Bangla E-mail Body to Subject generation using sequence to sequence RNNs

Conference Paper · November 2023

DOI: 10.1109/ICCCNT56998.2023.10306627

Md Abrar Hamim
University of Wollongong
8 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Samiul Alim
Daffodil International University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Afraz Ul Haque Rupak
Daffodil International University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Asif Iquebal Niloy
Daffodil International University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Bangla E-mail Body to Subject generation using sequence to sequence RNNs .

Moyin Talukder
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
moyin15-11827@diu.edu.bd

Afraz Ul Haque
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
afra15-13252@diu.edu.bd

Md Abrar Hamim
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
abrar15-11821@diu.edu.bd

Md. Samiul Alim
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
samiul15-11808@diu.edu.bd

Asif Iquebal Niloy
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
asif15-11895@diu.edu.bd

Md. Selim Reza
Computer Science and Engineering
Daffodil International University
Dhaka,Bangladesh
selim15-12020@diu.edu.bd

***Abstract*—In recent years, subject generation has emerged as one of the major challenges for deep learning and natural language processing (NLP). 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 subjects from those texts. Our model uses bi-directional RNNs in the encoding layer along with LSTMs and the decoding layer we used an attention model. Our model generates subjects using a sequence-to-sequence model. While developing this model, we encountered difficulties with text pre-processing, missing word counting, vocabulary counting, identifying unfamiliar words, word embedding, and other tasks. Our primary objectives in this model were to generate a subject and lessen its training loss of it. By crafting a succinct, fluent topic from an email body, we effectively decreased the training loss in our study trial to 0.001.**

Keywords—Bangla email, Email body to subject generation, sequence to sequence, RNN.

I. Introduction

Advancements in Natural Language Understanding (NLU) have seen the most recent models perform better than humans on several common tasks. An impactful subject line in the email helps get the job done. So, we proposed a natural language processing (NLP) model for creating an email subject by following the email body. As a result, the subject of an email is crucial for language comprehension, thinking, and the use of common-sense information like a person would. Any email subject can be generated with two distinct parts. Those techniques are extractive and abstractive. The majority of email topic generating systems use an extractive strategy that includes pulling the most crucial phrases or sentences from the given content first. So, they must be combined to create the subject. Second, in the abstractive technique, the email body is used to create a bottom-up subject, even when not all the words are 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 Bengali-speaking students. The approach we suggest in this study uses email text to generate the subject line. We collected a vast quantity of data and utilized it to train the model. After completing the program, the outcome was as expected. Textual work is always difficult. We must go through several steps in order to create a topic that is appropriate for an abstractive technique, including missing word counting, and vocabulary counting. As we know working in the Bangla language so we also needed word embedding, counting, text pre-processing, and using certain specific tokens for word encoder and decoder. In our approach, we use each of these phases. The sequence-to-sequence model we are using a two-layered and bidirectional RNN. To create a useful subject, the target body text is combined with two layers of RNN, each layer used LSTM, and the model which name was the Bahdanau attention model [1]. The phrases as input and encoder encode into a fixed length of vector, from which the decoder generates output a sequence. We have adapted the model for Bengali 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 several important variables. Deep learning techniques and models are explained in detail for the most part, but the description of major operations is more important. In this portion of our research article, we describe the experimental data set, data pre-processing, model training, and training loss. There are some benefits to using the sequence-to-sequence model for generating subjects from the Bengali email bodies. Some objectives are given below:

- Develop an efficient model to generate the subjects.
- Reduce the user's time to write emails subject.
- The new model is used in the Bengali language to generate subjects from the email body.

