

**BANGLA E-MAIL BODY TO SUBJECT GENERATION USING SEQUENCE TO  
SEQUENCE RNNs**

**BY**

**Moyin Talukder**

**ID: 183-15-11827**

**AND**

**Md. Samiul Alim**

**ID: 183-15-11808**

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

Supervised By

**Abdus Sattar**

Assistant Professor

Department of CSE

Daffodil International University

Co- Supervised By

**Dr. Md. Tarek Habib**

Assistant Professor

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**SEPTEMBER 2022**

## APPROVAL

This Project/internship titled “**Bangla E-mail Body to Subject generation using sequence to sequence RNNs**”, submitted by Moyin Talukder, ID No: 183-15-11827 and Md. Samiul Alim, ID No: 183-15-11808 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 September 2022.

### BOARD OF EXAMINERS



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

**Chairman**



**Raja Tariqul Hasan Tusher**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology

**Internal Examiner**



**Md. Sabab Zulfiker**  
**Senior Lectuer**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



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

**External Examiner**

## DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Abdus Sattar, Assistant Professor, Department of CSE Daffodil International University**. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**



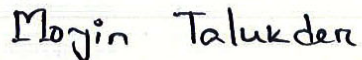
-----  
**Abdus Sattar**

Assistant Professor

Department of CSE

Daffodil International University

**Submitted by:**

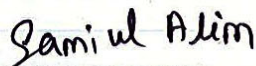


-----  
**Moyin Talukder**

ID: 183-15-11827

Department of CSE

Daffodil International University



-----  
**Md. Samiul Alim**

ID: 183-15-11808

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Abdus Sattar, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Natural language processing (NLP)” to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Pro. Dr. Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

## **ABSTRACT**

The development of subjects has become one of the major problems facing deep learning and natural language processing in recent years. A brief comment on a lengthy email body is condensed in the subject generation. Our goal is to develop a Bengali subject generator that is effective and efficient and can produce a clear and insightful subject from a given Bengali email body. To do this, we have gathered a variety of emails body, including educational, commercial, etc. and will use our model to generate subject from those texts. In the encoding layer of our model, bi-directional RNNs are employed, while the decoding layer makes use of LSTMs and an attention model. Our model generates subject using a sequence-to-sequence model. While developing this model, we encountered difficulties with text pre-processing, vocabulary and missing words counting, word embedding, detecting new terms and other tasks. Our primary objectives in this model were to generate a subject and lessen its train loss. In our study, we successfully reduced the train loss to 0.001 by producing a smooth concise subject from a provided email body.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE NO</b>
Board of examiners	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Problem Definition	2
1.4 Research Questions	2
1.5 Research Methodology	2
1.6 Research Objective	2
1.7 Report Layout	3
<b>CHAPTER 2: BACKGROUND</b>	<b>4-8</b>
2.1 Introduction	4
2.2 Related work	4-8
2.3 Bangladesh Perspective	8
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>9-17</b>
3.1 Introduction	9

3.2 Experiment Data Set	10
3.3 Data Pre-Processing	11
3.3.1 split data	11
3.3.2 Add Contractions	11
3.3.3 Stop word remove	11
3.3.4 Checking Subject and Body Purified	12
3.3.5 Data Preprocessing	12
3.4 Vocabulary Count & Word embedding	13
3.5 Model	13
3.5.1 Neural Machine Translation	13-14
3.5.2 RNN Encoder–Decoder	14-16
3.5.3 Sequence to Sequence Model	16-17
<b>CHAPTER 4: PERFORMANCE OF THE PROPOSED MODEL</b>	<b>18-20</b>
4.1 Training, Testing and the Validation of the model	18-19
4.2 Model efficiency	20
4.2.1 Training loss	20
<b>CHAPTER 5: RESULT COMPARISON AND DISCUSSION</b>	<b>21</b>
<b>CHAPTER 6: CONCLUSION AND FUTURE WORK</b>	<b>22</b>
<b>REFERENCES</b>	<b>23</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.1 System diagram	9
Figure 3.2 Collected Dataset	10
Figure 3.3.5 data preprocessing	13
Figure 3.5.2 rnn encoder decoder	16
Figure 3.5.3 sequence to sequence model	17
Figure 4.2.1 Train loss	20



## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 4.1 Sample result I	18
Table 4.1.1 Sample result II	19
Table 5.1 Comparison with some previous works	21

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

Conversation in natural language is a challenging subject in artificial language and summarization is a crucial aspect of it. Humans are easily capable of producing any Automatically send SMS or email. They choose the key words. from a predetermined body and produce the subject. within the computer system, the task is significantly more challenging. As a result of the subject is crucial for language comprehension, thinking, and use of common sense information like a person would. Any email body or content can be produced in two different ways: These methods are conceptual and extractive. Some methods of subject generation use an extractive strategy, which lengthens key phrases or lines from the content that is provided. Then they must be combined to create the subject. In the abstractive technique, the email body is used to create a bottom-up subject, even when not all of the words may be present. This indicates that an abstract method may create the topic for a specific email body on its own. Few works have been completed for Bangla subject generation. The approach we suggest in this study uses the email text to generate the subject line. We acquired a lot of information and trained the model using that information. After completing the program, the outcome was as expected. Textual work is always difficult. To construct a suitable subject with an abstractive approach, we must go through a number of phases, including text pre-processing, vocabulary and missing word counting, word embedding and counting and using certain specific tokens for word encoder and decoder. In our approach, we use each of these phases. We have utilized a bidirectional RNN with two layers to apply the sequence to sequence model. Bahdanau attention model [1] is used on the target body text and two layers of RNN each with LSTM to provide an effective subject. The encoder converts all of the input phrases into a fixed-length vector and the decoder uses that vector to produce an output sequence. We have modified this model for Bangla from its initial application in relevant text resolution for machine translation. To increase productivity and create a subject that is more fluent and effective, we have highlighted a number of important variables. Major procedures are described in depth, but the explanation of deep learning techniques and models is more crucial.

## **1.2 Motivation**

Bangladesh has recently seen an ICT revolution. Bangladesh has made significant progress in the area of information and communication technologies. The use of NLP in several languages is clear. But it's really constrained in Bangla. For this reason, we utilize the body of the Bangla email to produce the topic. After writing the email content in Bangla, users occasionally are unsure about what the subject line should be. Where It Will Aid in Subject Generation. And it is the primary reason.

## **1.3 Problem Definition**

The subject of emails is generated using an RNN model. The subject needs to be created. Additionally, it must consider how that model functions in practice. How the data functions How to produce the subject line even with a smaller train loss from the email body.

## **1.4 Research Questions**

Following are the key inquiries on which this thesis is focused:

- What difficulties have faced for collecting data?
- How to the training loss is very low?
- How Bengali user could be benefited?

## **1.5 Research Methodology**

We describe the Experiment Data Set, Data Pre-processing, and Model Training, Training loss in this portion of our research article. The performance of the proposed models will be discussed at the conclusion of this chapter.

