A Bengali Text Summarization using Encoder-Decoder Based on Social Media Dataset

BY
MINHAJUL ABEDIN RAHAT
ID: 162-15-7689

MD. TAHMID ALIE - AL - MAHDI
ID: 162-15-7692

FATAMA AKTER FOUZIA
ID: 162-15-7675

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

Supervised By

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2021

APPROVAL

This Project titled **“A Bengali Text Summarization using Encoder-Decoder Based on Social Media Dataset”,** submitted by Minhajul Abedin Rahat, Md. Tahmid Alie - Al – Mahdi and Fatama Akter Fouzia 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 B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 27/01/2021

BOARD OF EXAMINERS

____
Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____
Nazmun Nessa Moon **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Aniruddha Rakshit
____
Aniruddha Rakshit **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md Arshad Ali **External Examiner**
Associate Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Sheikh Abujar, Senior Lecturer, Department of Computer Science and Engineering** Daffodil International University. We 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:

Sheikh Abujar
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Rahat

(Minhajul Abedin Rahat)
ID: 162-15-7689
Department of CSE
Daffodil International University

Mahdi

(Md. Tahmid Alie - Al - Mahdi)
ID: 162-15-7692
Department of CSE
Daffodil International University

Fouzia

(Fatama Akter Fouzia)
ID: 162-15-7675
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First of all, we express our sincere gratitude and thanks to God that we were able to successfully complete our final year project.

We all are truly grateful & extend our best wishes to Sheikh Abujar, Assistant Professor, Department of CSE Daffodil International University. His unlimited tolerance, study control, faith, strong surveillance, valuable judgment, appropriate attitude, understanding of many things made our project achievable He has done everything he can to help us complete this project with appropriate knowledge.

We would like to express our sincere gratitude to our Daffodil International University CSE Head Professor **Dr. Touhid Bhuiyan** Sir and the staff of the university. We would like to thank our classmates for their cooperation in our work.

Finally, we would like to thank our parents for their support.

ABSTRACT

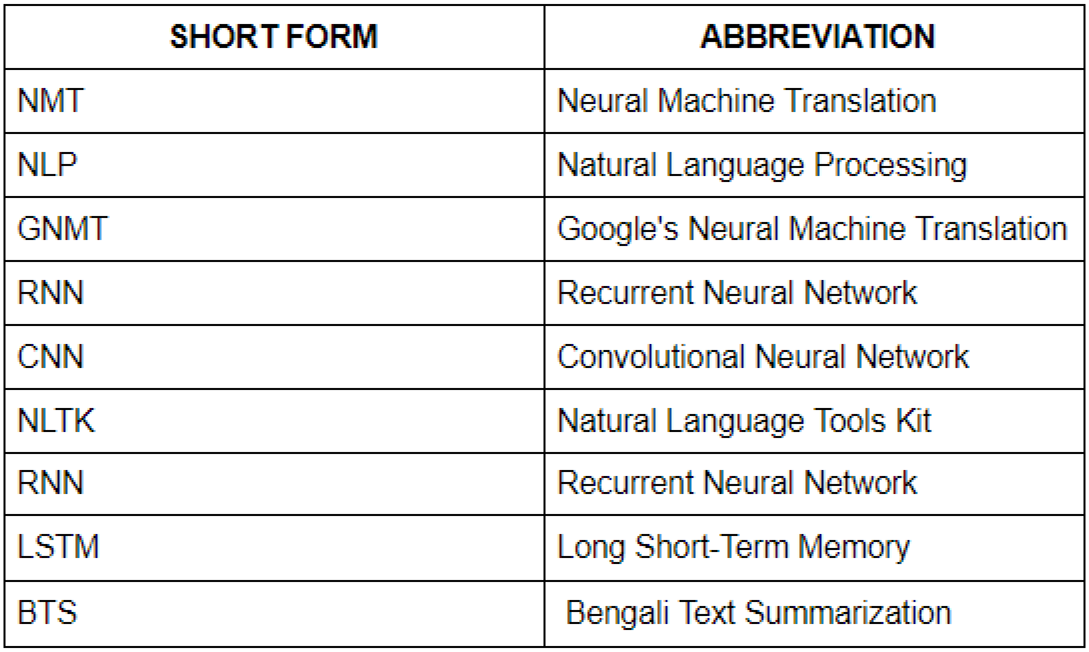
Text summarization defines artifices of reducing a long document to create a tale of the main aims of the original text. Due to the huge number of long posts nowadays, the value of summarization is produced. Reading the main document and getting a desirable summary, time and stress are worth it. Using Machine learning & natural language processing built an automated text summarization system can solve this problem. So, our proposed system will distribute an abstractive summary of a long text automatically in a period of some time. We have done the full analysis with the Bengali text. In our planned model we used a chain-to-chain models of RNN with LSTM in the encrypting layer. The structure of our model works applying an RNN decoder and encoder where the encoder inputs text documents and creative output as a short summary at the decoder. This system improves two things namely, summarization & establishing great performance with ignoble train loss. To train our model we use our dataset that was created from various online media, articles, Facebook, and some people's personal posts. The difficulties we face most here are Bengali text processing, limited text length, enough resources for collecting text.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION



SHORT FORM	ABBREVIATION
NMT	Neural Machine Translation
NLP	Natural Language Processing
GNMT	Google's Neural Machine Translation
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
NLTK	Natural Language Tools Kit
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BTS	Bengali Text Summarization

CHAPTER 1

Introduction

1.1 Introduction

A huge amount on data comes out digitally, so maintaining the original concept involves developing an emphasis system to reduce the longer text immediately. Typically, for output we use two types of text abbreviation techniques: extract and abstract summarizations. Extract summaries are created by removing some sentences from the main document. Different words are used to describe the contents of a document in abstract summary. Nowadays, text mining has become an attractive research field due to the large amount of text in social media. Currently, numerous strategies have been developed to summarize texts and place them in different zones. The NLP subdomain has a number of text reduction strategies that help reduce keywords.

We used some data from the dataset to train the model. To create our dataset, we collected data from various online media platform, Facebook, articles and some people's personal posts. We encountered various problems while collecting data so that we can get some real data. Sometimes difficulties arise to produce a proper abstractive text summary but here we are trying to get a better summary. Our advance steps for preprocessing the dataset, i.e cleaning data, using special tokens, counting missing words for decoder-encoder, word2vec, LSTM for input-output text.

We applied the sequence of sequence models with bi-directional RNN, including LSTM, to create the unlimited summary. In a decoder-encoder, the encoder encrypts the main text in a vector of a certain length and a short output of the decoder is generated. Our central object of on-screen text summarization is to minimize the total loss and improve efficiency to build an expressive summary.

1.2 Motivation

Text Summarization indicates summarizing a long text to get its main key component or words. The goal is to extract the main key component is the motivation behind summarization. Everything that is written may have a meaning if it is written in data structured form but it is irritating sometimes to read the whole text and sometimes it is time consuming also. It is sometimes difficult if the document is multiple in document size but hard to understand as a result it is quite difficult to extract the meaning. To reduce this kind of problem automatic text summarization helps to summarize the text and also counts the number of documents, words, frequently used words.

In this modern era, we spend our time on the internet reading books, web pages, newspapers but we feel bored after some time. The reason behind this is unstructured data and also dizzy meaning. That's why a text summarizer is needed. Obstructive text summarization summarizes text based on the text, it may summarize best on the text keyword or may not.

In this era everything we read is stored. Data plays an important role in this era. We cannot think without reading but it also creates problems. Everyday huge amounts of text are being generated from many different sources. It requires a lot of memory to store it. But if we summarize the text then remove the unnecessary parts from it and then save the core part then it will help us to reduce the size. That is the reason why text summarizer is a must thing.

Mother Language is every one’s comfort zone when it comes to reading and writing or expressing thoughts. In our Bangla language NLP resources are very low. NLP tools and technologies must be made.

