# Summarizing Text in Bangla using Natural language processing

**Submitted by**

**Mahadi Hasan Thanmay**

**191-35-2657**

Department of Software Engineering

Daffodil International University

**Supervised by**

**Md. Shohel Arman**

Assistant Professor

Department of Software Engineering

Daffodil International University

This Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering

# APPROVAL

This thesis titled on "**Summarizing Text in Bangla using Natural language processing**", submitted by **Mahadi Hasan Thanmay (ID: 191-35-2657)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

## BOARD OF EXAMINERS

------------------------------------------------------- **Chairman**

**Dr. Imran Mahmud**
**Head and Associate Professor**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

------------------------------------------------------- **Internal Examiner 1**

**Md. Khaled sohel**
**Assistant Professor**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

------------------------------------------------------- **Internal Examiner 2**

**Md. Shohel Arman**
**Assistant Professor**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Rimaz

------------------------------------------------------- **External Examiner**

**Rimaz Khan**
**Managing Director**
Tecognize Solution Limited

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Shohel Arman**, Assistant Professor, Department of Software Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Department of Software Engineering,
Daffodil International University

Submitted by:

Mahadi

Mahadi Hasan Thanmay
Id- 191-35-2657
Department of Software Engineering
Daffodil International University

## ACKNOWLEDGEMENT

First, I offer my deepest appreciation and gratitude to the All-Powerful God for His wonderful grace that helped me to successfully complete the final year thesis.

Md. Shohel Arman, Assistant Professor, Department of Software Engineering, Daffodil International University, is extremely appreciative of our gratitude. For the implementation of this thesis, our supervisor's expertise and enthusiasm in the subject of "Deep Learning" are crucial. This research could not have been completed without his inexhaustible patience, intellectual direction, continuous encouragement, constant and energetic supervision, constructive criticism, helpful suggestions, reviewing numerous subpar versions and fixing them at every level.

I would like to offer my deepest appreciation to Dr. Imran Mahmud, Head of the Software Engineering faculty at Daffodil International University, for his assistance in completing our research, as well as to the other faculty and staff of the SWE department.

Finally, I must respectfully recognize the consistent support and patience of our parents.

## Table of Contents

## CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATION

Table 1: list of abbreviation

| | |
|---|---|
| NLP | Natural Language Processing |
| RNN | Recurrent Neural Network |
| NLTK | Natural Language Tools Kit |
| LSTM | Long Short-Term Memory |
| GNMT | Google Neural Machine Translation |
| NMT | Neural Machine Translation |

# ABSTRACT

Someone can legitimately utilise materials as a means of articulating his feelings. Accordingly, realising the material's importance is fundamental. Reading these texts and deciphering their meanings can be difficult and time-consuming. The most efficient solution to this issue is to use a machine. When it comes to artificial intelligence's potential in the realm of language learning, the content outline is a vastly underexplored but promising field of research. Research efforts should prioritise developing a system to automatically summarise content. An important part of a large report may now be written much more quickly thanks to the content summary generator. In contrast to languages like English, Bengali does not have any summarising software. The fundamental goal of this study is to extend the range of Bengali linguistic resources and developments. An attempt at a computer-generated book summary in Bengali is the subject of this inquiry. The Bengali language section of this testing was quite challenging. So far, I have set the groundwork for an automatic summarising software in the Bengali language. The information is gleaned from people's typical online behaviour. A deep learning model was used to create the summarizer. The model affects the findings of the study because it takes into consideration the fact that a faster train shortens the time it takes to recover from a tragedy. My work has improved the efficiency with which our Bengali text summarizer and its related rundown model can summarise a book in a few short sentences.

## CHAPTER 1

### 1.1 INTRODUCTION

These days, everyone wants to be well-informed, but few have the time to read extensively in order to achieve this goal. The world needs real-time automated text summary to keep everyone informed in the least amount of time possible. In addition to helping, you decide which of various potential reading options is most relevant to your needs, text summaries can be used to rapidly determine whether or not a lengthy Bengali document is worth reading in its entirety. Bangladesh's rapid technological progress has led to widespread use of the Bengali language online, making Bengali Text Summarization a need. This proposed method aims to absorb Bengali-language articles and generate a shortened version that is faithful to the original's meaning.

The most common types of summarizations are extractive and abstractive. By excluding the sentences that contribute the least to the text's intended meaning, the Extractive technique produces a shorter and more accurate version of the material. To use an abstractive approach means to condense the source material while keeping the major points and ideas intact. The synthetic summary will have natural language flow and word choice. The most important concepts from a text are drawn out, and the sentences are streamlined for maximum clarity. The abstract content summarizer shown here includes a theoretical repository of content. The system is supplied news articles from the national daily "The Daily Prothom Alo," and from there a summary is generated; in pre-processing, the system tokenizes the extract and removes the stop words from the extract so that they do not affect the summary production. The programme stems the words to their fundamental forms after the stop words have been removed. Words that share a root are grouped together for this purpose. As I continue my research, I hope to develop a method by which I can provide an abstract summary of the material found in Bengali.

No machine can deliver perfectly reliable results every time, but the best possible results can usually be obtained. My automated abstract content summarizer has also looked like that. While no summary is guaranteed to be perfect, the most extreme reaction of a machine outline is acceptable for a summary of Bengali material.

## 1.2 MOTIVATION

Bangladesh is undergoing a digital revolution and developing more astounding technology; hence a stronger infrastructure is needed. As elsewhere, most services in Bangladesh are digital. As more individuals rely on digital media, the number of monthly newspaper subscribers has dropped. To keep up with modern life's fast pace, we need a system that can swiftly summarise our information so we're always well-informed.

A content summary condenses vast literature or data sets. A rundown focuses on the most important catchphrases and gives them context. It's difficult and time-consuming to read long material smoothly and find dynamic structural content. Sometimes we read a book but don't understand it. Many archives in a report make finding the theoretical challenging. A programmed content summarizer condenses content quickly. A computerised content summarizer can also identify which terms are most important, most frequently used, and comprehensive. Reading online articles, blogs, and news sites can be tiring. Why the data is disorganised and unclear. Another reason for using a book summary tool. A content outline approach determines the significance part from the given content archives, not an abstract content summarizer.

Today's society depends on information. Multiple sources of infinite content data. This much data requires a lot of storage space, which is expensive and causes placement issues. The summarizer minimises the record size and includes only the most significant details. Modern technology requires a computer-generated book summary. Everyone speaks Bengali. The client believes the present NLP asset is insufficient for this language. This is why NLP tools, technology, and developments are needed. This research aims to construct a computer-generated abstractive book summarizer for Bengali to find its natural language processing goldmine.

## 1.3 OBJECTIVE

In order to locate an appropriate title within the news piece. In order to provide output that is meaningful and relevant while also reducing the number of irrelevant words. In order to pinpoint the particular model that will work best for the Bengali news topic.

In this paper, we propose a system that would use innovative algorithms to generate summaries of Bengali texts automatically. We expect that our study will throw new light on the problem and pave the path for further research because there has been relatively little work on Bengali text summarization to date.