## **1.6 Research Objective**

Using a sequence-to-sequence model to create a subject from a Bangla email body has certain advantages. The list below includes a few goals;

- Create a productive model to generate the subject.
- Reduce the users time to write emails subject.
- New model used in Bangla language to generate subject from email body.

## **1.7 Research Layout**

Chapter 1: Introduction, motivation, problem definition, research question, research methodology, and discussion of research goals.

Chapter 2: Describes the background and current status of this study and related work from a Bangladeshi perspective.

Chapter 3: Introduction, experimental Datasets, data preprocessing, vocabulary size and word embeddings and research methodology model.

Chapter 4: Performance Considerations of Proposed Models.

Chapter 5: Focuses on comparing and discussing results.

Chapter 6: Describes the completion of this study and future work.

Chapter 7: All references used in this study can be found here.

## CHAPTER 2

### BACKGROUND

#### 2.1 Introduction

There hasn't been any equivalent study or effort done in Bangladesh that can accurately generate subject from the email body. The current state of the Bangla language and the use of NLP in Bangladesh's language sector serve as the background.

#### 2.2 Related Works

By allowing a model to automatically look for elements of a source phrase that are pertinent to predicting a target word without having to explicitly create these pieces as a hard segment, we propose to enhance this fundamental encoder-decoder architecture. We propose that the performance of this fundamental encoder-decoder design is constrained by the usage of a fixed-length vector. We achieve translation performance using our unique method that is comparable to the most cutting-edge phrase-based system while translating from English to French. Additionally, a qualitative examination shows that the alignments that the model discovered are in good agreement with human intuition [1].

We use Attentional Encoder-Decoder Recurrent Neural Networks to mimic abstractive text summarization and demonstrate that they deliver cutting-edge results on two independent corpora. We provide a variety of distinctive models that address crucial summarizing problems that the fundamental architecture does not adequately address, such modeling keywords, capturing the hierarchy of sentence to word structure and emitting words that are uncommon or unanticipated during training. Our research demonstrates that many of the models we've suggested help performance continue to improve. We also develop performance goals for future studies and suggest a new dataset made up of multi-sentence summaries [2].

On the Qualities of Neural Machine Translation, Encoder-Decoder Methods utilizing two models, a newly developed gated recursive convolutional neural network and an RNN encoder-decoder. We focus on analyzing the characteristics of neural machine translation. We show that the neural machine translation system does well enough on simple phrases sans unfamiliar terms but as the sentence's length and the quantity of unfamiliar words rise, the system's performance rapidly deteriorates. Additionally, we find that the recommended gated recursive convolutional network readily picks up on the syntax of a phrase [3].

It provide a generic end-to-end method for learning sequences that places the fewest constraints on the sequence structure. On the WMT-14 dataset, translations made by the LSTM on English to French translation get a BLEU score of 34.8. Long phrases presented no difficulties for the LSTM, as well as a phrase-based SMT system. On the same dataset, a BLEU system earns a score of 33.3 on a word-by-word basis. The 1000 hypotheses generated by the SMT system were reranked using the LSTM and its blue score rose to 36.5 nearly matching the prior state of the art. It acquired logical sentence and phrase structures that are responsive to syntax and mostly independent of the active and passive voice. The performance of the LSTM has been enhanced by word reversal in all source texts (but not target sentences) this is due to the fact that several transient dependencies are introduced between the source and destination phrases, this simplifies the optimization challenge [4].

The global method, the two types of attentional processes that are explored in this study are the global approach, which always attends to all origin phrases, and the local method, that only focuses at a selection of origin phrases at a time. We demonstrate the effectiveness of both methods using the WMT translation tasks between German and English across both languages. With local attention, we significantly outperform non-attentional systems that already employ well known strategies like dropout by 5.0 blue points. Our ensemble model, which included different attention architectures, delivered a new state-of-the-art result in the WMT'15 English to German translation issue with 25.9 blue points, a gain of 1.0 blue points over the previous record system backed by an n-gram ranker and NMT. [5].

We demonstrate that our approach is capable of producing logical, fluid paragraphs with several sentences, even complete Wikipedia articles. We demonstrate that it can extract pertinent factual information from reference documents, as evidenced by confusion, ROUGE scores, and human evaluations. We demonstrate that our approach is capable of producing logical, fluid paragraphs with several sentences, even complete Wikipedia articles. We demonstrate that it can extract pertinent factual information from reference documents, as evidenced by confusion, ROUGE scores, and human evaluations [6].

In this research, we suggest and put into practice an efficient method to deal with this issue. The output of a word alignment method is used to enrich the data used to train an NMT system, enabling it to emit the position of the relevant word in the source text for each OOV text in the destination phrase. A dictionary is used to translate each OOV word in a subsequent post-processing phase. The evaluations of the WMT'14 French translation contest show that our approach significantly outperforms a similar NMT system that does not employ this technique by up to 2.8 BLEU points [7].

A target recurrent language model is used to represent the creation of the translation, and a convoluted sentence model is used to simulate the conditioning of the source sentence. Second, we show that despite the lack of alignments, they are extremely sensitive to the original sentence's wording, grammar, and meaning. Additionally, we show that, when rescored n-best lists of translations, they are comparable to a cutting-edge system [8].

This paper introduces a new approach to the Translation problem by word segmentation techniques. Typical approaches use backing off to dictionaries strategy to deal with unknown words. They dealt with names, compounds, or cognates with compositional translation, phonological or morphological transformation, etc. However, in this study, classification depends on the byte pair encoding compression technique and employs straightforward character n-gram models to verify the NMT model improves over the typical method for WMT translation tasks. They demonstrated growth to be English-German and English-Russian by up to 1.1 and 1.3 blue respectively [9].

The NRM is a short-text response generator based on neural networks. It has been trained using a significant portion of one-round discussion information from a microblogging website. The response generation method in NRM is formalized as a decoding procedure based on the subconscious representation of the entered data, according to the general encoder-decoder architecture with recurrent neural networks used for both encoding and decoding (RNN). According to empirical research, NRM can create grammatically accurate replies to more than 75% of the input text surpassing state of the art models in the same environment [10].

By using an RNN encoder-decoder model, a set of symbols is converted into a fixed length vector representation. Encoder and decoder in the suggested model are trained in tandem to enhance the posterior probability of a target sequence provided a source sequence. An additional feature in the current log-linear model, provisional probabilities of word pairs, increases the effectiveness of a numerical system for machine translation [11].

For the first time, a graphical representations sentence rating function is presented by the writer for summarizing Bangla news documents. It is an Extractive based approach. They evaluated the final result by using the tool of the ROUGE evaluation package. They use included 3400 Bangla news documentations dataset from daily prothom alo newspaper. From this data set, they use 200 data randomly and use it in two groups. The suggested technique has average accuracy, recall, and F-measure scores of 0.60, 0.68, and 0.63, respectively [12].

