BANGLA NEWS CLASSIFICATION AND TEXT SUMMARIZATION BASED ON MULTI-TYPE TEXT

BY

MD. ASADUZZAMAN
ID: 191-15-12907

MD. RASEL
ID: 191-15-12729

NAZAM BEEN SADDI
ID: 191-15-12470

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

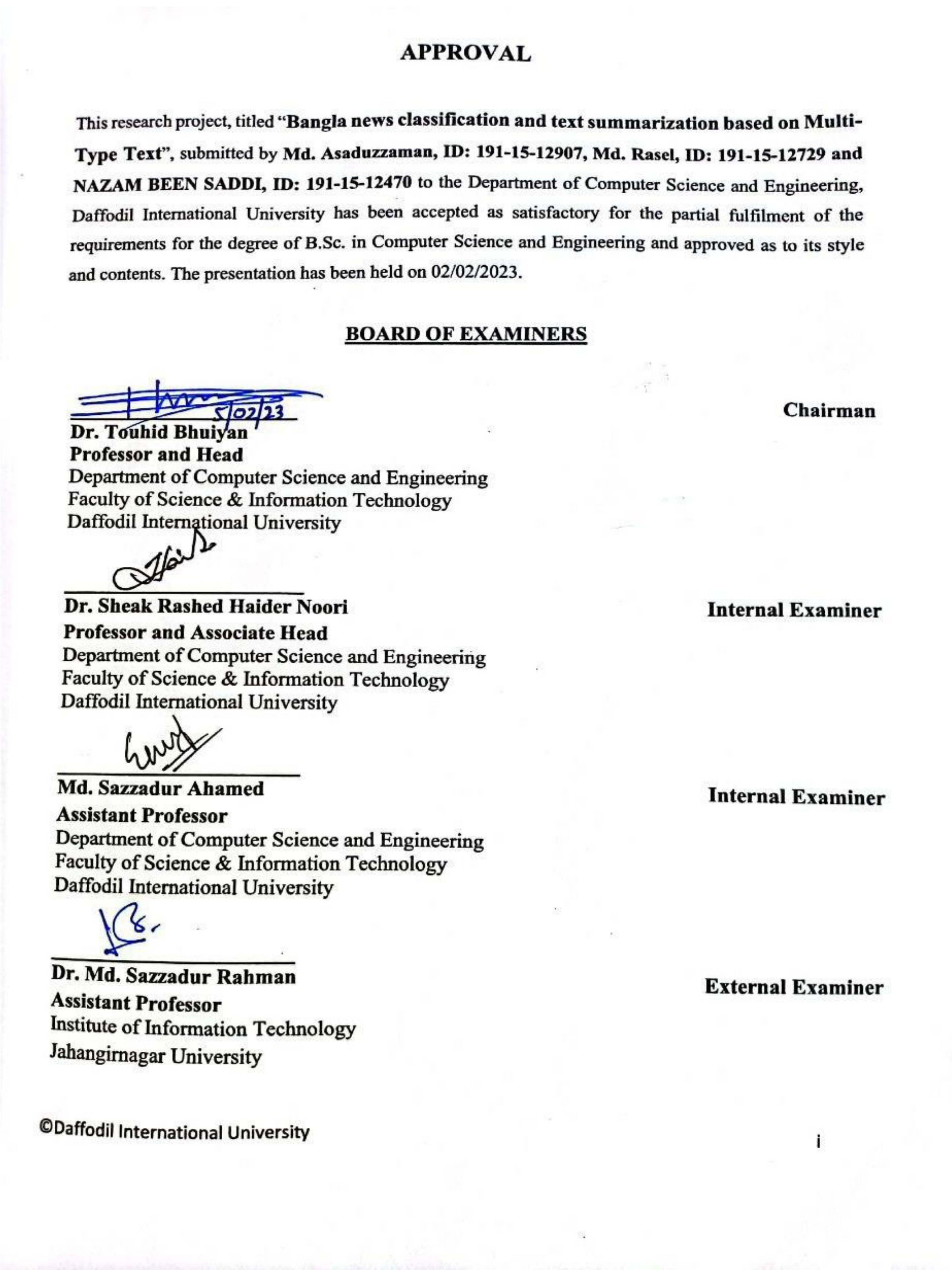
Supervised By

Sharun Akter Khushbu
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Johora Akter Polin
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
FEBRUARY 2nd 2023



APPROVAL

This research project, titled **"Bangla news classification and text summarization based on Multi-Type Text"**, submitted by **Md. Asaduzzaman, ID: 191-15-12907, Md. Rasel, ID: 191-15-12729 and NAZAM BEEN SADDI, ID: 191-15-12470** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 02/02/2023.

BOARD OF EXAMINERS

5/02/23

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

Chairman

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

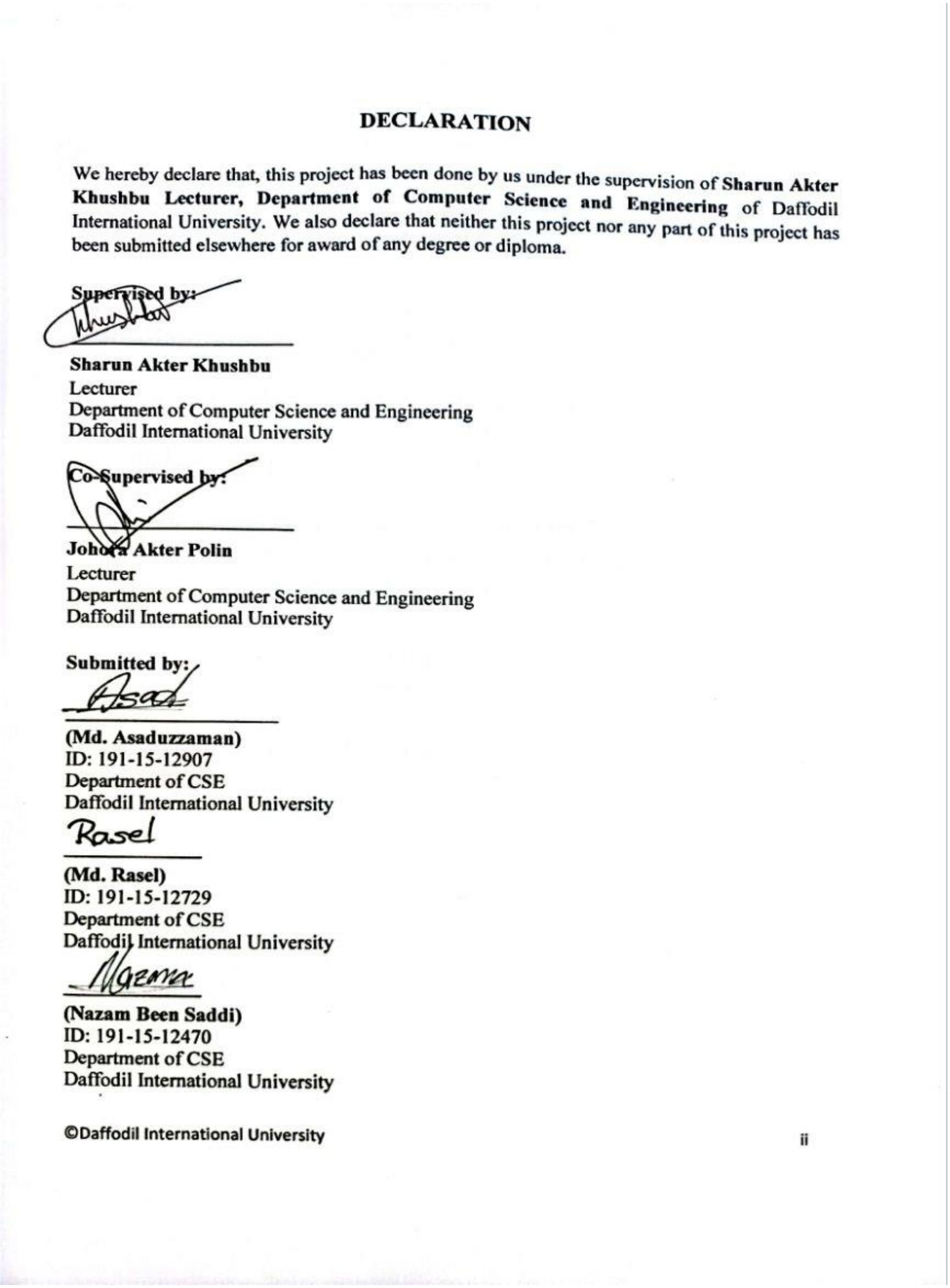
Internal Examiner

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Sazzadur Rahman
Assistant Professor
Institute of Information Technology
Jahangirnagar University

External Examiner



DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Sharun Akter Khushbu Lecturer, Department of Computer Science and Engineering** of Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Khushbu

Sharun Akter Khushbu
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Johora Akter Polin
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Asad

(Md. Asaduzzaman)
ID: 191-15-12907
Department of CSE
Daffodil International University

Rasel

(Md. Rasel)
ID: 191-15-12729
Department of CSE
Daffodil International University

Nazma

(Nazam Been Saddi)
ID: 191-15-12470
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 year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Shahrun Akter Khushbu, Lecturer,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor **Dr. Touhid Bhuiyan** Sir**,** Department of CSE, for his kind help to finish our project and also to other faculty member 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 patients of our parents.