## 1.4 EXPECTED OUTPUT

Since this is a research project, our main goal was to get a research article out into the field. Science is a process that is always going on. Many people do searches on narrow research topics to find a good answer. The creator then puts together the final products for the customers. Most research and tools for the Bengali language are made with an extractive content outline, not with an abstractive content breakdown. In the same way, many analysts and engineers don't want to share their data and tools with everyone. We aren't getting the most out of our exploration efforts right now because of this. The Bengali language also looks at how well you can summarize what you've learned. Researchers have done some preliminary work on a general overview of Bengali content in the past. The result wasn't good enough for a computer to use to make a summary of a Bengali book. The machine is important for a programmed framework to work. So, the machine has to learn on its own. So, the framework in which the learning model works could be a web app or a program with a wider range of uses. In this study, we talk about an AI technique for outlining abstract Bengali content and talk about two important steps toward making a model for automatically summarizing Bengali content.

## 1.5 CHALLENGES

It was more challenging to implement text summary in Bengali than in English, the language most often used around the world. Multiple English summarization efforts have led to widely available packages and tools for doing rapid preparation of test data. Because no equivalent library existed for Bengali, we had to manually build the necessary codes.

The algorithm also had to deal with the fact that words in Bengali literature don't always appear in their English-language derivatives; instead, the syntax of the Bengali language changes the words to fit the context of the sentence. Since the most significant way to alter an English sentence is to add a suffix or prefix, the lemmatization approach provided by the Natural Language Processing Toolkit (NLTK) written in Python can easily manage such little alterations. Because every Bengali word is reduced to its simplest form during the stemming process, the system has to be launched with a dedicated stemming class.

## 1.6 RESEARCH QUESTION

- Benefits of Bengali content summary are discussed.
- When working with NLP, what steps must be taken to ready Bengali content?
- How exactly does the Bengali content summary work?
- How far in advance can we guess the topic of upcoming Bengali works?
- Why do translations from Bengali to English sometimes use different phrases when summarising the same material?
- How can I learn more about the inner workings of the Bengali content rundown model?

# CHAPTER 2

## 2.1 LITERATURE REVIEW

A machine must interpret any language presentation. Machine translation assists a computer in understanding text and can be used to construct automated systems. NMT can also be used to perform machine translation. (Bahdanau, et.al.,2014) used their combined knowledge in their paper to improve how common encoding and decoding techniques are presented. The encoder receives a vector of content phrases as input, and the decoder returns the potential vector groups. Then, to make NMT even more successful, the Attention-Based Methods are introduced. Some issues with NMT include the fact that training and testing take more time and money, and that utilising a rare word in an arrangement would not produce a satisfactory outcome. (Wu Y., et.al.,2016) developed a GNMT framework to help people grasp NMT. Summarizing text material is the most studyable subfield of natural language processing. For a wide variety of languages, this topic has been the subject of countless studies. A great deal of study has gone into extractive summarization, but not so much has been done on abstractive summarization. In this section, we'll take a closer look at some outstanding examples of work in these areas.

For content-related issues, RNN provides a solution. The best outcomes for summarising abstract content are achieved through the mastery of grouping to arrangement with RNNs, according to (Nallapati et.al., 2016). A set of bits equal in length to the vector's information is used by the encoder, and the information is reassembled in the most semantically equivalent way by the decoder. There is a demand for enhanced abstract summaries of data. Continuous, real-time, real-world use of DRGD. Abstract material summaries have advanced uses, including facilitating education. To better understand how to summarise complex data, (Wang, et.al., 2018) proposed utilising a fortification system as a model. To create a more comprehensive overview, convolutional succession learning is employed here. The importance of learning in context is emphasised throughout this research. The use of the CNN is crucial to the procedure as a whole (Chen, et.al.,2018).

In 1958, Luhn proposed utilising phrase-level word frequency measures to evaluate sentences and finally choose the highest-ranking sentences for the summary (Luhn, et.al.,1958). Numerous efforts have been made in recent years to automate the task of summarising texts. Abstract summarization strategies that take advantage of both structural and semantic information were among them (Cheng, et.al.,2022). Effective and popular extractive summarising approaches include cluster-based methods, neural network summarization methods, graph-based methods, latent semantic analysis (LSA) methods, and fuzzy logic based, query-based methods (Song et.al., 2010). Though many researchers have looked into various approaches to summarising texts in English, very little has been done in other natural languages such as Bengali. This is the case despite the fact that summarization of texts is a vital use of NLP.

(Ilya Sutskever, et.al.,2014) present a method for detecting up sequences using multilayer LSTM. The encoder acts as a link between the information grouping structure and a content vector. Decoding occurs when a second person translates the grouping vector. As a result, the LSTM will not be surprised by the larger group. This would allow the request to group to be reversed (Sutskever, et.al., 2014). For such long lists, people typically use Wikipedia’s "extract content summary." This approach may also generate multi-report summaries from large datasets similar to the one being used. Lifeng Shang show how to generate a succinct overview of the content (El-Kassas, et.al., 2021). Only a few publications in recent years have attempted to summarise Bengali-language works (Das et.al., 2022). The summarization results obtained with the seq2seq learning model, which incorporates LSTM, are good. The "sequence to arrangement learning" technique is used in this research project to generate an abstract content index. Use the analysis tool to create an abstract summary of the text. For this aim, Bengali summaries of little bits of information and their English translations are used as data. When the encoding and interpreting component model is used, the abstractive content summary in Bengali improves.

While the practise of summarising texts in Bengali is not as widespread as it is in English, there has been a lot of interest in the area as a new area of study in recent years. One of the earliest attempts to summarise Bengali texts was published in 2004 by Islam (Islam, et.al.,2004) which can be viewed as the pioneering work in this field. Their suggested corpus-based search engine would look for the term in many texts and then summarise the papers that included it, making it possible to search through vast collections of documents using keywords. Uddin and khan (Ghosh, et.al., 2018) created a java summarizer that sorts sentences in a suitable order using a combination of the location methodology, the cue method, the title, and the frequency of terms. As a result, just the top 40% of the text with the highest ratings were presented. Sarkar (Barua, et.al., 2021) employed a variety of methods, including TF-IDF (Term Frequency - Inverse Document Frequency) positional value, and sentence length, to synthesise Bengali news items. His plan was to boil down a news story to its essentials, so providing greater context and meaning to the reader. He used reference summaries he had written for a total of 30 Bengali documents in his analysis.

## CHAPTER 3

## 3.1 RESEARCH METHODOLOGY

### 3.1.1 SYSTEM WORKFLOW

Whole thing is structured like. Within that section, we will simply go over each of the steps that will be used during process. The following is an overview of the entire research process, outlining its development as work progressed.