The Author Proposed a new method of automatic Bangla news text summarization by using the approach of pronoun replacement and an improved version of sentence ranking. The main parts of this approach were preprocessing the input document then word tagging after the replacement of pronoun, and sentence ranking. After finishing this replacement process the sentences are ranked by considering some words that are term and sentence frequency, numerical figures and title words. They didn't use any benchmark data set they collected 3000 news document instances from Daily Prothom alo newspapers. F-measure scores for ROUGE-1 result was 0.6003 The F-measure score for ROUGE-2 was found 0.5708 And pronoun replacement accuracy was 71.80% [13].



The Author Proposed a Bangla abstractive text summarization method to help the machine provide the summarization text from a body text. They used Recurrent Neural Network (RNN) as an approach to this project and the most useful RNN's model is LSTM and also used contextual tokens for better sequence understanding. By talking about text production using n-gram language modeling and building a recurrent neural network as the training model. They used their own dataset by collecting from social media. Drawbacks include the inability to create text without knowing the text's length and the required n-gram sequence, which makes the procedure time-consuming. Sometimes cannot give correct order sentences and also cannot generate random length of the text. Accuracy [14].

### **2.3 Bangladesh Perspective**

Bangladesh views the use of NLP models in Bangla and data collection as a major danger. It is incredibly challenging to get data. Where we gather it: Through several academic institutions, corporate offices, as well as from individual emails. As a result, we encountered several challenges. Following implementation, we experience very little training loss. Additionally, the topic is generated from the email text. And mostly the Bengalis will benefit from it.

# CHAPTER 3

## RESEARCH METHODOLOGY

### 3.1 Introduction

We have a large data collection that includes email body text and an equal number of email subject lines. Let's assume that the text's input sequence has  $D$  words. As a result, the words  $x_1, x_2, \dots, x_d$  are coming from a vocabulary with a size of  $V$ , which produced an output sequence that is comparable to  $y_1, y_2, \dots, y_s$  where  $S < D$ . That indicates that the subject sequence is shorter than the email body sequence's text description. Consider that the same language is used throughout the whole output sequence. In this part, we will illustrate our approach to create a subject generator from a Bangla email body. We sought to create a subject generator that can generate a suitable subject from a given email body because there haven't been many previous efforts made for Bangla subject generating. Tensorflow CPU version 1.15.0 was used to set up and train this model.

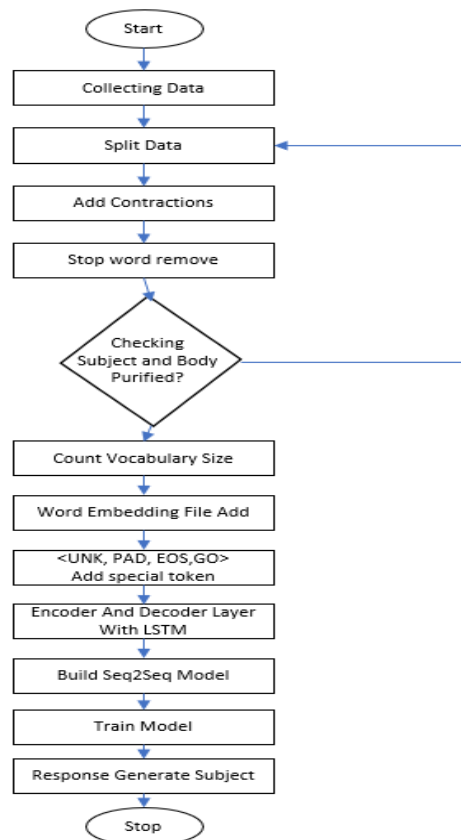


Figure 3.1 System diagram

## 3.2 Experiment Data Set

Each deep learning method requires a significant quantity of information. When compared to the size of the dataset, the results are clearly superior. We also need a substantial amount of information for the model. On the internet not enough datasets available though. We painstakingly gathered all the information for it. Consequently, we have acquired a range of data for our study, including commercial, personal, academic, and other types. In our dataset, just the two necessary columns—the emails' bodies and subjects—are present. It's typical for people to have a personal address in addition to an educational postal address, which is only for use by educational institutions. They are distributed by all levels of education, from primary to graduate. And we have personally gathered those emails from person to person. Commercial emails are advertisements that are sent to a user in an effort to raise awareness, promote interaction, or close a deal. The emails you send to subscribers who have chosen to receive your brand's promotional communications are known as commercial emails. They include emails intended to increase onboarding and user engagement, sales offer, newsletters, and announcements about new products. And we personally obtained the email from the corporate organization. Emails marked "personal" are sent from individuals rather than organizations. This indicates that we should send email from a personal address rather than a generic business or company email account. And we obtained this information through friends and family.

Subject	Body
মাইক্রোসফট ভার্সিয়াল হ্যাঁকাধন ২০২২	অত্যন্ত প্রত্যাশার সাথে, মাইক্রোসফট, মাইক্রোসফট ভার্সিয়াল হ্যাঁকাধন ২০২২, প্রযুক্তি উত্সাহী, প্রতিভাবান বিকাশকারী, প্রকৌশলী, ডিজাইনার এবং এপিএসি-এর উদ্যোক্তাদের জন্য এজুয়েটর-এআই ব্যবহার করে বাস্তব-বিশ্বের ব্যবসায়িক চ্যালেঞ্জগুলি সমাধান করার জন্য একটি চ্যালেঞ্জ, সর্বশেষ সংস্করণ চালু করার ঘোষণা দিতে পেরে আনন্দিত। আমাদের চারপাশে নতুন প্রযুক্তির সাথে আমরা এখন এই হ্যাঁকাধন সিরিজের চতুর্থ পুনরাবৃত্তির মাধ্যমে যে সমস্যার সমাধান করার কথা আগে কর্তব্য করতে পারিনি সেগুলি সমাধান করতে পারি। মাইক্রোসফট আপনাকে বিভিন্ন বিকাশমান শিল্পের জন্য উন্নত ডেটা বিশ্লেষণ এবং এআইতে উদ্ভাবনী সমাধান প্রদানের জন্য চ্যালেঞ্জ করছে। এখনই আপনার সমাধান জমা দিতে স্বতন্ত্রভাবে যোগ দিন বা ৪ সদস্য পর্যন্ত দলে দলে যোগ দিন। অংশগ্রহণকারীদের বিভিন্ন সমস্যা বিবৃতি মোকাবেলা করতে একাধিক এন্ট্রি জমা দিতে স্বাগত জানাই। প্রতিযোগিতাটি এ বছর এশিয়া জুড়ে ১৫ দেশের জন্য উন্মুক্ত। বাংলাদেশ, ভূটান, ক্রুনাই, কম্বোডিয়া, ইন্দোনেশিয়া, কোরিয়া, লাওস, মালয়েশিয়া, মায়ানমার, নেপাল, ফিলিপাইন, সিঙ্গাপুর, শ্রীলঙ্কা, থাইল্যান্ড, ভিয়েতনাম যদি আপনার স্বপ্ন হয় এপিএসিতে অ্যাডভান্সড ডেটা অ্যানালিটিক্স এবং এআই-এর স্বরূপে অবদান রাখা, তাহলে বিভিন্ন দেশ থেকে আপনার মতো অ্যাডভেব প্রিমিয়ার একটি ভিডিও সম্পাদনা সফ্টওয়্যার যা আপনাকে সহজেই একটি ভিডিও সম্পাদনা করতে এবং পরিবর্তন করতে সহায়তা করে। এটি চলচ্চিত্র নির্মাতা ভিডিও সম্পাদক এবং টেলিভিশন সম্প্রচারকদের জন্য একটি শীর্ষস্থানীয় অ্যাপ। এটি ভয়েস-ওভার রেকর্ডিং, মিডিয়া আমদানি এবং বিভিন্ন রেজোলিউশনের জন্য কাস্টম সিকোয়েন্স প্রিসেট তৈরি করার জন্যও ব্যবহৃত হয়। উদ্দেশ্য: ১. ইউটিউব ইন্টো তৈরি করতে প্রিমিয়ার ভিডিও এডিটিং প্রি-মেড টুল, মোশন গ্রাফিক্স এবং টেমপ্লেট ব্যবহার করে সহজেই নিম্ন তৃতীয় শ্রেণীর আপনার ভিডিও অনন্য করতে। ২. আপনার সৃজনশীলতা প্রকাশ করতে। ৩. ভিডিও সম্পাদনায় ব্যবহৃত পরিভাষা বোঝার জন্য। ৪. সিকোয়েন্স তৈরি সম্পাদনা এবং একত্রিত করতে। ৫. বিশ্ববিদ্যালয়, কাজ, বিক্রয়, অফিস এবং সমস্ত শিল্পের জন্য উপযুক্ত শো-এর জন্য উচ্চ-মানের পেশাদার ভিডিও তৈরি করা। এই কোর্স থেকে আপনি যা শিখবেন: ১. শুরু থেকে শেষ পর্যন্ত একটি ভিডিও সম্পাদনা প্রকল্প সম্পূর্ণ করার প্রক্রিয়া। ২. মিডিয়া আমদানি করা এবং বিভিন্ন রেজোলিউশনের জন্য কাস্টম সিকোয়েন্স প্রিসেট তৈরি করা। ৩. একাধিক সম্পাদনা ট্র্যাক জুড়ে ফুটেজ ছবি, সঙ্গীত এবং অডিও একসাথে সম্পাদনা করা। ৪. বিভিন্ন ভিডিও এডিটিং ফ্রেম রেট, রেজোলিউশন এবং আকৃতির অনুপাত। ৫. বিভিন্ন ধরনের ট্রানজিশন যোগ করা। ৬. পাঠ্য শিরোনাম, প্রভাব এবং নিম্ন তৃতীয় শিরোনাম তৈরি করা। ৭. রঙ সংশোধন এবং রঙ গ্রেডিং এর মৌলিক বিষয়। ৮. পটভূমি শব্দ এবং উন্নত অডিও সম্পাদনা সরান। ৯. শো মোশন ফুটেজের সাথে কীভাবে কাজ করবেন এবং কীভাবে রিয়েম্পের গতি বাড়ান। ১০. আরও দক্ষতার সাথে ভিডিও সম্পাদনা করতে কীবোর্ড শর্টকাট। ১১. অ্যাডভেব প্রিমিয়ার প্রো-এ ভিডিও সম্পাদনা করুন। ১২. পেশাদার ভিডিও সম্পাদনার পথ।
বুদ্ধ পূর্ণিমা ২০২২-এ বিজ্ঞপ্তি আজকের সিডিসি-ফিল্ম জবস চাকরির সতর্কতা	এটি জানানো যাচ্ছে যে ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটির সমস্ত একাডেমিক এবং প্রশাসনিক কার্যক্রম ১৫ মে ২০২২ তারিখে 'বুদ্ধ পূর্ণিমা' উপলক্ষে বন্ধ থাকবে। প্রিয় ছাত্র, ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) থেকে শুভেচ্ছা! আমরা আপনার জন্য কিছু কাজের সুযোগ খুঁজে পেয়েছি। অনুরোধ করে এখানে যুক্তি এবং উপযুক্ত চাকরির জন্য আবেদন করুন।
৭ দিন বর্ধিত   কমিশনাল ক্যারিয়ার ২.০: ফাউন্ডেশন অফ প্রোডাক্ট ডিজাইন   ডিআইইউ জিসিপিপি	প্রিয় উদ্ভোগ ডিআইইউ জিসিপিপি থেকে সালাম ও শুভেচ্ছা! আপনি ভাল করছেন আশা করি। আমরা সবাই জানি ডিআইইউ জিসিপিপি ক্যারিয়ার ২.০: ফাউন্ডেশন অফ প্রোডাক্ট ডিজাইন বিষয়ক একদিনের কর্মশালার আয়োজন করতে যাচ্ছে। কিছু সমস্যার কারণে নিবন্ধনের সময়সীমা মে ১৯, ২০২২ পর্যন্ত বাড়ানো হয়েছে এবং কর্মশালাটি ২১ মে ২০২২-এ অনুষ্ঠিত হবে। সুতরাং, এখনই নিবন্ধন করুন। পূরণ করে আপনার আসন বুক করুন।

Figure 3.2 Collected Dataset

## 3.3 Data Pre-Processing

### 3.3.1 Split Data:

A string is divided into a list using the split technique. We'll provide the separator by default any whitespace serves as a separator. After segmenting the text depending on the supplied separator the split function produces a list of strings. The benefits of utilizing a split function in Python are as follows. We might need to divide a long string into several shorter ones at some time. It is the complete opposite of unification, which joins two strings. In our dataset, we split it for further processing. Then we remove the whitespace by using re (regular expression) library. Our dataset is Bangla for this reason we will UNICODE for with whitespace.

### 3.3.2 Add Contractions:

Contractions are a unique type of word that combines two or more other words in an abbreviated shape, ordinarily with punctuation. Contractions can make the peruser feel like we're talking specifically to them and having a discussion. It makes a difference make our composing show up uncomplicated for everybody to get it and make sense of. Because contractions are shorter, it also means that they take up less space. Since of that, we'll frequently see them in notices where space is profitable. We are using "বি.দ্র ": "বিশেষ দ্রষ্টব্য", "ড.": "ডক্টর", "ডা.": "ডাক্তার", "ইঞ্জি.": "ইঞ্জিনিয়ার", "রেজি.": "রেজিস্ট্রেশন", "মি.": "মিস্টার", "মু.": "মুহাম্মদ", "মো.": "মোহাম্মদ" etc . As a result it will help more accurate model build.

### 3.3.3 Stop word remove:

"Stop words" usually refers to the most often used terms in a dialect. There isn't a comprehensive list of "Stop words" that all NLP tools use uniformly. "Stopwords" are any words, regardless of dialect, that add little to a sentence's meaning. They can be safely ignored without changing the sentence's meaning. These are some of the most prevalent short function terms for certain search engines. Such as "এই", "যেন", "কোন", "তাই", "এটি" and on. we also remove special character. Also, we remove "https://", Bangla digit, Bangla special character, and on. we also remove English words and unaccepted words.