1.3 Rational of the study

Bengali language is the sweetest language in this world. It has a great history Bengali language is the one & only language among the world for that people sacrificed their lives so that they can talk their own language. But in this modern world there are so many advanced tools & technologies available for linguistic research purposes but the Bengali language research tools or technologies are not advanced like any other languages. That’s why we should contribute to our language. If we work on text-based work most of the time use NLP tools & techniques to solve the problem. In the NLP branch there are so many core problems on that & one of the most important core problems is Text summarization. From a long sequence of text, we can make its quintessence throughout a text summarizer. Nowadays people have not much time to read the whole text so that reduces the time consumption. A fluent & error free text summary can help the people to understand the meaning of long text in a short time. for the other languages such as English, French, Spanish etc. there are so many dedicated tools & models are available. And on the other hand, there are also some Bengali models & technologies available in this NLP field but those are very limited which is not enough. That’s why we should increase the Bengali NLP research area. Preprocessing is the major problem for work on Bengali text. Unicode is the possible best solution for this problem. We can use the Unicode of those characters or symbols for handling this problem. For Bengali text NLTK library is not available. That’s the reason why Bengali tools do not give the expected performance and the result is we don’t get the accurate result as other languages. There is no other way to find the solution of this problem. Research is the only way to find the solution of this problem. Therefore, in our research work, we are trying to processes Bengali language & show how to process it. Also trying to make Bengali abstractive text summarizer. That helps us to reduce the text size & give the best possible summary of that social media text.

1.4 Research Questions

- How can we define BTS?
- How does BTS work?
- What are the advantages of BTS?
- What is the contradistinction between BTS & ETS?
- What is the process of preprocessing the Bengali text in Natural Language Processing?
- How can we work on BTS in the future?
- How Bengali text summarization Model works?

1.5 Expected Output

Our main interest is to get the paperwork out in the fields and we did it. The developer then develops tools for the users. Compressing the text in Bengali is a new study. So many analysis works have been produced in the past to reduce Bangla lessons. Here we are trying to create an automated system to reach our goal. An automated system is conditioned on the machine. So, you have to read the machine to learn our proposed model. Our purpose for our research is to create an abstract text abbreviator using our proposed method and to maintain a remarkable efficiency of this method. In this research paper, we were trying to discuss our thoughts on the implementation part to increase perfection and reduce total loss while preparing the model.

1.6 Report Layout

This report has 5 chapters in total. Chapter-1 covers a summary of the entire effort. So many sections like 1.1- Introduction, 1.2- Inspiration for the study, 1.3 Research Logic, 1.4- Questions, 1.5- Expected Output, 1.6- Report Outline Research. In the second chapter, discussion parts are 2.1- Introduction, 2.2- Literature review, 2.3- summary, 2.4- Problematic Area, 2.5- Challenge. From Chapter 3 discussion part is research method including sub-sections 3.1- Introduction, 3.2- subjects and tools, 3.3- Data preprocessing methods, 3.4- Arithmetical analysis, 3.5- executional necessities. The fourth part discusses the tests and paragraph 4.1- Introduction, 4.2- implementational fallouts, 4.3- Graphic analysis, 4.4- moral. In Chapter 5 covers sub-sections 5.1- Summary, 5.2- Final Conclusion, 5.3- recommendation 5.4- Involved future study. Situations were made at the end of all sections that helped our research work.

CHAPTER 2

Background Studies

2.1 Introduction

The process of shortening the text is to display the long text in a short form. Finding concise, comprehensible and clear abbreviations is the main purpose of shortening the text. Creating long text documents in short form is much more time consuming and expensive for people. So automated text shortening has made our job much easier. The only way to get started with machine learning.

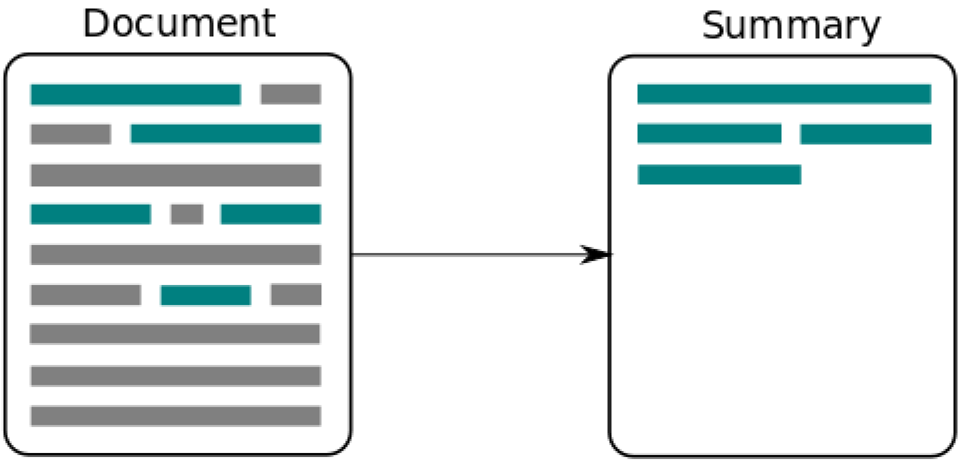


Figure 2.1.1: Summarization view

Daily we see large number of documents on internet including social media. Saving and retrieving information from social media becomes a multipart method for us at the same time we need space to store information. Thus, this strategy effectively solves the problem. Save in memory, abstract, key notes need to aid the necessary information. Two forms of text abbreviations: abstract & extract text. We have used the abstract process in our research

2.2 Literature Review

The process of shortening the text is the most researched subject. Much research has been done to summarize different languages. Large studies were conducted on abbreviated text abbreviations but there were few abstract text abbreviations. Here we are going to discuss some important work in this regard.

Some people want to make decisions based on peer analysis that are not effective for long. A portion of the NLP uses the POS tagging method using the Mau Colony algorithm [1] to shorten the Arabic text. Several strategies are suggested for shortening the text. Various researches the paper proposes and performs various methods to summarize them. Many papers have already been published for abbreviations in English, Arabic and other languages, but very few papers have been published on abbreviations in Bengali which may be a major reason for our research paper.

Bengali abstract text abbreviation [2] introduces a sequence of sequences to RNN using LSTM and the disadvantage of this method is overwhelmed. Stated NRM based on NNM encoding and decoding as neural network-based feedback for short-text abbreviations [3] indicate large amounts of data as NRM.

Encoder-decoders use specific neural machine agreements to create neural interfaces (Bahadanao et al 2014) [4]. Limited Vocabulary, An Unavailable Dataset of this paper are proposed for abbreviated RNN [5] English text contractions. some challenges they faced like text processing, loss reduction, vocabulary size. Increase performance and consume total losses is the main goal of this paper.

RASG [6] worked on a reader-based basis for intellectual summaries, Because of not formal and noisy comments reader comments are very challenging to link with models.
To overcome the challenges, they created a RASG, based on four parts. So, they want to use their large-sized dataset for future research. Overview [7] Try to build a powerful automated text compression system and deliver their goals both passive and abstract.

For arbitrary English and other language texts that combine strong NLP processing based on the equation "Abbreviation = Subject Identification + Explanation + Generation". These stages contain several individual modules. Our motive for our research is to create an abstract text summary and maintain the remarkable effectiveness of this method.

2.3 Research Summary

While doing this research, our team have thought of a Bengali's abstractive text summarization. This model is based on using deep learning. we have used our own dataset to utilize this model. Our team have collected ours Dataset from social-media. We, initially gathered Bengali's status, comments, pages, and groups posts from Facebook. Summary creation of each Bengali text is the next methods. The datasets have two attributes one of them is Bangla text and another is relevant summary. The total number of one thousand data with their relevant summary contains in the dataset. Preprocessing text is the precondition of creating a deep learning model. Preprocessing stage does text splitting and then insert Bengali contractions and delete stop words from the text. We have counted the vocabulary of the whole dataset after preprocessing. Word embedding plays a vital role in deep learning models.W2V provides a numeric value in the relevant vocab file. Pre trained w2v file is necessary for Bengali text which is currently online.