ABSTRACT

Text Summarization is a strategy of summarizing any passage, document or text automatically. Summarized text is nothing but the minimized form of the given text. There are many techniques available for the English text summarization but for Bangla there are few works exits. Our main purpose of this project is to generate an understandable and meaningful summary which is fluent and easy for the people. Because language is the main obstacle for communication. Text summarization can help reduce the time it takes to read and understand long texts by providing a condensed version of the content. Summarizing text can help clarify the main points and improve overall comprehension of the material. We have collected data from kaggle, a web platform and newspaper. To get an outline we've got to use our model. Our model is Sequence-to-Sequence supported bi-directional RNN with LSTM. Throughout this project we've got some problems like preprocessing, vocabulary count, missing words count, word embedding and so on. During this project, our main goal is to deduce the operating loss and build a fluent outline and build a higher method for Bangla text summarization. We tend to area units able to cut back the loss worth below 0.031. Our model accouracy is 97.69%. Our model is ready to make theoretical text summarization.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

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

CHAPTER 1
INTRODUCTION

1.1 Introduction

Bangla is not only a language, It's an emotion. We fight to achieve our mother language. A large amount of data or information has been a part of our regular life. Those data are mainly digital. Nowadays we have some issues like reading a lot of stuff like a large number of paras to understand something. Which could be articles, reviews, or newspapers. If we saw a newspaper, we saw a huge number of texts. If we want to understand the main news, we have to read the full article, which takes a lot of time.

With the help of LSTM & Sequence to Sequence algorithm, we present to you a summarization that could save time & easily understand the material or news. Through Natural Language Processing (NLP) we can understand the document to generate the summary. Text summarization is like a process where a large amount of text or data has been submitted and with the help of Machine learning (ML) it compresses the data and presents us with an understandable short form of the data which is called text summarization. Text summarization is not new in Machine learning (ML). In the English language, it is used vastly. But on the other hand, in the Bangla language, it is very difficult to summarize data or form. The main reason for this kind of issue is that the Bangla language has a large number of alphabets and grammatical rules.

Text summarization will be worn out in 2 ways: they're Extractive and theoretical. The Extractive text summarization is a lot of correct to handle massive amounts of knowledge and provide an associate degree correct outline. And it's done by looking for the foremost used words or common words and so shrinking them and presenting newly designed sentences. On the other hand, Abstractive text summarization clarifies the data and improves them, and makes a sentence [1][2]. Yes, that's correct! Extractive text summarization involves selecting and combining important sentences or phrases from the original text to create a summary, while abstractive text summarization involves generating new sentences and phrases that capture the meaning of the original text. Extractive summarization is generally easier to implement than abstractive summarization because it does not require the ability to understand the text and generate new sentences. However, extractive summarization can sometimes produce a summary that is choppy or lacks coherence, as it may not consider the overall structure and organization of the original text. Abstractive summarization, on the other hand, can produce a more coherent summary, but it is more difficult to achieve due to the complexity of natural language processing and the need to understand the meaning of the text. [1]. For better understanding, the summarization system underscores the needs of the developers designing [3].

On sequence-to-sequence model, it’s a class that mainly uses Recurrent Neural Network (RNN) to solve complex Language problems like Text Summarization. It gained a lot of popularity in the last few years. It provides good-quality summarization because of its improved technique which was proposed on sequence-to-sequence model. Most techniques of this mode follow three different categories: parameter inference, network structure, and decoding [4].

Long short-run Memory (LSTM) could be a style of continual neural network (RNN) that's capable of learning long-run dependencies in ordered knowledge. RNNs are a sort of neural network that are notably well-suited for process-ordered knowledge, like statistics, language text, and speech. LSTMs can learn long-run dependencies by employing a special style of a memory cell which will store data over long periods of your time and by selecting retain or forget data as required. This enables LSTMs to model dependencies that are for much longer than those shaped by ancient RNNs, which might solely learn dependencies over some time steps [9]. LSTMs have been applied to a wide range of tasks, including language translation, language modeling, speech recognition, and more. They are also widely used in natural language processing (NLP) tasks, such as language generation, machine translation, and text classification. One disadvantage of LSTMs is that they can be computationally expensive to train, particularly for very large datasets. Additionally, LSTMs can be difficult to optimize, as the learning process can be unstable. Despite these challenges, LSTMs have proven to be very powerful and effective at learning long-term

dependencies in sequential data, and they continue to be an important tool in the field of deep learning [8].

At the forefront of development is the ongoing language model, BERT (Bidirectional Encoder Representations from Transformers), which OpenAI pre-arranged using a sizable corpus of text. BERT achieves text framing by jumbling the data text and using it to generate a summarization yield using its coarse discourse depiction capabilities [27]. BERT is a viable device for this cycle, as its capacity to create fine-grained portrayals of messages decreases the size of the information for outline models and makes it conceivable to catch significant distance conditions between words that can be lost in a conventional, pack-of-words approach. It likewise enjoys the benefit of being more productive than different models, permitting synopsis errands to be finished quicker with fewer computational assets [28], [29]. BERT's benefits have permitted it to turn into an important device in the outline of extensive messages, as its fine-grained portrayals and capacity to catch significant distance conditions give a viable answer for some synopsis assignments that different models will be unable to do [27].

The OpenAI-made GPT-2 (Generative Pretrained Transformer 2) is a sizable language model that can be changed for a variety of NLP applications, including text frames. GPT-2 can provide a list by selecting the most important information from the given text and utilizing its language creation capabilities and comprehension setting [30]. In this manner, clients truly should include GPT-2 as a gadget to give an early phase to summarizing text rather than relying entirely upon its outcomes. GPT-2 is a valuable resource that can be used to create text summaries, allowing clients to save time while writing lengthy social event reports [31]. Rather than relying on the system to convey a full and exact rundown, clients should review and change GPT-2's outcome to ensure that the possible result resolves their issues [32]. In this capacity, GPT-2 should be viewed as a supplementary tool in the creation of outlines rather than a complete replacement for a traditional human rundown [31].

In the scope of NLP errands, Google and Carnegie Mellon's transformer-based language model XLNet perform better compared to BERT (another transformer-based language model). The facts show that XLNet outperforms BERT in specific NLP tasks. However, it is important to remember that the results obtained from any synopsis framework are only as good as the data it is prepared on [33], [34]. To sum up, XLNet is a strong outline device with a history of creating excellent outlines. Regardless, it is critical to consider various variables when selecting an outline model, including the presentation of the framework, the nature of the information, and the use of case-explicit goals, to ensure that the most appropriate synopsis model is chosen for a given task [35].

During the time spent grouping news article distributed in the Bangla language into preset classes, like economics, sports, entertainment, international, Science and Technology and so on, we additionally order the news dataset. This is frequently accomplished by utilizing AI calculations and regular language handling strategies, which examine the content of the reports and group them according to catchphrases, subjects, and topics. The order of Bangla news intends to improve the client experience for Bangla-speaking crowds and simplify it for perusers to find relevant data.

1.2 Objective

The objective of Bangla text summarization is to create a summary of a longer Bangla text document or passage that retains the most important information and main points from the original document. The goal is to provide a condensed version of the text that is easier to read and understand, while still conveying the most important information. Text summarization can be useful for a variety of applications, such as quickly understanding the main points of a document, creating a summary for a report or presentation, or reducing the amount of time needed to read and comprehend a large text.

Here are some common objectives of Bangla text summarization:

- To reduce the time needed to read and understand a lengthy text
- To make a document shorter so that it will be simpler to read and comprehend
- To rapidly comprehend a document's major points
- To summarize a text or chapter in preparation for a report or presentation
- To glean crucial details and essential points from a text
- To assist readers in recognizing the main points and defenses made in a text
- To make it possible for readers to get a sense of a piece without reading the whole thing
- To draw attention to the text's most crucial details so they can be examined or discussed in further detail.

1.3 Motivation

Firstly, we have to understand what is the main purpose of text summarization. We live in a period where we had to deal with a large amount of data daily. Like blogs, newspapers and many more. Text account is the method of generating a crisp and fluent outline of an extended text document. The goal of a text account is to form an outline that retains the foremost vital data from the first text, whereas reducing the length of the text and removing redundant or orthogonal data. There square measure many motivations for exploitation text account, including:

Time-saving: When we read a newspaper article, to understand the news we have to read the whole article to understand. It will take a lot of time. To save this waste of time we need text summarization to understand a short amount of time.

Information overload: With the abundance of knowledge obtainable on the web, it is overwhelming to sift through all of it. Text accounts will facilitate folks quickly get the most points from an oversized quantity of text.

Improved comprehension: A literate outline will facilitate folks to perceive the most points of a text by presenting them in a very clear and crisp manner.

Text reuse: Summaries are used as a start line for making new content, like articles or reports.

Translation: account is accustomed to produce shorter versions of texts that square measure easier to translate into different languages.

Sentiment analysis: account is accustomed to quickly perceive the sentiment of a text, like whether or not it's positive or negative.

1.4 Reasons behind the study

Bengali is the most beautiful language in the world. It has a fascinating past. The only language in the world for which individuals have given their lives in order to preserve it is the Bengali language. However, there are a lot of cutting-edge tools and technologies available in the current contemporary world for linguistic study, however, the approaches or methods used to study the Bengali language are less developed than those used to study other languages. We should participate in our language due to this. We can summarize its essence using a text summarizer. As a result, the time required to read the entire text is less in today's society. A concise & error-free text summary can assist readers in quickly grasping the meaning of a lengthy work. There are numerous specialized tools and models available for the various languages. In the NLP field, there are also a few Bengali methods and approaches, although they are extremely limited and not enough. We should therefore broaden the Bengali NLP research domain. Preprocessing is the biggest challenge while working with Bengali text. Unicode might be the best solution to this problem. We can use the Unicode representations of particular characters or symbols to address this problem. Bengali text cannot be accessed in the NLTK library. Because of this, Bengali tools do not work as intended, and as a result, the results are not as accurate as for other languages. There isn't any other way to resolve this issue, the only method for resolving this problem is through research. As a result, we are attempting to process Bengali languages and explain how to do so in our research project. We are trying to create an abstract text summarizer for Bengali. This makes it possible for us to summarize the news and offer the most accurate summarization of the documents.

1.5 Research Problems

- What is the Bangla Text Summarization technique?
- What is meant by text summarization?
- What is summarization method?
- How can we define Bangla Text Summarization?
- What is the process of Bangla Text Summarization?
- What benefits does Bangla Text Summarization offer?
- How do Bangla text summarization and English text summarization different from one another?
- How is the Bengali text pre-processed using Natural Language Processing?
- How can we continue to improve Bangla Text Summarization in the future?
- How does the model for summarizing Bengali text work?
- Which model is best for text Summarization?
- What is the purpose of Bangla news classification?