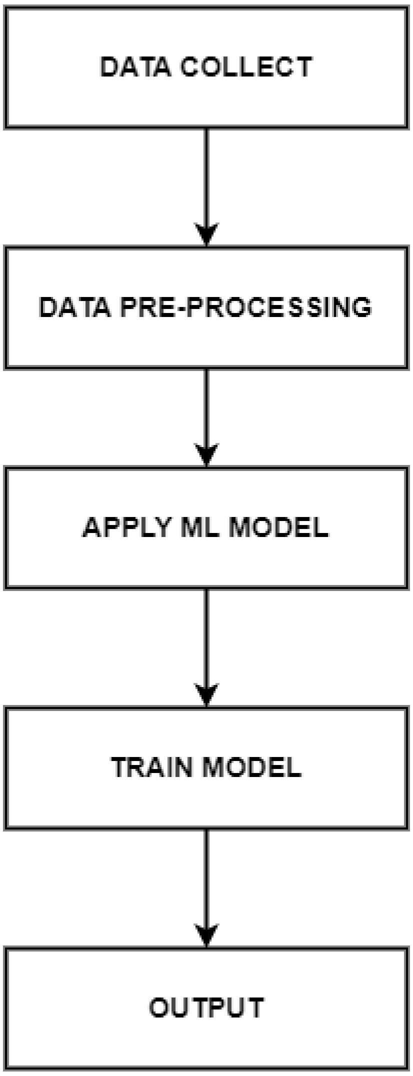


Figure 3.1.1: System Workflow

### 3.1.2 DATA COLLECTION

This dataset was imported from an online Kaggle and consists of a collection of texts with their human-generated summaries. It uses the abstractive approach to summarization. Texts for text summarizing have included pieces from "The Daily Prothom Alo" and "Kaler Kantho," two of Bangladesh's most widely circulated dailies, among others. Two news stories from the Daily Prothom Alo website are provided.

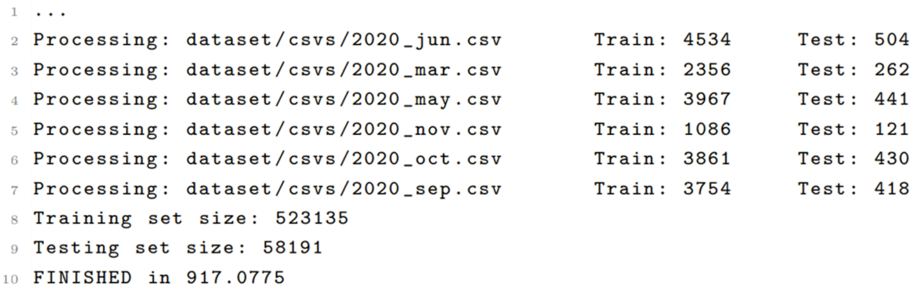
Table 2: Content and title length quantiles for tokenized articles

| Quantile | Content Length | Title Length |
|---|---|---|
| min | 1 | 3 |
| 25% | 169.0 | 7.0 |
| mean | 369.0 | 8.7 |
| 75% | 464.0 | 11.0 |
| max | 14739.0 | 651.0 |

### 3.1.3 PREPROCESSING

Fundamental preprocessing activities must be carried out before the actual model construction can begin. Using unorganized or cleansed text data could have serious consequences. Here, we remove any extraneous elements from the text that aren't essential to solving the problem at hand.

After the dataset was moved, all of its content should have been tokenized. Because of that, the long text has been summed up in one word. Which is helpful if you want to build your dataset around a few key features, like a lack of messy data and a lot of silly, empty limits. It also adds to the language of the dataset, which is important for issues related to natural language processing. This language can be used to find important data structure word embedding report. Some words are used over and over again in every language. Another thing is that Bengali doesn't have a lot of ways to shorten words. It refers to a word's shortened form or the process of making a shortened form of a term. The machine can't understand the short form of the book because it doesn't give the full meaning of the word. You won't hear those words very often. These words are used on signs, posters, and announcements, among other things. Using the standard way to say something takes away from the drug's amazing or terrible quality. The standard pronunciation we use includes white space, outside spacing, English characters, highlighted content structures, and leaving Bengali digits out of the text itself. In NLP, it is a common practice to make it clear what stop words mean. Most of the time, stop words are used to get rid of words that don't mean much. For written English to not have the "keep word" problem, NLTK users must do work in the library to come up with the necessary "out" words. But the Bengali stop word can't be taken out of the library right now. So, I started by getting together every Bengali stop word I could find online.



```
...
Processing: dataset/csvs/2020_jun.csv       Train: 4534     Test: 504
Processing: dataset/csvs/2020_mar.csv       Train: 2356     Test: 262
Processing: dataset/csvs/2020_may.csv       Train: 3967     Test: 441
Processing: dataset/csvs/2020_nov.csv       Train: 1086     Test: 121
Processing: dataset/csvs/2020_oct.csv       Train: 3861     Test: 430
Processing: dataset/csvs/2020_sep.csv       Train: 3754     Test: 418
Training set size: 523135
Testing set size: 58191
FINISHED in 917.0775
```

Figure 3.1.3: Train test of slicing through a log

We do the splitting for you when it comes to CSV files. As we said before, we put together all the goods that happened in the same month and year. By splitting files at the file level, you can make sure that both the training and testing sets have the same number of newly released files at different times. 90% of the data is used to train the deep learning model, and the other 10% is used to test it. Listing 3 shows that there are 520 examples in the training set and 60,000 examples in the testing set.

Table 3: Text Cleaning (Preprocessing)

