TEXT ANALYSIS FOR BENGALI TEXT SUMMARIZATION USING DEEP LEARNING

BY

ABDULLAH AL MUNZIR
ID: 152–15–5731
&
MD. LUTFOR RAHMAN
ID: 152-15-5516

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

Supervised By

Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
MAY 3, 2019

APPROVAL

This Project titled "**Text analysis for Bengali text summarization using deep learning**", submitted by Abdullah Al Munzir, ID No: 152-15-5731 & Md. Lutfor Rahman, ID No: 152-15-5516 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 May 3, 2019.

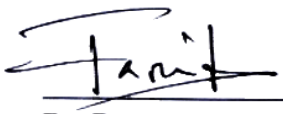
BOARD OF EXAMINERS

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

Internal Examiner

Dr. Md. Ismail Jabiullah
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

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



Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

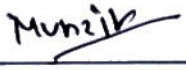
DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Sheikh Abujar, Lecturer, 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:

Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Abdullah Al Munzir
ID: 152 – 15 – 5731
Department of CSE
Daffodil International University

Md. Lutfor Rahman
ID: 152 – 15 – 5516
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We have given our efforts to this thesis. However, it would not have been possible without the kind support and help of many individuals. We would like to express our deepest appreciation to all those who provided us the possibility to complete this report.

At first, we express our heartiest thanks and gratefulness to almighty Allah for His divine blessings which allowed us to complete this thesis successfully.

A special gratitude we give to our supervisor, Sheikh Abujar, Lecturer of CSE department, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our thesis especially in writing this report. His endless patience, scholarly guidance, constant and energetic supervision, constructive criticism, valuable advice has made it possible to complete this thesis.

Furthermore, we would also like to acknowledge with much appreciation the crucial role of our department head, Professor Dr. Syed Akhter Hossain, who provided us with his precious time and kind help to finish this thesis. We also give our deepest thanks to all the faculty members and staff of CSE department of Daffodil International University.

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

ABSTRACT

Text summarization is an approach by which the size of one or more document is shorten and the shorten passage presents the core information of the document. In this modern era of information technology, we are over flooded with online data which raised the necessity of summary of the original text. Many methods have already implemented for English text and the effort for Bengali text are gaining alongside. In this paper we propose an extractive text summarization technique based on a deep learning model of Recurrent Neural Network (RNN). Our method is to classify the sentences as significant or not for the summary. We have used Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) for the backpropagation method. Between them we found Long Short-Term Memory (LSTM) more promising and we achieved average F1 scores- 0.63, 0.59, 0.56 for Rouge-1, Rouge-2 and Rouge-3 in some respects.

Keywords: Deep neural network, Supervised learning, Sequence classification, Bengali text, Extractive summary.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	II
Declaration	III
Acknowledgements	IV
Abstract	V
List of Figures	VIII
List of Tables	IX
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Outcome	3
1.5 Report Layout	3
CHAPTER 2: BACKGROUND	4-6
2.1 Introduction	4
2.2 Related Works	4
2.3 Research Summary	6
2.4 Scope of the Problem	6
2.5 Challenges	6
CHAPTER 3: RESEARCH METHODOLOGY	7-12
3.1 Data Collection	7
3.2 Dataset Creation	7
3.3 Proposed Model	8
3.4 Preprocessing	10
3.5 Vector Encoding	10
3.6 Model Settings	11
3.7 Train, Validation & Test	11
3.8 Summary Generation	11
3.9 Additional Model	12

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION 13-16

4.1 Results 13

4.2 Qualitative Analysis 14

4.3 Comparative Analysis 15

CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH 17

5.1 Summary 17

5.2 Conclusion 17

5.3 Recommendation 17

5.4 Implication for Further Research 17

REFERENCES 18-19

PLAGIARISM REPORT 20

LIST OF FIGURES

FIGURES		PAGE NO
Figure 3.1	Work flow of dataset creation	8
Figure 3.2	Architecture of our model	9
Figure 3.3	Preprocessing sample of a sentence	10
Figure 4.1	Comparison between result of GRU-RNN & LSTM	15
Figure 4.2	Comparative analysis of average F1 score of three models on the basis of Rouge-1 And Rouge-2.	16

LIST OF TABLES

TABLE		PAGE NO
Table 3.1	Example of the collected data	7
Table 3.2	Example of the labeled data	8
Table 3.3	A snip of pre trained word embedding	11
Table 3.4	Some sentences with predicted probabilities	11
Table 4.1	Performance of our model in different Rouge scores	13
Table 4.2	Average scores in different Rouge	13
Table 4.3	Example of our model generated summary	14

CHAPTER 1
INTRODUCTION

1.1 Introduction

There is an ocean of data available Online and increasing rapidly. People are searching through search engines for their coveted information. Sometimes they are getting what they precisely required and sometimes they are overwhelmed by similar types of duplicate information. For a seeker of selective information, it's really hard and time consuming to go through all the related documents available on the internet. Automatic text summarization can be very efficient for these types of circumstances. Anyone can go through a summary of a large document and ensure whether that is effective for him or not [1]. Lately it has become a part and parcel for opinion monitoring, indexing efficiency, recommendation of news and blogs [2]. Computer generated summaries are free from bias and have the ability to dislodge a large amount of selective information [1]. Automatic text summarization suffers from loss of information often but it has become necessary to ascertain the huge amount of data [3].

There are different types of summary can be prefaced based on the operation. Summary can be of two types depending on the input document. When the summary is generated from a single document it can be described as single document text summarization adversely if the input is similar types of two or more documents it is called multi document text summary.

A summary can also be described by its nature such as an extractive summary is created by drawing out the most significant points from a document without modification. On the contrary abstract summary represents the main gist of an article by reproducing significant part from that article [4]. An extractive summary does not include any text that are not part of the original article but an abstractive summary can add new word or sentence into the summary relevant to the source text. Thus most of the cases Implementing extractive summarization method is easier than abstractive summarization [5].