- How is Bangla news classified into different categories?
- What are the common categories used in Bangla news classification?
- What techniques and algorithms are used in Bangla news classification?
- How does Bangla news classification benefit readers and media organizations?
- What are some of the challenges in implementing Bangla news classification effectively?

1.6 Expected Outcome

Our main goal was to place the required documentation out in the fields, which we accomplished. The builder then creates tools for users. A recent study looks at text compression in Bengali. There have already been numerous analyses works created in the past to lessen Bangla lessons. We're attempting to build an automation process to accomplish our goal. The machine is conditioned by an automated system. So, in order to understand our proposed model, you must read the machine. The objective of our study is to generate an abstract text abbreviation using our developed model while maintaining the method's remarkable efficiency. While preparing the model, strive to achieve perfection and reduce total loss.

1.7 Report Layout

There is a total of five chapters in our report. Throughout the chapter-1 we an overview of the whole project. Several sections like 1.1- Introduction, 1.2- Objective, 1.3- Motivation, 1.4- Reasons behind the study, 1.5-Research Questions, 1.6- Expected Outcome, 1.7- Report layout. The discussion sections in the second chapter are 2.1- Introduction, 2.2- Literature review, 2.3- Research summary, 2.4- Challenges. The research method, including its subsections, is covered in Chapter 3 they are 3.1- Introduction, 3.2- Research subject and intermediary,3.3-Data collection procedure, 3.4- Data fetching and data pre-processing, 3.5- Arithmetical analysis, 3.6- Executional requirements. The tests and paragraph are covered in the fourth part 4.1-Introduction, 4.2- Implementational results, 4.3- Descriptive Analysis, 4.4- moral. The fifth chapter discusses the subsections 5.1- Summary of the study, 5.2- Conclusion, 5.3- recommendations 5.4- Further study. At the conclusion of every section, situations were created that aided our research.

CHAPTER 2

BACKGROUND STUDIES

2.1 Introduction

Text summarization is the process to shorten a large document or passage into short form. Making the document more readable, summarize, and comprehensive is the main theme of text summarization. Shortening the long text into the short form takes more time and costlier for the people. After reading the full document, sometimes we are unable to understand and catch the main theme of the document. But text summarization makes it easy for us and converts the task in a more efficient way.

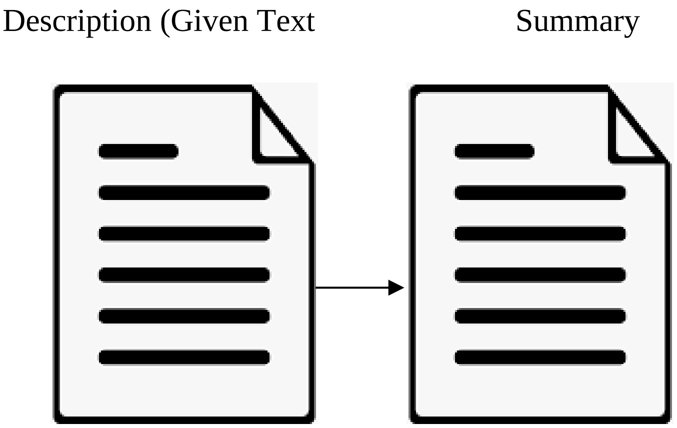


Figure 2.1.1: Summarization View

We collect our data from different sources like the internet, various news channel websites, Kaggle, etc. Getting and saving the information is a multipart process for us at the same time we need a space for storing it but text summarization solves the problem. There are two techniques, extractive and abstractive, to deal with the problem of text abbreviation. We use the abstractive way to summarize.

2.2 Literature Review

In the Bangla language, there are many grammatical rules. That's why it can't be analyzed as other languages [1]. The main reason to choose this method is an approach to Bangla text summarization. The primary benefit of deep learning is that it allows computers to perform tasks that were previously impossible or too difficult for traditional algorithms. This makes it possible to solve complex problems that require a high level of understanding, such as understanding natural language and recognizing facial expressions. Additionally, deep learning algorithms require significantly less human intervention than traditional algorithms, making them more efficient and cost-effective [5]. To improve the performance of a deep learning model, hyperparameter tuning can be used. Hyperparameter tuning involves adjusting the values of parameters such as the learning rate, number of layers in each layer to optimize the model's performance [18]. Other techniques such as regularization, data augmentation, and transfer learning can also be used to improve the model's performance. Finally, it is important to use the right evaluation metric to measure the performance of the model [6]. For text summarization, we use the Extractive text summarization, LSTM, and Sequence-to-Sequence model [16].

Extractive text summarization it's formed in three phases. They are: Analyzing text, scoring sentences, and Summarizing the high-scoring sentences. Then creating representation is the major step for the document. Mainly it splits the text and creates it as a paragraph, then sentence and tokens. Then the process is word removal [14]. Then the second step is to determine the sentences which are important or not for the documents and extend the information to different topics with the help of sentence scoring. The source should be measured if the sentence is understandable or not. Previous steps send combinations to the last steps and the last steps combine the sources and generate a summary [2] [17] [21].

Sovereign representations are the easiest system derives like primary representation or Sovereign representation of the text, which have summarized and then dig-in the importance of the content depending on this representation [15]. Some of the greatest summarization methods hinge on this topic of representation and this class of approaches exhibits a powerful variation in sophistication and illustration power [3]. Here we explain fifteen (15) sentences with the help of a scoring method. Every method (15) is implemented [1]. The input gate controls the flow of data into the cell, permitting it to buy and select the elements of the input sequence to recollect. The output gate controls the flow of data out of the cell, permitting it to buy and select the elements of the input sequence to output. The forget gate controls the flow of data among the cell, permitting it to buy selection forgetting elements of the input sequence [7].

By controlling these gates, LSTMs can learn long-term dependencies and generate output sequences over an extended period. The gates in an LSTM are composed of a sigmoid layer and a pointwise multiplication operation. The sigmoid layer outputs a number between 0 and 1, which is then multiplied by the previous state or the output from the previous layer [8]. This multiplication operation is used to control how much information is passed through the gates. The forget gate controls how much of the previous state is forgotten, while the input gate controls how much of the new input is added to the state [11] [13]. This allows the LSTM to find long-range dependencies and create predictions supported by discourse data from previous time steps. Overall, LSTM square measure is a strong tool for modeling sequent knowledge, and square measure is widely utilized in a range of tongue processes (NLP) and speech recognition tasks [9].

A sequence-to-sequence model, additionally called a seq2seq model, could be a sort of neural specification designed to method consecutive information. It's usually used for tasks like computational linguistics, language report, and dialog systems. The seq2seq model consists of 2 components: an associate degree encoder and a decoder. The encoder takes a sequence of an input file, like a sentence in linguistic communication, and converts it into a fixed-length illustration, known as a latent vector or context vector [11]. The decoder then takes this latent vector associate degrees and generates an output sequence, like a translated sentence or an outline. The seq2seq model is trained to maximize the probability of the output sequence given the input sequence. It will do this by minimizing a loss operation, like cross-entropy loss, that measures the distinction between the anticipated output sequence and therefore the ground truth sequence. The seq2seq model has become in style for linguistic communication process tasks as a result of it will effectively capture

the semi-permanent dependencies between words in an exceeding sentence, permitting it to get coherent and correct translations or summaries [10].

TensorFlow provides a comprehensive set of tools for structure and planting machine literacy models. It includes a library of algorithms, pre-trained models, and tools for training and planting models. It also provides APIs for Python, C++, Java, Go, and other programming languages. TensorFlow also supports distributed computing, allowing druggies to train models on multiple machines in parallel [12]. TensorFlow is designed to be flexible and effective, allowing inventors to make and emplace machine literacy models on a variety of platforms, including desktop, mobile, and pall. It provides a range of tools and libraries for structure, training, and planting machine literacy models, including support for training on multiple GPUs and distributed systems. It also provides a range of pre-trained models and datasets for inventors to use in their operations [13].

2.3 Research Summary

Our team considered the abstractive text summarization of a Bengali when carrying out this research. We have created this model using deep learning. To use this model, we utilized a personal dataset. We have collected our data from Kaggle and the online newspapaer. The following techniques include summarizing each Bengali text. The datasets contain two attributes: one is a useful summary and the other is the description. The dataset comprises a total of 21969 data along with their related summaries. The foundation for building a deep learning model is preprocessing text. Text is split during the preprocessing stage, after which Bengali contractions are added, and stop words are removed. After preprocessing, we counted the vocabulary across the entire dataset. Deep learning models heavily rely on word embedding. In the necessary vocabulary file, W2V offers a quantitative value. W2V files which have already been trained are required for Bengali text that is now online.

Relying on the attention model, we constructed a chain-and-chain model. This model uses a bi-directional LSTM cell with an encoder and decoder. Word vectors are the encoder's input, and word vectors relating to those words are the decoder's outputs. A sign that is identified as a unique sign, like PAD, UNK, or EOS, is required to pass the sequence. More than five hours were spent training the model. The machine itself then provided us an acknowledgment that was significant.

We also classify the dataset news in the process of categorizing news article written in the Bangla language into predefined categories, such as economics, sports, entertainment, international, Science and Technology etc. This is typically done using natural language processing techniques and machine learning algorithms, which analyze the content of the news articles and assign them to appropriate categories based on keywords, topics, and themes. The goal of Bangla news classification is to make it easier for readers to find relevant information and to improve the overall user experience for Bangla-speaking audiences.

2.4 Challenges

There is no organized data in Bengali. Everyone is there, yet they are all dispersed or disorganized. The biggest obstacle for this research, however, is data collecting. The required dataset might have been used before. The newspaper dataset is one such example; however, it cannot be used in other research projects. Consequently, this study project needs a brand-new dataset. The removal of words from the text might also be a problem. English has a built-in library that can be used to eliminate stop words from texts. However, there is none for Bengali, making it a significant difficulty.