| Raw Text | Cleaning |
|---|---|
| চলতি বছরে জানুয়ারির প্রথম সপ্তাহে অরুণাচল প্রদেশের টুটিং-এ গোপনে চিনের রাস্তা তৈরির পরিকল্পনা ভেস্তে দিয়েছে ভারত। ভারতীয় সেনাবাহিনীর দাবি, এই নিয়ে ইতিমধ্যে ক্ষমাও চেয়েছে তারা।<br>নিজস্ব প্রতিবেদন : সম্প্রতি অরুণাচলে অনুপ্রবেশের জন্য ক্ষমা চেয়েছে চিন। ফোর্ট উইলিয়ামে সেনা দিবসের অনুষ্ঠানের ফাঁকে সোমবার এমনটাই জানালেন ইস্টার্ন কমান্ডের প্রধান, লেফটেন্যান্ট জেনারেল অভয় কৃষ্ণ। তিনি বলেন, "দেশের পশ্চিম সীমান্তে অনুপ্রবেশের ঘটনা যেমন ভারতীয় সেনাবাহিনীকে ভাবাচ্ছে, তেমনই এবার পূর্ব সীমান্তে অনুপ্রবেশের সংখ্যা বৃদ্ধিও হালকাভাবে নিচ্ছে না ভারত।" চলতি বছরে জানুয়ারির প্রথম সপ্তাহে অরুণাচল প্রদেশের টুটিং-এ গোপনে চিনের রাস্তা তৈরির পরিকল্পনা ভেস্তে দিয়েছে ভারত। ভারতীয় সেনাবাহিনীর দাবি, এই নিয়ে ইতিমধ্যে ক্ষমাও চেয়েছে তারা। অন্যদিকে, নেপালে চিনা পরিকাঠামো ব্যবহার করে ইন্টারনেট পরিষেবা চালু হওয়ার ঘটনাকেও হালকাভাবে নিতে চাইছে না ভারত। বিষয়টি যথেষ্ট উদ্বেগের বলেও মনে করছেন লেফটেন্যান্ট জেনারেল অভয় কৃষ্ণ। | চলতি বছরে জানুয়ারির প্রথম সপ্তাহে অরুণাচল প্রদেশের টুটিং এ গোপনে চিনের রাস্তা তৈরির পরিকল্পনা ভেস্তে দিয়েছে ভারত ভারতীয় সেনাবাহিনীর দাবি এই নিয়ে ইতিমধ্যে ক্ষমাও চেয়েছে তারা নিজস্ব প্রতিবেদন সম্প্রতি অরুণাচলে অনুপ্রবেশের জন্য ক্ষমা চেয়েছে চিন ফোর্ট উইলিয়ামে সেনা দিবসের অনুষ্ঠানের ফাঁকে সোমবার এমনটাই জানালেন ইস্টার্ন কমান্ডের প্রধান লেফটেন্যান্ট জেনারেল অভয় কৃষ্ণ তিনি বলেন দেশের পশ্চিম সীমান্তে অনুপ্রবেশের ঘটনা যেমন ভারতীয় সেনাবাহিনীকে ভাবাচ্ছে তেমনই এবার পূর্ব সীমান্তে অনুপ্রবেশের সংখ্যা বৃদ্ধিও হালকাভাবে নিচ্ছে না ভারত চলতি বছরে জানুয়ারির প্রথম সপ্তাহে অরুণাচল প্রদেশের টুটিং এ গোপনে চিনের রাস্তা তৈরির পরিকল্পনা ভেস্তে দিয়েছে ভারত ভারতীয় সেনাবাহিনীর দাবি এই নিয়ে ইতিমধ্যে ক্ষমাও চেয়েছে তারা অন্যদিকে নেপালে চিনা পরিকাঠামো ব্যবহার করে ইন্টারনেট পরিষেবা চালু হওয়ার ঘটনাকেও হালকাভাবে নিতে চাইছে না ভারত বিষয়টি যথেষ্ট উদ্বেগের বলেও মনে করছেন লেফটেন্যান্ট জেনারেল অভয় কৃষ্ণ |

### 3.1.4 IMPLEMENTATION

We shall present our entire study process here. When it comes to the methods used to find answers, each piece of academic work is different. The implementation section will show all the methods used to complete the study. Within this approach section, we will talk about utilizing models and briefly describe each section. For a visual representation of all the steps involved, see the flowchart below:

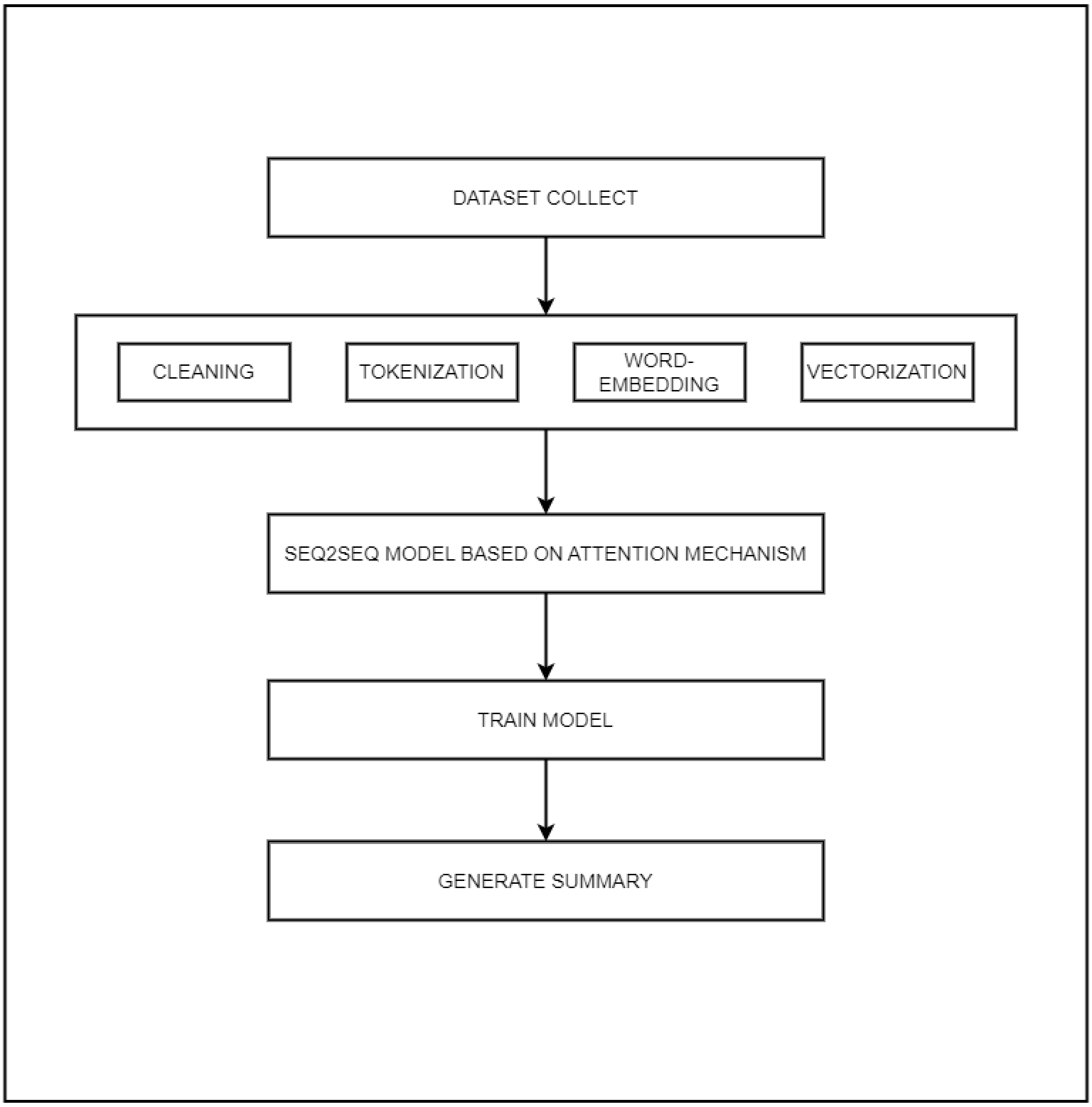


Figure 3.1.4.1: Implementation overview of the workflow

## ATTENTION MECHANISM

The attention mechanism is a recent Deep Learning technique that has proven particularly useful for Natural Language Processing applications such as machine translation, image captioning, dialogue generation, and so on. It is a method of improving the performance of the RNN model, which is used to encode and decode data sequences. In this paper, I'll attempt to explain how attention works and how it influences how we categorize text.

The Encoder-Decoder paradigm has a simple flaw: it encodes the input sequence into a constant-length vector before decoding the output at each time step. People believe that decoding long sequences is difficult because the neural network cannot handle sentences that are significantly longer than those in the training corpus. The model attempts to guess the following word by searching for a set of locations in the source sentence that contain the most important information. The model predicts the next word based on context vectors linked to these source locations and all of the target words that have already been created. When the attention model performs an output time step, it also generates a new context vector that can be used to filter the input sequence.