Based on the Content of an article summary can be separated into two types such as indicative and informative summary [4]. An indicative summary states briefly the primary idea about the whole text without explaining the content whereas an informative summary describes the context of the input data partly [6].

In this paper we focus on the single document text summary. The sentences of each document is patterned as vector based on feature extracted from text. These features are dependent on the nature of our input document. The sentences are taken as a classification problem if the sentence belongs to the summary then it is scored as 1 otherwise scored as 0. Whenever a new document is given to the summary every sentence gets a score according to the pre-trained pattern [7].

1.2 Motivation

The main objective of this paper is to introduce a compressed form of an article using deep learning method. RNN is a powerful model which operates very efficiently on text sequence. Several extractive and abstractive text summarization approach have been carried out on Bengali text throughout recent years as far our knowledge.

Many methods have been conducted for Bengali text summarization such as word scoring, sentence ranking, Graph scoring [8], cluster based method, TF-IDF method. On the best of our knowledge no neural network based approach was conducted for Bengali text document. There are approximately 150 national, regional and online Bengali newspapers in Bangladesh [9]. Most of them are covering various news everyday which is producing a large amount of data carrying on same topic. These data require summarization for future reader and analyzer.

We have followed a simple approach which is known as sequence classification. The work presented in [7], motivated us to conduct our summarization process based on neural network. In their work the sentences were labeled as crucial or not for the summary. Their approach indicated that the trained summary is intended to learn the pattern which helps in contributing summary. Whenever a new article passed on, the trained pattern will classify the sentences by giving score 0 to 1 which will produce a summary of certain sentences.

1.3 Rationale of the Study

There are a good number of researches going on English text summarization and the result are getting higher time by time. English text summarization is very much on the similar stage as human generated summarization. Bengali texts summarization is way backward compared to English text summarization as it faces some challenges of proper

linguistic tools, efficient stemming and lemmatizing, miscellaneous structure of Bengali text, unable to understand the context etc.

1.4 Outcome

Our research work aims at

1. Classify each sentence of the newspaper article whether they are part of the summary or not.
2. Generate summary which are nearly comparable to human generated summary.
3. Comparing the performance of our model with other previously recognized Bengali text summarization approach.

1.5 Report Layout

In this chapter we have discussed about the introduction of Automatic text summarization, motivation, rationale of the study and the outcome of the thesis. Later followed by the report layout.

In chapter 2, we will discuss about the background of our research topic.

In chapter 3, we will discuss about the methodologies employed in our study.

In chapter 4, we will discuss about the obtained results and discussion.

In chapter 5, we will discuss about the conclusion and future works.

CHAPTER 2
BACKGROUND

2.1 Introduction

Text summarization is one of most valuable parts of Natural Language processing. The first approach of text summarization was introduced in 1950. Since then several method was evaluated and suggested. Earlier on some simple techniques like position of word or sentences, frequency of words, terms from user queries and key phrases [1] are used. Bengali text summarization approach is also affected by those techniques.

2.2 Related Works

The very first work of automatic text summarization was carried out by Luhn [10] in 1958 based on term frequency and the approach was extended by Baxendale [11] by comprising the cue words and position of the sentences in the document. These valuable contributions laid the foundation of computerized text summarization and from then researchers are eagerly contributing in this arena of Natural Language Processing.

In the paper of Sarker [4] as usual preprocessing and stemming are performed at first, all the sentences are ranked based on features like thematic terms, positional value and the length of the sentences. Thematic terms were considered if the TFIDF value of a term was greater than a predefined threshold. The top ranked k sentences were considered as desired summary.

A neural attention architecture was proposed by cheng and Lapata [12] for extracting words and sentences. Their encoder aims to deduct the variant of neural attention of the input article as uninterrupted sentence features and decoder extracts sentences based on the applied attention.

A sequence classification based Neural network model is also proposed by Nallapati et al. [13]. They have treated the sentences of the document as binary form depending upon their existence in the summary which is very similar to our proposed method. They have used GRU-RNN based neural network model and we found LSTM-RNN more favorable.

Some enhanced features can be found on the proposed approach of Verma and Nidhi [3]. Extra features like number of proper nouns, number of numerals, Term Frequency-Inverse Sentence Frequency (TF-ISF), sentence to centroid similarities are extracted for

efficient abstract summarization. Highest TF-IDF scored sentence is considered as the centroid sentence and then cosine similarity between the highest TF-ISF scored and every sentence is calculated which is termed as Sentence to cosine similarity. The produced feature matrix was used as output to a two layers Restricted Boltzmann Machine (RBM) hence an enhanced feature vector is produced. The enhanced feature vector values are added to produce a score for every sentence. Then the sentences are sorted in a descending order and the most efficient sentences was selected for the summary.

A Recursive Neural Network application for multi document summarization has come out from the approach of Cao et al. [14].They have developed a hierarchical regression process for the sentence ranking task. They have conducted their research on multi document summarization datasets the DUC 2001, 2002 and 2004 and showed that their proposed method exceeds R2N2 state-of-the-art extractive summarization approaches.

Akter et al. [8] proposed a summarization approach for Bengali single and multiple text document. They have used K means clustering method for the candidate summary. The centroids of the clusters are considered by the highest scored sentences. Sentence is scored by the TF-IDF value of each word. If any cue word is detected in the sentences, then the score of the sentence is increased by 1.

A document classification task is also inaugurated by Isonuma et. al. [15] for single document summary. They have evaluated their neural network based model on the documents of two financial based news publisher. Convolutional Neural Network(CNN) is used for sentence embedding from word embedding because of its efficiency on sentence level classification problem. Another neural network based model LSTM-RNN is used for extracting summaries from the document.