II. LITERATURE REVIEW

In Bangladesh, no comparable research or effort has been made to appropriately derive subjects from email bodies. Now, the Bangla language and the use of NLP (Natural Language processing) in the Bangla language business give context. An individual neural network it may translate simultaneously is often connected to encoders and decoders. The simple encoder and decoder will work better when given a predetermined amount of text to input and output [1]. Utilizing a 3-layer neural network, the scattered representation is combined. To create headlines, the group suggests modifying an encoder-decoder recurrent neural network technique [2]. A strategy for using the Linear discriminant analysis (LDA) model to categorize news. For each topic, the team determined the average number of words. The following formulae are used to get the average topic for each word. The subject of the document is the one with the highest likelihood value [3]. The frequency of swinging pronouns in the output summary has been decreased using pronoun replacement. The next stages are carried out in that specific order: preprocessing the input material, word tagging, pronoun substitution, and sentence rating. We have gone with ordering sentences based on sentence frequency as the pronoun replacement [4]. The Viterbi algorithm's decoding parameters influence the headlines that are produced. Using Expectation Maximization methods and headline generation, their parameter can be learned from the training data. To create a cohesive and grammatically correct headline, we are discussing how to identify the appropriate selection of keywords and how to combine them properly [5]. We have developed a bi-directional encoder, which entails the simultaneous reading of the input sequence by two GRU networks. The Seq2Seq model is built Recurrent Neural Network (RNN). For translation purposes, Both the encoder and decoder employ the Recurrent Neural Network (RNN), which has a single hidden layer with a size of 150 [6]. we have built a system that generates headlines for newspaper articles using linguistically motivated heuristics. We have demonstrated the value of choosing words in sequence from a newspaper piece when creating headlines [7]. Short summaries of written news articles are created by the authors using a parse-and-trim method. Requiring just one topic resulted in the highest ROUGE scores when testing written news. Topiary can be used for broadcast news, according to experts. [8]. We suggest and investigate a novel end-to-end technique for swiftly creating email responses that may be utilized with Inbox by Gmail [9]. The hybrid pointer Artificial intelligence (AI) generation network used by the attention-based seq2seq can develop words that are outside of a person's lexicon. It keeps the conceptual accuracy and coherence of the source content while producing text output of a manageable size. We mostly used "BANSData," a popular Bengali dataset that is freely accessible [10]. A graph-based model has been developed to merge numerous linked sentences. It uses an unsupervised abstractive text summarizing method which is the POS tagger and previously trained model for language. The collection contains 139 instances of abstractive document-summary pairings authored by humans [11]. Bangla text documents are pre-processed via tokenization, stop-words, and stemming before being turned into a summary. To make the summary more precise and understandable, the frequency of Words and positional value of sentences are employed. Use cue words and the document's skeleton to make the summary more accurate to the substance of the document [12]. Proposed a steaming method for Bangla texts that uses almost two lakh POS tags. Also, they proposed a redundancy removal method to remove redundancy from the summarized sentences. By assessing recall, accuracy, and f-score based on the Rouge metric, the performance of the suggested methodology is evaluated. [13]. RNN's goal-driven approach to eliciting personal preferences and its capacity to identify when the recommendation dialogue can end without compromising the quality of the solution is a powerful combination [14]. Suggested method for summarizing Bangla news documents outperforms all other methods already in use. A graph-based sentence scoring feature is included in this suggested method. The outcome is assessed using a common summary assessment tool called ROUGE, and it is discovered that the suggested method outperforms them all [15]. Based on sentence connections, sentiment analysis, and keyword scoring, the model generates a summary text. Empirical validation with other comparable systems shows that it can be used as a different approach to deal with this issue [16]. Using Seq2Seq Recurrent Neural Networks, we developed a comprehensive abstractive headline creation method in this study. Bangla uses data from more than 5,14,108 filtered full-text Bangla news items [17]. Although the frequency summarizer retrieves a summary more quickly, it is of higher quality. Less quality improvement in the sentence similarity technique is the disadvantage. Instead of abstraction summarization, the study discusses single document extraction-based summaries. We are illustrated by removing stop-words, Tokenization, Lemmatization, and Removing Duplicate Sentences [18]. These analyses perform analysis by tokenizing every word, removing stop-words, and stemming words. The statistical output accuracy is good enough for that are above 4 out of 5. But the final document's outcome is in no way satisfactory [19]. We have developed an extractive document summarizer using the method of sentence clustering approach. Data clustering is the process of identifying organic groupings or clusters within multidimensional data based on similarity metrics. The goal functions were optimized using a discrete differential evolution technique [20].

III. METHODOLOGIES

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 x1, x2,..., xd are coming from a vocabulary with a size of V, which produced an output sequence that is comparable to y1, y2,..., ys 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 creating 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.

