A. Mallick, Md. Saiful Islam, and Marium-E-Jannat, “Bengali Abstractive News Summarization (BANS): A Neural Attention Approach,” *Advances in Intelligent Systems and Computing*, pp. 41–51, Dec. 2020, doi: 10.1007/978-981-33-4673-4_4.

[5] S. Abujar, A. K. M. Masum, Md. Sanzidul Islam, F. Faisal, and S. A. Hossain, “A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN,” *Innovations in Computer Science and Engineering*, pp. 509–518, 2020, doi: 10.1007/978-981-15-2043-3_55.

[6] Tan, Jiwei et al. “From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach.” *International Joint Conference on Artificial Intelligence* (2017).

[7] Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Veselin Stoyanov and Luke Zettlemoyer (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. CoRR, abs/1910.13461.

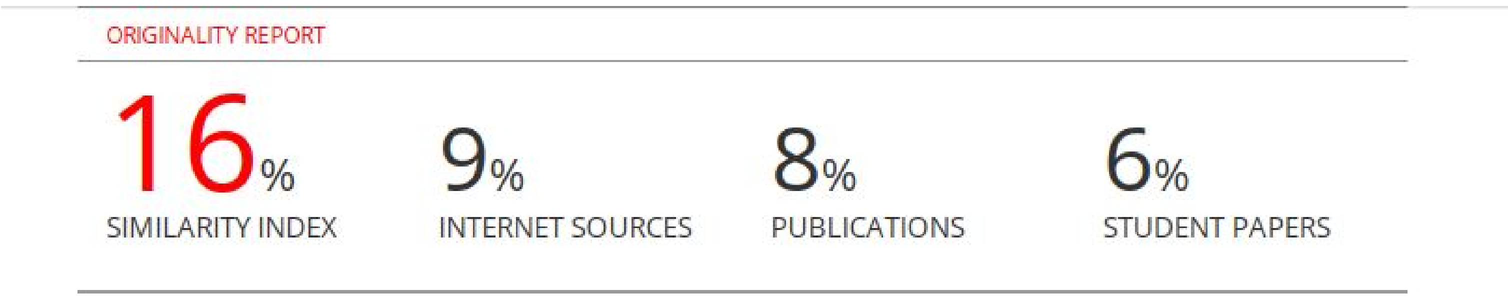
[8] J. Vig, “A Multiscale Visualization of Attention in the Transformer Model,” *ACLWeb*, Jul. 01, 2019. https://www.aclweb.org/anthology/P19-3007 (accessed Jan. 09, 2022).

[9] S. Abujar, A. K. M. Masum, Md. Sanzidul Islam, F. Faisal, and S. A. Hossain, “A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN,” *Innovations in Computer Science and Engineering*, pp. 509–518, 2020, doi: 10.1007/978-981-15-2043-3_55.

[10] N. Dhar, G. Saha, P. Bhattacharjee, A Mallick, and M.S. Islam et al., “Pointer over Attention: An Improved Bangla Text Summarization Approach Using Hybrid Pointer Generator Network” 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021

[11] Manning C. Understanding human language: Can NLP and deep learning help?.InProceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval 2016 Jul 7 (pp. 1-1).

[12] Sen O, Fuad M, Islam MD, Rabbi J, Hasan MD, Baz M, Masud M, Awal M, Fime AA, Fuad M, Hasan T. Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods. arXiv preprint arXiv:2105.14875. 2021 May 31.



ORIGINALITY REPORT

16%	9%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS