A Deep Learning Approach For Bengali Text Summarization

BY

AMIT KUMER SORKER

ID: 192-25-797

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

Supervised By
Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By
Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

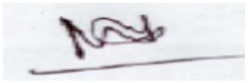
JULY 2020

APPROVAL

This Thesis titled "**A Deep Learning Approach for Bengali Text Summarization**", submitted by Amit Kumer Sorker to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Ismail Jabiullah — **Internal Examiner**
Professor
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

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of**, Abdus Sattar, Assistant Professor, Department of CSE** Daffodil International University. I 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:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Amit Kumer Sorker

ID:192-25-797
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

From the outset, I unequivocal appreciation to god for his ideal blessing to makes me possible to complete the master thesis effectively.

I might want to thanks my supervisor **Abdus Sattar** sir for his legitimate exhortation to finish this great research work for the Bengali language. His help and guidance gave me the fortitude to finish this exploration venture precisely. He served every one of us related assets and important data to do this examination for the Bengali language. I additionally thank our co-supervisor **Sheikh Abujar** sir bolsters us to finish this work. I am really an appreciation to our specialty head, **Dr. Syed Akhter Hossain** sir for his significant help to do such sorts of research work in the Bengali Language. Likewise, as to thank other employees and the staff of our area of expertise for their backings.

ABSTRACT

Anyone can depict his state of mind with the assistance of content. Thusly understanding the significance of the content is significant. In some cases, it is difficult to comprehend the significance of those writings, close by this is additionally tedious. The machine is the most ideal approach to take care of this issue. As a piece of AI, the content outline is an enormous field of research in characteristic language preparing. Assemble programmed content summarizer is the principle centring purpose of all examination. Content summarizer produces an essential part of an enormous report in a brief timeframe. Programmed content summarizer for different dialects has made already however not for the Bengali language. Expanding the devices and innovation of Bengali language is the fundamental objective of this exploration. In this exploration work, I have attempted to fabricate a programmed book summarizer for the Bengali language. However, working with the Bengali language was an extremely testing piece of this examination. Yet, until the end, I have made a base for programmed content summarizer for the Bengali language. The dataset utilized is gathered from an online web-based life. The profound learning model is utilized to make the summarizer. In the model, train time lessen the misfortune is legitimately influence the investigation result. I have diminished the preparation loss of our rundown model for Bengali content summarizer and which are competent to create a short book outline.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
NLTK	Natural Language Tools Kit
NLP	Natural Language Processing
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
NMT	Neural Machine Translation

CHAPTER 1

Introduction

1.1 Introduction

In the field of content synopsis, there are two classes. Abstractive and Extractive content synopsis. Abstractive content summarizer contains a theoretical of the content archive. Essentially giving unique is the portrayal of the fundamental thought of the content however here summarizer doesn't rehash the first sentences. Here is the primary test to finding the essence of the content in common language handling. The most extreme number of research work is hung on the extractive content outline. Concentrate catchphrase and locate the most continuous words from the content is the fundamental thought of extractive content rundown. Yet, create another word or sentences dependent on content is the most testing stuff. This isn't required to have the word in the giving abstractive rundown is additionally present in the first set. There is a lot of abstractive book rundown investigate work has done already in various dialects. In this time, I have attempted to construct an abstractive content summarizer for Bengali text applying profound learning calculations.

Bengali is one of the most utilized dialects on the planet. Expanding the instruments and innovation for this language is significant. In this way, the examination region of the Bengali language needs an extension. A programmed framework message should be prepared. NLP apparatuses and library especially assists with preparing any sorts of content. Working in the Bengali language to construct a programmed framework is troublesome contrasted with different dialects. Since some NLP libraries are not worked for the Bengali language along these lines, all methods and libraries are applying by crude coding. My exploration work can give an abstractive content synopsis to Bengali content. No machine gives 100% exact outcomes each time however greatest time a palatable outcome can be gotten. My programmed abstractive content summarizer has likewise resembled that. All the produced rundowns are not 100% precise but rather the most extreme reaction of machine outline is palatable for Bengali content synopsis.

1.2 Motivation

The content summarize is a quicker method to abridge a long book or content record. Concentrate the principle catchphrases and make an important significance of relating content is the fundamental thought of rundown. Perusing long content smoothly and furthermore locate the fundamental dynamic structure content is extremely difficult and tedious. In some cases when we read along with book however can't comprehend the importance. In the event that the report size is numerous archives, it is significantly harder to locate the theoretical. Thusly, programmed content summarizer is an instrument that can assist with condensing the content inside a brief timeframe. Programmed content summarizer additionally distinguishes the complete reports, words, all-out frequented word and which words are significant for the content record.

Today we invest a ton of energy in web-based life, perusing pages, news stories and online journals yet after some time may fill exhausted. Reason for unstructured information and indistinct significance. That is likewise a reason which intends to require a book summarizer. Abstractive content summarizer is content outline approach which finds the significance part from the given content archives, that is not compulsory if containing rundowns words present or not in the first content reports.

Information is one of the most important things in this cutting edge world. Consistently countless content information delivered from various sources. This colossal number of information needs tremendous memory space which is generally expensive and makes an issue in putting spaces. In this manner, summarizer makes a rundown of those long content and decrease the size of the record and put just centre data by evacuating the superfluous content. That is the reason a programmed book summarizer is required for present-day innovation.

Bengali is our language. NLP asset for this language isn't adequate for the client. That is the reason we have to manufacture NLP assets and devices and advances. So the fundamental focal point of this examination is building a programmed abstractive book summarizer of the Bengali language to arrive at the Bengali NLP treasure.