Topic based opinion summary for Bengali document is carried out by Das and Bandyopadhyay [16]. For distinguishing the sentiment information, they have used an annotation tool which annotate sentence for summary by pointing out the root words. Annotator spots the sentiment words according to their Part of Speech (POS) categories. K-means clustering is used for combining topic-sentiment. Finally, for Selecting the sentence of summary theme based relational graph is used and page rank algorithm is used for information recovery. Their approach is efficient for theme detection.

2.3 Research Summary

A lot of contribution have been made for English text summarization. Many rule based and machine learning algorithm based Extractive and abstractive text summarization has been implemented for English texts. Deep learning models are the new player in this area. More accurate results have been achieved by using these deep learning models. For Bengali text summarization, all the efforts were rule based up to now. Extractive and abstractive text summarization for Bengali documents have been implemented throughout the recent years but they are not effective so far. We tried to implement Bengali text summarization in an extractive way.

2.4 Scope of the Problem

No deep learning based approach was conducted for Bengali text summarization as far as our knowledge. We have tried our best for summarizing Bengali single document text summarization using deep learning. Machine generated English text summarization has already reached to human generated text summarization. Due to difficulty of processing the Bengali text, the summarization of Bengali text did not reach that level yet. Efficient stemming and lemmatizing will be a revolution for Bengali text summarization problem.

2.5 Challenges

The main problem we faced for this work is limitations of data. Bengali data is very hard to collect. Processing the Bengali texts is another challenging matter for us. There is no rich annotated Bengali text summarization corpus. Besides LSTM as well as deep learning models need high specifications of hardware components. These models also require a large amount of time to operate.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Data Collection

It is very challenging to work on Bengali data analysis due to the unavailability of Bengali dataset. We have used a dataset of 200 Bengali news article with 3 sets of summary for every article from [17]. The article was on random topics. To make the dataset compatible for our proposed model, we needed the best one summary for each documents so we manually sorted out the best summary from available 3 sets. And all the articles and the corresponding summaries were saved on separate .txt file with UTF-8 encoding for final dataset creation. Table 3.1 represents an example of collected data.

Table 3.1 Example of the collected data

Article: রাজধানীর বনশ্রীতে দুই ভাইবোনের রহস্যজনক মৃত্যুর ঘটনায় এখনো মামলা হয়নি। শিশুদের বাবা মামলা করবেন বলে জানিয়েছে পরিবার। দুই শিশুর লাশের ময়নাতদন্ত হয়েছে। তাঁদের গ্রামের বাড়ি জামালপুরে লাশ দাফন করা হবে। খাবারের নমুনা পরীক্ষার ফলাফল এখনো পাওয়া যায়নি। শিশুদের বাবা আমান উল্লাহর বন্ধু জাহিদুল ইসলাম আজ মঙ্গলবার বেলা সোয়া ১১টার দিকে প্রথম আলোকে এসব কথা জানিয়েছেন। রামপুরা থানার ভারপ্রাপ্ত কর্মকর্তা (ওসি) রফিকুল ইসলাম বলেন, এখনো মামলা হয়নি। পরিবারের পক্ষ থেকে আজ মামলা হতে পারে। জিজ্ঞাসাবাদের জন্য চায়নিজ রেস্তোরাঁর ব্যবস্থাপক, কর্মচারী, পাচককে থানায় নেওয়া হয়েছে। চায়নিজ রেস্তোরাঁ থেকে আগের দিন আনা খাবার গতকাল সোমবার দুপুরে গরম করে খেয়ে ঘুমিয়ে পড়ে নুসরাত আমান (১২) ও আলভী আমান (৬)। এরপর তারা আর জেগে ওঠেনি। অচেতন অবস্থায় হাসপাতালে নেওয়া হলে চিকিৎসকেরা তাদের মৃত ঘোষণা করেন। পরিবারের অভিযোগের ভিত্তিতে পুলিশ জিজ্ঞাসাবাদের জন্য ওই রেস্তোরাঁর মালিককে ওই দিনই থানায় নিয়ে গেছে। নুসরাত ভিকারুননিসা নূন স্কুল অ্যান্ড কলেজের পঞ্চম ও আলভী হলি ক্রিসেন্ট স্কুলে নার্সারি শ্রেণির শিক্ষার্থী।
Summary: রাজধানীর বনশ্রীতে দুই ভাইবোনের রহস্যজনক মৃত্যুর ঘটনায় এখনো মামলা হয়নি। শিশুদের বাবা আমান উল্লাহর বন্ধু জাহিদুল ইসলাম আজ মঙ্গলবার বেলা সোয়া ১১টার দিকে প্রথম আলোকে এসব কথা জানিয়েছেন। চায়নিজ রেস্তোরাঁ থেকে আগের দিন আনা খাবার গতকাল সোমবার দুপুরে গরম করে খেয়ে ঘুমিয়ে পড়ে নুসরাত আমান (১২) ও আলভী আমান (৬)। পরিবারের অভিযোগের ভিত্তিতে পুলিশ জিজ্ঞাসাবাদের জন্য ওই রেস্তোরাঁর মালিককে ওই দিনই থানায় নিয়ে গেছে।

3.2 Dataset Creation

Since our summarization model is extractive basis with supervised learning, we need article where every sentence will be labeled as 1 or 0. 1 means this sentence belongs to summary and 0 means it isn't. So First of all, we split the sentences of the articles and the summaries. Then converted the articles comparing with the corresponding summaries that sentences with labeling format. We have done this conversion by simple python program with the help of jellyfish 0.7.1 library [18]. We saved the whole labeled

articles into a .csv file with UTF-8 encoding for further processing. Table 3.2 represents some of labeled sentences.

Table 3.2 Example of the labeled data

Sentence	Label
রাজধানীর বনশ্রীতে দুই ভাইবোনের রহস্যজনক মৃত্যুর ঘটনায় এখনো মামলা হয়নি	1
খাবারের নমুনা পরীক্ষার ফলাফল এখনো পাওয়া যায়নি	0
শিশুদের বাবা মামলা করবেন বলে জানিয়েছে পরিবার	0

The workflow of dataset creation is depicted in Figure 3.1 where the inputs are unannotated documents and summaries and outputs are labeled articles.