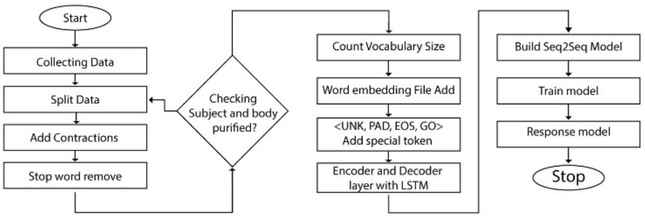
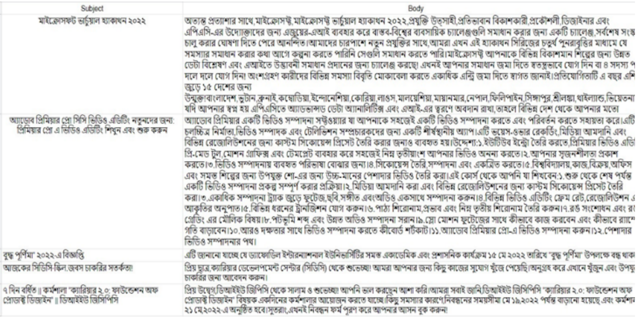


Fig. 1. System diagram

A. Experiment Data Set

Each deep learning method requires a significant quantity of information. When compared to the size of the dataset, the results are superior. We also need a substantial amount of information for the model. On the internet, not enough datasets are 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 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 offers, newsletters, and announcements about new products. And we obtained the email from the corporate organization. Emails marked "personal" are sent from individuals rather than organizations. This indicates that we should send emails from a personal address rather than a generic business or company email account. And we obtained this information through friends and family. In our dataset, we have 4232 emails.



Subject	Body
মাইক্রোসফট ভার্চুয়াল হ্যাকাথন ২০২২	অত্যন্ত প্রত্যাশার সাথে,মাইক্রোসফ্ট,মাইক্রোসফ্ট ভার্চুয়াল হ্যাকাথন ২০২২,প্রযুক্তি উত্সাহী,প্রতিভাবান বিকাশকারী,প্রকৌশলী,ডিজাইনার এবং এপিএসি-এর উদ্যোক্তাদের জন্য এজুয়্যের-এআই ব্যবহার করে বাস্তব-বিশ্বের ব্যবসায়িক চ্যালেঞ্জগুলি সমাধান করার জন্য একটি চ্যালেঞ্জ,সর্বশেষ সংস্ক চালু করার ঘোষণা দিতে পেরে আনন্দিত।আমাদের চারপাশে নতুন প্রযুক্তির সাথে,আমরা এখন এই হ্যাকাথন সিরিজের চতুর্থ পুনরাবৃত্তির মাধ্যমে যে সমস্যার সমাধান করার কথা আগে কল্পনা করতে পারিনি সেগুলি সমাধান করতে পারি।মাইক্রোসফ্ট 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 প দলে দলে যোগ দিন! অংশগ্রহণ কারীদের বিভিন্ন সমস্যা বিবৃতি মোকাবেলা করতে একাধিক এন্ট্রি জমা দিতে স্বাগত জানাই।প্রতিযোগিতাটি এ বছর এশি জুড়ে ১৫ দেশের জন্য উন্মুক্তঃবাংলাদেশ,ভুটান,ব্রুনাই,কম্বোডিয়া,ইন্দোনেশিয়া,কোরিয়া,লাওস,মালয়েশিয়া,মায়ানমার,নেপাল,ফিলিপাইন,সিঙ্গাপুর,শ্রীলঙ্কা,থাইল্যান্ড,ভিয়েতনা যদি আপনার স্বপ্ন হয় এপিএসিতে অ্যাডভান্সড ডেটা অ্যানালিটিক্স এবং এআই-এর ত্বরণে অবদান রাখা,তাহলে বিভিন্ন দেশ থেকে আপনার মতো
অ্যাডোব প্রিমিয়ার প্রো সিসি ভিডিও এডিটিং নতুনদের জন্য: প্রিমিয়ার প্রো এ ভিডিও এডিটিং শিখুন এবং শুরু করুন	অ্যাডোব 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 ইন্ট্রো তৈরি করতে,প্রিমিয়ার ভিডিও এডি প্রি-মেড টুল,মোশন গ্রাফিক্স এবং টেমপ্লেট ব্যবহার করে সহজেই নিম্ন তৃতীয়াংশ 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 এবংঅডিও একসাথে সম্পাদনা করুন।৪.বিভিন্ন ভিডিও এডিটিং ফ্রেম রেট,রেজোলিউশন এ আকৃতির অনুপাত।৫.বিভিন্ন ধরনের ট্রানজিশন যোগ করুন।৬.পাঠ্য শিরোনাম,প্রভাব এবং নিম্ন তৃতীয় শিরোনাম তৈরি করুন।৭.রঙ সংশোধন এবং রং গ্রেডিং এর মৌলিক বিষয়।৮.পটভূমি শব্দ এবং উন্নত অডিও সম্পাদনা সরান।৯.স্লো মোশন ফুটেজের সাথে কীভাবে কাজ করবেন এবং কীভাবে র‍্যাম্প গতি বাড়াবেন।১০.আরও দক্ষতার সাথে ভিডিও সম্পাদনা করতে কীবোর্ড শর্টকাট।১১.অ্যাডোব প্রিমিয়ার প্রো-এ ভিডিও সম্পাদনা করুন।১২.পেশাদার ভিডিও সম্পাদনার পথ।
বুদ্ধ পূর্ণিমা' ২০২২-এ বিজ্ঞপ্তি	এটি জানানো যাচ্ছে যে ড্যাফোডিল ইন্টারন্যাশনাল ইউনিভার্সিটির সমস্ত একাডেমিক এবং প্রশাসনিক কার্যক্রম ১৫ মে ২০২২ তারিখে 'বুদ্ধ পূর্ণিমা' উপলক্ষে বন্ধ থাকা
আজকের সিডিসি-স্কিল.জবস চাকরির সতর্কতা!	প্রিয় ছাত্র,ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) থেকে শুভেচ্ছা! আমরা আপনার জন্য কিছু কাজের সুযোগ খুঁজে পেয়েছি।অনুগ্রহ করে এখানে খুঁজুন এবং উপযু চাকরির জন্য আবেদন করুন।
৭ দিন বর্ধিত ‖ কর্মশালা "ক্যারিয়ার ২.০: ফাউন্ডেশন অফ প্রোডাক্ট ডিজাইন" ‖ ডিআইইউ জিসিপিসি	প্রিয় উদ্বেগ,ডিআইইউ জিপিসি থেকে সালাম ও শুভেচ্ছা! আপনি ভাল করছেন আশা করি।আমরা সবাই জানি,ডিআইইউ জিসিপিসি "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 কর্মশাল ২১ মে,২০২২-এ অনুষ্ঠিত হবে।সুতরাং,এখনই নিবন্ধন ফর্ম পূরণ করে আপনার আসন বুক করুন!