1.3 Rational of the study

History of the Bengali language is exceptionally rich. Today a huge number of individuals utilize the Bengali language as their local language. Be that as it may, in this cutting edge period, the apparatuses and innovation of Bengali language are not rich like different dialects. Along these lines, we have to build innovations for this language. The vast majority of the content-related issues can be explained by NLP apparatuses and strategies. Content outline is a central issue of NLP. A book summarizer contains the theoretical of a long arrangement of content. It causes human to comprehend the significance of a long book effectively with a familiar and mistake-free synopsis. Most and significant NLP procedures as of now work for different dialects, for example, English, French, Chines and so on. Be that as it may, for Bengali content, a couple of models have been constructed which isn't sufficient. Subsequently, the examination region of Bengali NLP should be expanded. The primary hindrance for Bengali content is preprocessing. The machine can't see a portion of the Bengali characters and images. To deal with this issue needs to utilize the Unicode of those characters or image. NLTK library isn't accessible for Bengali content. That is the reason Bengali devices don't perform precisely as like as different dialects. Research is the best way to give an answer to these sorts of issues. Along these lines, in this examination work, we attempt to tell the best way to forms Bengali language and make abstractive content summarizer for the Bengali language. That encourages us to lessen the size of the report and give a familiar synopsis in shot time.

1.4 Research Questions

- What is Bengali language text summarization?
- How Bengali content synopsis functions?
- What are the advantages of Bengali content summarization?
- What are the contrasts among Bengali and English content summarization?
- How to preprocess Bengali content in NLP?
- What are the future works of Bengali content outline?
- How Bengali content rundown Model functions?

1.5 Expected Output

Since this is an examination venture, our primary concern was to distribute an exploration paper in a related field. Research work constantly a nonstop procedure. Numerous individuals investigation explicit research themes to locate a proficient arrangement. At that point, the designer builds up the devices for the clients. The most extreme number of research work and devices are created utilizing extractive content outline in the Bengali language yet not in an abstractive content rundown. Likewise, numerous analyst and engineer are not intrigued to give their dataset and assets to everybody. Therefore, some exploration work is not, at this point been utilized. In Bengali language content synopsis is another examination theme. Some exploration works are done in past for Bengali content outline. In any case, the outcome was insufficiently agreeable for making a programmed Bengali book synopsis. A programmed framework is reliant on the machine. Accordingly, the machine needs to learn. At that point, the learning model is working in the backend of a framework, for example, a web or versatile application. In this exploration, we present an AI technique for abstractive Bengali content outline and tell to important strides on the best way to fabricate a model for programmed Bengali content synopsis.

1.6 Report Layout

In this report have an aggregate of 5 parts. Section 1 contains a review of the entire research work. It has a few areas, for example, 1.1 Introductions of the work, 1.2 Motivation of this exploration, 1.3 Rational Study of the pursuit, 1.4 Research Questions, 1.5 Expected Output and 1.6 Reports Layout of the examination. In Chapter 2 we have talked about Background Studies of the exploration and its subsections are 2.1 Introductions, 2.2 Related works, 2.3 Research Summary, 2.4 Scope of the Problem, 2.5 Challenges. In Chapter 3 we have talked about the entire Research Methodology with subsections 3.1 Introduction, 3.2 Research Subject and Instrumentation, 3.3 Data assortment strategy, 3.4 Statistical Analysis of Datasets, 3.5 Implementation Requirements. In Chapter 4 Experiment and Results of the examination are talked about and the subsection is 4.1 Introduction, 4.2 Experimental Results, 4.3 Descriptive Analysis, 4.4 Summary. Part 5 contains the Conclusion and future works of the exploration with the subsections 5.1 Summary of the Study, 5.2 Conclusion, 5.4 Implication for Further Study. End of all segment given the references which helped us in our exploration work.

CHAPTER 2

Background Studies

2.1 Introduction

Content summarization is a procedure which makes a short perspective on long content or long content record. Content record has a long succession of content. Discovering short, smooth and justifiable synopsis is the fundamental spotlight of content summarization. Making a synopsis of a long book archive for a human is tedious and exorbitant. Therefor programmed content summarization is an extraordinary method to support a human. AI approaches are the best way to build up a programmed framework. Programmed content summarizer causes us to lessen the size of the record and spare memory space which likewise decreases the expense of room.

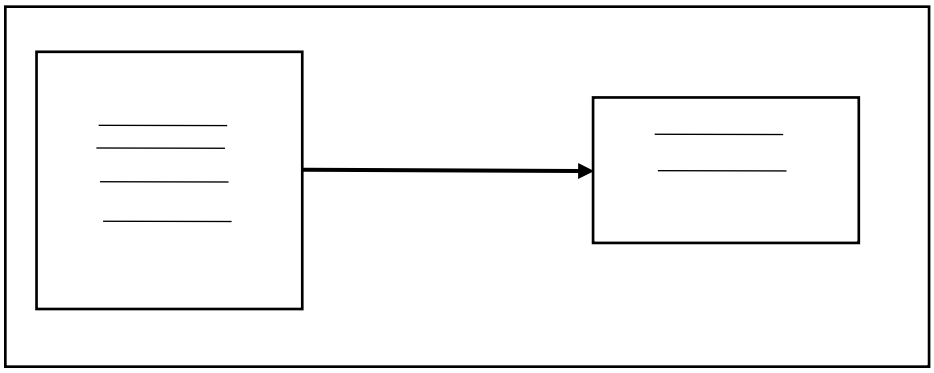


Figure 2.1.1: Demo figure of text summarization.

Consistently an immense number of archive forms on the web, for example, Facebook posts. An enormous number of clients finding or search data structure a record in time. Be that as it may, sparing and discover the data from that archive is an intricate procedure and furthermore need a tremendous space to store those measure of data. Along these lines, the outline method is utilized to tackle this issue with a proficient, quick and compelling way. In memory, spare the theoretical, principle words and sentences of the content archive at that point required they serve the data rapidly. There are two sorts of content synopsis: Abstractive and Extractive content rundown. The abstractive content outline contains unique of the long content which isn't required to introduce or not present the content archive. Where extractive content rundown contains the fundamental words, periods of the first content record to condense content archive.

2.2 Related Work

Text content summarization is the most researchable point in normal language handling. Many research work has been done in this field for various language. Significant research was hung on extractive content summarization yet not many in the abstractive content summarization. This segment will be examined about some honorable works in these fields.