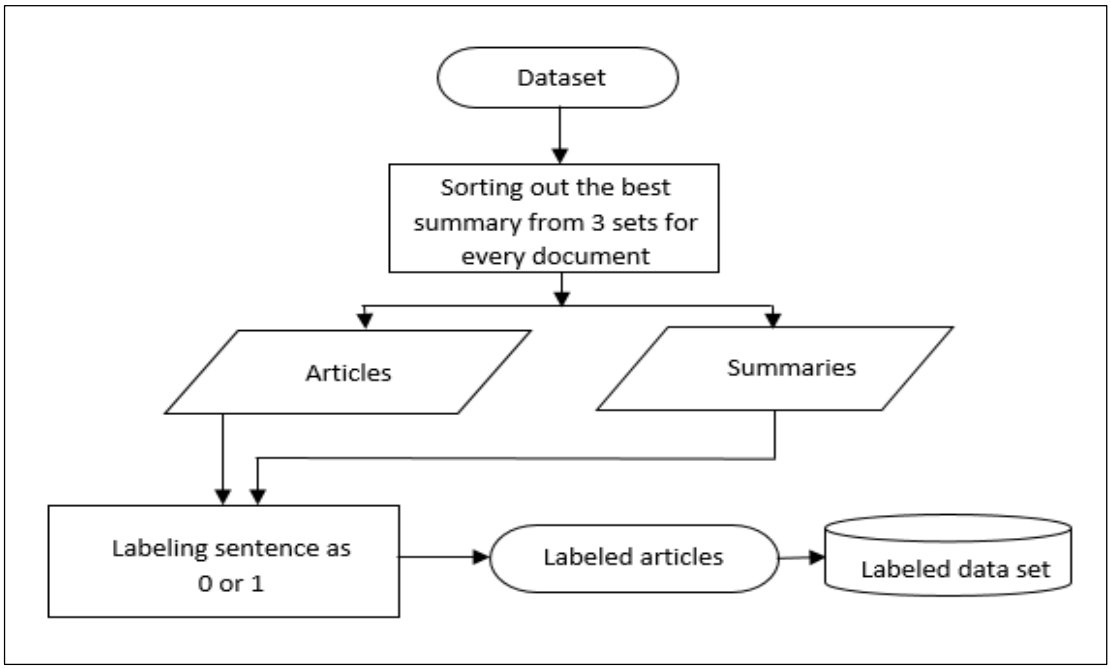


Figure 3.1 Work flow of dataset creation

3.3 Proposed Model

In this work, we have built a model of summarizing Bengali single document by sequences classification. Where all the sentences of the documents will be binary classified after visited all the sentences. The reason behind the binary classification is to ensure the membership of that sentence to the summary or not. The base of this model is Long Short-term Memory (LSTM), an architecture of deep neural network. The LSTM network achieved very high performances along with solving gradient vanishing or gradient exploding problems [19]. The LSTM architecture has three gates in every cell along with a single memory cell.

Simply the mathematical computation of each cell can be represented as shown below:

$K = [h_{t-1}, k_t]$ ..... (1)

$o_t = \sigma(w_o \cdot K + b_o)$ ..... (2)

$f_t = \sigma(w_f \cdot K + b_f)$ ..... (3)

$i_t = \sigma(w_i \cdot K + b_i)$ ..... (4)

$h_t = o_t * \tanh(c_t)$ ..... (5)

$c_t = f_t * c_{t-1} + i_t * \ tanh(w_c \cdot K + b_c)$ ..... (6)

Here w is weighted matrices, $o \rightarrow$ output gate, $f \rightarrow$ forget gate and $i \rightarrow$ input gate. k_t is input at current time step. b stands for biases and c_t means cell state. σ represents sigmoid function.

Our model is structured with several sequential layer as Embedding layer, LSTM layer and Dense layer. Every sentence will be feed as a sequence of words. The sequences of words will be converted as word vector by the current embedding of Embedding layer. The Embedding layer gives the output of fixed length vector encoded sequences to the next LSTM layer. LSTM layer starts calculation process for every sequential input in individual cell through three gates and memory cell. It produces output to the next-Dense layer. Outputs from the previous layer are converted by the dense layer, basis on sigmoid activation function for binary classification. Through Training and validation, the model builds a parameter and by this, measures sentence weight in testing time for predicting the membership in summary.

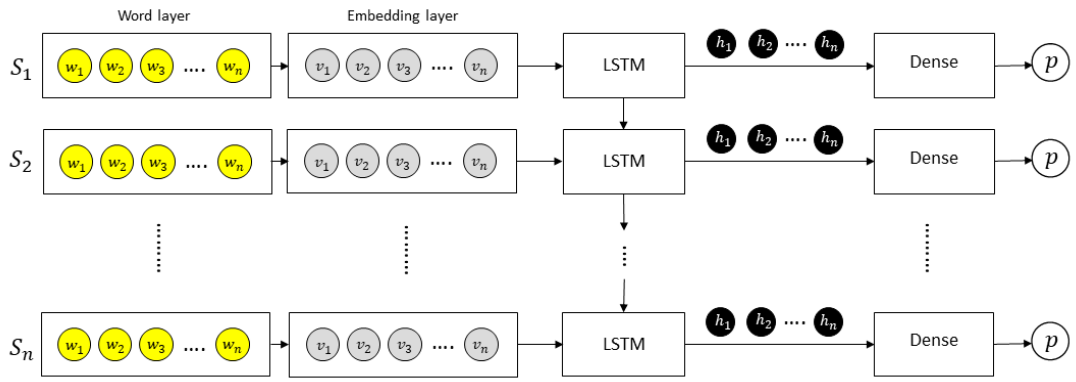


Figure 3.2 Architecture of our model

Figure 3.4 depicts our model where *w* defines the words of sentences, *v* denotes the word vectors of every sentence and *h* stands for hidden vectors. *p* is the probability of every sentence.