### 3.3.4 Checking Subject and Body Purified:

In this portion we keep a function that will check and also purify the subject & body text. Like it will lowercase the data, remove punctuation, numbers, unnecessary space, replace punctuation repeats, remove emoticons, contractions and emojis. But if it not works well then it will repeat the same procedure from the split body and then will continue.

### 3.3.5 Data Preprocessing:

We have carried out a few steps for data pre-processing. First, we introduced contractions to the email body's content as well as its subject. There are several contractions available, including "বি. দ্ব. ", "উ. ", "মো. " etc. As a result, we have eliminated them and replaced them with their complete forms. After that, we cleansed the texts. That indicates all the unnecessary characters have been eliminated. Regular expression has been utilized to exclude those extraneous components from the passages. After that, we got rid of stop words.

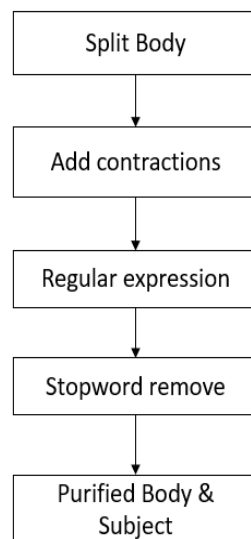


Figure 3.3.5 data preprocessing

### **3.4 Vocabulary Count & Word Embedding:**

The similarity of words affects their meaning as much as frequency. Therefore, we must count the total amount of words in the subject and body of the cleaned-up emails. Word keeps track of the words when we type a document. Pages, paragraphs, lines, and characters are all counted by Word. To find out how many words, pages, characters, paragraphs or lines are in a document look at the status bar. The vocabulary in this work is being counted. Following the vocabulary count (20011), we evaluated word frequency. For example, we tested the term “ভার্চুয়াল, and the frequency of this word was 8. In a word embedding, which is a learned representation for text, words with the same meaning are represented identically. This approach of encoding words and documents may be responsible for one of the significant developments in deep learning for challenging natural language processing problems. The phrase "word embedding" refers to the act of expressing individual words as real valued vectors in a specified vector space. Since each word is allocated to a distinct vector and the vector values are learned similarly to a neural network the technique is frequently referred to as deep learning. We used a pre-trained word to vector file to improve the model. Word to vector file "bn w2v model" was utilized. Where we get the word embedding length (497405).

### **3.5 Model:**

In deep learning, there are several models and various model types that are employed for various purposes. For text modeling, the Longest Short Term Memory (LSTM) will be quite helpful. while we are dealing with text. For a machine to learn about text sequence, machine translation is crucial. Encoders and decoders like Google Translate are used by all translators. A text string is translated from one language to another by the translator.

#### **3.5.1 Neural Machine Translation:**

Translation from one language into another can be done via neural machine translation. Encoders and decoders are commonly used in machine translation to convert one language to another. The encoder uses the input sequence, while the output sequence is predicted and shown by the decoder.

It uses a target sentence  $x$  to increase the posterior probability of  $x$  in neural machine translation.  $arg(\max_p(x|y))$  is used if  $y$  is the source sentence.

### 3.5.2 RNN Encoder–Decoder:

Recurrent Neural Networks are used in the encoder-decoder paradigm to solve sequence-to-sequence prediction issues. Though it was first created to solve machine translation issues, it has also succeeded in solving related sequence to sequence prediction issues including text summarization and question answering. The encoder is a neural network with four convolutional layers that, like the DQN, have an identical design. Each layer is followed by an ELU activation function. After that, the result is flattened to produce a flat, 288-dimensional vector. Numerous projects use the encoder-decoder. It serves as the foundational tenet of translation software. The neural network that powers Google Translation contains it. As a result, it is utilized for Computer Vision as well as NLP activities and word processing!

Cho et al. [11] introduce the first two levels of the RNN encoder-decoder design. Later, Bahdanau et al. [1] exacerbated this. The only use for these encoder and decoder models was machine translation. The RNN Encoder-Decoder is made up of two Recurrent Neural Networks (RNNs), one of which acts as an encoder and the other as a decoder. The encoder transforms a variable length origin sequence into a fixed length vector, while the decoder transforms a variable length destination sequence back into the vector representation. The two RNN layers of this neural network were used. A phrase's fixed length is included in the encoder, while its output sequence is contained in the decode. In order to keep the target word sequence's greatest posterior probability, the RNN network's two layers are jointly trained. a covert gadget that improved memory development and capacity. To assess the likelihood that a Bangla sentence will match its matching Bangla sentence, we train our model.

Tables 1 & 2 include the input words for the model if the encoder received the target Input phrase as  $X = (x_1, \dots, x_{Tx})$ . where context vector  $c$  is present, so

$$h_t = (t, h_{t-1}) \dots \dots \dots (1)$$

and

$$c = (\{h_1, \dots, h_{Tx}\})$$

where  $h_t$  represents the state that was concealed at time  $t$ . Context vector created from the concealed state sequence is denoted by the symbol  $c$ . A non-linear function is  $f$  and  $g$ .

If the Response subject of tables 1 and 2 are the expected word sequence  $\{y_1, \dots, y_T\}$  that the decoder anticipated, therefore the likelihood will be,

$$p(y) = \prod_{t=1}^T p(y_t | \{y_1, \dots, y_{t-1}\}, c) \dots\dots\dots(2)$$

Where,  $(y_1, \dots, y_T)$ . A model for conditional probability is now presented by,

$$p(y_t | \{y_1, \dots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \dots\dots\dots(3)$$

Where,  $g$  = nonlinear function,  $y_t$  = output of probability,  $s_t$  = secret state.

$$c_i = \sum_{j=0}^T a_{ij} h_j \dots\dots\dots(4)$$

Bi-directional RNNs were employed. That is made up of Recurrent Neural Networks that run both forward and backward. The hidden state of a forward recurrent neural network is  $(h_1 \rightarrow, \dots, h_T)$  and the forward Recurrent Neural Network sequence order is  $(x_1 \text{ to } x_T)$ . Backward Recurrent Neural Network sequence order is  $(x_T \text{ to } x_1)$  and secret state is  $(h_1 \leftarrow, \dots, h_T)$ . So,

$$h_j = [h_j \rightarrow; h_j \leftarrow] \dots\dots\dots(5)$$

Where,  $h_j$  = Summary of anticipating and following words.