## ENCODER & DECODER

When dealing with input and output sequences of varying lengths, the Encoder-Decoder architecture is typically employed. Let's take a look at it from a text summarization perspective. An extensive word-string serves as input, and the resulting output is a shortened version of the input. In most cases, it's best to use a form of a Recurrent Neural Network (RNN), such as a Gated Recurrent Unit (GRU) or a Long Short-Term Memory (LSTM) for the encoding and decoding stages. This is because they solve the vanishing gradient problem, allowing them to capture long-term dependencies. Training and inference are two distinct options for setting up the Encoder-Decoder.

If we assume that x is a neutral arrangement of sentences, then x has the highest probability of being the next word in the word vector sequence. Given that y is the original sentence order, the probability that,

$arg\ (maxyp(x|y))$……………………. (1)

One-directional RNNs and Bi-directional RNNs are the two main varieties of RNNs. Both the information and the output of a unidirectional RNN are linked to each other in a linear fashion. A bidirectional RNN consists of two layers, each with two different directions of rotation, or "bearings" [9]. Each has two possible directions: forward and backward. These are put to use in an effort to solve the problem of machine interpretation. For this project, we used an RNN with two layers. Since we used RNN for Bengali, the encoder's contribution is a constant measure of the language's length. Information provided by a decoder allows for the calculation of a yield sequence that is dependent on that information. In this case, likelihood calculation lies at the heart of the basic computation. If X is the full data sequence, with values of (x1, x2, x3,..., xn), and c is the configuration vector, then the grouping is,

$ht = f(xt - ht{-}1\ )$ ……………………. (2)

And c,

$c = q(\{h1, \ldots .\ .\ hTx\})$……………………. (3)

Take the hypothetical case when Decoder expects a yield succession of y = {y1....., yTy}. After that, the most likely response or synopsis will be,

$p(y) = \prod p(y \vee \{y1, \ldots , yt{-}1\}, c)\ T\ t{=}1$ ……………………………………. (4)

$p(y \vee \{y1, \ldots , yt{-}1\}, c) = g(yt{-}1, st\ , c)$…………………………………… (5)

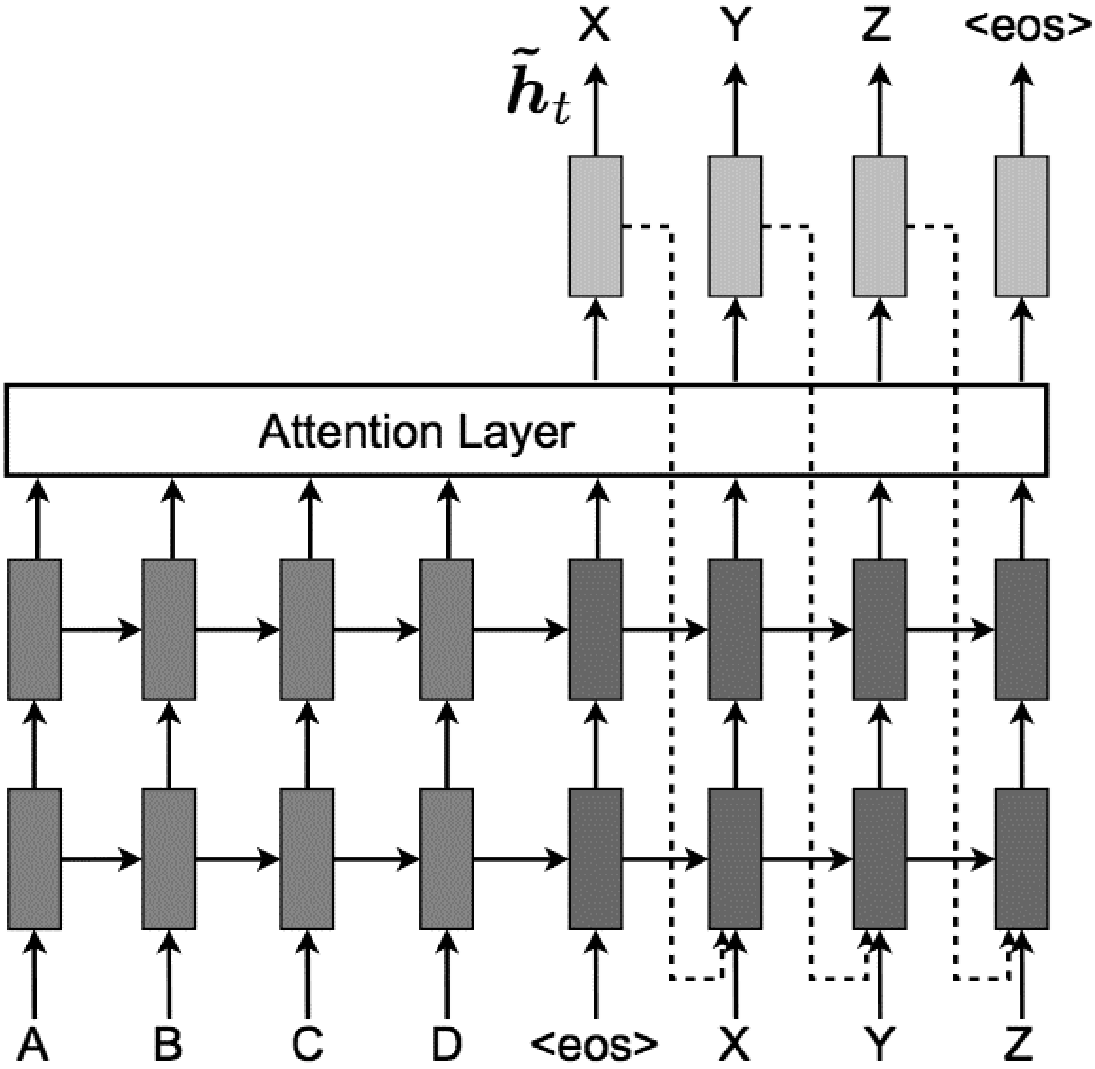


Figure 3.1.4.2: Encoder & Decoder of Attention Layer

### SEQ2SEQ

The model for Seq2Seq is built by the LSTM cell. First, the vector file is utilised to form the input of the word. In the vector file, each connected word is assigned a numerical value. These encoded data are treated in the same way that encoder input is. To store the sequence value, the encoder employs a Long Short-Term Memory (LSTM). To identify their relative places, a token was used at the beginning and end of each sequence in this scenario. The code specified some distinct sequences. All of those distinct tokens have a function in the encoder and decoder, where they are used to process the sequence. Use <END> to indicate the last input character. The encoder will instantly stop processing the input sequence when it hits the <END> token. The decoder then deciphers the sequence and generates output that matches the input. Stop decoding when you reach the end of the output sequence, which is represented by the <END> character.