Generating the summary is another challenging task after gathering the dataset. Furthermore, dealing with the Bengali text is a constant challenge. To construct the text for the model's input during the processing stage, some raw coding is necessary. Assume that each punctuation mark requires Unicode in order to be removed from the text using raw code Training the large data is also time consuming because it takes much time and more memory space. Our dataset text size is big and sequence is also big so implementation summary is challenging and process time is huge.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

We mention that our entire study process in this part. Every research project is different from the perspective of the methods used to solve it. The methodology includes every strategy that was used in the study project. This methodology section includes a brief summary of each component as well as a discussion of applying models.

The following flow chart illustrates the entire work process.

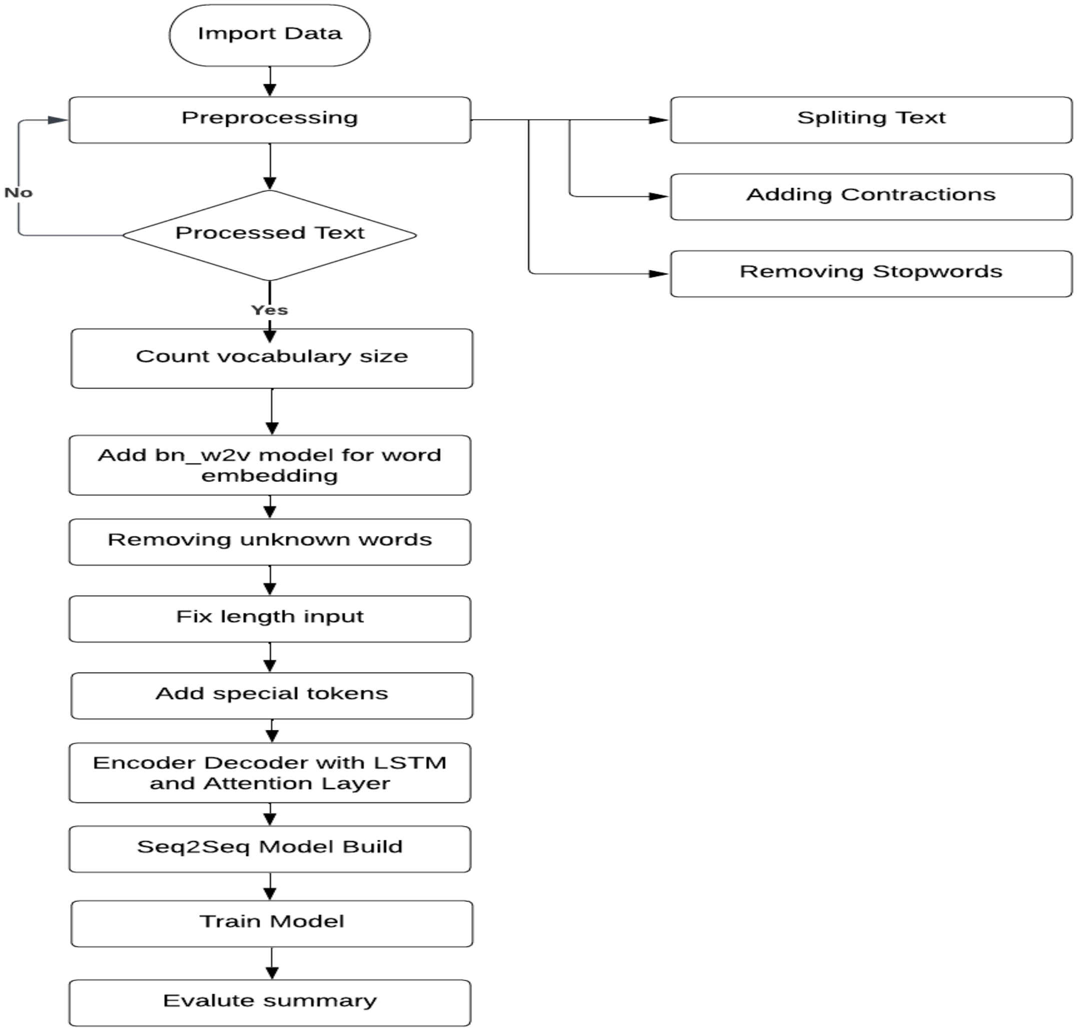


Figure 3.1.1 Text-summarization process flow diagram

In this research paper, deep learning has been applied to summarize the text in this research article. The deep learning algorithms have also been applied, in line with the type of research that they were developed for. RNN is a deep learning technique used to address text-related issues.

Each deep learning model needs a good dataset to find an entirely self-executing system. Therefore, the dataset must be ready and preprocessed before applying the method. Each component of the technique is addressed in brief. All areas will be followed once the study has been done.

Providing the nobility with a thorough description of the process is vital to improve efficiency. Understanding the work is aided by mathematical equations, a graphical representation of the model, and a narrative. For continued research, a clear justification of the technique is necessary. The entire piece appears to be a framework. Our methodology section includes a detailed discussion of all the key phases. The main section's subheadings help readers understand the model's overview and intended use.

The web is gigantic and persistently refreshed. The number of Bangla news websites has grown rapidly in the computer age, and each one has its own unique plan and framework for organizing news. The association and grouping's assortment cannot always address the issues of every client. It is a troublesome task to dispose of this inconstancy and classify the new things as per client desire. In this exposition, we give a technique for assisting clients with finding reports that are pertinent to a specific classification. We utilize naive Bayes, SVM, RF, Gradient Boosting, adaboost and KNN classifiers to classify the items in Bangla news datasets.

The following flow chart illustrates the entire work process for news classification.

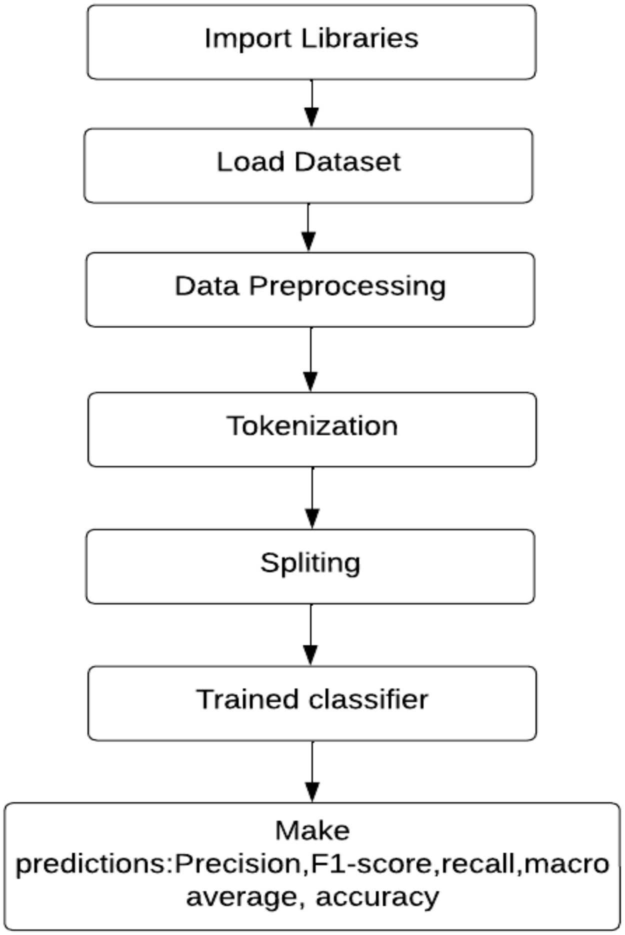


Figure 3.1.2 News classification process

3.2 Research subject and intermediary

The title of our research topic is " Bangla news classification and text summarization based on Multi-Type Text ". It is a significant field of research in Bengali NLP. This research study includes a brief explanation of the theoretical and conceptual approach used to create an abstractive text summary in Bengali. In a deep learning model, a strong PC with a GPU and other tools is important.

Below is a list of the equipment needed for this model.

Hardware and software	Development Tools
Intel Core i5-8250U CPU @ 1.60GHz 1.80 GHz	Windows 10
1 TB HDD	Python 3.7
Google Colab, Jupyter Notebook	TensorFlow Backend Engine (1.15.0)
RAM 12GB	Pandas
	NLTK

Table 1: Software and Tools

3.3 Data Collection Procedure

For bangla news classification and text summarization data collection procedure is very hard. There no such data set we would like to use that. For our work we collect our data in Kaggle and other online newspaper and then crete data set we would like.

3.4 Data fetching and data pre-processing

Data preprocessing is an effective way to clean our own dataset. Because without a good dataset we cannot achieve the desired output. Our dataset contains different kinds of news data such as sports news, economic and financial news, political news, news of different contents. Obtaining Bengali information has several drawbacks, such the way Bengali content is organized. Our dataset includes text content that has been sorted and summarized.

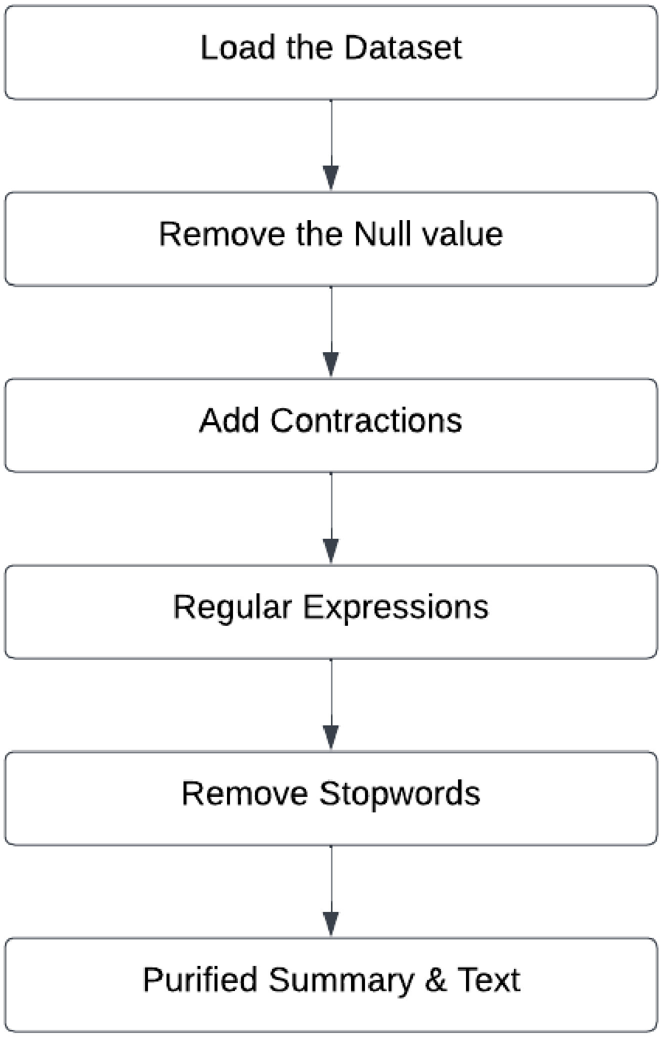


Figure 3.3.1 Dataset preprocessing

Before models are formed, processing data is a significant task. Preprocessing the data requires a variety of processes. Data preprocessing for Bangla is really challenging. The initial step in preprocessing is to eliminate extraneous words and spaces. Our shorthand for embedding it here.