3.4 Preprocessing

To analysis textual data in machine learning it should be properly processed for more accurate results. In the initial state our data was very noisy with punctuation marks, numeric values and huge number of stop words. The presence of those type of noise confuses the model to make decision. In the first instance we removed the punctuation marks from all the sentences of our data set. Then removed numeric and non-alphabetic characters and finally removed stop words from [20] after tokenization. Figure 3.3 depicts preprocessing sample of a sentence.

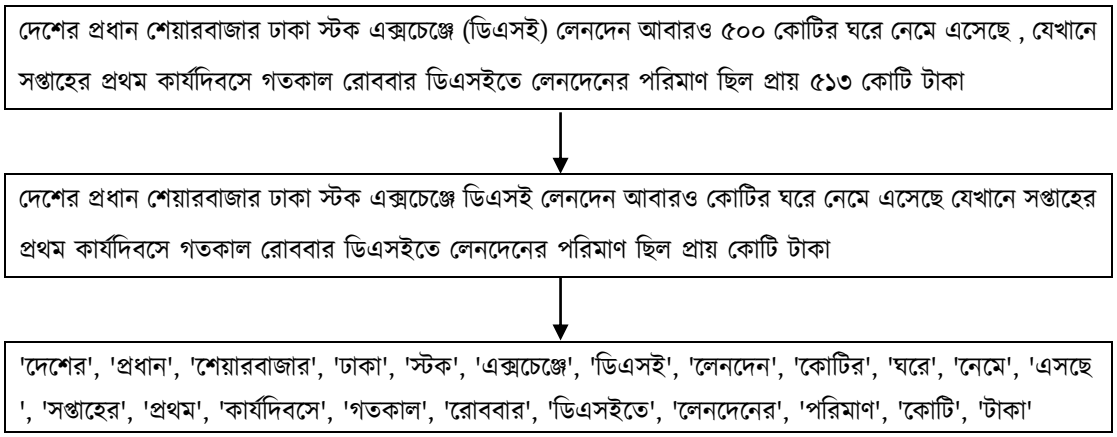


Figure 3.3 Preprocessing sample of a sentence.

3.5 Vector Encoding

Vector encoding of the words is done by word Embedding. The Natural Language Processing algorithms of deep learning cannot perceive the textual data. Here word embedding has a great role to make it sensible for neural network by defining vector of integer for every word uniquely. It is actually 2D vector for each vocabulary where vocabularies indicate rows and columns represent corresponding integers. The larger size of embedding illustrates more accurate relations between words and produces better output in textual analysis. Any rich pre-trained embedding for Bengali language isn't available in the web till now. So we have used our own word embedding of about 7500 vocabularies with 256 embedding dimension. Table 3.3 illustrate a snip of pre trained word embedding in next page.

Table 3.3 A snip of pre trained word embedding.

বছর	0.0713	-1.6886	2.0468	-1.023	-0.616	-2.621	1.62177	0.1077	-1.46	-1.030
করার	-1.4230	1.41716	1.9259	1.00332	-1.185	-1.2140	1.35827	0.48385	-1.300	0.57008
সময়	-1.8481	0.1628	-1.2672	-0.8246	0.2001	1.1979	0.61423	-0.0566	-0.551	0.92260
একজন	-0.6051	-0.5536	0.65958	0.8611	-1.332	0.04647	-0.0566	-0.5514	1.8330	0.64281
থাকে	-0.2871	-0.2179	0.1711	-1.4828	-0.422	0.53509	0.47058	0.26869	-1.483	1.46605
পর্যন্ত	2.14932	-0.7410	0.138951	1.93817	-1.483	-0.7888	1.46605	-1.4693	-0.592	-1.5522

3.6 Model settings

Due to balance our data we added padding to the sentence on the basis on maximum sentence length. We set this length as 60 words. Padding makes the sentences to the equal length. We set 128 neurons in the LSTM layer and drop out value as 0.5 for reducing the overfitting.

3.7 Train, Validation & Test

From preprocessed 200 documents containing about 2300 sentences, we used 90% for train our model. 5% used for validating the result during training. The overall value accuracy was 67% after several epochs. Rest of 5% has used to taste our model performance. After completing the training, model gives the probability of every sentence of given document. From that probabilities the summary will be generated in next process.

3.8 Summary generation

The given result from the model is actually the percentage of probability for every sentence that ensures their presence in the summary. We extracted 5 most probable sentences and made summary by maintaining chronological order of original documents. The summary length is depending upon user choices. The example of sentences with predicted probabilities is given in Table 3.4.

Table 3.4 some sentences with predicted probabilities.

Sentences	Probabilities
মুন্সিগঞ্জ থেকে পুরান ঢাকার ভিক্টোরিয়া পার্ক এলাকায় মেয়ের বাসায় বেড়াতে এসেছিলেন মুক্তি যোদ্ধা মো. হাবীবুল্লাহ	0.853
পথে শ্রীনগরে বাসের ভেতরে তাঁকে অচেতন অবস্থায় পাওয়া যায়	0.399

তাঁকে উদ্ধার করে প্রথমে স্থানীয় হাসপাতাল ও পরে ঢাকার মিটফোর্ড হাসপাতালে ভর্তি করা হয়	0.602
চিকিৎসাধীন অবস্থায় গতকাল শনিবার তাঁর মৃত্যু হয়	0.745
মৃতের ভাতিজা নহর হোসেন বলেন, হাবীবুল্লাহ মুন্সিগঞ্জ জেলা আওয়ামী লীগের উপদেষ্টা এবং ইছাপুর ইউনিয়ন পরিষদের চারবার নির্বাচিত চেয়ারম্যান ছিলেন	0.561
৩১ জানুয়ারি তিনি বাসে করে ঢাকার উদ্দেশে রওনা হওয়ার পর অচেতন হয়ে পড়েন	0.717

3.9 Additional model

We also built an additional model by GRU-RNN where GRU stands for Gated Recurrent Unit for comparative analysis [21]. It is also an updated version of Recurrent Neural Network. Where LSTM has three gates, it has two gates named update gate & rest get. It also gave an optimum performance but more less than our proposed model. The result of this model will be compared with our model in comparative analysis section in the next chapter.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results

For experimenting and evaluating our model we used 10 news article documents. Every document has a human generated summary. To evaluate our model, we have used Rouge score measure of Rouge-1, Rouge-2 and Rouge-3, that is actually calculated from Recall, Precision & F1. The experimental result of the model is described in Table 4.1.