We have built a chain 2 chain model based on the attention model. This model has an encoder and decoder which is used with bi-directional LSTM cell. input of the encoder is word vector and related word vector in the decoder’s outputs. passing the sequence needs a token which is recognized as a special token like PAD’s, UNK’s, EOS’s. We trained the model for more than 5 hours. Then we got a relevant acknowledgment from the machine itself.

2.4 Challenges

Organized Bengali data is unavailable. Each and every one is present in an unsorted or unorganized way. However, the collection of the data is the main challenge for this research. The dataset needed might be already used. As an instance, the newspaper dataset is available but it can't be used as other research work. So, a completely new dataset is required for this research work.

After collecting the dataset, making the summary is another challenging work. And always working with the Bengali text is a challenge. However, in the processing step, some raw coding is needed to create the text as an input of the model. Let's assume, while removing punctuation from the text, for each punctuation Unicode is needed and raw code can remove it. And the second problem can be stopping word removal from the text. In the English language, there is a built-in library to remove stop words from the text. But for the Bengali language there is none so it is a great challenge. And a large vocabulary is another challenge. A large vocabulary can be gathered from a large data set and a large vocabulary helps to generate an optimal summary.

CHAPTER 3
Research Methodology

3.1 Introduction

In this section we will show the total methodology of our research work. From the view of solving techniques every research work is unique. All the approaches that have been applied in the research work are included in the methodology. This methodology part discusses applying models along with a short description of every individual part. A flow chart given below for show the total work processes:

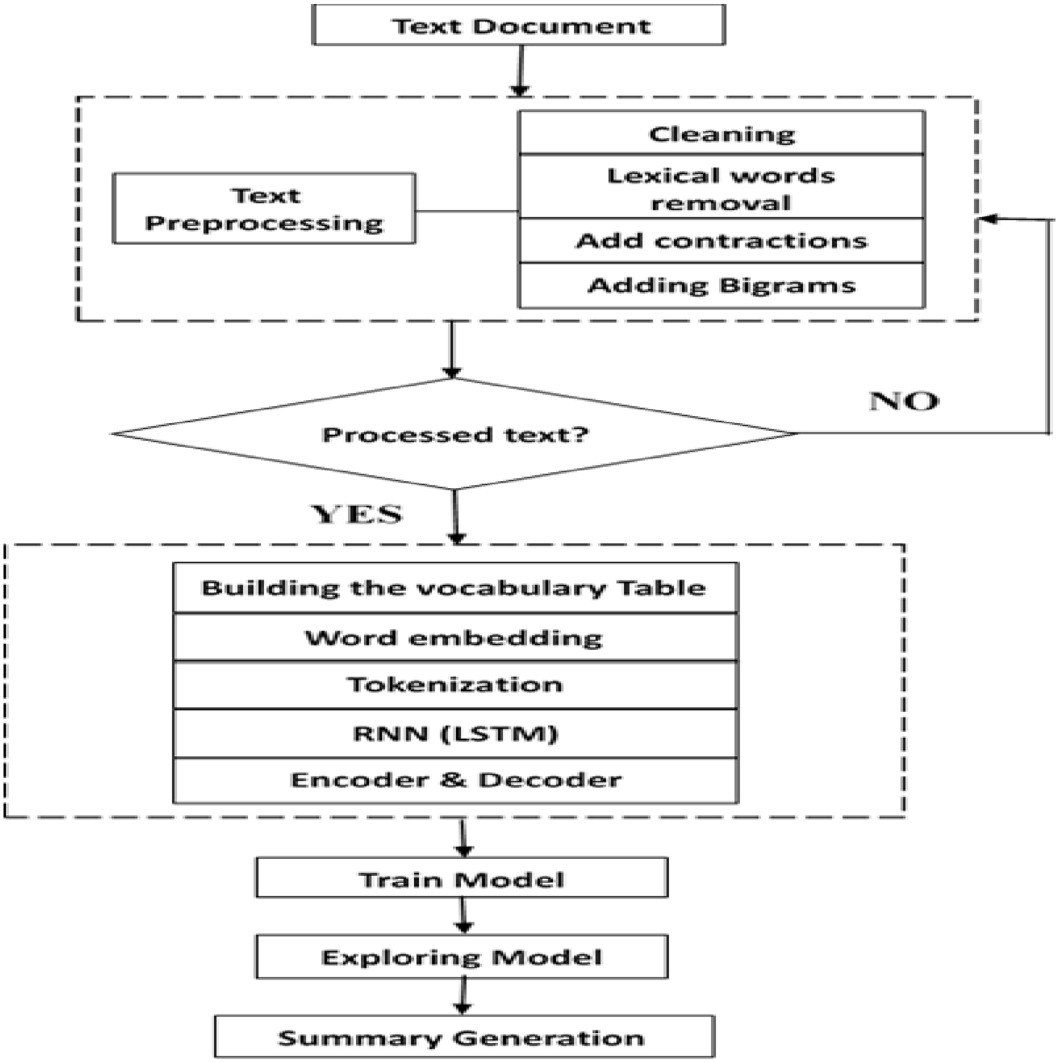


Figure 3.1.1 Workflow for text summarization

In this research paper, deep learning has been used to summarize the text. According to the type of research containing the deep learning algorithms have also been used. RNN is a method in deep learning that is used for solving text related problems.

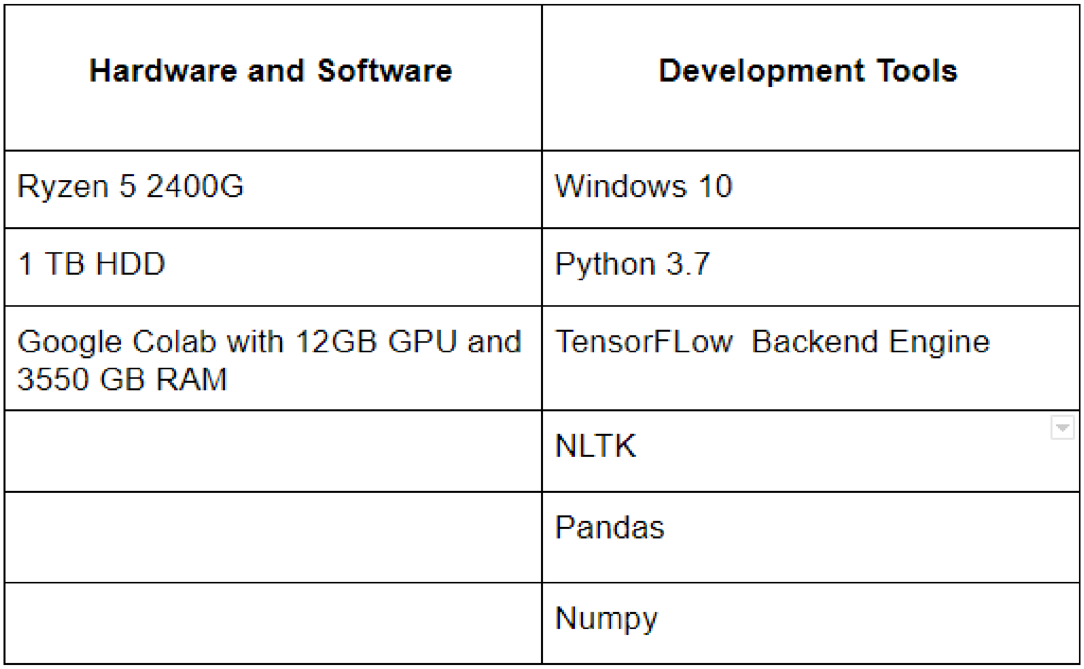
A good dataset is mandatory in every deep learning model to discover an absolutely self-executing system. So, before using the algorithm, the dataset needs to be prepared and preprocessed. Every part of the methodology is briefly discussed. When the research is completed, all sections will be followed.

To increase the efficiency and to give the nobility a rich explanation of the methodology is necessary. Mathematical equations along with the graphical overview of the model with the description help to understand the work. However, a good explanation of the methodology is required for further research. The total work seems like a framework. All the major steps are elaborately discussed in our methodology section. Some subsections of the main section aid to realize the summary of the model along with the purpose of using it.