Here are a few samples of Bengali contractions:

Reduced form	Extended Form
বি.দ্র	বিশেষ দ্রষ্টব্য
ড.	ডক্টর
মি.	মিস্টার
ডা.	ডাক্তার
ইঞ্জি:	ইঞ্জিনিয়ার
রেজি:	রেজিস্ট্রেশন
মু.	মুহাম্মদ
মো:	মোহাম্মদ
প্রো:	প্রোপাইটোর
হিবি.	হিসাববিজ্ঞান
১লা:	পয়লা
২রা:	দোসরা
৪ঠা:	চৌঠা
১৬ই.	ষোলোই
২১শে.	একুশে

Table 2: List of Contractions

We have deleted less important words and Bengali numbers that we don’t need. Following all the procedures, the text has been cleaned.

3.4.a Conflicting Issues

Due to our use of a huge dataset. So, there are a lot of ammount of words. We have made a large vocabulary so that every word is connected to another. A small summary is generated with long text.

3.4.b Vocabulary Counting

The model has owned a lexical set and the necessity for a vocabulary set to distinguish comparisons between the two description and the summary is a lot acceptable here. That’s why we must count the vocabulary. Considering the word count, “মেসি” has been used 809 times. The per-trained "bn w2v model" has been used as a vector file.

3.4.c Cleaned-up text & summary

Text pre-processing is must for the better accuracy. After processing the text, it looks so much clean and usable. In the pre-processing steps, we have removed the stop words, extra space, English letter, punctuation, abbreviations, Bengali digits, etc. Some text samples of pre-processing in Table 3 are listed under.

Initial Text	Clean Text
ইংল্যান্ড টেস্ট দলে ব্যাটসম্যানের সংকট নতুন নয়। ভারতের বিপক্ষে লর্ডস টেস্টে হারের পর নতুন করে এই আলোচনা সামনে আসছে। বিশেষ করে ইংল্যান্ডের দুই ওপেনার ররি বার্নস ও ডম সিবলিকে নিয়ে হচ্ছে তুমুল সমালোচনা। রান করা যেন ভুলেই গেছেন দুজন।দুই ওপেনারের টানা ব্যর্থতায় ব্যাটিং অর্ডারের একমাত্র ফর্মে থাকা ব্যাটসম্যান জো রুট আছেন চাপে। ইংল্যান্ড দলের এই ব্যাটিং সমস্যার সমাধান কী হতে পারে, সেটা জানালেন পাকিস্তানের সাবেক উইকেটকিপার-ব্যাটসম্যান কামরান আকমল। তিনি তরুণ ব্যাটসম্যান জ্যাক ক্রলিকে একাদশে দেখতে চান।নিজের ইউটিউব ..	ইংল্যান্ড টেস্ট দলে ব্যাটসম্যানের সংকট নতুন নয় ভারতের বিপক্ষে লর্ডস টেস্টে হারের পর নতুন করে এই আলোচনা সামনে আসছে বিশেষ করে ইংল্যান্ডের দুই ওপেনার ররি বার্নস ও ডম সিবলিকে নিয়ে হচ্ছে তুমুল সমালোচনা রান করা যেন ভুলেই গেছেন দুজন দুই ওপেনারের টানা ব্যর্থতায় ব্যাটিং অর্ডারের একমাত্র ফর্মে থাকা ব্যাটসম্যান জো রুট আছেন চাপে ইংল্যান্ড দলের এই ব্যাটিং সমস্যার সমাধান কী হতে পারে সেটা জানালেন পাকিস্তানের সাবেক উইকেটকিপার ব্যাটসম্যান কামরান আকমল তিনি তরুণ ব্যাটসম্যান জ্যাক ক্রলিকে একাদশে দেখতে চান নিজের ইউটিউব চ্যানেলে .

Table 3: Sample of Text Preprocessing

3.5 Arithmetical study

We have 21969 data. But in sample dataset we use 1072 data. All the data separated into two subgroups such as Description and Summary.

Description	Summary
আফগানিস্তানের ক্ষমতা চলে গেছে তালেবানদের হাতে। ক্ষমতার পালাবদলের পর আফগানদের জীবন অনিশ্চয়তায়। দেশটির বিভিন্ন খেলার কী হবে, প্রশ্ন উঠেছে সেটি নিয়েও। রশীদ খান তাঁর পরিবার নিয়ে উৎকণ্ঠা জানিয়েছেন। তালেবানদের পক্ষ থেকে অবশ্য আগেই আশ্বস্ত করা হয়েছিল, তাদের ক্ষমতা দখল ক্রিকেটের ওপর প্রভাব ফেলবে না’ বরং তারাই দেশটিতে ক্রিকেট এনেছে মনে করিয়ে বলা হয়েছিল, বরং উন্নতিই হবে ক্রিকেটের।আপাতত দেশটির ক্রিকেটের ওপর তেমন প্রভাব পড়েনি সরাসরি। নতুন ব্যাটিং কোচ হিসেবে সাবেক শ্রীলঙ্কা ওপেনার আভিস্কা গুণাবর্ধনেকে ব্যাটিং কোচ হিসেবে নিয়োগও দিয়েছে আফগানিস্তান ক্রিকেট বোর্ড। আজ কাবুলে শুরু হয়েছে জাতীয় দলের অনুশীলনও।টুইটারে গুনাবর্ধনের নিয়োগের ঘোষণা আনুষ্ঠানিকভাবে দেওয়ার পর অবশ্য পরে মুছে ফেলা হয়েছে টুইটটা। তবে ক্রিকবাজ বলছে, পাকিস্তান সিরিজের জন্য নিয়োগ দেওয়া হয়েছে গুনাবর্ধনেকে।	আফগানিস্তানের ক্ষমতা চলে গেছে তালেবানদের হাতে ক্ষমতার পালাবদলের পর আফগানদের জীবন অনিশ্চয়তায় দেশটির বিভিন্ন খেলার কী হবে প্রশ্ন উঠেছে সেটি নিয়েও রশীদ খান তাঁর পরিবার নিয়ে উৎকণ্ঠা জানিয়েছেন

Table 4: Dataset Sample

1) vocabulary size- 30,213k.
2) Unique words are 10,321k.
3) 497405-word embedding.
4) 87% of the word used for model.
5) Text All-out length is 800 words & total summaries of the length is 250.
6) Total number of words in headlines are 427402
7) Total number of UNKs in headlines are 13020
8) Percent of words that are UNK 3.05%

3.6 Executional requirements

3.6.a. Problem discussion

The dataset's rundown and depiction inputs are both impartial. Commonly, a result or outline is more limited than a portrayal (text). Consider W the exact word count of the dataset's feedback chain depiction (text). The word size should be V, and the info chain begins at 1..2..3. On the other side, the result chains like y1, y2,... and y, where S>W, are created by the information chain. It suggests that the rundown's chain isn't significantly longer than a text report. Each connection in the chain is created by words.

3.6.b. Vocabulary and Word embedding

The importance of words isn't just relying on frequency but also rely on word likeness. Therefore, we have to calculate the entire vocabulary from filtered description (text) and summary. Following the vocabulary count, we look up the word circumstance. like as, we have sampled word used “মেসি” and the incident was 809. One significant and usable pre-trained model was discovered which is used for our research purposes for the improvement of our model, we used a pre-trained Bengali W2V model which was downloaded from the internet. There are various Word embedding pre-trained models in other languages, but the Bengali language only has a small number of word embedding files, most of which are inappropriate for research. We set up the vocabulary size and calculate those words which happened 25 or more times in our dataset. The model was studied from scratch rather than applying a pre-trained word embedding model.

3.6.c. RNN Encoder & Decoder

RNN Encoder and Decoder are the combination of two Recurrent Neural Network which does the task as an encoder and decoder pair. In the encoder phase, it produces the input sequence and generating the encoder state. By using the encoder state summarize the information into the input sequence. Decoder generates the output sequences by using encoder states.

Encoder

The encoder is given as contributing one course book of a news structure at a time. Each word's initial goes through an implanting level that changes the word into an ordered portrayal. This characterized portrayal is additionally coordinated by utilizing a multi-facet brain network with the resigned layers created after entering the previous word, or all 0's for the main word in the reading material. Despite the fact that GRU enjoys a benefit in preparation time, our model uses LSTM rather than GRU because LSTM is easier to tune boundaries and has a more grounded hypothetical assurance. h.