Following the completion of the encoding phase, the sequence must wait for an instruction before going to the decoder. In this case, a token is utilised to tell the decoder that a specific encoding sequence is about to be inserted. Some text or words in the sequence are left out. We must determine the identification of each link in the chain. So we settled with the more mysterious one-of-a-kind token. As soon as a new token is identified in the sequence, it is added to the text as a normal token. The train runs in time increments. Sequences of roughly the same length had to be grouped in batches. As a result, we made use of a device known as a token.

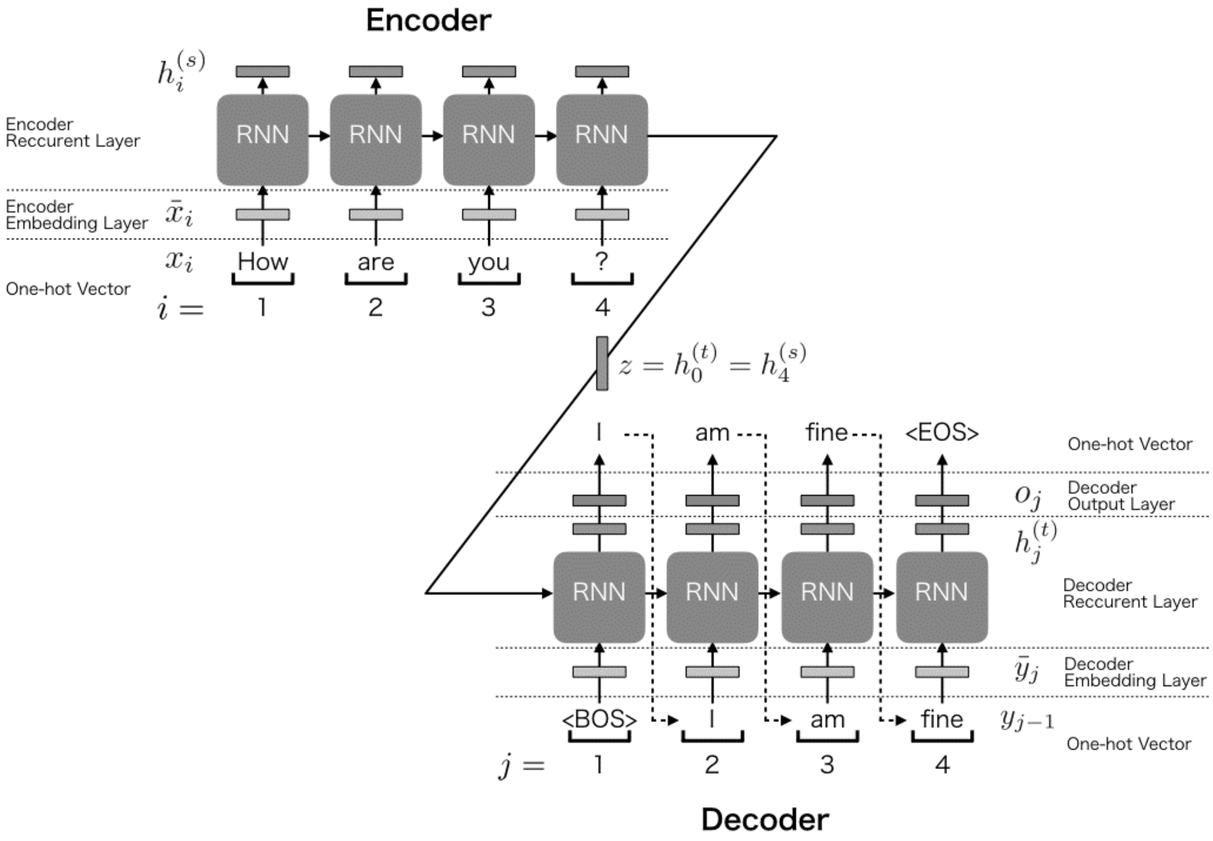


Figure 3.1.4.3: Encoder & Decoder of Seq2Seq

# CHAPTER 4

## 3.1 EVALUATION AND RESULTS

### 3.1.1 INTRODUCTION

Abstract content summarization is one of the most challenging problems in natural language processing. If certain components of the material are absent or are not necessary for the reaction, the machine will naturally shorten it. As a result, it is difficult to find a concise outline. In such a content summarizer, the computation of probability is used frequently. Due to the fact that the machine generates results based on the worst-case scenario. As words are added to the model during the train phase, their weights are calculated to indicate the likelihood of a response to the outline. The content data must be prepared after preprocessing in order for the model to build machine competence. Every model of deep learning incorporates a training backend motor. TensorFlow 2.10.0 was utilized as the backend work engine for this project. To characterize the train, it is necessary to characterize certain crucial parameters. Age, cluster size, learning velocity, and the number of layers is examples of variables. Based on these qualities, preparation can begin. It is imperative to shorten the train's delay. We employed the "Adam" optimizing agent for this evaluation to reduce error and improve the model. A well-rounded model can produce outstanding results when tested. Preparing data for the deep learning model requires a robust computer. GPU excels in this scenario. We refrained from utilizing GPU work to create our model during this inquiry. This is achieved by first training my model on a direct computer. This wastes a great deal of time and effort in model development and does not produce sufficient summarization results. We use Google Colab to complete the training of the model. This allows the user complete GPU control. This will reduce the train's travel time. The parameter estimation code for this experiment is provided below.

### 3.1.2 RESULTS

We developed a system to summarize English content first, and then adapted that to Bengali. Under specific scenarios, each model can be advantageous. Its goal was to lessen the impact of whatever features we might lose. There was a decrease in the learning model's error rate. There must be a reduction in the loss function for all chain information to be useful. Through training, we have accumulated the loss function. At the end of the training period, the loss function is summed. A major initial loss is incurred by the model. Recently, though, progress in reducing losses has stagnated. The value of the weight is 0.08. The "train" and "test" portions of each of our datasets are distinct. Right now, you can choose between 800 training data sets and 200 test data sets.

The approximate yield of the machine is correct. Everyone is aware that no machine can consistently guarantee a faultless output. In addition, our pre-built model yields acceptable outcomes, albeit not for all qualities. Its responses to the content are inconsistent. However, the content's significance is mirrored by the fact that it received the most responses overall.

I was able to cut the failure rate from 0.009 to 0.002 with minimal practice. I store the model in a file called "model.ckpt" for the purpose of evaluating the outcomes. I then schedule a TensorFlow session to reload the chart that was previously saved. Then, outline the content and overview information at random to confirm that everything is accurate. Then on, I capitalize on the contribution of the model—the motivation of a succession—by translating it into industrial jargon in order to encourage a succession.