Any language displaying is reliant on machine interpretation. Machine Translation encourages a machine to comprehend the handled content and assists with making a programmed framework. NMT is another way to deal with machine interpretation. Bahdanau et al [1] present another methodology for NMT. They utilized jointly figuring out how to adjust and improve the exhibition of typical encoder and decoder strategies. Vector of the content sentences is worked input the encoder where the decoder produces the conceivable yield of the vector groupings. After that expanding the effectivity of NMT, the Attention-Based [2] techniques are presented in later. NMT has a few restrictions, for example, preparing and testing are increasingly costly and can't give a decent outcome to an uncommon word in the arrangement. Wu Y et al [3] introduced a GNMT framework for diminishing the intricacy of NMT.

RNN is a method to take care of content-related issues. Nallapati et al [4] demonstrate grouping to arrangement getting the hang of utilizing RNN's give beating result for abstractive content synopsis. Encoder contains the fixed length of the information succession of the vector and decoder produces the most related grouping of the encoded arrangements. Improve the presentation of abstractive content synopsis DRGD is utilized [5] in ongoing time. Support learning is the advanced use in the abstractive content rundown. Wang et al [6] proposed a model utilizing fortification system for the abstractive content rundown. This time they use convolutional succession learning for abstractive content synopsis. Relevant learning of content is the fundamental thought of this work. The entire procedure is needy the CNN [7].

Ilya Sutskever et al [10] present a strategy for seq2seq picking up utilizing multilayer LSTM. One maps the information grouping structure the objective content vector which is the encoder. Another translates the grouping vector which is the decoder. Thusly, the LSTM has no unpredictability in working with the long grouping. Turning around the request for grouping is conceivable in this procedure.

For the rundown of long haul content, for example, Wikipedia extractive content outline utilized broadly. Subside J. Liu et al [11] present a decoder rather than encoder and decoder model which can produce a familiar outline from the long content succession. Additionally, this technique can produce multi-report synopsis from a huge and resembled dataset. Lifeng Shang et al [12] give a technique for the short content rundown. In this recent years few Bengali text summarization works are done [18-23]. Seq2seq learning model with LSMT is provide good result for the summarization. In this examination work, we have utilized succession to arrangement learning for making abstractive content rundown. Follow the consideration instrument for making the abstractive content synopsis. Utilizing dataset contains the short Bengali content and their relating synopsis. After the utilizing of encoding and interpreting component model give a decent outcome to the Bengali abstractive content synopsis.

2.3 Research Summary

My analysis, I have presented a strategy for Bengali abstractive content synopsis. I manufacture a model utilizing profound learning. To assemble this model i have utilized not my own dataset and collected from online. Dataset has gathered structure web-based life. From the outset gather Bengali status, remark, page and gathering posts from Facebook. At that point make a synopsis of every Bengali content. In this way, the dataset contains two segments, one is Bengali content and another is their comparing rundown. The all outnumber of 200 information with their rundown in the dataset. Before making a profound learning model they have preprocessed the Bengali content. In the preprocessing stage, from the start split the content and afterwards include Bengali compressions and expel prevent words from the content. In the wake of preprocessing, i have checked the jargon of entire information. Word inserting is significant for profound learning model. Word vector assists with sparing the related jargon in a record with a numeric worth. I utilized a pre-prepared word vector document for Bengali content which is accessible in on the web. I manufacture a grouping to arrangement model dependent on consideration model. In this model encoder and decoder is utilized with Bi-directional LSTM cell. Word vector is the contribution of the encoder and significant word vector in the decoder is the yield of the model. An encoder and decoder to pass the grouping need a token which is referred to as an uncommon token, for example, PAD, UNK, EOS and so forth. In the wake of pronouncing and characterize

all capacity and library, i train the model for over 3 hours. At that point, I found a decent reaction from the machine.

2.4 Scope of the problem

NMT is another way to deal with machine interpretation. It relies upon the encoder and decoder. The encoder gives a grouping which has a fixed length and the decoder creates the right arrangement structure the encoder succession. For that in machine interpretation, NMT gives a decent outcome to short length grouping of content [8]. Since the content outline is new research in Bengali NLP diverse method is created step by step. This exploration work utilizes NMT for Bengali content rundown dependent on short Bengali content succession. In dataset content length isn't enormous yet sufficient for rundown however the content outline is the short length inside sentences. This examination reason we utilized NMT model with encoder and decoder. So short content is the reaction to a decent outcome for content summarizer. In calculation additionally reactions outline for a short succession. Be that as it may, it is hard to deal with the long content successions and outline them. In this manner, there is another territory of research to assemble a long or any length content synopsis in Bengali content summary.

2.5 Challenges

Organized Bengali information isn't accessible. All information are available in an unstructured manner. Thusly, information assortment is a test for this exploration. Some paper dataset is accessible however a portion of the exploration work has nearly done utilizing this dataset. In this way, I need another dataset to finish this examination work. After the assortment of the dataset, the content information make an outline of that content is another difficult work. Working with Bengali content is continually testing. The Library of NLTK is accessible for Bengali content preparing. In this manner, in preprocessing step need to crude coding to set up the content as a contribution of a model. Assume, while expelling accentuation from the content, need Unicode of every accentuation and evacuate it by crude coding. Another issue is preventing word expel from the content. For different dialects like English have a form in the library to expel prevent words from the content.

CHAPTER 3

Research Methodology

3.1 Introduction

In this area, I will talk about the entire system of the thesis work. Each exploration work has a one of a kind explaining procedure. Applying all methodologies are remembered for the technique part. Here give an itemized conversation of applying utilizing model with a short portrayal of every individual pieces of the system. In my exploration, i have utilized profound learning model for content summarization. The profound learning calculations are utilized after the sort of research contain. RNN is utilized for taking care of the content related issue in profound learning. Each profound learning model needs a decent dataset to locate an exact programmed framework. Before applying the calculation dataset should be gathered and preprocessed. In full part every segment of philosophy are examined exclusively. Given all segment are followed when the exploration work in finishing. A superior clarification of technique increment the proficiency for work and give the respectability. Scientific condition and graphical perspective on the model with their portrayal is assisting with understanding the entire work. In this manner, further research and expanding the exploration documented great clarification of system is required.

Entire work resembles a structure. All means of the procedure are quickly talked about in this part. Sub segment of some center segments are assists with understanding the essence of the model with it reason for utilizing. Working progression of entire research work is given beneath which give a short perspective on complete research work.

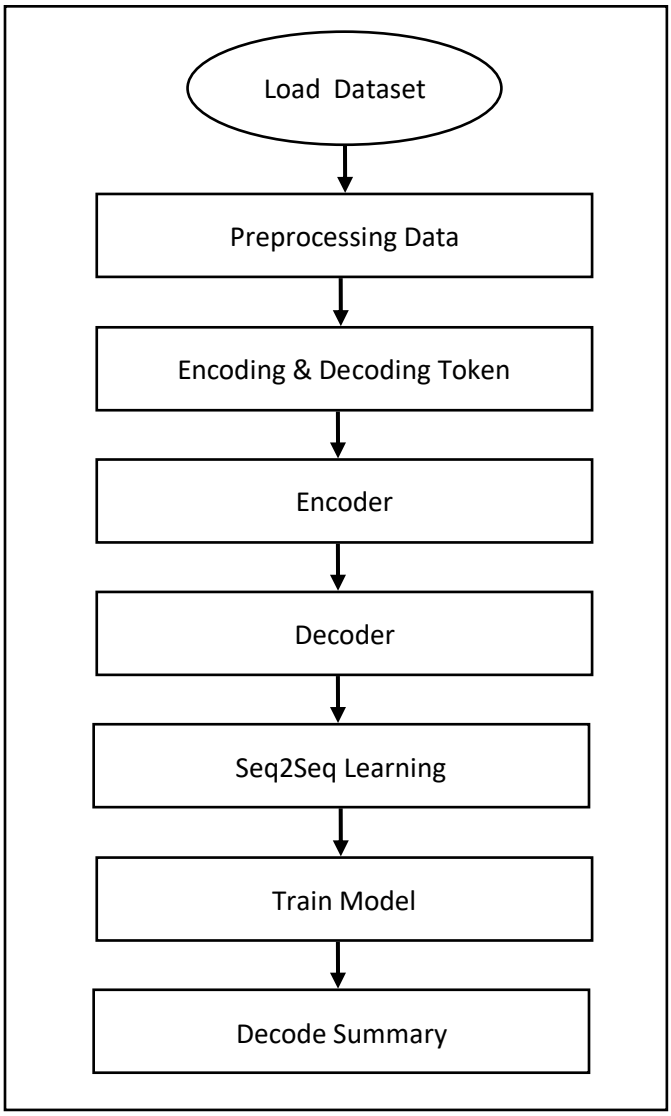


Figure 3.1.1 Content summarization process flow.

3.2 Research Subject and Instrumentation

I proposition subject name is "A Deep Learning Approaches For Bengali text summarization". This is a significant research region in Bengali NLP. I have examined the way toward making an abstractive content synopsis in Bengali with the calculated and hypothetical procedure first to now. A profound learning model needs high design pc with GPU and other instruments. Presently a rundown is given underneath of the necessary instrument for this model.

Hardware and Software:

- Intel Core i3 7th generation with 4GB RAM

- 1 TB HDD

Development Tools:

- Windows 10
- Python 3.7
- Keras
- Tensorflow Backend
- Pandas
- Numpy

3.3 Data Collection

I used collected data from online. All data is accumulated from virtual life, for instance, singular status, Blog posts, page posts and assembling posts. There is some complexity to accumulate data from online life for security issues. For that, they accumulate data using manual procedures. The total of 200 posts is accumulated structure web-based life. The dataset is utilized in making Bengali abstractive synopsis [19]. By then dataset necessities to preprocess and convey an ideal book for the model. The basic steps of the preprocessing stage in the figure 3.3.1 are discussed region adroit in underneath.

3.3.a) Split

In the wake of moving the dataset whole substance ought to have been tokenized. Those are divided the long substance to a lone word. Which is helpful for clean substance and empty silly limit structure the dataset. Also, it makes the language of the dataset which is noteworthy for NLP issues. This language is used to find significant information structure word embedding report.

3.3.b) contractions

Each language has tightening influences of specific words. Also, the Bengali language has relatively few compressions for some word. It suggests the short sort of a word or short making technique out of a word. The machine doesn't fathom the short sort of a book for that need to portray the full significance of the short kind of the word. Those words are not utilized typically. These words are utilized in a couple of regions, for example, notice, billboard and so on.

3.3.c) Remove expression

The standard enunciation is used to empty the remarkable character or unfortunate character to remove from the substance. Oust space, whitespace, English character, highlight structure content, Bengali digit from content is the rule use of the standard enunciation in our assessment.

3.3.d) Remove stop words

The clear stop word content is a common strategy in NLP. Inconsequential word removes from the substance is the rule usage of stop word. In NLTK work in library produce to oust keep word from the English substance. Nevertheless, there is no library open for Bengali stop word to oust. As such, from the beginning, i assemble all stop word in the Bengali language from on the web. There is an all dwarf of 393 stop words and a short time later, I inserted it into a record for extra usage.

3.3.e) Clean content & summary

After completing the previous steps, the sequence of the text and summary will look clean. Here all text and summary have not any punctuation or any extra space. All words look in a sequential order. Both clean text and summary are inserted into two different lists. Those lists are used input sequence of the summarization model. In table 1 an example of text preprocessing is given below.

Table 1: content processing example

Raw content	Clean content
বাসায় নোকিয়া ফোনে 'সাথিয়া' গানের মিউজিকে রিংটোন না দিলে তুমি নাইন্টিজ কিড না... 'গোলাম' দেখে জিহ্বায় ম্যাচের কাঠি নেভানোর চেষ্টা না করলে তুমি এই দলের বহুত বাইরে... আশেপাশের বর্তমানে মুরুব্বি বড় ভাই স্থানীয় অন্তত কেউ একজনের 'তেরে নাম' স্টাইলে চুল কাটা মনে করতে না পারলে তোমার নাইন্টিজ কমিউনিটির সাবসক্রিপশান ক্যান্সেল।	বাসায় নোকিয়া ফোনে সাথিয়া গানের মিউজিকে রিংটোন না দিলে তুমি নাইন্টিজ কিড না গোলাম দেখে জিহ্বায় ম্যাচের কাঠি নেভানোর চেষ্টা না করলে তুমি এই দলের বহুত বাইরে আশেপাশের বর্তমানে মুরুব্বি বড় ভাই স্থানীয় অন্তত কেউ একজনের তেরে নাম স্টাইলে চুল কাটা মনে করতে না পারলে তোমার নাইন্টিজ কমিউনিটির সাবসক্রিপশান ক্যান্সেল

3.4 Statistical Analysis

1. The all out number of information 200. 200 information have 3 subsections, for example, Post type, Text and Summary. A short perspective on our dataset is given beneath in table.

Table 2: Different types of data in the dataset

Types	Content	Summary
Personal post	জীবনে অনেক সময় বিপদও আশির্বাদে রুপান্তর হয় কারন অনেক সময় আর টাকা খরচ করে মানুষ চেনা না গেলেও বিপদে সহজেই শত্রু বন্ধু চেনা যায়।	বিপদে সহজেই শত্রু বন্ধু চেনা যায়।
Group post	রাস্তাঘাট সাধারণ মানুষের চলাফেরার বদলে হরতাল,অবরোধ, সমাবেশ নামক নাটক প্রর্দশন এর মঞ্চ বানিয়ে দেওয়া হোক। সাধারণ মানুষ উড়াও এর মাধ্যমে আকাশ পথে চলাফেরা করবে।	মানুষ উড়াও এর মাধ্যমে আকাশ পথে চলাফেরা করবে।
Page post	স্বপ্নীল একটা সূচনার পর মাইকেল ওয়েনের ক্যারিয়ার নিয়ে ছিল অনেক স্বপ্ন, অনেক আশা। কিন্তু কিছু ভুল সিদ্ধান্ত, সাথে দুর্ভাগ্যজনক সব ইনজুরি – সব মিলিয়ে প্রত্যাশার কাছাকাছিও তিনি যেতে পারেননি।	ভুল সিদ্ধান্ত সাথে দুর্ভাগ্য।

2. Total 5k present vocabulary for the dataset.
3. Total 500k pre-train word embedding is used for this work.
4. 4.5k single and unique word present in the dataset.
5. In the model total 80% word is used for model implementation.
6. Maximum content size is 288 and summary size is 19.
7. Microsoft excel file is used for save the dataset.

3.5 Implementation Requirements

3.5.a) Mathematical discussion

In the dataset, the quantity of the content and their outline is equivalent. By and large, the content has a long length yet the outline has a short length contrasted with one another. Presently consider P contains the quantity of expressions of information content arrangement of the dataset. Accordingly $x1$, $x2$,…. is input arrangement and which is originating from the jargon need to estimate V. That creates the yield arrangement, for example, y1, y2,… , yd, here S>P. That implies the grouping of the outline is not exactly the information content record. Notice that all arrangement is originating from a comparative jargon.

3.5.b) Word embedding

Tallying the jargon is significant for word inserting. My tally jargon dependent on the dataset. At that point, the jargon is utilized in a word implanting. For word, inserting needs a word vector. Here i utilized a pre-prepared Bengali word vector record which was gathered structure on the web. Word to vector record contains a numeric worth related word. It spares the expressions of all related word in a square. At that point utilize the worth when working. Those vector of the word is utilized as the contribution of the model at that point give related word which will be the yield of the model. In this way, the arrangement of grouping learning is effectively finished its activity. We have utilized "bn_w2v_model" which is accessible online for everybody.

3.5.c) Encoder & Decoder

After the development of machine interpretation, a profound learning calculation makes an extraordinary achievement in the Artificial Intelligence field. All content related issue are given precise yield in the profound learning model. RNN is the most usable calculation in profound learning. It works all the more effectively in any content related issue. Each RNN are made by LSTM cell. LSTM cell resembles a momentary memory. Encoder and decoder are utilized in LSTM cell. The information content is go in the encoder where each info is word vector grouping. The decoder takes the information grouping and produces the yield of the content from the significant content organize.

On the off chance that we consider x is an objective arrangement of sentences than the most extreme likelihood of the word vector succession will be x. Here y is the source arrangement of the sentences then likelihood will be,

$$arg\left(max_y p(x|y)\right) \ldots\ldots (1)$$

There are two kinds of RNN, one-directional and another is Bi-directional. One directional RNN has info and yield each is associated with others in a consecutive way. Bi-directional RNN has two layers [9] with two bearings. One is the forward bearing and another is reverse way. Those are utilized to take care of the machine interpretation issue. In our work, we utilized two layered RNN. Since we utilized RNN for the Bengali language at that point fixed the length of Bengali is the contribution of the model which is conveyed by the encoder. Decoder gives the related succession of the yield relies upon the info. Here the fundamental computation is working dependent on the likelihood figuring. On the off chance that X is the all-out info succession where $X = (x_1, x_2, x_3 \ldots\ldots x_n)$ and in the event that c is the setting vector, at that point the grouping will be,

$$h_t = f(x_t - h_{t-1}) \ldots\ldots (2)$$

And c,

$$c = q(\{h_1, \ldots\ldots h_{T_x}\}) \ldots\ldots (3)$$

Suppose, $y = \{y_1 \ldots\ldots, y_{T_y}\}$ is yield succession anticipated by Decoder. At that point the reaction or gave synopsis likelihood will be,

$$p(y) = \prod_{t=1}^{T} p(y \vee \{y_1, \ldots, y_{t-1}\}, c) \ldots\ldots (4)$$

$$p(y \vee \{y_1, \ldots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \ldots\ldots (5)$$

3.5.d) Seq2Seq model

Seq2Seq model is created by LSTM cell. Firstly, the input of the word is formed from the vector file. In the vector file, each related word has an embedded value. Those embedded values are worked like the input of the encoder. The encoder saves the sequence value in short memory which is LSTM. Here each sequence used a token to identify the end and start point of the sequence. In

the program, we defined some special sequence. All of those special tokens are used for working in handling the sequence in the encoder and decoder. <END> is used to identify the end of the input sequence. In the encoder when the sequence of the input ends the <END> token automatic discard the sequence. Then the sequence will go to the decoder to decode the sequence by providing related output. End of the decoder that means when the output sequence ends the <END> token stop the decoder. Figure 3.3.2 shows the working process of the encoder and decoder. After the end of the encoding, the sequence needs an instruction to enter the decoder. Here we use token to give the instruction of encoding sequence to enter the decoder. In the text sequence, some the text or word are not replaced. All of that sequence need to identify. Therefore, we used a special token which means an unknown token. When an unknown token is found in the sequence it will be added 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 token.

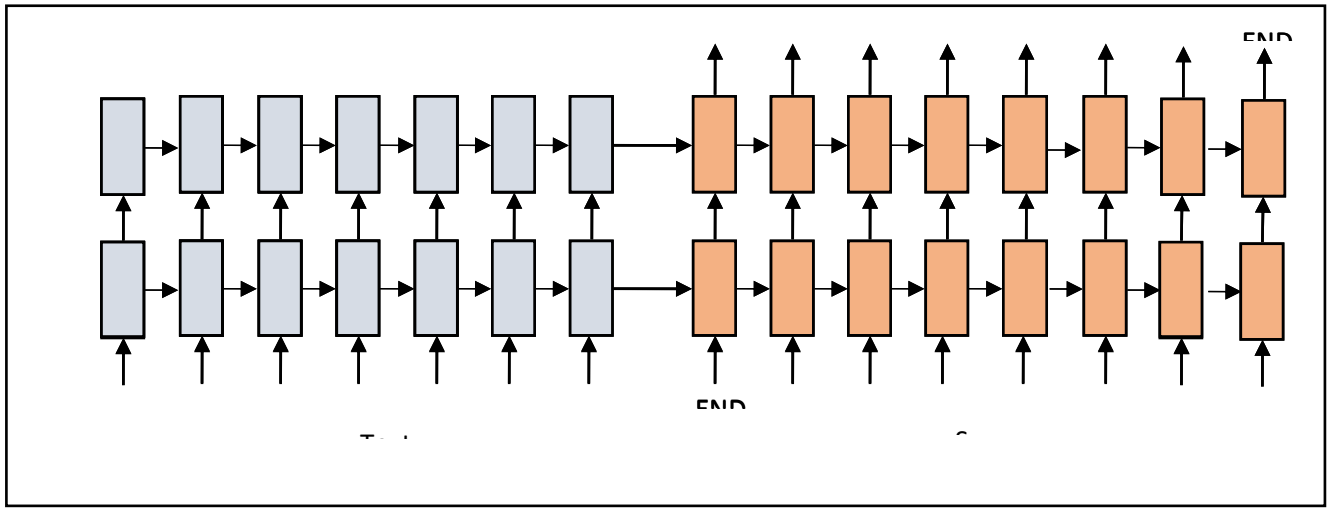


Figure 3.3.2: Seq2Seq model for content summarization

The above figure describes the Seq2Seq model process of the sequence. Also given a view of how a long sequence can provide a probable output sequence. In the encoder, Bengali text is the

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

Abstractive content summarize is a perplexing issue in NLP. The machine can abridge the content naturally that isn't obligatory for reaction rundown are available or not present in the content. Along these lines, locate a precise outline is troublesome. Likelihood computation is significant for this content summarizer. Since the machine gives yield based on the most extreme likelihood. In the model load of each word are learning in the train time at that point figure the likelihood reaction the outline dependent on the related word weight. After the preprocessing, the content information needs to prepare for the model to gain proficiency with the machine. For preparing, every profound learning model has a backend motor. This trial i utilized TensorFlow 1.15.0 for a backend work enginee. For the train, some essential parameter esteem needs to characterize. For example, ages, cluster size, learning rate, number of layer and so on. Those parameter preparing is reliant. Lessen the misfortune in train time is significant. In this test, we have utilized "Adam" streamlining agent to diminish the misfortune and advance the model. An all-around prepared model can give a hitter yield in the test time. High setup pc needs to prepare information in the profound learning model. GPU is exceptionally useful for that.

This investigation we have not to work in GPU to prepare our model. For that, i train my model right off the bat in direct pc. That sets aside enormous effort to prepare the model and the given yield isn't acceptable for summarizer content. At last, we train the model utilizing google colab. Which gives free GPU administration to the client. That is absolutely first and decrease train time. Presently the estimation of the parameter is given beneath which is utilizing this trial.

- ✓ Set the batch size = 8
- ✓ Hidden unit = 50
- ✓ Total epochs = 100
- ✓ Define the model Learning rate= 0.01
- ✓ Data validation=0.1

4.2 Experimental Results

The machine gives yield about the real yield. Everyone realizes that no machine gives a 100% exact yield. Additionally, our prepared model gives a decent outcome however not for all qualities. At times it reactions wrong content comparing to the content. Yet, the greatest number of reaction words is like the importance of the content.

I train my model in 70 ages at that point decrease the misfortune which is 0.008. For checking the yield, I spare the model in a record whose name is "model.ckpt" document. At that point i make TensorFlow meeting for reloading the chart which was spared in past advances. At that point characterize the content and synopsis information outline haphazardly to check their rundown. From that point onward, i convert the in incentive to jargon for the succession which was the contribution of the model.