3.2 Research subject and intermediary

The name of our research topic is "A Bengali Text Summarization using Encoder-Decoder Based on Social Media Dataset". In Bengali NLP it is a major research area. This research work has a brief description about the method of making an abstractive text summarization in Bengali along with the theoretical and conceptual method. A high configuration PC with GPU and other instruments is significant in a deep learning model. A list of required instruments for this model is showed below.

Table 1: Software and Tools



Hardware and Software	Development Tools
Ryzen 5 2400G	Windows 10
1 TB HDD	Python 3.7
Google Colab with 12GB GPU and 3550 GB RAM	TensorFLow Backend Engine
	NLTK
	Pandas
	Numpy

3.3 Data fetching and data preprocessing

We have 1027 Bangla data from various social media and we have also added a summary in each. Since we know that the better the dataset, the better it can be summarized, so we have created a huge Bengali dataset in this way. After we have collected the posts, we have manually created a summary of each post through the machine. 3.3.1 The steps required in the preprocessing stage of Figure are discussed below.

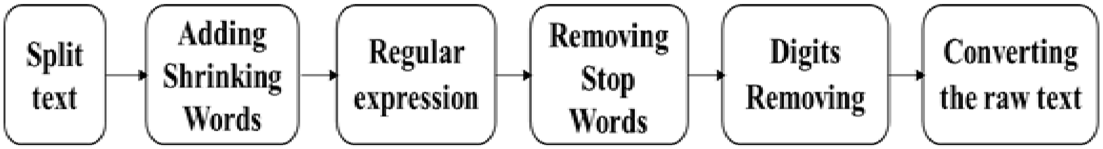


Figure 3.3.1 Dataset preprocessing

Data processing is a big job before models are created. Different steps are required to preprocess the data. It is very difficult to preprocess Bangla data. The first thing we do in preprocessing is to remove unwanted words as well as space. Our abbreviation to embed it here. Some examples of Bengali contractions are-

Table 2: Contraction List

SHORT FORM	FULL FORM
ডাঃ	ডাক্তার
মিঃ	মিস্টার
প্রোঃ	প্রোপাইটার
ইঞ্জিঃ	ইঞ্জিনিয়ার
মু.	মুহাম্মদ
মো.	মোহাম্মদ

We have removed the most unnecessary words and Bengali numbers that we need. Lastly, we lematerialized the words by deleting the words with the same meaning. After all the steps we have clean the text.

3.3.a Problem Contention

Since we have worked with a huge dataset here, there are a huge number of words here. We have created a huge vocabulary where each word is linked to another. A summary of very few words has been made here with long input text.

3.3.b Vocabulary Counting

Model has a lexis set and the need for a vocabulary set to find similarities between the text description and the output is much higher here. So, we should calculate the vocabulary. The word “বিশ্বাস” has been used 143 times since the word count. We have taken the pre-trained "bn_w2v_model" as the vector file.

3.3.c Purified text & summary

Next finalizing each of the steps, text look clear, there is no punctuation or extra space anywhere in the abbreviations. Both the clear text and the abstract are sorted by two separate lists. Examples of text preprocessing in Table 1 are specified under.

Table 3: Text Preprocessing Example

Original Text	Clean Text
জীবনে অনেক মানুষ আসবে যাবে, অনেক মানুষ কষ্ট দিবে।একগাদা স্বপ্ন দেখিয়ে আবার ভোরের আলোর সাথে মিলিয়ে যাবে !!কারো আসা যাওয়ায় তোমার হাত নাই।কেউ চলে গেলে তাই মন খারাপ করার কোনো কারন নেই !! প্রত্যেকটা বিচ্ছেদ তোমাকে শক্ত করে গড়ে তুলবে ... প্রত্যেকটা আঘাত তোমায় আরো পাকাপোক্ত করে গড়ে তুলবে !! ডোন্ট বি স্যাড, রোজ রোজ জীবনে নতুন মোড় আসে। কালকের সূর্যটার অপেক্ষা করতে থাকো,এক একটা নতুন ভোর হাজারটা নতুন স্বপ্নের জন্ম দেয়। নতুন একটা ভোরে নতুন করে সবকিছু শুরু করতে হয়। কালকের ভোরটা তোমার, শুধুই তোমার, একান্ত। ব্যক্তিগত একটা ভোরের অপেক্ষায় বাঁচতে শেখো !!	জীবনে অনেক মানুষ আসবে যাবে অনেক মানুষ কষ্ট দিবেএকগাদা স্বপ্ন দেখিয়ে আবার ভোরের আলোর সাথে মিলিয়ে যাবে কারো আসা যাওয়ায় তোমার হাত নাই।কেউ চলে গেলে তাই মন খারাপ করার কোনো কারন নেই প্রত্যেকটা বিচ্ছেদ তোমাকে শক্ত করে গড়ে তুলবে প্রত্যেকটা আঘাত তোমায় আরো পাকাপোক্ত করে গড়ে তুলবে ডোন্ট বি স্যাড রোজ রোজ জীবনে নতুন মোড় আসে কালকের সূর্যটার অপেক্ষা করতে থাকো এক একটা নতুন ভোর হাজারটা নতুন স্বপ্নের জন্ম দেয় নতুন একটা ভোরে নতুন করে সবকিছু শুরু করতে হয় কালকের ভোরটা তোমার শুধুই তোমার একান্ত ব্যক্তিগত একটা ভোরের অপেক্ষায় বাঁচতে শেখো

3.4 Arithmetical study

1) Data 1027. 1027 data divided into 3 subgroups for example Post Type, Text and Summary.

Table 4: Dataset Sample

Post Type	Text	Summary
Page	ক্লাসে সবচেয়ে দূর্বল ছেলেটি কাল সমাবর্তনে এসেছিল সবার চেয়ে হাই পজিশনের জব নিয়ে। বারবার প্রেমে ব্যর্থ হওয়া মেয়েটি এসেছিল একটি সুন্দর ছোট্ট পরিবার নিয়ে। কারো কাছে পাত্তা না পাওয়া, তোকে দিয়ে কিছু হবে না বলা ছেলেটিই সবচেয়ে সুন্দর বউ নিয়ে এসেছে। পড়াশোনার খরচ যোগাতে টিউশন করে হাত খরচ চালানো মেয়েটি কাল গাড়ি দিয়ে ক্যাম্পাসে এসেছিল। ক্লাসের সবচেয়ে সাক্সেস্ফুল ছেলেটি ডিপ্রেশনে ভুগছে জব না পাওয়ায়। ডিপার্টমেন্টের হার্টথ্রোব মেয়েটির চোখে নিচে কালি বিয়ে হচ্ছে না বয়স হয়ে গেছে।এভাবেই সময়ের সাথে বদলে যায় মানুষের জীবনে ইকুয়েশন। আসলে সমাবর্তনের মাধ্যমে শিক্ষা জীবনের শেষ হলেও সফলতা ও ব্যর্থজীবনের হিসাব গণনা শুরু হয়ে এখান থেকেই।তাই ঘৃণা, হিংসা, কম্পিটিশন বাদ দিয়ে জীবনটাকে বাচা উচিত সম্পূর্ণ স্বাদ ও ভালবাসা নিয়ে। কখন জীবনের কোন মোড় দেখায় কোন নিশ্চয়তা নেই, তাই কোন মুহূর্তের জন্য যাতে আফসোস না থাকে।	সমাবর্তনের মাধ্যমে শিক্ষা জীবনের শেষ হলেও সফলতা ও ব্যর্থজীবনের হিসাব গণনা শুরু হয়ে এখান থেকেই

Group	আমি সুখ পাই না, লেদারিংয়ের সাথে দু: খ পাইলস। অনেক দিন হয়েছে, আমি যত্ন করে ক্লান্ত। আপনাকে অনুসন্ধানের প্রক্রিয়ায়, আমি নিজের সাথে যা করছি তা বরখাস্ত করা হচ্ছে। আমি নিজেও পাই না, আমি নিখোঁজ হয়ে ক্লান্ত। আমার হাসি আর নেই, এই আত্মা মরে যাচ্ছে। অশ্রুও এখন শুকনো, আমি কাঁদতে ক্লান্ত হয়ে পড়েছি। আমি আর কিছু চাই না, আমার মনে এটি প্রদর্শন আপনি। আল্লাহ আমার চিৎকারের সাক্ষী, আমি প্রার্থনা করে ক্লান্ত। এই চোখ আর ভিজা হয় না, হৃদয় ও মনও বিতর্ক করছে না। আপনি যে পথে চলে গেছেন সেখানে দাঁড়িয়ে, আমি অপেক্ষা করতে করতে ক্লান্ত। আমি টুকরো টুকরো টুকরো টুকরো হয়েছি আমার হৃদয় ক্ষমা করছে না। শুধু আমার পালানোর সন্ধান, আমি বেঁচে থাকতে ক্লান্ত হয়ে পড়েছি। শুরুর দিনগুলিতে আপনি সেখানে ছিলেন, কোথায় আমার শেষ। আমি জীবনের বোঝা বহন করতে পারি না, আমি প্রেজেন্ডিং করে ক্লান্ত হয়েছি	জীবনের প্রতি ক্লান্ততা.