Fig. 2. Collected Dataset

B. Data Pre-Processing

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, etc. As a result, we have eliminated them and replaced them with their complete forms. After that, we cleaned 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. We described data pre-processing processes below:

- 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 following are some advantages of using a split function in Python. We might need to divide a long string into several shorter ones at some time. It is the opposite of unification, which joins two strings. In our dataset, we split for further processing. Then we remove the whitespace by using "re" (regular expression) liberty. Our dataset is Bangla for this reason we will UNICODE with whitespace.

- Add Contractions: A special kind of term called a contraction combines two or more additional words into a condensed form, usually with punctuation. Contractions can make the per-user feel like we're talking specifically to them and having a discussion. It makes a difference and makes our composing show up uncomplicated for everybody to get and make sense of. Because contractions are shorter, it also means that they take up less space. Since that, we'll frequently see them in notices where space is profitable. We are using "বি.দ্র ", "ড.", "ডা.", "ইঞ্জি:", "রেজি:", "মি.", "মু.", "মো." etc. As a result, it will help a more accurate model build.

- Stop-Word Remove: "Stop-words" often allude to the foremost common words in a conversation of any language. There's no all-inclusive list of "Stop-words" utilized by all NLP devices. Any dialect's "stop-words" are words that add little to a sentence's meaning. Without affecting the sense of the statement, they can be safely disregarded. These are some of the most prevalent, succinct function terms for various search engines, such as "এই", " ওকে", " কিন্তু", "তাই", " তবুও " and on. we also remove the special character. Also, we remove "https://", Bangla digits, Bangla special characters, and on. we also remove English words and unaccepted words.

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

C. 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. Look at the status bar to see how many pages, paragraphs, words, characters, or lines a document has. The vocabulary in this work is being counted. Following the vocabulary count (20011), we evaluated word frequency. For example, we tested the term, and the frequency of this word was 8. In word embedding, a learned representation of text, words with the same meaning are represented in the same way. One of the key advancements in deep learning for difficult natural language processing issues may have been made possible by this method of encoding words and documents. The process of storing certain words as the true meaning of a vector space is known as word embedding. Since each word is assigned to a different vector and the vectors are learned similarly to a neural network, the process is commonly referred to as "deep learning." To enhance the model, we applied a word-to-vector file that was previously learned. Word to vector file "bn w2v model" was utilized. Where we get the word embedding length (497405).