Already i have made a calculated capacity to offer the reaction response legitimately. The calculated capacity is the reaction to the synopsis dependent on the likelihood. The likelihood esteem is determined by the weight esteem and installing estimation of the content. Two example yield of our outcome is given beneath in table 3 and table 4. Each table unique content contains the crude information which was gathered structure on the web. The first outline is given by a human to relating content. Subsequent to preprocessing of crude content, the content is changed over into unadulterated content which is Input words. Last factor reaction word was given by machine after the preparation and learning.

Table 3: Example of sample model response 1.

Content:	সুযোগ পাইলে কেউই হাত ছাড়া করে না,কিন্তু সমস্যা টা হল একটা যায়গাতেই ,নিজে সুযোগ পাইনাই তাই অন্যকেও পাইতে দিবো না,অপ্রত্যাশিত ভাবে যদি সে পেয়েই যায় তাহলে মূল্যহীন করে দিবো (চিন্তাধারা)!!!
Summary:	নিজে সুযোগ পাইনাই তাই অন্যকেও পাইতে দিবো না।
Model Summary:	নিজে সুযোগ পাইনাই অন্যকে দিবো না।

Table 4: Example of sample response 2.

Content	ব্যস্ত নগরীর শত শত অট্টালিকার ভিড়ে,তুমার হাসি আজও আমার কানে বাজে।অভিমানে আছ আজ বহুদূরে,খুজে ফিরে তুমাকে আমারই চারিপাশে।তুমার কথা ভাবতে গিয়ে হয়েছে কত না আধার রাত,আমি আর নিস্পাপ ঐ চাদ সারারাত কথা বলে করেছি পার।।
Summary:	আমি আর নিস্পাপ চাদ সারারাত কথা বলে করেছি পার।
Model Summary:	আমি আর নিস্পাপ সারারাত কথা বলে।

4.3 Descriptive Analysis

Before the creation of Bengali content synopsis model, i have made a model for English content outline. The two models give a decent outcome for various content. Outline model is made to decrease the loss of capacity in the model. The misfortune work assesses the model. It diminishes the blunder for learning model. Misfortune work diminish is significant for consecutive information.

4.4 Summary

This area examined the investigation of our model. What's more, the reaction of the machine to make a rundown. All are talked about in upper quickly in subtleties with the evil presence of machine reaction outline**.**

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Outline is the errand of gathering a bit of content to a shorter variant, decreasing the size of the underlying content while simultaneously protecting key educational components and the significance of substance. Since manual content outline is a period costly and for the most part arduous errand, the automatization of the assignment is increasing expanding prevalence and in this way comprises a solid inspiration for scholarly research.

There are significant applications for content outline in different NLP related undertakings, for example, content characterization, question replying, lawful writings synopsis, news rundown, and feature age. Additionally, the age of rundowns can be incorporated into these frameworks as a middle stage which assists with decreasing the length of the record.

In the large information time, there has been a blast in the measure of content information from an assortment of sources. This volume of content is a limitless wellspring of data and information which should be successfully summed up to be valuable. This expanding accessibility of reports has requested comprehensive research in the NLP territory for programmed content rundown. Programmed content synopsis is the assignment of creating a succinct and familiar outline with no human assistance while safeguarding the significance of the first content report.

5.2 Impact on Environment

Programmed content summarizer for different dialects has been made beforehand however not for the Bengali language. Expanding the apparatuses and innovation of Bengali language is the principle objective of this exploration. In this examination work, we've attempted to manufacture a programmed book summarizer for the Bengali language. Albeit, working with the Bengali language was an extremely testing piece of this exploration. Be that as it may, until the end, we have made a base for programmed content summarizer of the Bengali language.

Robotized Text Summarization is a procedure of summing up any archive or content consequently. Summed up content is the compact type of the given content. In Natural language preparing

numerous content rundown strategies are accessible for English language, yet just a couple for Bangla language. Bangla is one of the most educated and utilized language everywhere throughout the world. The greater part of the content rundown methods are executed in two unique manners, known as abstractive or extractive methodology. This paper manage the outline of Bangla content dependent on extractive strategy. Another effective extractive synopsis strategy is proposed in this work. The other rundown apparatuses created for Bangla language appears very little fitting from application perspective. The proposed examination models are relevant for Bangla content rundown. In the proposed approach, fundamental extractive rundown is applied with new proposed model and a lot of Bangla content investigation rules got from the heuristics. Each Bangla sentences and words from unique content is examined appropriately with Bangla sentence bunching technique. This work proposed another kind of sentence scoring forms for Bangla content outline. In the assessment of this procedure, the framework reflects great exactness of results, contrasting with that of the human produced summed up result and other Bangla content rundown instruments.

5.3 Ethical Aspects

It is testing, since when we as people sum up a bit of content, we for the most part read it altogether to build up our comprehension, and afterward compose a rundown featuring its primary concerns. Since PCs need human information and language capacity, it makes programmed content outline an extremely troublesome and non-minor undertaking.

5.4 Sustainability

Programmed content outline is an energizing examination zone with a few applications on the business. By gathering huge amounts of data into short, synopsis can help numerous downstream applications, for example, making news digests, report age, news outline, and feature age. There are two conspicuous sorts of outline calculations.

CHAPTER 6

Conclusion and Future Work

6.1 Summary of the Study

I entire undertaking is identified with the Bengali NLP. In this undertaking, i have manufactured a profound learning model for Bengali abstractive content outline. That is extremely useful for making a programmed Bengali book synopsis. I have finished this task inside the 3 months. The entire task is separated into certain parts. The entire synopsis of the task is given beneath with bit by bit.