1) vocabulary size- 2,50,000k.

2) Unique words are 60321k.

3) 507400-word embedding.

4) 95% of the word used for model.

5) Text All-out length is 121 words & total summaries of the length is 12.

3.5 Executional requirements

3.5.a. Problem discussion

The text input and the summary are the same in the data set. Generally, summary has the shortest length considering text. Suppose D has the number of words of the input chain text containing the dataset. Thus $x1$, $x2\ldots$ is input chain the vocabulary has to be size V. That produces the result chains such as y1, y2,…,yd, here S>D. It indicates the chain of the summary is not much as a text document. All chains are produced from vocabulary itself.

3.5.b. Vocabulary and Word embedding

Casting up the vocabulary plays a vital role for word embedding. Vocabulary is counted based on a dataset. Which is a prerequisite in word embedding. If we want to embed it needs a W2V file. We used a pre-trained Bengali W2V which was collected from the internet. W2V file contains an arithmetic value corresponding to the word. It instantly saves the related words. The value is required while working. The vectors words required as an input which is used in the model and it gives words which are the output of our model. Therefore, the chain to chain learning easily done its job.

3.5.c. RNN Encoder & Decoder

When machine translation was invented, it generates a boundless revolutionary in this AI area. Typescript type model does work pretty well with expected outcome in DLM RNN are best and suitable system. LSTM consisted by individual RNN. Here encoder-decoder are being used by LSTM. In this paper decoder part incomes, input order and given outcome based on input line.

We know 2 forms of RNN, one - directional & bi - directional. RNN contains an input and an output and each is connected to one another. There are 2 layers of Bi-directional [9]. One is forward & backward. Here, 2 layered RNN is being used. Translator provides the related output conditional on input.

For our input sequence, each word is well-defined as x_i where i is the collection sequence and the unseen state h_{t-1} and the input vector x_t. The enclosed states hi are estimated by this formula:

$$\boldsymbol{h_t = f(w^{(hh)} h_{t-1} + w^{(hx)} x_t)} \text{.......................................}(1)$$

The encoder input $\boldsymbol{x = x_1, \dots, x_{tx}}$ into a constant c. Each time period t the RNN is updated by

$$h_t = f(\boldsymbol{x_t}, h_{t-1}) \text{..}(2)$$

And

$$C = q\ (\{h1, \dots\dots, htx\}) \text{...}\ (3)$$

Where c = unseen part, f and q are nonlinear part. Particularly, we can estimate the probability translation for the decoder using X sequence

$$p(y) = \prod_{t=1}^{T} \quad p(y \mid \{y_1, \dots, y_{t-1}\}, c) \text{...s.}(4)$$

Where $\boldsymbol{y}$ = ($\boldsymbol{y1}$, … . , $\boldsymbol{yTy}$). Conditional statement, e.g. $\boldsymbol{P}$ = ($\boldsymbol{y1}$, … . , $\boldsymbol{yt}|\boldsymbol{x1}$, … . , $\boldsymbol{xt}$).

$$\boldsymbol{h_t = f\ (h_{t-1}, y_{t-1},} c) \text{..}(5)$$

for conditional probability is as

$$\text{p}(\boldsymbol{y_t} \mid \{\boldsymbol{y_1, \dots\dots, y_{t-1}}\}, \text{C}) = \text{g}(\boldsymbol{y_{t-1}, s_t, C}) \text{..............................}(6)$$

Here context vector $\boldsymbol{Ci}$. $\boldsymbol{Ci}$ Then calculated as a weighted sum

$$c_i = \sum_{j=0}^{T} \quad a_{ij} h_j \text{...}(7)$$

Suppose, input sequence ($x_T\ to\ x_{T_x}$), also ($h_1\ to\ h_{T_x}$) is the hidden state. The hidden state ($h_{T_x}\ to\ h_1$) thus,

$$h_j = [\overrightarrow{h_{jT}}; \overleftarrow{h_{jT}}]^T \quad \text{...} \quad (8)$$

Here, h_j= predicted summary

$$e_{ij} = a(s_{i-1}, h_j) \text{ ……………………………………………….(9)}$$

Now RNN encoder-decoder assumed lower,

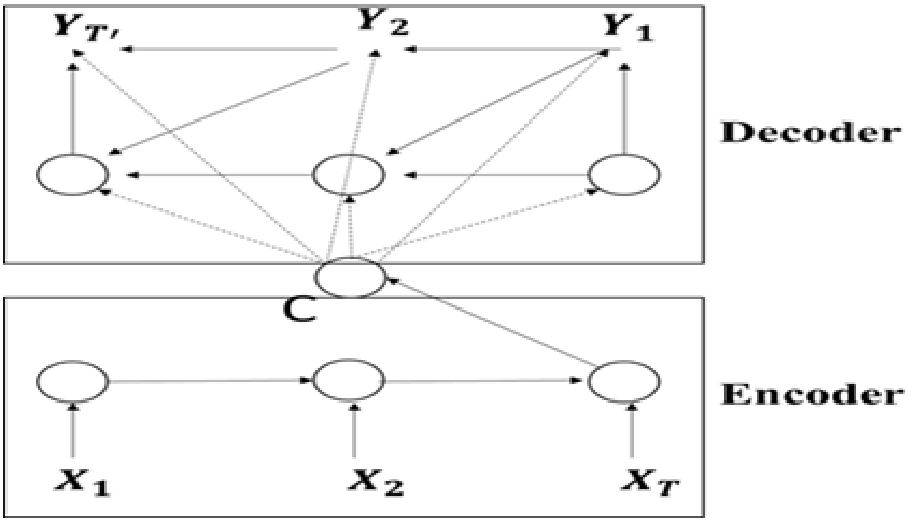


Figure 3.5.1 RNN encoder-decoder

3.5.d. Sequence to Sequence Learning

The Seq2Seq model is created by an LSTM cell. Firstly, the input of the word is generated from the vector file. In the vector file, each related word has an embedded value. Embedded values worked like the input of the encoder. The encoder saves the sequence value in short memory which is called LSTM. Here each sequence used a token to identify the end and start point of the sequence. In the program, we defined some special sequences such as <PAD>, <EOS>, <GO>, <UNK> etc. All of those special tokens are used for working in managing the sequence in the encoder and decoder.

<EOS> is used to indicate the end of the input sequence. In the encoder when the sequence of the input ends the <EOS> token automatic dis-select the sequence. Then the sequence will go to the decoder for decoding the sequence by giving related output. When the output sequence ends the <EOS> token stops the decoder. Figure 3.5.2 is the working process of the encoder and decoder. After the end of the encoding, the sequence needs a guide to enter the decoder.

Here we use a <GO> token to give the instruction of the encoding sequence to enter the decoder.