Table 4.1 Performance of our model in different Rouge score

Documents no.	Rouge-1			Rouge-2			Rouge-3		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Doc-1	0.73	0.48	0.58	0.66	0.43	0.52	0.64	0.41	0.50
Doc-2	0.73	0.69	0.71	0.69	0.65	0.67	0.67	0.63	0.65
Doc-3	0.82	0.81	0.82	0.78	0.77	0.77	0.76	0.74	0.75
Doc-4	0.53	0.58	0.55	0.51	0.56	0.53	0.49	0.53	0.51
Doc-5	0.47	0.38	0.42	0.38	0.30	0.33	0.37	0.29	0.33
Doc-6	0.55	0.61	0.58	0.511	0.57	0.53	0.48	0.53	0.50
Doc-7	0.5	0.45	0.47	0.42	0.39	0.40	0.40	0.36	0.38
Doc-8	0.69	0.70	0.69	0.67	0.69	0.68	0.65	0.67	0.66
Doc-9	0.85	0.76	**0.80**	0.82	0.73	**0.77**	0.80	0.72	**0.76**
Doc-10	0.69	0.66	0.67	0.62	0.59	0.61	0.58	0.56	0.57

Table 4.1 Illustrates the recall, precision and F1 score in different Rouge measures of some documents. Where we got highest F1 scores- 0.80, 0.77 & 0.76 for Rouge-1, Rouge-2 and Rouge-3 consecutively. The average scores of those document's summary has represented in Table 4.2.

Table 4.2 Average scores in different Rouge

Rouge	Recall (avg.)	Precision (avg.)	F1 (avg.)
Rouge-1	0.66	0.61	0.63
Rouge-2	0.61	0.56	0.59
Rouge-3	0.59	0.54	0.56

4.2 Qualitative Analysis

We have used very little size of data (only 200 documents) yet any deep learning model needs huge amount of data for best performance. The more training gives more accurate result. In this case, though our model has trained with comparatively less amount of data, it gives a surprising level of performance. We chose documents having fixed five length of sentences in summary for test set to maintain the length ration of model generated and referenced summaries. A document of human generated summary and our model generated summary has displayed in Table 4.3, where our model generated summary is very close to the reference summary.

Table 4.3 example of our model generated summary

Document: মুন্সিগঞ্জ থেকে পুরান ঢাকার ভিক্টোরিয়া পার্ক এলাকায় মেয়ের বাসায় বেড়াতে এসেছিলেন মুক্তিযোদ্ধা মো . হাবীবুল্লাহ। পথে শ্রীনগরে বাসের ভেতরে তাঁকে অচেতন অবস্থায় পাওয়া যায়। তাঁকে উদ্ধার করে প্রথমে স্থানীয় হাস পাতাল ও পরে ঢাকার মিটফোর্ড হাসপাতালে ভর্তি করা হয়। চিকিৎসাধীন অবস্থায় গতকাল শনিবার তাঁর মৃত্যু হয়। মৃতের ভাতিজা নহর হোসেন বলেন, হাবীবুল্লাহ মুন্সিগঞ্জ জেলা আওয়ামী লীগের উপদেষ্টা এবং ইছাপুর ইউনিয়ন পরিষ দের চারবার নির্বাচিত চেয়ারম্যান ছিলেন। তিনি পরিবারের সঙ্গে মুন্সিগঞ্জের সিরাজদিখানের শিয়ালদি এলাকায় থাকতে ন। ৩১ জানুয়ারি তিনি বাসে করে ঢাকার উদ্দেশে রওনা হওয়ার পর অচেতন হয়ে পড়েন। খবর পেয়ে স্বজনেরা শ্রীনগ রের শনবাড়ি এলাকা থেকে তাঁকে উদ্ধার করেন। তিনি কীভাবে অচেতন হয়ে পড়লেন, তা আর জানা যায়নি। তবে তাঁ র পাকস্থলীতে 'বেনজোডায়াজিপিন' নামক রাসায়নিকের উপস্থিতি পাওয়া গেছে। তাঁর দুই ছেলে ও এক মেয়ে রয়েছে। নহর বলেন, প্রাথমিকভাবে অজ্ঞান পার্টির খপ্পরে পড়ার ধারণাই তাঁরা করছেন। তবে রাজনৈতিক প্রতিদ্বন্দ্বীদের কেউ তাঁকে বিষ প্রয়োগে হত্যা করে থাকতে পারে।
Model generated: মুন্সিগঞ্জ থেকে পুরান ঢাকার ভিক্টোরিয়া পার্ক এলাকায় মেয়ের বাসায় বেড়াতে এসেছিলেন মুক্তি যোদ্ধা মো. হাবীবুল্লাহ। চিকিৎসাধীন অবস্থায় গতকাল শনিবার তাঁর মৃত্যু হয়। মৃতের ভাতিজা নহর হোসেন বলেন, হা বীবুল্লাহ মুন্সিগঞ্জ জেলা আওয়ামী লীগের উপদেষ্টা এবং ইছাপুর ইউনিয়ন পরিষদের চারবার নির্বাচিত চেয়ারম্যান ছিলে ন। ৩১ জানুয়ারি তিনি বাসে করে ঢাকার উদ্দেশে রওনা হওয়ার পর অচেতন হয়ে পড়েন। খবর পেয়ে স্বজনেরা শ্রীনগ রের শনবাড়ি এলাকা থেকে তাঁকে উদ্ধার করেন।
Human generated: মুন্সিগঞ্জ থেকে পুরান ঢাকার ভিক্টোরিয়া পার্ক এলাকায় মেয়ের বাসায় বেড়াতে এসেছিলেন মুক্তি যোদ্ধা মো. হাবীবুল্লাহ। মৃতের ভাতিজা নহর হোসেন বলেন, হাবীবুল্লাহ মুন্সিগঞ্জ জেলা আওয়ামী লীগের উপদেষ্টা এবং ইছাপুর ইউনিয়ন পরিষদের চারবার নির্বাচিত চেয়ারম্যান ছিলেন। ৩১ জানুয়ারি তিনি বাসে করে ঢাকার উদ্দেশে রওনা হওয়ার পর অচেতন হয়ে পড়েন। তিনি কীভাবে অচেতন হয়ে পড়লেন তা আর জানা যায়নি। তবে তাঁর পাকস্থলীতে বেনজোডায়াজিপিন নামক রাসায়নিকের উপস্থিতি পাওয়া গেছে।