D. Model

In deep learning, there are several models and various model types that are employed for various purposes. For text modeling, the LSTM (Longest Short-Term Memory) 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. For this model, the splitting ratio is 80% for training and 20% for testing.

Identify applicable funding agency here. If none, delete this text box.

- 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 y_p (x|y)) is used if y is the source sentence.

- RNN Encoder-Decoder: The encoder and decoder paradigm uses RNNs to address seq2seq prediction difficulties. Although it was first developed to address problems with machine translation, it has also been successful in addressing problems with related sequence-to-sequence prediction, such as 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 the 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. A variable-length origin sequence is converted into a fixed length vector by the encoder, and a variable length destination sequence is converted back into the vector representation by the decoder. 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. 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 $X = (x_1, \dots\dots x_{T_x})$where context vector c is present, so

  $$h_t = f(x_t, h_{t-1}) \qquad (1)$$

  and

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

  where h_t equation represents that was concealed at time t. The context vector created from the concealed state sequence is denoted by the symbol c. And the non-linear function is f and g.
  If the Response subject of tables 1 and 2 is the expected word sequence $(y_1, \dots\dots y_{T_x})$ that the decoder anticipated, therefore the likelihood will be,

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

  (2)

  Where $(y_1, \dots\dots y_{T_0})$ The following is a conditional probability in our model,

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

  (3)

  Where, g = nonlinear function, y_t = output of probability, s_t= secret s ate.

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

  Bi-directional RNNs were employed. That is made up of Recurrent Neural Networks that run both forward and backward. A forward recurrent neural network's hidden state is $(\overrightarrow{h_1}, \dots\dots, \overrightarrow{h_{T_x}})$ and the following is the forward RNN (Recurrent Neural Network) order: (x_1 to x_{Tx}). Backward RNN sequence order is (x_{Tx} to x_1) and the secret state is $(\overleftarrow{h_1}, \dots\dots, \overleftarrow{h_{T_x}})$ So,

  $$h_j = [\overrightarrow{h_{jT}}\ ;\ \overleftarrow{h_{jT}}]^T \qquad (5)$$

  Where h_j= Summary of words that are anticipated and followed.
  Here, a_{ij} = is the SoftMax of e_{ij} This demonstrates how exceptional function 'i' is normalized and input position 'j' lines up with the (o/p) output at the location,

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

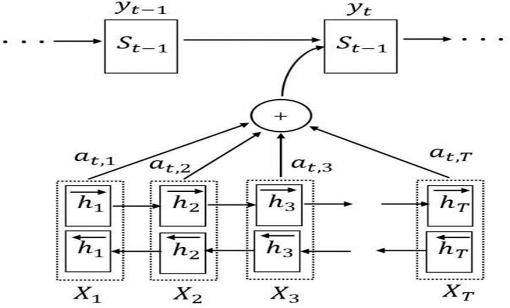


Fig. 3. RNN encoder-decoder

- 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 Seq-to-seq is utilized for natural language processing. Applications include text summarization, conversational modeling, picture captioning, and language translation. Seq-to-seq is an RNN-based model for encoder and decoder. It may be used as a guide for translating and communicating automatically. 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 particular instance of a group of letters combined into a meaningful semantic unit for processing in a particular text. A type is a group of tokens having the same character arrangement. Since tokens are the basic building blocks of Natural Language, the token level is where most of the 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 in 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.