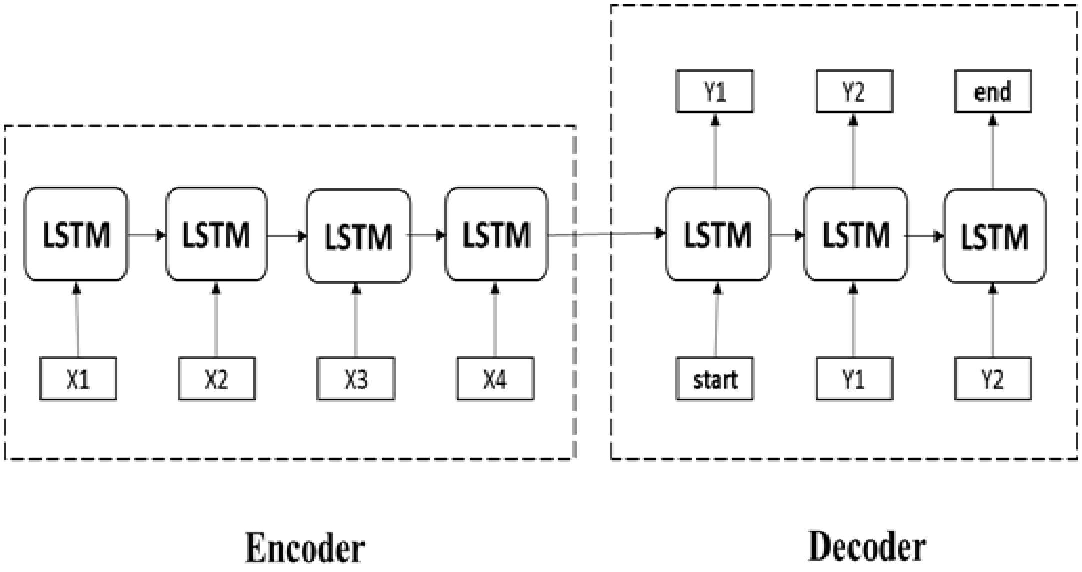


Figure 3.5.2 Chain to chain model

In the text sequence, some of the text or word are not replaced. All of that sequence needs to identify. Therefore, we used a special token <UNK> which means an unknown token. When an unknown token is set-up in the sequence it will be added to the <UNK> token in the text. In the train, the time sequence is divided into the batch. In a batch size similar length of the sequence needed to be together. Thus, we used a token which is known as a <PAD> token.

CHAPTER 4

Experimental Outcome Overview

4.1 Introduction

Abstract document abbreviation is the grim problem in the world of NLP. It is much more difficult for people to extract a borough text and summary from themselves. Since machine tries its best to give output based on its potential. After preprocessing, the machine needs to be trained to learn with the data model. For each training, the model has a backend engine. In this test, we completed the work with TensorFlow. Nearly initial values have been set. Such as epoch, keep probability, run size, batch size, learning rate, number of layers, etc. We have managed to reduce the time to train data. Here, using optimizer is the "Adam" for model adjustment. Higher configuration PC data training needs to be made easier.

Finally, we use Google Collab to train the model. Its ability to work is much faster and reduces time. Parameter Value is-

Table 5: Value of the Parameter

Parameter	Value
Epoch	150
Keep probability	0.70
Run size	256
learning rate	0.001
batch size	32

4.2 Implementational results

Almost the actual output is generated by the machine. It is known to everyone that no machine can give 100% accuracy. And thus, the trained model also provides a better outcome but it’s not give a better outcome for all values. And very often it also reacts towards false texts in terms of the actual text. But it keeps words similar to the meaning of the text most of the time or maximum in number. Our model is trained in 150 epochs and also can reduce the loss which is 0.0006. To check the output a file named "model.ckpt" has been saved in the model. Then, to reload the graph that was saved in previous steps we have created a TensorFlow session. After that the text and summarized data frame has been defined randomly to check the summary. And at last, the values into vocabulary have been converted to sequences and it was used as the input value for the model. In the past a logistic function was created to provide the response answer rationally. Based on the probability the rational function is the repercussion to the summary. And this value is calculated by the weight value and embedding value of the text. In table 6 we provide two sample outputs of our result. Each table contains the raw data of the original text and it was collected from online. A human was the provider of original summary for corresponding text. Input data was preprocessed. After training and learning the machine could generate the final variable response word.

Table 6: Sample example one of the response summary

Original Text:	এখনকার ছেলে-মেয়েরা সম্পর্কে জড়ালো সর্বপ্রথম যে জিনিস টা হারায় সেটা হলো তার নিজের ব্যক্তিত্ব।যা ১০০ ভাগের ভিতরে ৬০ ভাগ ছেলেই করে থাকে।আর যদি মেয়েদের হিসাব করা হয় তাহলে দেখা যাবে ১০০ ভাগে ৪০ ভাগ।নিজের ব্যক্তিত্বকে সম্মান করতে শিখুন।একটা মেয়ে কিংবা একটা ছেলের জন্য সম্পর্ক টিকিয়ে রাখার জন্য নিজের ব্যক্তিত্ব হারাবেন না।ভালো থাকুন। সবাইকে ভালো রাখুন।প্রতিদিনের মতো আজকের দিনটাও যেন সৃষ্টিকর্তা অনেক নিয়ামত দিয়ে অতিবাহিত করে।আমিন।
Original Summary:	নিজের ব্যক্তিত্বকে সম্মান করতে শিখা।
Input Words:	এখনকার ছেলে মেয়েরা সম্পর্কে জড়ালো সর্বপ্রথম যে জিনিস টা হারায় সেটা হলো তার নিজের ব্যক্তিত্ব যা ভাগের ভিতরে ভাগ ছেলেই করে থাকে আর যদি মেয়েদের হিসাব করা হয় তাহলে দেখা যাবে ভাগে ভাগ নিজের ব্যক্তিত্বকে সম্মান করতে শিখুন একটা মেয়ে কিংবা একটা ছেলের জন্য সম্পর্ক টিকিয়ে রাখার জন্য নিজের ব্যক্তিত্ব হারাবেন না ভালো থাকুন সবাইকে ভালো রাখুন প্রতিদিনের মতো আজকের দিনটাও যেন সৃষ্টিকর্তা অনেক নিয়ামত দিয়ে অতিবাহিত করে আমিন
Response Summary:	ব্যক্তিত্বকে সম্মান করতে শিখা।

Original Text:	কাউকে 'ভুলে যাওয়ার' ট্রাই করা উচিত না! তাহলে মনে পড়তেই থাকবে! এরচেয়ে বেটার যে যেভাবে আছে,সেভাবে থাকুক। মনের উপর প্রেশার দেওয়ার কি দরকার? পড়ে থাকুক। আপনি আপনার কাজ করে যান । বিজি থাকেন। দেখবেন 'আগের মতো মনে পড়তেসে না! জোর করে ভুলে যাওয়া যায় না। ভুলে যেতেও সময় লাগে! 'ভুলে যাবো ভুলে যাবো' বলে নরমাল জিনিস টা কে অতি সিরিয়াস বানানোর দরকার নাই। নরমাল ভাবেই থাকুক। দেখবেন আসলেই মনে পড়বে না খুব বেশি।
Original Summary:	জোর করে না ভুলার চেষ্টা করার দরকার নেই , একদিন এমনেই ভুলে যাবেন।
Input Words:	কাউকে 'ভুলে যাওয়ার' ট্রাই করা উচিত না! তাহলে মনে পড়তেই থাকবে এরচেয়ে বেটার যে যেভাবে আছে,সেভাবে থাকুক মনের উপর প্রেশার দেওয়ার কি দরকার পড়ে থাকুক আপনি আপনার কাজ করে যান বিজি থাকেন। দেখবেন 'আগের মতো মনে পড়তেসে না জোর করে ভুলে যাওয়া যায় না ভুলে যেতেও সময় লাগে 'ভুলে যাবো ভুলে যাবো' বলে নরমাল জিনিস টা কে অতি সিরিয়াস বানানোর দরকার নাই নরমাল ভাবেই থাকুক দেখবেন আসলেই মনে পড়বে না খুব বেশি
Response Summary:	জোর করে ভুলার চেষ্টা করার দরকার নেই , এমনেই ভুলে যাবেন।

4.3 Descriptive Analysis

We have made a mode for text summarization of English before creating the model for Bengali. Each model generates good outputs for different scenario. It is made so that we can reduces the function loss. Errors are reduced for the learning model. Loss function reduction is necessary vital for chain data. While training we have added up the loss function. After completing the train time, the final loss function is counted. Initially, the model has a high loss. But at last, the losses are gradually decreasing. Weight value which is 0. 008.Out data is divided in two segments one is train and other one is test. Finally, we have got 800 data for training and 200 for testing.