Here,  $a_{ij}$  = is soft max of  $e_{ij}$  This shows how input position  $j$  aligns with output at location and normalizes exponential function  $i$ ,

$$e_{ij} = (s_{i-1}, h_j) \dots\dots\dots(6)$$



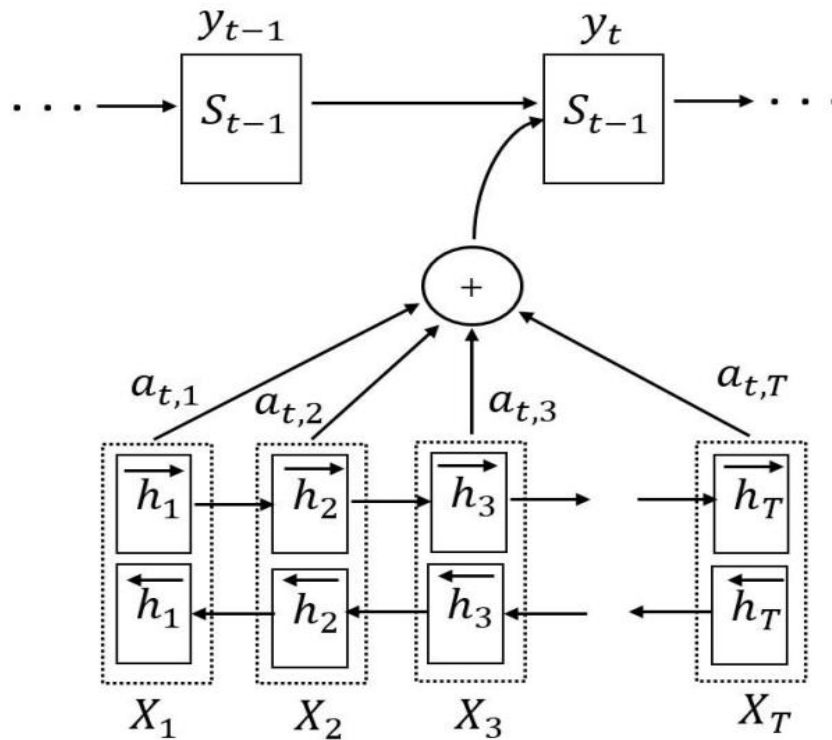


Figure 3.5.2 rnn encoder-decoder

### 3.5.3 Sequence to Sequence Model:

A Seq2Seq model is one that creates another sequence of items (words, letters, time series, etc.) from a single sequence of objects. A collection of words is used as the input for neural machine translation, while a set of translated words is used as the output. A series of machine learning techniques called Seq2seq is utilized for natural language processing. Applications include text summarization, conversational modeling, picture captioning, and language translation. Seq2Seq is a kind of RNN-based encoder-decoder model. It can serve as a model for automated communication and translation. The complete model may be divided into two compact sub-models. The first sub-model is known as [E] Encoder, whereas [D] Decoder is the name of the second sub-model. Like all RNN systems, [E] accepts raw input text data. Finally, [E] generates a neuronal representation. The model has a problem when dealing with long phrases since the output sequence is heavily dependent on the context created by the hidden information in the encoder's

desired outputs. There is a high probability that long sequences may lose the actual context by they reach their conclusion. Encoder and decoder using LSTM cells are a part of any sequence-to-sequence model. We utilized a word embedding file in our subject generating approach. The vocabulary size of this file that would be utilized as model input was then measured. A token is a specific instance of a sequence of letters that are put together as a meaningful semantic unit for processing in a certain document. The class of all tokens with the same character sequence is known as a type. Since tokens are the basic building blocks of Natural Language, the token level is where the majority of processing of the textual content takes place. Tokenization is the initial stage in textual data modeling. To create tokens, the corpus is segmented. Utilizing the tokens described below, develop a vocabulary as the following stage. The vocabulary in the corpus is the collection of distinctive tokens. Remember that the top K Frequently Occurring Words or each unique token in the corpus may be used to form a vocabulary.

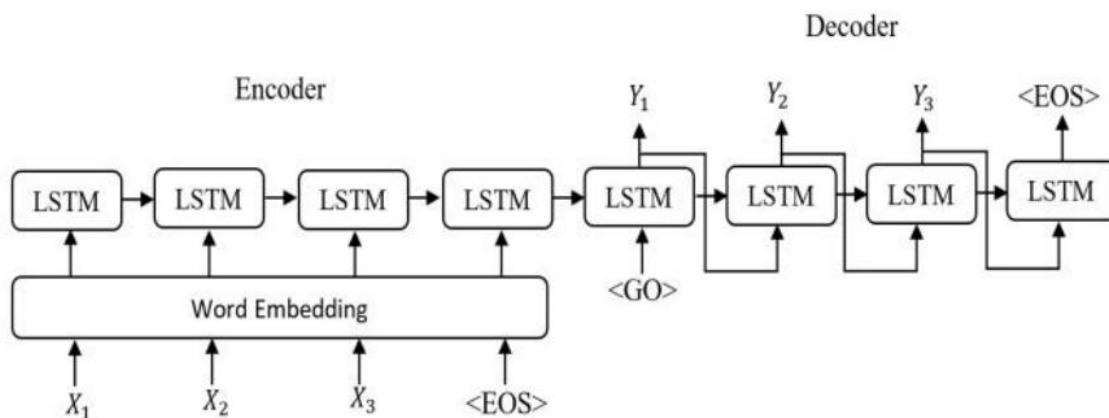


Figure 3.5.3 sequence to sequence model

This model have incorporated certain special vocabulary cues, such as  $\langle \text{UNK} \rangle$ ,  $\langle \text{PAD} \rangle$ ,  $\langle \text{EOS} \rangle$ ,  $\langle \text{GO} \rangle$ . There are various limitations on vocabulary. Some words are still in use. UNK token is used in place of the words. Each sentence in a batch added by a PAD token is the same length. The end of the sequence that alerts the encoder when it receives input is included in the EOS token. The GO token instructs the decoder to begin the output sequence procedure. We replace the vocabulary and add UNK during the data preparation stage. We applied sequence translation on the data that contains words before choosing GO and EOS. This sequence's mode x is the encoder's input sequence, while mode y is the produced output or response output sequence.

## CHAPTER 4

### PERFORMANCE OF THE PROPOSED MODEL

#### 4.1 EXPERIMENT AND OUTPUT:

Tensorflow 1.15.0 and the sequence-to-sequence model were employed. Machine will be able to create a subject once finished training. We will select an input sentence from the dataset and randomly determine the subject length to generate the subject. We have utilized an attention-based encoder for the parameter. We utilized Adam Optimizer to determine the learning rate for each parameter using the following values: epoch= 70, batch size= 2, rnns size= 256, learning rate= 0.001, maintain probability= 0.75. Use the standard gradient descent optimizer for quicker convergences.