Decoder

Based on the grouping to-arrangement worldview, our concept divides the decoder segment into two distinct modes: replication mode and creation mode. The words in our feedback arrangement have clear definitions on the off chance that the decoder expected word grouping, with "I" representing the assortment succession, "H" for the concealed state, and "V" for the information vector. The encased states are assessed using this recipe.

$$hi = (()) \mp ())\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The encoder input was entered as=…., together into fixed c. The RNN is changed by every period of time t.

$$ht = f(ht - L)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

And

$$\mathrm{C}= q(\{hl\ldots, htx\})\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where f and q are non - linear parts, while c is the unseen component. Specifically, by using the X sequence, we may determine the probabilistic translations for the decoder.

$$() = \prod = \mathrm{l}(\{\mathrm{l}, \ldots., -\mathrm{l}\})\ldots\ldots\ldots\ldots (4)$$

Where = (, … . ,). Conditional statement, e.g. = (, … . , | , … . ,). = (− , −, c)………………………………………………(5)

for conditional probability is as

$$p(|\{\,,\ldots.,-\},C) = g(-))\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Here context vector.

Then calculated as a weighted sum= $\sum\ = 0$ ……………………………………(7)

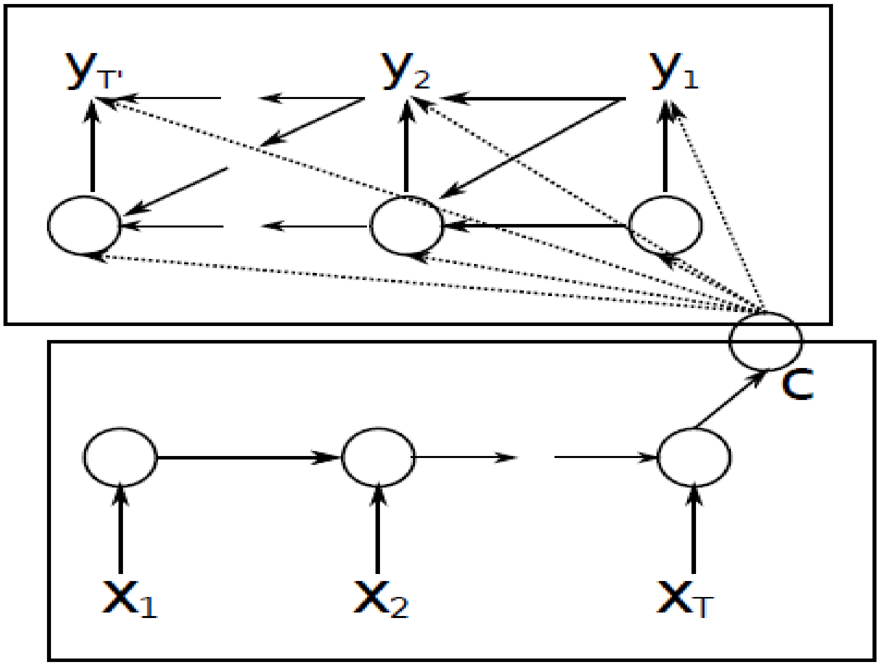
Suppose, input sequence (h), also (h1 h) is the hidden state. The hidden state (h h1) thus,

h=[h;h]……………………..……………… (8)

Here, h=generated summary = (-1,h)……………………………………………(9)

Now RNN encoder-decoder assumed lower,

Decoder



Encoder

Figure 3.5.1 RNN encoder and decoder

3.6.d. Sequence to Sequence Learning

For a few regular language handling (NLP) applications, like title age for machine interpretation, text rundown, and sound acknowledgment, seq2seq models have basically been carried out. A model of grouping-to-succession is created. In such an idea, a bi-directional LSTM cell is utilized along with an encoder and a decoder. The most reasonable and notable RNN strategy in profound learning It settles text-related issues all the more effectively. Each LSTM cell that makes up a RNN LSTM cells function similarly to transient memory. In LSTM cells, encoders and decoders are utilized. The important text grouping is utilized to create the result after Encode has passed the information texts. A token is used to verify the beginning and end of each succession.

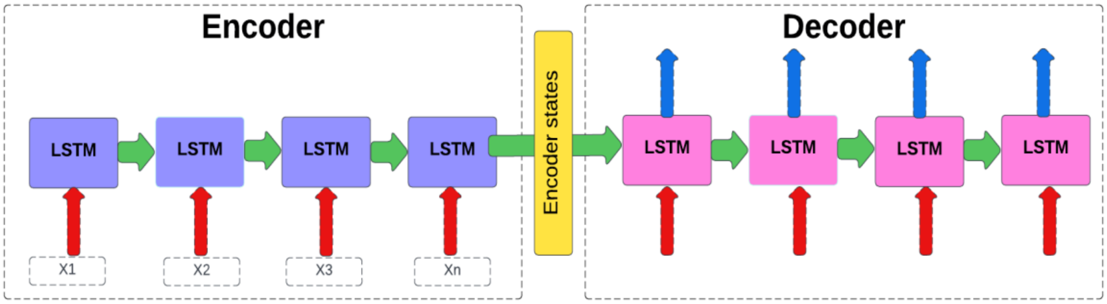


Figure 2: Sequence to Sequence model

In this paper, we have utilized a few successions like <PAD>, <EOS>, <GO>, <UNK>, and so on. These kinds of unique tokens are used to work in dealing with the arrangement in the encoder and decoder. <GO> tokens start the course of result succession in the decoder. In the information preprocessing stage, we have added <UNK> and supplanted the jargon. In this way, we utilized the exceptional token <UNK> which implies an obscure token. At the point when an obscure token is set up in the succession, it will be added to the <UNK> token in the text. Before our preparation information, we selected <GO> and <EOS> in information that contains words that would be utilized for arrangement interpretation.

3.6.e BERT

BERT is an assortment of pre-prepared transformer encoders that might be joined to makea contextualized language model that can be applied to an assortment of downstreamapplications. Since BERT was prepared as a veiled language model, we used it toproduce a portrayal for each sentence in our strategy. We then changed the information. arrangement and embeddings of BERT to empower synopsis In spite of the fact that BERT draws onThe transformer design and its objectives are one of a kind and require special preparation. It haphazardly covers 12%.to 17% of the preparation phrases in one phase. The other stage takes an info sentence and an up-and-comer sentence, endeavoring to foresee ifThe applicant sentence properly follows the information sentence. Indeed, even with a sizablenumber of GPUs, the preparation time frame for this activity required a few days. Accordingly, Google made accessible for circulation two BERT models, one of which had 110 millionfactors, and the other had 340 million.

We have chosen various layers for embeddings with the default pre-trained BERT model. The N x E matrix required for clustering is produced by using the [cls] layer of BERT, where N is the number of sentences and E is the embeddings dimension. However, the output of the [cls] layer does not always result in the best embedding representation for sentences. The basic BERT implementation for the newspaper summarization service makes use of the "huggingface" company's pytorch-pretrained-BERT library.

3.6.f GPT-2

The goal of text summarizing is to produce a clear, succinct summary while preserving the core information and message [36]. The most important information in a document must be included in a high-quality summary, which should also be shorter than the original. Although many different strategies have been put out for the automatic summary generation task [37], the NLP community still finds it to be a difficult task [38]. GPT-2 has achieved amazing performances and has a remarkable capacity for language representation. Based on its exceptional capacity for language representation, GPT-2 has achieved amazing performances. The OpenAI GPT-2 model just uses the transformer's decoder portion. GPT-2 creates one token at each time step using a beginning token that has been predetermined as input. The created token is increased with each token that is generated. When a text length criterion specified by pre-defined the model's generation procedure is finished. As indicated, a token [CLS] is added at the start of each article and a token [SEP] is placed at the conclusion of each article and summary. More specifically, [CLS] is utilized as a symbol to combine features from a single article, and [SEP] is used in our model to represent the end of an article or a summary.

3.6.g XLNET

Text summarization demands a group ensures of the words, sentences, and passages throughout the entire article than other downstream NLP tasks. In this paper, we present a model under the generic sequence to sequence pattern to utilize the full information of the entire articles. with a layered Transformer decoder and a recurrent XLNet encoder Evey part could acquire more contextual information than information that is only contained within the current sequence because the entire article is encoded repeatedly until the last part of the article. This produces both the representation and the memory, the content of which is information from the previous time step. To achieve this, we first divide the entire article into a number of identically sized pieces in accordance with the maximum length. The GPU's capabilities, particularly its memory, place a limit on the duration.

We fully utilized XLNet's memory mechanism during the training phase, feeding the segments into the model one at a time to create a hidden state and memory. Here are the model's settings: label smoothing with a factor of 0.015 and dropout with a rate of 0.01 are both employed. There are 12 heads, 768 hidden units, and feedforward layer in each decoder. Beam search with size 5 is employed for producing. We did not take any action to prevent the repetition and lack of vocabulary problems.

3.6.h Sentence Classifier

Naive Bayes

D. Lewis initially suggested and used the Guileless Bayes (NB) probabilistic characterization technique for the gig of text ordering [7]. In a probabilistic system, it is established by the Bayes hypothesis. To gauge the likelihood of classifications given a text, the principal idea is to use the consolidated probabilities of words and classes. The model's mistaken assumption of the word freedom makes it such. Since the Credulous Bayes classifier doesn't utilize word mixes as indicators, its computation is undeniably more proficient than the remarkable intricacy of Non-Guiltless Bayes procedures because of word freedom's effortlessness. When we use the Credulous Bayes classifier in the Text Arrangement issue, we use the following condition:

$$p(class \mid documentation) \; = \; p(class).p \; (document \mid class) \; p(doc.) \ldots\ldots\ldots (1)$$

where *p(class | document)* is the probability that a given report D belongs to a given document C. P(doc.) is the likelihood of a report, we can see that p(doc.) is a consistency divider to each computation, so we can overlook it. P (class) is the likelihood of a class (or classification). We can process it based on the number of archives in the classification separated by report numbers across all classifications. *p(doc. | class)* addresses the likelihood of a record given class, and archives can be represented as sets of words, so *p(doc. | class)* can be written as: p(doc. | class) = Y I p(wordi | class) (2)

$$p(doc. \mid class) \; = \; Y \; I \; p(wordi \mid class) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Thus, we can modify equation (2) as follows:

$$p(class|document) = p(class).(wordi|class) \; Y \; I \; p$$

Support Vector Machine

Joachims utilized the help vector machine, an interesting characterization technique, to sort texts. In light of the most reduced primary gamble idea, this is a strong and managed learning test. This method fosters a hyperplane during the preparation phase to recognize positive and negative information. Then, at that point, it orders new examples by directing where everyone should be situated on the hyperplane. SVC is used for the arrangement work in this review. Furthermore, we chose straight as the piece type because we expected to move our preparation tests directly to the higher layered space. The main part of prior research that utilized SVM to group text utilized all of the text's terms without considering their appropriate pertinence. Then again, a few explorations on the selection of catchphrases for text classification have been completed. As was recently demonstrated, this exploration often focused on the model for watchword determination, utilizing set- or catchphrase-based procedures. These methodologies additionally disapprove of the time-intricacy of examination and the shortfall of standard datasets. Also, the SVM is the most widely recognized grouping technique utilized in research [11].