4.4 Moral

In this section we talked about how we experimented with our model, what are the outputs and how it creates a summary. Everything is talked about in detail.

CHAPTER 5

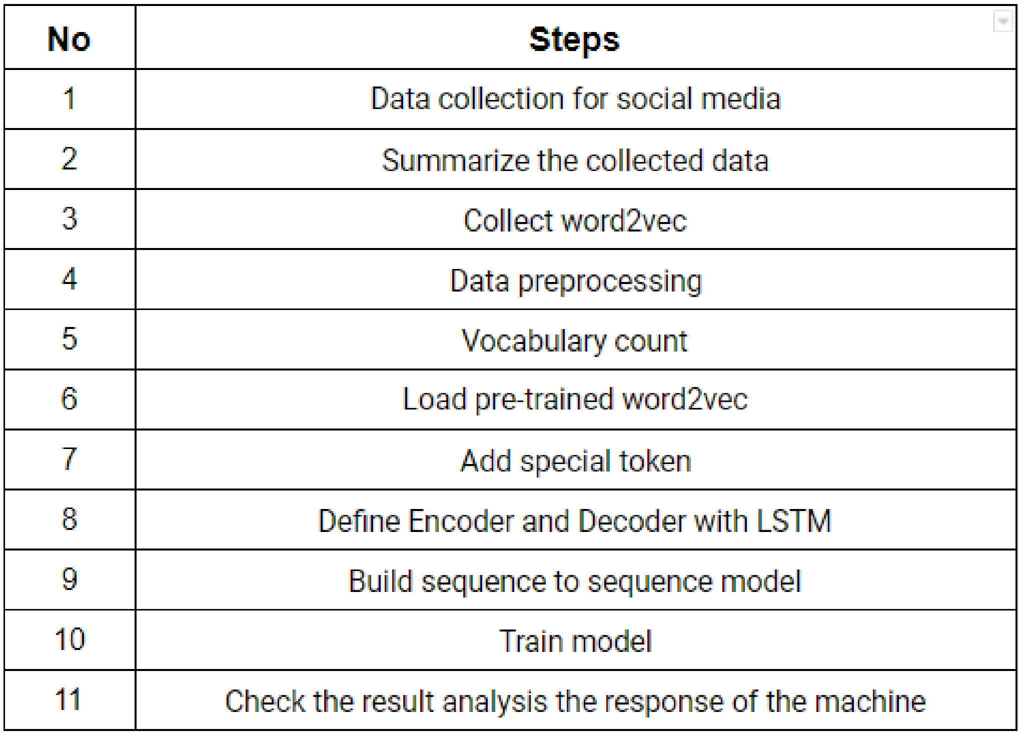
Summary and Future Work

5.1 Summary of the Study

Our Project is based on the Bengali NLP. In this project, we are trying to build up a Bangla abstract text summarization model using deep learning. For making an automatic Bengali text summarization this model is very helpful. The whole project has taken almost 5 months to be completed. The research & project work has several portions.

The whole summary of the project is given below with step by step.

Table 7: Project Summary



No	Steps
1	Data collection for social media
2	Summarize the collected data
3	Collect word2vec
4	Data preprocessing
5	Vocabulary count
6	Load pre-trained word2vec
7	Add special token
8	Define Encoder and Decoder with LSTM
9	Build sequence to sequence model
10	Train model
11	Check the result analysis the response of the machine

For further research, our proposed model possibly helps out our NLP research area to cut the long sentence using specific models for abstract lessons and compress Bangla text.

5.2 Conclusion

Core alarm of our paper is to further enrich NLP researching area. We create a model for Bangla text. The need for abbreviations is huge due to the number and length of online information. As it takes less time to read those short summaries, they also give us the whole text and an idea. Much work has been done in the past on summarizing the text. But we are trying to be more specific in our research. Although in the end we have encountered many obstacles in order to shorten the text we are trying to overcome those obstacles.

We used LSTM RNN with bi-directional to create the model but some errors still remain but we know that the machine never gives 100% efficiency. However, our build-ups have given the most accuracy to the loss of model training. Faced with some limitations such as limited text length, ample space for summaries but in the end, it was given the best, understandable, fluent and efficient summaries.

5.3 Recommendations

For the further steps of our research, we are trying to increase the dataset and the summary of that dataset to improve the performance of the model. For the Bengali summarization we will be trying to construct different models for the summarization that may be helpful to recognize the better performance. Our work is now only with the short sequence but need a better summarizer for Bengali text which are long sequences.

Some of the recommendations are stated below for text summarization.

- We need to understand the abstraction of the big text
- Reduce time for reading large texts
- Reduce the size of the text to keep the original idea
- Automatically summarizes text retrieval system

5.4 Further study

The model also has some limitations. As every research work changes every moment thus, we have also a plan how we would change in the future

- Dataset would be big
- More sequences/chains should be added
- More research should be done.
- No limitations length of text.
- Needs to update from TensorFlow version 1.15 to 2.4

We have the plan to make a web as well as mobile application using ai after our research is completed. This app will also provide Text Summarization of Bengali language.

REFERENCES

[1] Alhasan, Ahmad, and Ahmad T. Al-Taani. "POS tagging for Arabic text using bee colony algorithm." Procedia com- puter science 142 (2018): 158-165.

[2] Talukder, Md Ashraful Islam, et al. "Bengali abstractive text summarization using sequence to sequence RNNs." 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019.

[3] Lifeng Shang, Zhengdong Lu, Hang Li "Neural Responding Machine for Short-Text Conversation". Association for Computational Linguistics (ACL 2015)

[4] Dzmitry Bahdanau, K.Cho, Y.Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In- ternational Conference on Learning Representation (ICLR), 19 May 2014.

[5] Masum, Abu Kaisar Mohammad, et al. "Abstractive method of text summarization with sequence-to-sequence RNNs." 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019.

[6] Abujar, Sheikh, et al. "An Approach for Bengali Text Summarization using Word2Vector." 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019.

[7] Gao, Shen, et al. "Abstractive text summarization by incorporating reader comments." Proceedings of the AAAI Con- ference on Artificial Intelligence. Vol. 33. 2019.