Having spent a few hours training our model on the given dataset, the following machine response is shown as positive:

Table 4.1 sample result I

Original Body	সুপ্রিয় শিক্ষার্থীবৃন্দ, ডিআইইউ জিসিপিসি এর পক্ষ থেকে সালাম এবং শুভেচ্ছা। নতুন সূর্য, নতুন প্রাণ। নতুন সুর, নতুন গান। নতুন উষা, নতুন আলো। নতুন বছর কাটুক ভাল। কাটুক বিষাদ, আসুক হর্ষ। শুভ হোক নববর্ষ। পঙ্কিলতা পেরিয়ে পল্লব ও মুকুলে মুখরিত হয়ে রুদ্রতেজ ও তাপ নিয়ে আসে নববর্ষ। বৈশাখ তাই জাগে প্রাণ-উৎসবের আনন্দে। পহেলা বৈশাখ একইদিকে যেমন বাংলা নববর্ষ, বাঙালি জাতির ঐতিহ্য, কৃষ্টি, সংস্কৃতি, উৎসব, পার্বণ। সেই সাথে হিন্দু, মুসলিম, বৌদ্ধ, খৃষ্টান, পাহাড়ী, নৃতাত্ত্বিক জনগোষ্ঠী সকলে একসূত্রে গাঁথা। এটাই একমাত্র উৎসব যেখানে সকল বাঙালি মিলেমিশে একাকার হয়ে প্রাণ খুলে উৎসব পালন করে। একেই বলে সত্যিকারের বাঙালি উৎসব। বিদায় ১৪২৮, স্বাগত ১৪২৯! ডিআইইউ জিসিপিসি এর পক্ষ থেকে সবাইকে বাংলা নববর্ষ আগমনের অফুরান শুভেচ্ছা। "শুভ নববর্ষ"
Original Subject	শুভ নববর্ষ ১৪২৯
Input Body	সুপ্রিয় শিক্ষার্থীবৃন্দ, ডিআইইউ জিসিপিসি এর পক্ষ থেকে সালাম এবং শুভেচ্ছা। নতুন সূর্য, নতুন প্রাণ। নতুন সুর, নতুন গান। নতুন উষা, নতুন আলো। নতুন বছর কাটুক ভাল। কাটুক বিষাদ, আসুক হর্ষ। শুভ হোক নববর্ষ। পঙ্কিলতা পেরিয়ে পল্লব ও মুকুলে মুখরিত হয়ে রুদ্রতেজ ও তাপ নিয়ে আসে নববর্ষ। বৈশাখ তাই জাগে প্রাণ-উৎসবের আনন্দে। পহেলা বৈশাখ একইদিকে যেমন বাংলা নববর্ষ, বাঙালি জাতির ঐতিহ্য, কৃষ্টি, সংস্কৃতি, উৎসব, পার্বণ। সেই সাথে হিন্দু, মুসলিম, বৌদ্ধ, খৃষ্টান, পাহাড়ী, নৃতাত্ত্বিক জনগোষ্ঠী সকলে একসূত্রে গাঁথা। এটাই একমাত্র উৎসব যেখানে সকল বাঙালি মিলেমিশে একাকার হয়ে প্রাণ খুলে উৎসব পালন করে। একেই বলে সত্যিকারের বাঙালি উৎসব। বিদায় ১৪২৮, স্বাগত ১৪২৯! ডিআইইউ জিসিপিসি এর পক্ষ থেকে সবাইকে বাংলা নববর্ষ আগমনের অফুরান শুভেচ্ছা। "শুভ নববর্ষ"
Response Generate Subject	শুভ নববর্ষ ১৪২৯

Table 4.1.1 sample result II

Original Body	<p>প্রিয় ছাত্র, ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) থেকে শুভেচ্ছা! আপনি জেনে আনন্দিত হবেন যে ডিআইইউ-এর ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) ডিআইইউ-এর শিক্ষার্থীদের জন্য ১৪ অক্টোবর, ২০১৮ (রবিবার) ক্যারিয়ার প্ল্যানিং-এর ৬ষ্ঠ কর্মসূচির আয়োজন করতে চলেছে। কর্মশালাটি শিক্ষার্থীদের জন্য নিম্নলিখিত নির্দেশিকা প্রদান করবে যারা তাদের ভবিষ্যত পেশাগত পেশার জন্য নিজেদের প্রস্তুত করতে চায়: ১. একজন শিক্ষার্থীর একাডেমিক শৃঙ্খলা, ফলাফল, পারিবারিক পটভূমি, ঘাড় এবং মনোভাব, দক্ষতা, চিন্তাভাবনা এবং প্রত্যাশা ইত্যাদির সাথে সামঞ্জস্য রেখে কোন ধরনের ক্যারিয়ার উপযুক্ত হবে; ২. সম্ভাব্য ক্ষেত্রগুলি কী কী যা একজন শিক্ষার্থী তার পেশা হিসেবে বেছে নিতে পারে; ৩. একটি মর্যাদাপূর্ণ ক্যারিয়ার গড়তে একজন শিক্ষার্থীকে কী ধরনের প্রস্তুতি নিতে হবে; ৪. এই বিষয়ে প্রয়োজনীয় প্রস্তুতি নেওয়ার জন্য সিডিসি কীভাবে ডিআইইউ শিক্ষার্থীদের প্রয়োজনীয় সহায়তা প্রদান করতে পারে।</p>
Original Subject	<p>কর্মজীবন পরিকল্পনা কর্মশালা   আপনার ক্যারিয়ারকে সঠিক দিকনির্দেশনা দিন (সংশোধিত)</p>
Input Body	<p>প্রিয় ছাত্র ক্যারিয়ার ডেভেলপমেন্ট সেন্টার সিডিসি থেকে শুভেচ্ছা আপনি জেনে খুশি হবেন যে ডিআইইউ এর ক্যারিয়ার ডেভেলপমেন্ট সেন্টার সিডিসি ডিআইইউ এর শিক্ষার্থীদের জন্য অক্টোবর বৃহস্পতিবার দুপুর এ কর্মজীবন পরিকল্পনার উপর কর্মশালার তম প্রোগ্রামের আয়োজন করতে যাচ্ছে কর্মশালাটি শিক্ষার্থীদের জন্য নিম্নলিখিত দিকনির্দেশনা প্রদান করবে যারা তাদের ভবিষ্যত পেশাগত পেশার জন্য নিজেদের প্রস্তুত করতে চায় একজন শিক্ষার্থীর একাডেমিক শৃঙ্খলা ফলাফল পারিবারিক পটভূমি ঘাড় এবং মনোভাব দক্ষতা চিন্তাভাবনা এবং প্রত্যাশা ইত্যাদির সাথে সামঞ্জস্য রেখে কোন ধরনের ক্যারিয়ার উপযুক্ত হবে সম্ভাব্য ক্ষেত্রগুলি কী কী যা একজন শিক্ষার্থী তার পেশা হিসেবে বেছে নিতে পারে একটি মর্যাদাপূর্ণ ক্যারিয়ার গড়তে একজন শিক্ষার্থীকে কী ধরনের প্রস্তুতি নিতে হবে এই বিষয়ে প্রয়োজনীয় প্রস্তুতি নেওয়ার জন্য সিডিসি কীভাবে ডিআইইউ শিক্ষার্থীদের প্রয়োজনীয় সহায়তা প্রদান করতে পারে</p>
Response Generate Subject	<p>কর্মজীবন পরিকল্পনা কর্মশালা   আপনার ক্যারিয়ারকে সঠিক দিকনির্দেশনা দিন</p>

## 4.2 Model efficiency

### 4.2.1 Training loss:

The model's mistake on the training set is measured as "training loss". Remember that the dataset that was partially utilized to train the model is comprised of the training phase. The training loss is computed based on the sum of errors for each instance in the training set. In other words, loss is a measurement of how effectively a model for a given situation was predicted. If the model's forecast is true, the loss is zero; otherwise, it is larger. The goal of the modeling process is to identify a collection of weights and biases that on average have minimum loss across all cases. The value of the objective function that you are minimizing is called the train loss. Based on the precise objective function of your training data, this result might be either positive or negative. Over the complete training dataset, the training loss is determined.