Random Forest

As they collaborate, this technique generates an infinite number of choice trees. The foundations of this technique are choice trees. The pre-handling stage characterizes the hubs of the choice trees that make up irregular backwoods [7]. The best component is the haphazardly picked subset of highlights after many trees have been developed [8], [22]. Another idea that is made using the choice tree technique is to deliver a choice tree. These trees make up the irregular woodland, which is utilized to sort new articles from an information vector. Every choice tree assembled is utilized for grouping. Assume we give the tree votes to that class, and then the irregular woodland picks the grouping that has the best number of votes among every one of the trees in the backwoods.

Ventures for arbitrary timberland calculation:

Stage 1: From the preparation information, pick K arbitrary variables of interest.

Stage 2: Using these K pieces of information, construct a decision tree.

Stage 3: Before rehashing Stages 1 and 2, pick the number NTree of trees you need to build.

Stage 4: Anticipate the worth of y by making every last one of the NTree trees for another data point of interest and relegate another information point normal across totally anticipated y values.

KNN (K Nearest Neighbor)

The primary objective of this technique is to keep tantamount articles near each other (see Fig. 3). An informational collection's element vectors and class names are utilized by this model [9]. KNN keeps a data set of all occasions and uses a closeness metric to classify new models. To address text in K-closest neighbors, a spatial vector with the documentation S = S (T1, W1; T2, W1;... Tn, Wn) is used. Using the preparation texts, which are not completely settled for each text, the texts with the highest comparability are chosen. The classes are then settled using K neighbors. Select neighbor number K at the start. Consider the K-closest neighbors of the new information point in light of the Euclidean distance. Play out a count of the quantity of interesting data among the K nearest neighbors for every class. In the aftermath of counting, relegate the new information point where we counted the most neighbors.

Gradient Boosting

A troupe AI approach called "inclination-based help" is utilized to sort news. Different models are prepared to utilize a slope classifier in an additive, moderate, and consecutive way. Generally, support is a technique that transforms poor students into good students by reducing errors (most commonly, Choice trees). Solid students are made by consolidating a lot of frail students. Angle support will likewise improve the model's presentation with higher accuracy.

AdaBoost Algorithm

The "helping" group of calculations incorporates "adaptive boosting" [1]. This sort of student zeros in on additional, mistakenly grouped examples during preparation, changes the example conveyance, and rehashes this cycle until the feeble classifier has gone through a specific number of preparation stages, so all in all, learning is finished. Finish the circle, then. AdaBoost has been altered to such an extent that all preparation tests start with a similar weight. The following classifier will be prepared to utilize information from the earlier classifier's mix-up, improving the probability that the following powerless classifier will pick the right blunder test and diminishing the probability that it will pick a matched example. AdaBoost can focus on the examples that have been erroneously classified thanks to this strategy. Try not to stress over overfitting for AdaBoost. However, the AdaBoost approach is defenseless against peculiar and loud information. This suggests that assuming there is more commotion in the information, the AdaBoost calculation will invest more energy handling it, diminishing its effectiveness.

CHAPTER 4

EXPERIMENTAL OUTCOME OVERVIEW

4.1 Introduction

The terrible issue in the field of NLP is the shortening of abstract documents. People find it considerably more challenging to pull a borough text and summary from within themselves. Since the machine does its best to produce according to its capabilities. The computer needs to be trained to learn using the data model after pre-processing. The model contains a backend engine for each training. We used TensorFlow to complete the task in this test. Initial values are almost complete. Examples include epoch, keep probability, run size, batch size, learning rate, number of layers, etc. The amount of time needed to train data has been decreased. The "Adam" in this case for model adjustment is the optimizer. Making higher configuration PC data training easier is necessary. Finally, we train the model via Google Colab. It works considerably more quickly and cuts down on time.

Value of parameter is-

Parameter	Value
Number of Epoch	100
Batch size	64
Learning rate	0.008
Keep probability	0.75
Run size	256
Number of layers	3

Table 5: the parameter's value

4.2 Implementational results

We all know very well that it is not possible to give the 100% for a model. But Almost the real output is produced by the machine. Subsequent to making the model capacity we have to prepare our model. We fit the model with the present and next words. We used 80 percent data for the training and 20 percent data for the test. Hence, we have 21,969 but we import 1072 data in google notebook. So, 933 data used for training and another 140 data used for the teste have used TensorFlow sequence to sequence model. When we stop training we'll be able to create a machine's own summary. We take epoch=100, batch size=64, rnn's size=256, learning rate=0.008, keep probability=0.75 and we used Adam Optimizer, which calculated the learning rate of each parameter. For faster converges use a vanilla gradient descent optimizer. Train model right around 12 hours gives better accuracy with a loss of 0.031. Our model accuracy is 97.69%. A file called "model.ckpt" has been saved in the model so that the output can be checked. We then created a TensorFlow session in order to reload the graph that was previously saved. After that the description (text) and summarized data frame were then defined at random to ensure that the summary was accurate.

Original Description:	তখন মাত্রই পৃথ্বী শর তাণ্ডব থেকে রক্ষা পেয়েছে শ্রীলঙ্কা। ওপেন করতে নেমে ৯টা চার মেরে মাত্র ২৪ বলে ৪৩ করে দলের রান পাঁচ ওভারেই ৫৭-তে নিয়ে গিয়েছিলেন তরুণ এই ওপেনার। আভিষ্কা ফার্নান্দোর হাতে ক্যাচ বানিয়ে ধনঞ্জয় ডি সিলভা যখন শকে ড্রেসিংরুমে ফেরালেন, হয়তো একটু হাঁপ ছেড়ে বেঁচেছিল শ্রীলঙ্কা। হয়তো ভেবেছিল, কিছুক্ষণের জন্য তাণ্ডবের হাত থেকে রক্ষা পাওয়া গেল।অমনটা যদি ভেবেও থাকেন, কী ভুলই না ভেবেছিলেন ডি সিলভা! তিন নম্বরে নেমেছিলেন ইশান কিষান, ভারতের জার্সিতে যাঁর ওয়ানডে অভিষেক হয়েছে গতকাল। কিন্তু ডি সিলভার ভাবনায় ছেদ ঘটার কারণ ঘটল পরের বলেই! কিষানের খেলা প্রথম বল ছিল সেটি। ডাউন দ্য উইকেটে এসে লং অন বাউন্ডারির ওপর দিয়ে ডি সিলভাকে বিশাল এক ছক্কা মেরে ওয়ানডেতে নিজের আবির্ভাবের কথা জানান দিয়েছেন ভারতীয় ব্যাটসম্যান।এখন জানা গেল, ওয়ানডেতে নিজের প্রথম বলে ছক্কা মারার এই ঘটনাটা কাকতালীয় নয়। ....

Original summary:	শুধু প্রথম বলে ছক্কা মারাই নয়, ১৬তম ভারতীয় ক্রিকেটার হিসেবে ওয়ানডে অভিষেকে হাফসেঞ্চুরি করার কীর্তিও গড়েছেন ইশান কিষান। গতকাল আবার কিষানের জন্মদিনও ছিল। এক দিনে কিষানের প্রাপ্তির যেন কোনো শেষ নেই!
Input text (Description):	তখন মাত্রই পৃথ্বী শর তাণ্ডব থেকে রক্ষা পেয়েছে শ্রীলঙ্কা ওপেন করতে নেমে টা চার মেরে মাত্র বলে করে দলের রান পাঁচ ওভারেই তে নিয়ে গিয়েছিলেন তরুণ এই ওপেনার <UNK> ফার্নান্দোর হাতে ক্যাচ বানিয়ে ধনঞ্জয় ডি সিলভা যখন শকে ড্রেসিংরুমে ফেরালেন হয়তো একটু হাঁপ ছেড়ে বেঁচেছিল শ্রীলঙ্কা হয়তো ভেবেছিল কিছুক্ষণের জন্য তাণ্ডবের হাত থেকে রক্ষা পাওয়া গেল অমনটা যদি ভেবেও থাকেন কী ভুলই না ভেবেছিলেন ডি সিলভা তিন নম্বরে নেমেছিলেন কিষান ভারতের জার্সিতে যাঁর ওয়ানডে অভিষেক হয়েছে গতকাল কিন্তু ডি সিলভার ভাবনায় ছেদ ঘটার কারণ ঘটল পরের বলেই কিষানের খেলা প্রথম বল ছিল সেটি ডাউন দ্য উইকেটে এসে লং অন বাউন্ডারির ওপর দিয়ে ডি সিলভাকে বিশাল এক ছক্কা মেরে ওয়ানডেতে নিজের আবির্ভাবের কথা জানান দিয়েছেন ভারতীয় ব্যাটসম্যান এখন জানা গেল ওয়ানডেতে নিজের প্রথম বলে ছক্কা মারার এই ঘটনাটা কাকতালীয় নয় ব্যাট করতে নামার সময় সতীর্থদের বলেকয়ে এসেই আন্তর্জাতিক ওয়ানডেতে নিজের প্রথম বলে ছক্কা হাঁকিয়েছেন কিষান সনিকে দেওয়া সাক্ষাৎকারে ইশান কিষান নিজেই এই <UNK> ফাঁস করেছেন নিজের জন্মদিন উপলক্ষে নিজেকে এভাবেই উপহার দেওয়ার চিন্তা করেছিলেন সদ্যই তেইশে পা দেওয়া এই ব্যাটসম্যান ক্রিজে নামার আগেই সবাইকে বলে এসেছিলাম আমি প্রথম বলে ছক্কা মারব বল যেখানেই পিচ করুক না কেন …………..
Response summary:	শুধু প্রথম বলে ছক্কা মারাই নয় তম ভারতীয় ক্রিকেটার হিসেবে ওয়ানডে অভিষেকে হাফসেঞ্চুরি করার কীর্তিও গড়েছেন ইশান কিষান গতকাল আবার কিষানের জন্মদিনও ছিল এক দিনে কিষানের প্রাপ্তির যেন কোনো শেষ নেই মাত্র শেষ ইতিহাসে এক রান