Stage 1: Data assortment structure internet based life

Stage 2: Summarize the gathered information

Stage 3: Collect word2vec

Stage 4: Data preprocessing

Stage 5: Vocabulary tally

Stage 6: Load pre-preparer embedding

Stage 7: Add uncommon token

Stage 8: Set Encoder & Decoder

Stage 9: Build succession to grouping model

Stage 10: Model training

Stage 11: Check the outcome investigation the reaction of the machine

This model will help our Bengali NLP research community to build a full dependent automatic abstractive text summarization and further research of Bengali text summarization. Now i will discuss the future work and conclusion of this research work.

6.2 Conclusion

The principal worry of this exploration work is creating and expanding the Bengali NLP inquire about territory. I have utilized the Bengali content as the contribution of our model and produced a synopsis of that grouping will be Bengali content which is the yield of the model. From the outset, i assemble the model for English content then i make this model for Bengali content. Ordinarily, encoder and decoder are working comparably for the two messages also. Our dataset isn't huge for Bengali content. In any case, the machine gives astounding reactions to this dataset. This model is worked for the Bengali short content synopsis age. I have characterized the grouping length and outline length. The machine can produce the rundown following this fixed length. This is the fundamental restriction of my model. On the off chance that the succession length is over the model doesn't work appropriately. This is another examination scope for Bengali content rundown. Bengali content preprocessing is somewhat troublesome on the restrict to another dialect. In this manner, the preprocessing library needs to work for Bengali content. Word to vector is another significant piece of these sorts of issue. Solid word to vector needs to deliver for taking care of the content related issue. All things considered, no machine gives a completely precise outcome. Each machine has some constraint in their working documented. Essentially, my summarizer model additionally has a few impediments. In any case, the primary concern is that the model can create an abstractive synopsis for the Bengali language. This is an accomplishment for my Bengali NLP recorded which supportive for future research work.

6.3 Recommendations

In the following phase of my work, i will expand the dataset and their synopsis for improving the model execution. I will attempt to assemble another model for content synopsis which will assist us with finding the best entertainer for the Bengali content outline. I am working just the short arrangement yet for long content grouping, a summarizer is required in the Bengali language. A few suggestions for content outline is given beneath,

- Understand the theoretical of long content
- Reduce the understanding time
- Reduce the report size with sparing the main data
- An automatic framework to separate data

6.4 Implication for Further Study

Some constraint is introduced in this is model, for example, work for restricted succession, the dataset isn't sufficient. Be that as it may, the model is worked for the future turn of events. Since any examination work is a constant procedure. Thusly, this model will be created step by step for the Bengali language. To locate a legitimate arrangement any works need more research. At that point, all examination locate a legitimate answer for a particular issue. Thus, inquire about work needs the future to actualize or improvement. The future actualize is subject to the restrictions of the past work. Understanding the confinements of the past work assists with making a proficient framework. In this work, the future work will be expanding the dataset of the Bengali content. Refreshing the model and set up the model of any sorts of content length. That implies the model won't reliant on the content length. The model is intricate and working in TensorFlow 1.15 rendition. Be that as it may, need to change over the code in refreshed renditions. Subsequent to finishing research the model needs to convey. In this way, making an application like web and portable application is significantly depending on the eventual fate of computerized reasoning. Consequently, i have built up an application for programmed Bengali abstractive content synopsis.

REFERENCES

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

[2]. Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

[3]. 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.

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

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

[6]. 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).

[7]. Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

[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]. Sutskever et al "Sequence to Sequence Learning with Neural Networks". Conference on Neural Information Processing Systems (NIPS,2014).

[11]. Peter J. Liu et al. "Generating Wikipedia by Summarizing Long Sequences". International Conference on Learning Representation (ICLR), 2018.

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

[13]. H. T. Le and T. M. Le, "An approach to abstractive text summarization," 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), Hanoi, 2013, pp. 371-376.

[14]. I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," 2012 Seventh International Conference on Computer Engineering & Systems (ICCES), Cairo, 2012, pp. 132-138.

[15]. D. Sahoo, A. Bhoi, and R. C. Balabantaray, "Hybrid Approach To Abstractive Summarization," Procedia Computer Science, vol. 132, pp. 1228–1237, 2018.

[16]. Ganesan, K., Zhai, C., Han, J. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In Proc. of Coling 2010, pages 340–348.

[17]. Lloret, E., Palomar, M. 2011. Analyzing the Use of Word Graphs for Abstractive Text Summarization. In Proc. of IMMM 2011.

[18]. Masum, A.K.M., Abujar, S., Talukder, M.A.I., Rabby, A.S.A. and Hossain, S.A., 2019, July. Abstractive method of text summarization with sequence to sequence RNNs. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[19]. Talukder, M.A.I., Abujar, S., Masum, A.K.M., Faisal, F. and Hossain, S.A., 2019, July. Bengali abstractive text summarization using sequence to sequence RNNs. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[20]. Abujar, S., Masum, A.K.M., Islam, M.S., Faisal, F. and Hossain, S.A., 2020. A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN. In Innovations in Computer Science and Engineering (pp. 509-518). Springer, Singapore.

[21]. Masum, Abu Kaisar Mohammad, Sheikh Abujar, Raja Tariqul Hasan Tusher, Fahad Faisal, and Syed Akhter Hossain. "Sentence Similarity Measurement for Bengali Abstractive Text Summarization." In 2019 10th

International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2019.

[22]. Abujar, S., Masum, A.K.M., Mohibullah, M. and Hossain, S.A., 2019, July. An Approach for Bengali Text Summarization using Word2Vector. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[23]. Abujar, Sheikh, Abu Kaisar Mohammad Masum, SM Mazharul Hoque Chowdhury, Mahmudul Hasan, and Syed Akhter Hossain. "Bengali Text generation Using Bi-directional RNN." In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2019.

Plagiarism Report:

Bengali Text Summarization

ORIGINALITY REPORT

4%	2%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Sheikh Abujar, Mahmudul Hasan, M.S.I Shahin, Syed Akhter Hossain. "A heuristic approach of text summarization for Bengali documentation", 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017 Publication	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	projectcare.com.ng Internet Source	<1%
5	www.duo.uio.no Internet Source	<1%
6	www.hajjelectronics.com Internet Source	<1%