The performance statistic of your model that is human interpretable is called train error. Typically, it refers to the proportion of training instances that the model incorrectly predicted. Always a number between 0 and 1, this is. The same data used to train the model and determine its error rate are used to compute training error.

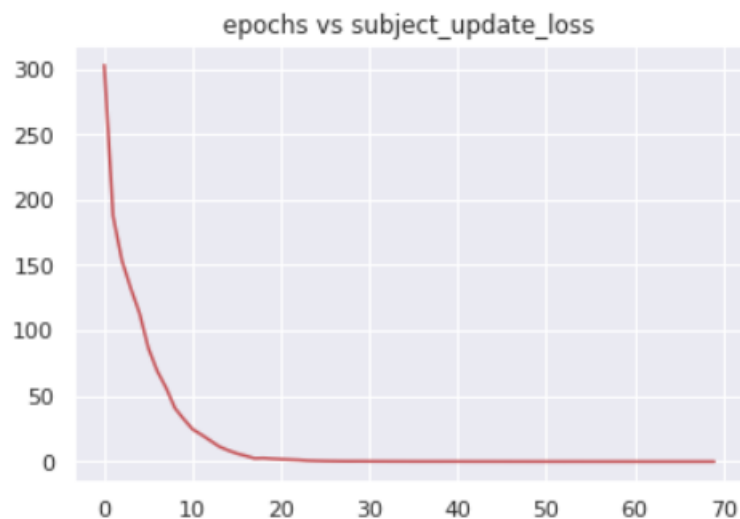


Figure 4.2.1 Train loss

## CHAPTER 5

### RESULT COMPARISON AND ANALYSIS

In comparison to other papers, this work provides us with the best accuracy. Some researchers did work that was almost identical to this study. Table 1 shows some connections between our work and some earlier text summarizing research.

Table 5.1 Comparison with some previous works

Work	Algorithm	Train loss
Bengali abstractive text summarization using sequence to sequence RNNs	RNN	0.008
Bangla E-mail Body to Subject generation using sequence to sequence RNNs	RNN	0.001

Based on the above table, we can see that several researchers have used various sorts of algorithms and have obtained a range of train losses, but the minimum train loss overall is 0.001.

## **CHAPTER 6**

### **CONCLUSION AND FUTURE WORK**

Through this study, demonstrated how to use LSTM encoding and decoding to create a model for creating a Bangla to Bangla topic from a given email content. No model including ours, can forecast the outcomes with hundred percent accuracy. However, our model can offer the most precise anticipated subject. Due to various flaws in our model, we were able to create a topic that is clear, relevant, and fluent while also lowering the training loss. The dataset used in our research experiment was the main constraint. We had to generate our own dataset because there wasn't one already available online. It's challenging to develop a dataset for generating subject, and we are aware that deep learning algorithms provide results that are much better the larger the dataset. As a result, we continue to gather data and expand our collection. Another drawback is that our model could only generate subject with a certain number of words. We'll work to expand it so that it can generate subject from emails body with an unlimited number of words. Additionally, Bangla language lemmatization and better words to vector are not readily available. Future efforts will be made to address these issues in the hopes of creating a stronger subject generation model for Bangla language.

## REFERENCE

- [1] Dzmitry Bahdanau et al. “Neural Machine Translation by Jointly Learning to Align and Translate”. International Conference on Learning Representation (ICLR), 19 May 2014.
- [2] Ramesh Nallapati, Bowen Zhou, et al “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond”. The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 26 Aug 2016.
- [3] K.Cho, B .van Merriënboer, 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), 7 oct 2014.
- [4] Sutskever et al “Sequence to Sequence Learning with Neural Networks”. Conference on Neural Information Processing Systems (NIPS,2014).
- [5] M. Luong, H. Pham, Christopher D. Manning “Effective Approaches to Attention-based Neural Machine Translation”. Conference on Empirical Methods in Natural Language Processing (EMNLP 2015).
- [6] Peter J. Liu et al. “Generating Wikipedia by Summarizing Long Sequences”. International Conference on Learning Representation (ICLR), 2018.
- [7] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba “Addressing the Rare Word Problem in Neural Machine Translation”. Association for Computational Linguistics (ACL, 2015).
- [8] Kalchbrenner et al (2013) “Recurrent continuous translation models”. In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics.
- [9] Rico Sennrich, Barry Haddow, Alexandra Birch “Neural Machine Translation of Rare Words with Subword Units”. Association for Computational Linguistics (ACL, 2016).
- [10] Lifeng Shang, Zhengdong Lu, Hang Li “Neural Responding Machine for Short-Text Conversation”. Association for Computational Linguistics (ACL 2015)
- [11] Cho, K. et al. (2014) Learning Phrase Representations using RNN Encoder Decoder for Statistical Machine Translation. Proceeding of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [12] Ghosh, P.P., Shahariar, R. and Khan, M.A.H., 2018. A Rule Based Extractive Text Summarization Technique for Bangla News Documents. International Journal of Modern Education & Computer Science, 10(12).
- [13] Haque, M., 2018. A new approach of bangla news document summarization (Doctoral dissertation, University of Dhaka).
- [14] Jahan, B., Khatun, M., Zabu, Z.A., Hoque, A. and Rayhan, S.U., 2021. Construction of an Automatic Bengali Text Summarizer Using Machine Learning Approaches. Journal of Data Analysis and Information Processing, 10(1), pp.43-57.



# BANGLA E-MAIL BODY TO SUBJECT GENERATION USING SEQUENCE TO SEQUENCE RNNs

*AS*  
*11/09/22*

## ORIGINALITY REPORT

<b>12%</b>	<b>8%</b>	<b>7%</b>	<b>6%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

## PRIMARY SOURCES

<b>1</b>	<b>www.researchgate.net</b> Internet Source	<b>4%</b>
<b>2</b>	<b>Md Ashraful Islam Talukder, Sheikh Abujar, Abu Kaisar Mohammad Masum, Fahad Faisal, Syed Akhter Hossain. "Bengali abstractive text summarization using sequence to sequence RNNs", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019</b> Publication	<b>2%</b>
<b>3</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>2%</b>
<b>4</b>	<b>trendingarxiv.smerity.com</b> Internet Source	<b>1%</b>
<b>5</b>	<b>Submitted to Liverpool John Moores University</b> Student Paper	<b>&lt;1%</b>
<b>6</b>	<b>Submitted to Napier University</b> Student Paper	<b>&lt;1%</b>