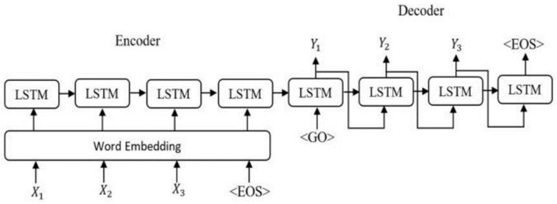


Fig. 4. Sequence to sequence model

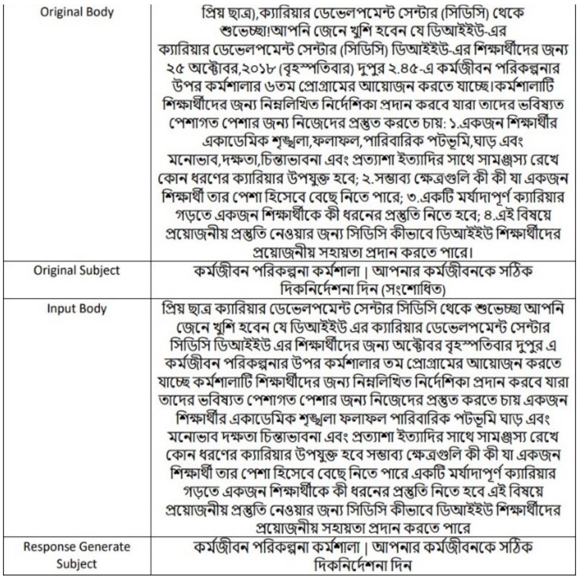
This model has incorporated certain special vocabulary cues, such as<UNK>, <PAD>, <EOS>, and <GO>. There are various limitations to vocabulary. Some words are still in use. The “UNK” token is used in place of the words. Each sentence in a batch added by a “PAD” token is the same length. Included is the conclusion of the sequence that notifies the encoder when it gets input 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, and the produced output or response output sequence is represented by mode y.

IV. EXPERIMENT AND OUTPUT

TensorFlow 1.15.0 and the sequence-to-sequence model were employed. The machine will be able to create a subject once finished training. To produce the subject, we will randomly choose an input sentence from the dataset and decide how long it should be. And for the parameter, we used an attention-based encoder. Using the following values: epoch=60, batch size=2, rnn size=256, learning rate=0.001, maintain probability=0.75, we used Adam Optimizer to calculate the learning rate for each parameter. Use the standard gradient descent optimizer for quicker convergences.

After several hours were spent on the provided dataset to train our model, the following machine response is shown as positive:

TABLE I: Sample result



Original Body	প্রিয় ছাত্র),ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) থেকে শুভেচ্ছা!আপনি জেনে খুশি হবেন যে ডিআইইউ-এর ক্যারিয়ার ডেভেলপমেন্ট সেন্টার (সিডিসি) ডিআইইউ-এর শিক্ষার্থীদের জন্য ২৫ অক্টোবর,২০১৮ (বৃহস্পতিবার) দুপুর ২.৪৫-এ কর্মজীবন পরিকল্পনার উপর কর্মশালার ৬তম প্রোগ্রামের আয়োজন করতে যাচ্ছে।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 ধরণের 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।
Original Subject	কর্মজীবন পরিকল্পনা কর্মশালা | আপনার কর্মজীবনকে সঠিক দিকনির্দেশনা দিন (সংশোধিত)
Input Body	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 নির্দেশিকা 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 ধরণের 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
Response Generate Subject	কর্মজীবন পরিকল্পনা কর্মশালা | আপনার কর্মজীবনকে সঠিক দিকনির্দেশনা দিন

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. Based on the total number of mistakes for each occurrence in the training set, the training loss is calculated. To put it another way, the loss is a gauge of how well a model for a certain circumstance was anticipated. 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 training 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 humanly 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.

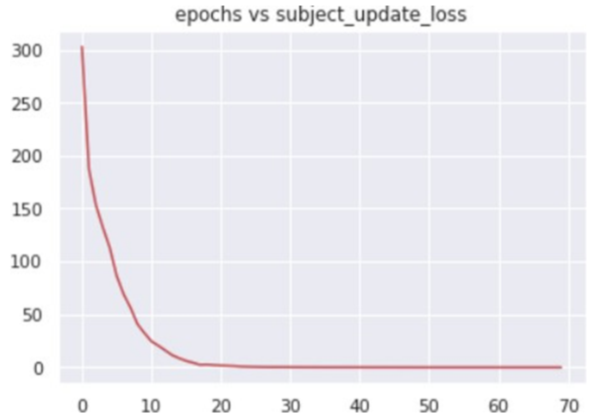


Fig. 5. Train loss

In contrast to other articles, this work has the highest level of accuracy. Some researchers did work that was almost identical to this study. TABLE II shows some connections between our work and some earlier text summarizing research. 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. Still, the minimum train loss overall is 0.001.