Since our model is supervised learning based, its performance is highly dependent on the accuracy of training data. If the summaries of the dataset are prepared by the experts, the model will predict close to their cogitation. By overcoming data size and quality limitations, it achieved a very satisfied results.

4.3 Comparative Analysis

The result of our proposed model has already presented in the previous section, now first of all we want to make a comparison between our model (LSTM) and GRU-RNN model that we built initially. The average F1 score in Rouge-1, Rouge-2 and Rouge-3 has been compared through bar diagram in Figure 4.1, below:

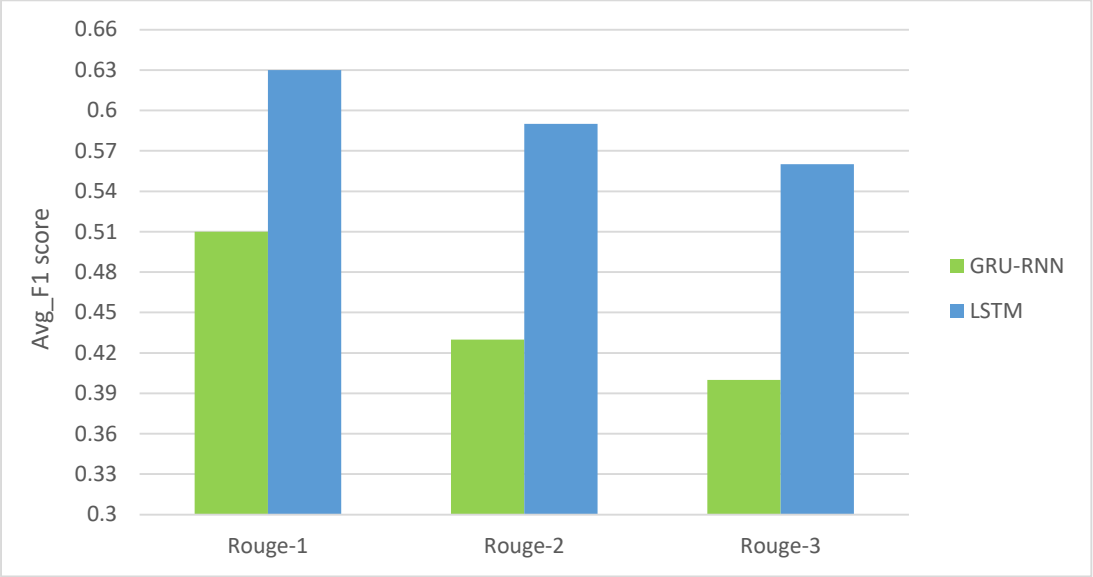


Figure 4.1 Comparison between the result of GRU-RNN and LSTM

Figure 4.1 signifies the performance of our model (using LSTM) is much better than GRU-RNN based model. In Rouge-1 the score of LSTM is 23% better than GRU-RNN where 37% and 40% increase in Roue-2 and Rouge-3 respectively.

Secondly, we have compared the result with two existing Bengali text summarization models [22,23] implemented in previous couple of years. The same dataset [17] has been used to evaluate those methods and our model too. A comparison result of average F1 measure based on Rouge-1 and Rouge-2 has been delineated in Figure-4.2. Models 1 denoted in [22] and model 2 denoted in [23].

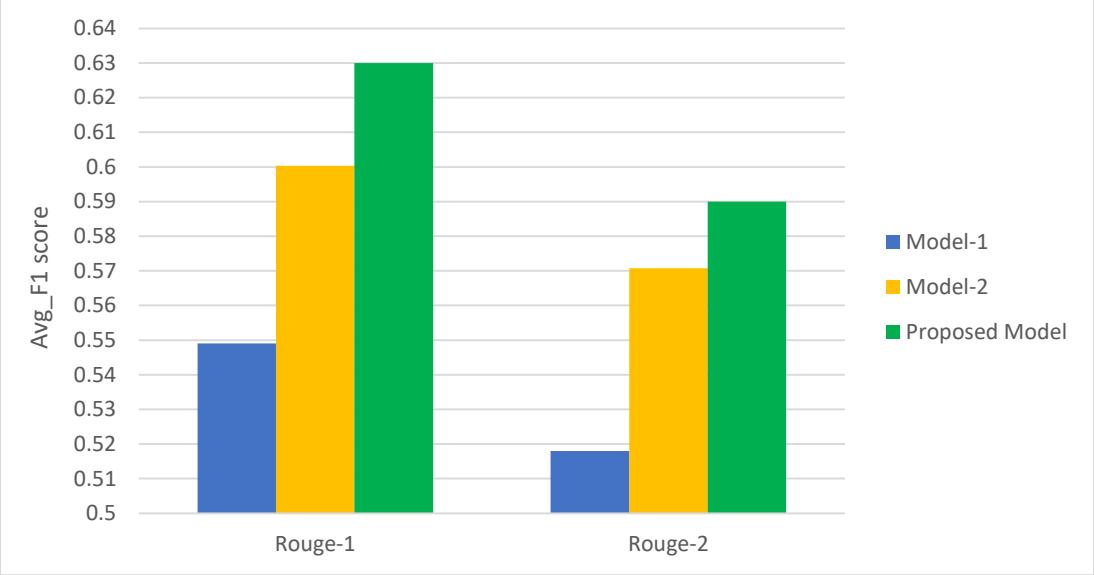


Figure 4.2 Comparative analysis of average F1 score of three models on the basis of Rouge-1 And Rouge-2.