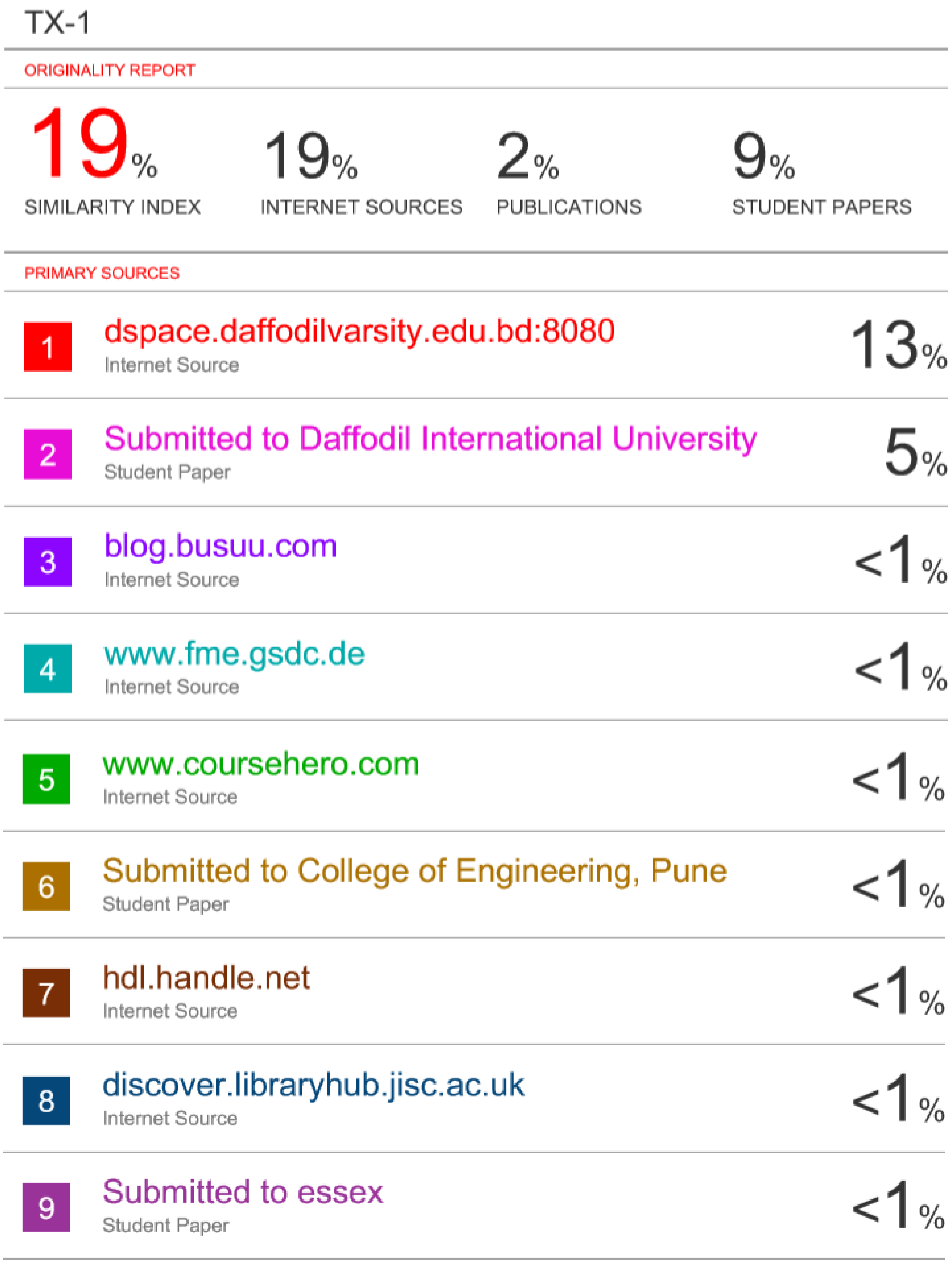
[8] K.Cho, B .van Merrienboer, D.Bahdanau, Y.Bengio " On the Properties of Neural Machine translation: EncoderDecoder Approaches". Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8),7oct 2014.

[9] Cho, K. et al. (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proceeding softhe2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)

[10] Abualigah L., Bashabsheh M.Q., Alabool H., Shehab M. (2020) Text Summarization: A Brief Review. In: Abd Elaziz M., Al-qaness M., Ewees A., Dahou A. (eds) Recent Advances in NLP: The Case of Arabic Language. Studies in Computational Intelligence, vol 874. Springer, Cham

[11] Qaroush, Aziz, et al. "An efficient single document Arabic text summarization using a combination of statistical and semantic features." Journal of King Saud University-Computer and Information Sciences (2019).

[12] Padmakumar, Aishwarya, and Akanksha Saran. Unsupervised Text Summarization Using Sentence Embeddings. Technical Report, University of Texas at Austin, 2016.



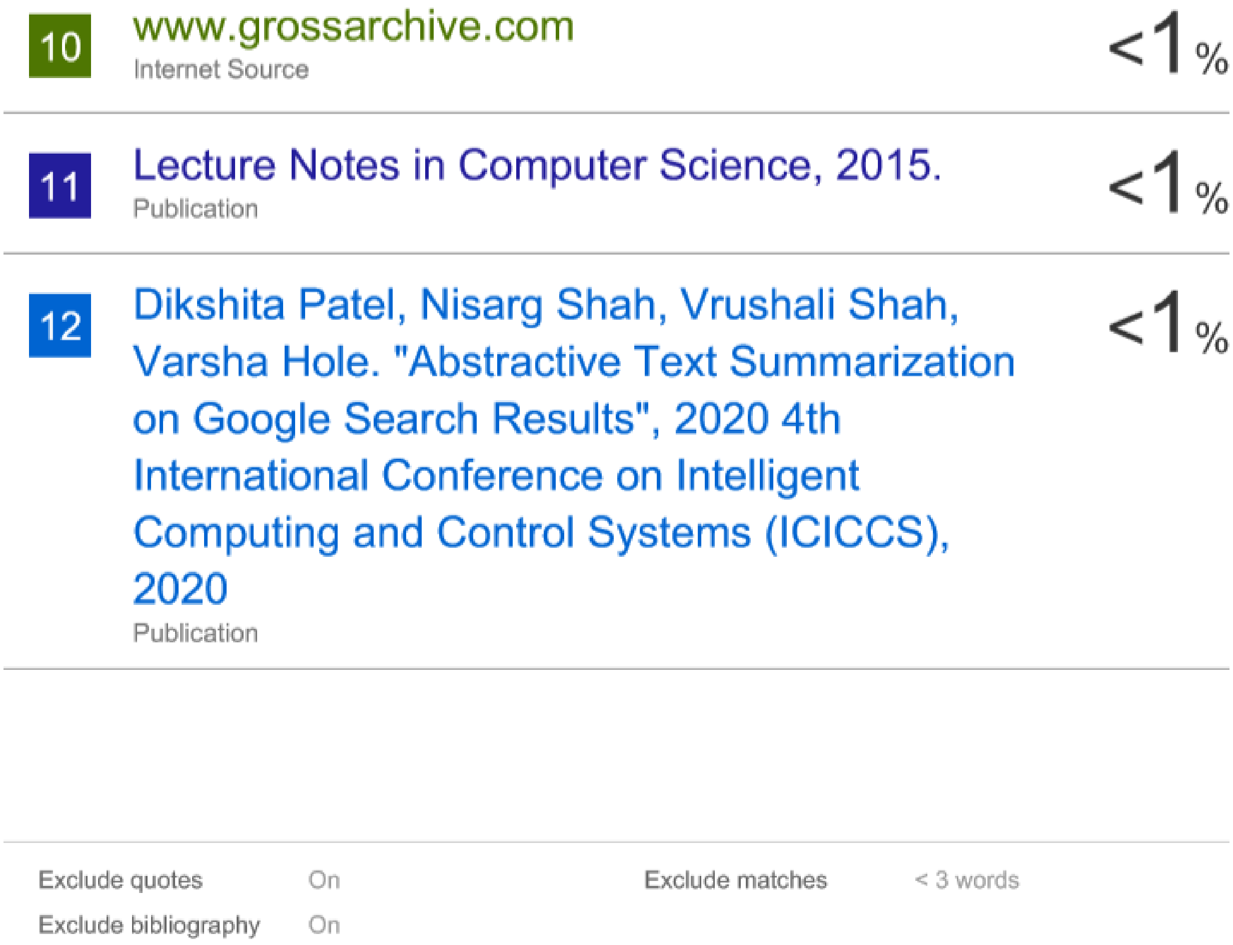
TX-1

ORIGINALITY REPORT

19%	19%	2%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	13%
2	Submitted to Daffodil International University Student Paper	5%
3	blog.busuu.com Internet Source	<1%
4	www.fme.gsdc.de Internet Source	<1%
5	www.coursehero.com Internet Source	<1%
6	Submitted to College of Engineering, Pune Student Paper	<1%
7	hdl.handle.net Internet Source	<1%
8	discover.libraryhub.jisc.ac.uk Internet Source	<1%
9	Submitted to essex Student Paper	<1%



10	www.grossarchive.com Internet Source	<1 %
11	Lecture Notes in Computer Science, 2015. Publication	<1 %
12	Dikshita Patel, Nisarg Shah, Vrushali Shah, Varsha Hole. "Abstractive Text Summarization on Google Search Results", 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		