TABLE II. Comparison with some previous works

Work	*Algorithm*	*Train loss*
Summarizing abstract Bengali text with sequence-to-sequence RNNs	RNN	0.008
Bangla E-mail Body to Subject generation using seq2seq RNNs	RNN	0.001

V. CONCLUSION

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 a 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 primary 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 subjects, and we know that deep learning algorithms provide much better results with a larger dataset. As a result, we continue to gather data and expand our collection. Another drawback is that our model could only generate subjects with a certain number of words. We'll work to expand it so that it can generate subjects from the email 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 the Bangla language.

REFERENCES

[1] Sriranga, G.L., Likitha, P., Meghana, B. and Jayanthi, N., EFFICIENT TEXT SUMMARIZER USING POINT TO GENERATOR TECHNIQUE.

[2] Liedtka, D.J. and Hankamer, D.M., Headline Generator.

[3] Al Helal, M. and Mouhoub, M., 2018. Topic modeling in Bangla language: An LDA approach to optimize topics and news classification. Computer and Information Science, 11(4).

[4] Jahan, B., Mahtab, S.S., Arif, M.F.H. and Siddiqi, I., An Automated Bengali Text Summarization Using Lexicon Based Approach.

[5] Mondal, A.K., Maji, D.K. and Karnick, H., 2013. Improved algorithms for keyword extraction and headline generation from unstructured text. First Journal publication from SIMPLE groups, CLEAR Journal.

[6] Erraki, M., Youssfi, M., Daaif, A. and Bouattane, O., 2020, October. NLP Summarization: Abstractive Neural Headline Generation Over A News Articles Corpus. In 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS) (pp. 1-6). IEEE.

[7] Dorr, B., Zajic, D. and Schwartz, R., 2003. Hedge trimmer: A parse-and-trim approach to headline generation. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

[8] Zajic, D., Dorr, B. and Schwartz, R., 2005. Headline generation for written and broadcast news. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

[9] Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., Young, P. and Ramavajjala, V., 2016, August. Smart reply: Automated response suggestion for email. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 955-964).

[10] Dhar, N., Saha, G., Bhattacharjee, P., Mallick, A. and Islam, M.S., 2021, December. Pointer over attention: An improved Bangla text summarization approach using hybrid pointer generator network. In 2021 24th International Conference on Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

[11] Chowdhury, R.R., Nayeem, M.T., Mim, T.T., Chowdhury, M., Rahman, S. and Jannat, T., 2021. Unsupervised abstractive summarization of Bengali text documents. arXiv preprint arXiv:2102.04490.

[12] Abujar, S., Hasan, M. and Hossain, S.A., 2019. Sentence similarity estimation for text summarization using deep learning. In Proceedings of the 2nd International Conference on Data Engineering and Communication Technology (pp. 155-164). Springer, Singapore.

[13] Ullah, S., Hossain, S. and Hasan, K.A., 2019, September. Opinion summarization of Bangla texts using cosine similarity-based graph ranking and relevance-based approach. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE.

[14] McSherry, D., 2005. Incremental nearest neighbor with default preferences. In Proceedings of the 16th Irish conference on artificial intelligence and cognitive science (pp. 9-18).

[15] 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).

[16] Islam, M., Majumdar, F.N., Galib, A. and Hoque, M.M., 2020. Hybrid text summarizer for Bangla document. Int J Comput Vis Sig Process, 1(1), pp.27-38.

[17] Amin, R., Sworna, N.S., Liton, M.N.K. and Hossain, N., 2021, August. Abstractive Headline Generation from Bangla News Articles Using Seq2Seq RNNs with Global Attention. In 2021 International Conference on Science & Contemporary Technologies (ICSCT) (pp. 1-5). IEEE.

[18] Sarkar, A. and Hossen, M.S., 2018, December. Automatic Bangla text summarization using term frequency and semantic similarity approach. In 2018 21st International Conference of Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[19] Abujar, S., Hasan, M., Shahin, M.S.I. and Hossain, S.A., 2017, July. A heuristic approach of text summarization for Bengali documentation. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.

[20] Aliguliyev, R.M., 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications, 36(4), pp.7764-7772.

[21] 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).

[22] Jahan, B., Mahtab, S.S., Arif, M.F.H. and Siddiqi, I., An Automated Bengali Text Summarization Using Lexicon Based Approach.

[23] Haque, M., 2018. A new approach of Bangla news document summarization (Doctoral dissertation, University of Dhaka).

[24] Abujar, S., Masum, A.K.M., Islam, 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.

[25] Nallapati, R., Zhou, B., Gulcehre, C. and Xiang, B., 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv preprint arXiv:1602.06023.