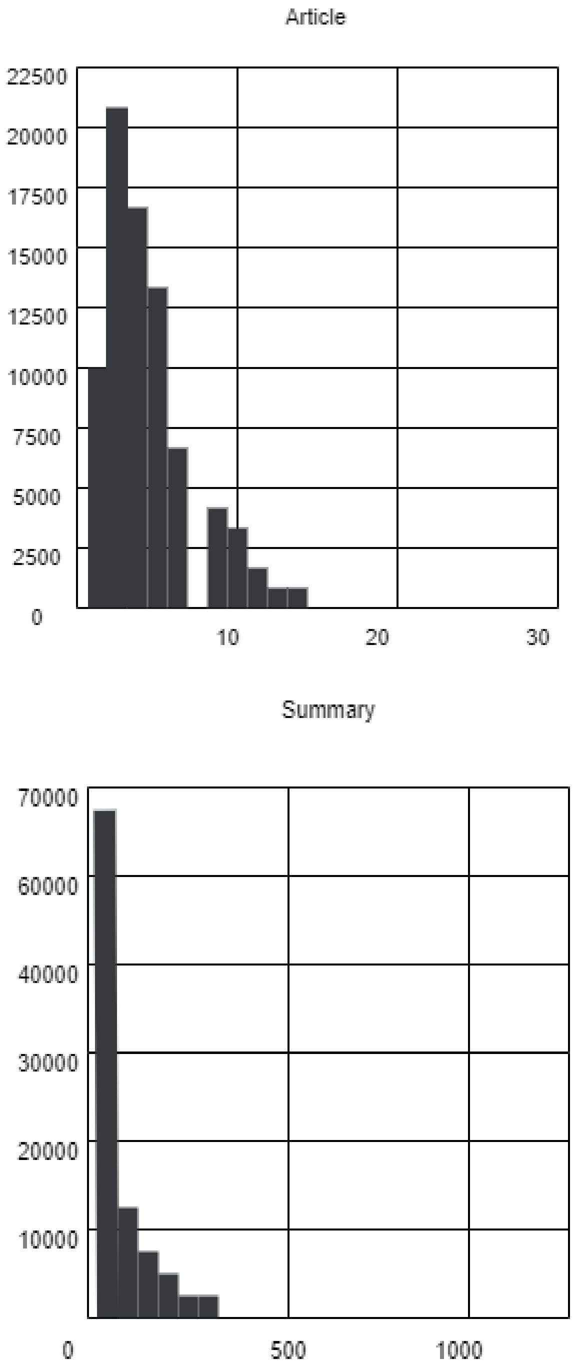


Figure 3.1.2: Summary and text graph comparison

### 3.1.3 TEST ARTICLES

Table 4: Article and summary results of 1

| Article | Summary |
|---|---|
| দ্য ভিঞ্চি কোড, এঞ্জেলস অ্যান্ড ডিমনসএর পর এবার ড্যান ব্রাউনের দ্য ভিঞ্চি কোড সিরিজের তৃতীয় উপন্যাস ইনফার্নো থেকেও সিনেমা নির্মাণের ঘোষণা দিয়েছেন রন হাওয়ার্ড। বরাবরের মতো প্রফেসর রবার্ট ল্যাংডনের ভূমিকায় দেখা যাবে অস্কারজয়ী অভিনেতা টম হ্যাঙ্কসকে। আর সিনেমায় গুরুত্বপূর্ণ আরেকটি চরিত্রে দেখা যাবে ভারতীয় অভিনেতা ইরফান খানকে। | জানুয়ারিতে এলসি খোলা কমেছে |

Table 5: Article and summary results of 2

| Article | Summary |
|---|---|
| ব্রিটিশ অ্যাকাডেমি অফ ফিল্মস অ্যান্ড টেলিভিশন আর্টস বা বাফটা অ্যাওয়ার্ডসের আসর বসছে ফেব্রুয়ারিতে। মনোনয়নের দিক থেকে শীর্ষে রয়েছে আত্মজীবনীমূলক সিনেমা দ্য থিওরি অফ এভরিথিং এবং কমেডি সিনেমা দ্য গ্র্যান্ড বুদাপেস্ট হোটেল। সিনেমা দুটি চারটি প্রধান ক্যাটাগরিতেই মনোনয়ন পেয়েছে। | বাফটা মনোনয়নে এগিয়ে দ্য গ্র্যান্ড বুদাপেস্ট হোটেল |

Table 6: Article and summary results of 3

| Article | Summary |
|---|---|
| চিরপ্রতিদ্বন্দ্বীর মুখোমুখি হওয়ার আগে আদর্শ প্রস্তুতি হয়নি কোনো দলের। তবু অস্ট্রেলিয়ায় গত কয়েক মাস তো ক্রিকেটের মধ্যে রয়েছে ভারত। সে কারণে তাদের পাকিস্তানের চেয়ে খানিকটা এগিয়ে রাখছেন ইনজামামউল হক। ৯২র বিশ্বকাপ দিয়ে বিশ্বমঞ্চে নিজের সামর্থ্যের প্রথম জানান দেওয়া এই কিংবদন্তির দৃষ্টিতে ভারতপাকিস্তান মহারণে পার্থক্য গড়ে দিতে পারে আরেকটি বিষয় টস | ইনজামামের কলাম: ভারতপাকিস্তান ফাইনালের আগের ফাইনাল |

Table 7: Article and summary results of 4

| Article | Summary |
|---|---|
| ২০১৯ বিশ্বকাপে সরাসরি জায়গা করে নেওয়ায় কাগজেকলমে সুযোগ পেলেও আফগানিস্তান ও আয়ারল্যান্ডের জন্য তেমন কোনো সুখবর নেই। যে সময়ের মধ্যে র‍্যাঙ্কিংয়ের ভিত্তিতে বিশ্বকাপের শীর্ষ আটটি দল নির্বাচন করা হবে, তার আগে এ দুটি দলকে বেশি ম্যাচ খেলার নিশ্চয়তা দিতে পারছে না আইসিসি। | আফগানিস্তান ও আয়ারল্যান্ডের জন্য সুখবর নেই |

Table 8: Article and summary results of 5

| Article | Summary |
|---|---|
| পরম প্রার্থিত বিশ্বকাপ জয়ের আনন্দ নিয়ে অবসরে এখন শচীন টেন্ডুলকার। সেবারের ম্যান অব দ্য টুর্নামেন্ট যুবরাজ সিংয়ের জায়গা নেই জাতীয় দলে। টেন্ডুলকারের ওপেনিং পার্টনার বীরেন্দর শেবাগ কিংবা ফাইনালে ৯৭ রানের অনবদ্য ইনিংস খেলা গৌতম গম্ভীর বিস্মৃতির অতলে। জহির খান, হরভজন সিংরাও নেই হিসেবে। কপিলস ডেভিলসএর পদাঙ্ক অনুসরণে ২০১১ বিশ্বকাপ জিতেছিল যে ভারতীয় দল, এবারের স্কোয়াডে তার প্রতিচ্ছবি যে সামান্যই | ভারত: ফেভারিট নয় শিরোপাধারীরা |

Table 9: Article and summary results of 6

| Article | Summary |
|---|---|
| অবশেষে ইংলিশ প্রিমিয়ার লিগে গোল পেলেন মারিও বালোতেল্লি। ইতালির এই ফরোয়ার্ডের এতদিন পর পাওয়া গোলটি লিভারপুলের জন্য হয়ে গেল মহামূল্যবান। তার এই গোলের কল্যাণেই টটেনহ্যাম হটস্পারের বিপক্ষে জয় নিয়ে মাঠ ছাড়তে পেরেছে অল রেড নামে পরিচিত দলটি। | অবশেষে বালোতেল্লির গোল, লিভারপুলের জয় |