The above comparison (in Figure 4.2) gives a clear conception about the performance of our proposed model. The average F1 scores are 3-5% higher than the nearest highest score based on the Rouge-1 & Rouge-2. We also calculate Rouge-3 that demonstrate the accuracy more precisely. Our Rouge-3's average F1 score is 0.56, that is very good score for the Bengali single document summarization till now.

CHAPTER 5
SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary

In this work, we proposed a straight forward deep neural network of sequence classification based sentence extractive summarizer model for Bengali single document. There are many research works on automatic summarizing technique for English and other languages but very few number for Bengali language. As far our study most of the Bengali summarizer are rule based and context specific. Comparing English, Bengali is a complex language for its grammatical configuration, complex alphabet, sentence structure and more. So generally it is quite difficult to set traditional algorithm for summarization.

5.2 Conclusion

In this instance we introduced Deep neural network with Bengali text summarizer. This is very new approach to this field. Where there is need of thousands to millions set of data for supervised training, we have very limited data. In spite of those inadequacy we have achieved satisfactory performance.

5.3 Recommendation

If this model will train with large amount of quality data set, the accuracy and performance will be surely increased. There can be used properly annotated dataset to increase accuracy. By the domain specific train set, the prediction will be more precise for the document under that specific domain.

5.4 Implication for Future Research

We used only one LSTM layer, by increasing the number of LSTM layer, the result could be more improved. There is an extra option for improving result by use of larger size of pre trained word embedding. Another scope is optimized the model for multi documents summarization.

REFERENCES

[1] Hinge, Sonam, and Sheetal Sonawane. "Cluster Based And Graph Based Methods Of Summarization: Survey And Approach." *International Journal of Computer Engineering and Applications* 10.II: 25-34.

[2] Hahn, Udo, and Inderjeet Mani. "The challenges of automatic summarization." *Computer* 33.11 (2000): 29-36.

[3] Verma, Sukriti, and Vagisha Nidhi. "Extractive Summarization using Deep Learning." *arXiv preprint arXiv:1708.04439* (2017).

[4] Sarkar, Kamal. "Bengali text summarization by sentence extraction." *arXiv preprint arXiv:1201.2240* (2012).

[5] Abujar, Sheikh, et al. "A heuristic approach of text summarization for Bengali documentation." *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017.

[6] Lin, C. Y., and E. Hovy. "Automated Text Summarization and the SUMMARIST System." *Proceedings of the TIPSTER Text Program* (1998): 197-214.

[7] Fattah, Mohamed Abdel, and Fuji Ren. "Automatic text summarization." *World Academy of Science, Engineering and Technology* 37 (2008): 2008.

[8] Akter, Sumya, et al. "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm." *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2017.

[9] Bangla Newspaper List of all Online Bangladeshi Newspaper https://www.24livenewspaper.com/bangla-newspaper Accessed on 29 march 2019.

[10]Luhn, Hans Peter. "The automatic creation of literature abstracts." *IBM Journal of research and development* 2.2 (1958): 159-165.

[11] Baxendale, Phyllis B. "Machine-made index for technical literature—an experiment." *IBM Journal of research and development* 2.4 (1958): 354-361.

[12] Cheng, Jianpeng, and Mirella Lapata. "Neural summarization by extracting sentences and words." *arXiv preprint arXiv:1603.07252* (2016).

[13] Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[14] Cao, Ziqiang, et al. "Ranking with recursive neural networks and its application to multi-document summarization." *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[15] Isonuma, Masaru, et al. "Extractive summarization using multi-task learning with document classification." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[16] Das, Amitava, and Sivaji Bandyopadhyay. "Topic-based Bengali opinion summarization." *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010.

[17] Bangla Natural Language Processing Community, http://bnlpc.org/research.php, accessed on 01 march, 2019

[18] jellyfish 0.7.1 python library. https://pypi.org/project/jellyfish, Accessed on 20 march, 2019

[19] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[20] Bengali Stop words list. https://www.ranks.nl/stopwords/bengali, Accessed on 10 march, 2019

[21] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[22] K. Sarkar, "A keyphrase-based approach to text summarization for English and Bengali documents," *International Journal of Technology Diffusion (IJTD)*, vol. 5, no. 2, pp. 28-38, 2014.

[23] Haque, Md Majharul, Suraiya Pervin, and Zerina Begum. "Automatic Bengali news documents summarization by introducing sentence frequency and clustering." *2015 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015.

PLAGIARISM REPORT

TEXT ANALYSIS FOR BENGALI TEXT SUMMARIZATION USING DEEP

ORIGINALITY REPORT

20%	15%	11%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Kookmin University Student Paper	1%
4	bnlpc.org Internet Source	1%
5	"Natural Language Processing and Chinese Computing", Springer Nature, 2018 Publication	1%
6	core.ac.uk Internet Source	<1%
7	Avik Sarkar, Md. Sharif Hossen. "Automatic Bangla Text Summarization Using Term Frequency and Semantic Similarity Approach", 2018 21st International Conference of Computer and Information Technology (ICCIT),	<1%