Table 10: Article and summary results of 7

| Article | Summary |
|---|---|
| বর্ণবাদের নিষেধাজ্ঞার কারণে প্রথম চার বিশ্বকাপে ছিল না তারা। পরের ছয় বিশ্বকাপে তাদের অবধারিত উপস্থিতি। আর অনিবার্যভাবে প্রতিবারই দুর্ভাগ্যের খাড়ায় কাটা পড়ে বিদায় ভাগ্যদেবী যদি খানিকটা মুচকি হাসতেন, তাহলে ১৯৯২ থেকে শুরু করে প্রতিটি বিশ্বকাপেই জেতার সম্ভাবনা ছিল দক্ষিণ আফ্রিকার। | দক্ষিণ আফ্রিকা: অপবাদ ঘোচানোর সুযোগ |

Table 11: Article and summary results of 8

| Article | Summary |
|---|---|
| নিশ্চয়তা কেবল অনিশ্চয়তায় পাকিস্তান ক্রিকেট দলের বিশ্লেষণে এই বাক্য যথেষ্ট। আর তা সর্বকালের, সর্বযুগের জন্য প্রযোজ্য। হোক সেটি ইমরানের খানের কিংবা ওয়াসিম আকরাম, ওয়াকার ইউনুস, ইনজামামউল হকের দল। এবারের মিসবাহউল হকের পাকিস্তানের ক্রিকেটীয় চরিত্রের উপসংহারও সেটি। | পাকিস্তান: নিশ্চয়তা কেবল অনিশ্চয়তায় |

Table 12: Article and summary results of 9

| Article | Summary |
|---|---|
| বিশ্বকাপের মূল লড়াইয়ে নামার আগে মানসিক ও শারীরিকভাবে প্রস্তুত হওয়ার ওপর গুরুত্ব দিচ্ছেন বাংলাদেশ অধিনায়ক মাশরাফি বিন মুর্তজা। পাকিস্তান ও আয়ারল্যান্ডের বিপক্ষের দুটি প্রস্তুতি ম্যাচের পর অধিনায়কের মনে হচ্ছে, মানসিক শক্তির ঘাটতির কারণে তার সতীর্থরা মাঠে সামর্থ্যের অনুবাদ করতে পারছেন না। | বাংলাদেশ দলে মানসিক শক্তির ঘাটতি |

# CHAPTER 5

## 5.1 CONCLUSION

Learning Bengali is more relevant than ever as we approach the Information Age. The use of a text summarization system can help you save time, energy, and data. The majority of summaries are extractive and abstractive. The development and extension of Bengali NLP research is the primary focus of this investigation. I have used Bengali material as our model's input and written up a brief summary of the material that would emerge from that clustering: Bengali. First, I put together a template for English-language material, and then I create this one for Bengali. In general, the encoding and decoding processes for the two messages are equivalent. The amount of Bengali text in our database is limited. Regardless, the machine's responses to this data set are impressive. The age of the short text summary has been accounted for in this methodology, which has proven effective in Bengali. I have defined both the clustering and outlining lengths. With this predetermined duration in mind, the machine may generate the rundown. In fact, this is the main limitation of my methodology. If the cycle time is too long, the model will not function properly. Here we have yet another potential test case for the Bengali content inventory. Pre-processing Bengali content for translation into another language might be challenging. The pre-processing library, in this case, must be compatible with Bengali text. Converting text to a vector is also a crucial part of these problems. To fix the content problem, reliable word to vector conversion must be provided. In the end, no machine can guarantee proper accuracy. It is known that there are limitations to the operation of each machine. In essence, my summarizer model has its own set of limitations. Above all else, it is essential that the model be able to generate an abstract summary of the Bengali language. This is a victory for my Bengali NLP recordings, which will be useful in my future studies.

## 5.2 FUTURE WORK

When the dataset is too little to employ the model without restrictions, like in the case of limited succession, for example. That being said, we've already accounted for the most likely future result in our model. Being that inspections are always an ongoing process. As a result, this structure will be gradually adopted by the Bengali language. There has to be much more research done to create a legal framework for any efforts. When that occurs, investigation always reveals the best answer. Therefore, questioning one's work is crucial to its eventual realisation and growth. What can be achieved in the future is constrained by the limitations set by past efforts. Understanding the shortcomings of prior work helps in developing a robust framework. Extending the size of the current dataset to include more Bengali text is the next phase of this work. Automatic length-agnostic model updates and fresh starts. Due to this, the model will be independent of the length of the material. Although it is a sophisticated model, it can be executed on the most recent version of TensorFlow (1.15). But for more recent updates, you'll need to change the code. The results of the study must be conveyed by the model after they have been analysed. Therefore, the growth of applications like online and portable applications is profoundly affected by the future of computerised reasoning. This is why I've written some code to generate Bengali-language summaries of abstract texts automatically.

## CHAPTER 5

## 6.1 REFERENCES

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Barua, A., Sharif, O., & Hoque, M. M. (2021). Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation. Procedia Computer Science, 193, 112-121.

Chen, D., Zhang, S., Ouyang, W., Yang, J. and Tai, Y., 2018. Person search via a mask-guided two-stream cnn model. In Proceedings of the european conference on computer vision (ECCV) (pp. 734-750).

Das, A., & Saha, D. (2022). Deep learning based Bengali question answering system using semantic textual similarity. Multimedia Tools and Applications, 81(1), 589-613.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679.

Ghosh, P. P., Shahariar, R., & Khan, M. A. H. (2018). A Rule Based Extractive Text Summarization Technique for Bangla News Documents. International Journal of Modern Education & Computer Science, 10(12).

Islam, M. T., & Al Masum, S. M. (2004, December). Bhasa: A corpus-based information retrieval and summariser for Bengali text. In Proceedings of the 7th International Conference on Computer and Information Technology.

Li, Piji, et al. "Deep recurrent generative decoder for abstractive text summarization." arXiv preprint arXiv:1708.00625 (2017).

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

Sarkar, K. (2012, August). An approach to summarizing Bengali news documents. In proceedings of the International Conference on Advances in Computing, Communications and Informatics (pp. 857-862). ACM.

Song, W., & Park, S. C. (2010). Latent semantic analysis for vector space expansion and fuzzy logic-based genetic clustering. Knowledge and Information Systems, 22(3), 347-369.

Cheng, W., Hu, P., Wei, S., & Mo, R. (2022). Keyword-guided abstractive code summarization via incorporating structural and contextual information. Information and Software Technology, 150, 106987.

Uddin, M. N., & Khan, S. A. (2007, December). A study on text summarization techniques and implement few of them for Bangla language. In Computer and information technology, 2007. iccit 2007. 10th international conference on (pp. 1-4). IEEE.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. 2016 Sep 26.

Wang, Li, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization." arXiv preprint arXiv:1805.03616 (2018).