Table 6: Sample of the response summary

Class	Precision	Recall	F1-score
Economics	1.00	1.00	1.00
Entertainment	1.00	1.00	1.00
International	0.94	1.00	0.97
Science and Technology	1.00	0.75	0.86
Sports	0.90	1.00	0.95
Accuracy			0.96

Table 7: Naive Bayes Classifier

In our datasets we have applied Multinomial Naive Bayes. In the wake of carrying out this calculation, we catagorized the outcomes into 4 parameters. It is seen from Table that a precision of 0.96 is acquired. For example, the precision of economics class acquired is 1.0, recall is 1.0, F1-score is 1.0. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 1.0, 0.95 F1-score. In international, a precision of 0.94 is acquired, recall 1.0, 0.97 F1-score are acquired. For sports class, precision acquired is 0.90, recall 1.0, F1-score of 0.95. For science and technology class, a precision of 1.0 is acquired, recall 0.75, F1-score 0.86 is acquired.

Class	Precision	Recall	F1-score
Economics	1.00	1.00	1.00
Entertainment	1.00	0.82	0.90
International	1.00	0.81	0.90
Science and Technology	0.58	0.88	0.70
Sports	0.90	1.00	0.95
Accuracy			0.96

Table 8: Random Forest Classifier

The accuracy of RF classifier is 0.96 is acquired. For example, the precision of economics class acquired is 1.0, recall is 1.0, F1-score is 1.0. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 0.82, 0.90 F1-score. In international, a precision of 1.0 is acquired, recall 0.81, 0.90 F1-score is acquired. For sports class, precision acquired is 0.90, recall 1.0, F1-score of 0.95. For science and technology class, a precision of 1.0 is acquired, recall 0.75, F1-score 0.86 is acquired.

Class	Precision	Recall	F1-score
Economics	1.00	1.00	1.00
Entertainment	1.00	0.91	0.95
International	0.89	1.00	0.94
Science and Technology	1.00	0.77	0.86
Sports	0.90	1.00	0.95
Accuracy			0.94

Table 9: SVM Classifier

The accuracy of SVM classifier is 0.94 is acquired. For example, the precision of economics class acquired is 1.0, recall is 1.0, F1-score is 1.0. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 0.91, 0.95 F1-score. In international, a precision of 0.89 is acquired, recall 1.0, 0.94 F1-score is acquired. For sports class, precision acquired is 0.90, recall 1.0, F1-score of 0.95. For science and technology class, a precision of 1.0 is acquired, recall 0.77, F1-score 0.86 is acquired.

Class	Precision	Recall	F1-score
Economics	1.00	1.00	1.00
Entertainment	1.00	1.00	1.00
International	1.00	0.94	0.97
Science and Technology	0.88	0.88	0.88
Sports	1.00	0.95	0.90
Accuracy			0.96

Table 10: Gradient Bosting Classifier

The accuracy of Gradient Bosting classifier is 0.96 is acquired. For example, the precision of economics class acquired is 1.0, recall is 1.0, F1-score is 1.0. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 1.0, 1.0 F1-score. In international, a precision of 1.0 is acquired, recall 0.94, 0.97 F1-score is acquired. For sports class, precision acquired is 1.0, recall 0.95, F1-score of 0.9. For science and technology class, a precision of 0.88 is acquired, recall 0.88, F1-score 0.88 is acquired.

Class	Precision	Recall	F1-score
Economics	0.18	0.67	0.29
Entertainment	1.00	0.64	0.78
International	0.67	0.25	0.36
Science and Technology	0.10	0.10	0.1
Sports	0.90	1.00	0.95
Accuracy			0.48

Table 10: Adaboost Classifier

The accuracy of Adaboost classifier is 0.48 is acquired. For example, the precision of economics class acquired is 0.18, recall is 0.67, F1-score is 0.29. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 0.64, 0.78 F1-score. In international, a precision of 0.67 is acquired, recall 0.25, 0.36 F1-score is acquired. For sports class, precision acquired is 0.90, recall 1.0, F1-score of 0.95. For science and technology class, a precision of 0.1 is acquired, recall 0.1, F1-score 0.1 is acquired.

Class	Precision	Recall	F1-score
Economics	0.18	0.67	0.29
Entertainment	1.00	0.64	0.78
International	0.67	0.25	0.36
Science and Technology	0.1	0.1	0.10
Sports	0.90	1.00	0.95
Accuracy			0.48

Table 12: KNN Classifier

The accuracy of KNN classifier is 0.48 is acquired. For example, the precision of economics class acquired is 0.18, recall is 0.67, F1-score is 0.29. Likewise, for entertainment, the values obtained are 1.0 precision, recall is 0.64, 0.78 F1-score. In international, a precision of 0.67 is acquired, recall 0.25, 0.36 F1-score is acquired. For sports class, precision acquired is 0.90, recall 1.0, F1-score of 0.95. For science and technology class, a precision of 0.1 is acquired, recall 0.1, F1-score 0.1 is acquired.

Algorithms	Accuracy
Gradient Boosting	0.96
Adaboost	0.476
Random Forest	0.94
SVM	0.88
Naive Bayes	0.48
KNN	0.96

Table 13: Comparison Accuracy of Classifier

After analysis all the algorithms, we have come to know that which algorithms is best for our dataset. We have applied six different algorithms as like Naive Bayes, SVM, Random Forest, Gradient boosting etc. After comparing among those algorithms, we got that Gradient boosting and Naive Bayes delivered best results and the accuracy according to the table is 0.96. Random forest and SVC are comparatively better because their accuracy is 0.88 and 0.94 which is less then Gradient and NB algorithms. KNN and AdaBoost made us disappointed hence those results are below acceptable. By implementing them, we got the accuracy around 0.48. So Gradient Boosting and Naive Bayes is best for our Bangla news classification.

4.3 Descriptive Analysis

For text Summarization, we have built a model for Bengali. Our model produces a better output for various scenario. We have created so that we may deduct the function loss. Learning model is used for reducing the errors. Loss function deduction is very much important for chain data. During the data training, we added the loss function. When the training is ended up we counted the loss. Earlier during the training, our model produces a hing loss. After some epoch, losses are decreased at alarming rate. Learning value which is 0. 008. Generally, we have divided our data into training and test set. That's why we have got 933 for training and 140 data for testing.

4.4 Moral

Throughout this chapter we have talked about how we built our model, how it generates the Summarization and the desired outputs. We have tried to explain in details.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

Nowadays, everything is turned into modern and smart. Where time is also a crucial element. Depending on technology our life becomes faster. In our daily life, general people are fond of reading news papers, News articles, short stories, online articles, etc. When they get enough time, they usually spend their time reading those articles or books. But all the time they don't have enough time to read the full article or news or stories. Then they wish if they got a short summary about that topic. This text summarization will help them to ease their reading and they get the short summarization instantly which is amazing to them. All classes of people get the benefits of this work. Such as businessmen, students, news readers, journalists etc.

5.2 Impact on Student's life

A student always is searching for knowledge to know new things. That's why they have spent most of their time reading new things. They read their daily readings and they also read many novels, articles, Story books, Newspaper. Sometimes they try to get the main theme within a short time. Suppose, a student has to go outside or go for an exam or any emergency work but he or she has not had much time to read the full passage or article or any text. In that case, text summarization helps him/her in an efficient way with a meaningful short summary. As they get the desired summary within a short time, they have benefited in many ways. Sometimes while reading the online article or any text, they have felt bored and anxious. In Facebook, sometimes we have seen some large articles on important matters but due to a short time we avoided it. Now students don't need to unread anything as it is large. They can easily summarize it and get their results.

5.3 Ethical issue

The collection and use of data raise numerous ethical and moral concerns about its use, sharing, and even protection. It also raises issues of privacy ethics, particularly the appropriateness of publicly sharing some information on these sites. In this research paper, we will look at the moral, social, and ethical issues that arise from the sharing and use of this data. Additionally, it raises concerns about privacy ethics, particularly the propriety of disclosing some information publicly on these sites. In this research paper, we will look at the moral, social, and ethical issues raised by the use and sharing of this data. People's privacy was compromised as a result of the unethical use of text, which also has an impact on information security and physical security.

CHAPTER 6

SUMMARY AND FUTURE PLAN

6.1 Summary of the Study

This project basically based on Natural language processing. In our project, we have made a model using deep learning for Bangla abstractive text Summarization. This model has given tremendous output for the text summarization. From the beginning to end, we have tried our heart and soul to complete this project and it took 7 months. We have done our project by several steps and options. The project's complete overview and step-by-step instructions are provided below.

No	Steps
1.	Collecting data from kaggle, online platform and Newspaper
2.	Marge all collected data into one
3.	Collect the Word2Vec
4.	Preprocessing the data
5.	Count Vocabulary size
6.	Loaded the pre-trained word2vec
7.	Adding special token
8.	Encoder and Decoder with LSTM
9.	Created the Sequence to Sequence model
10.	Trained the model
11.	Analyzing result based on the machine response
12.	Getting the summarization

Table 14: Project overview

Our suggested approach may aid future study in our NLP field by compressing Bangla language and shortening long sentences utilizing particular models for abstract lessons.

6.2 Conclusion

In recent years we get a large amount of data, and we have to understand the data quickly. That's why we need text summarization to summarize the large amount of data to understand in a short time. On text summarization we can use abstractive and extractive text summarization. Abstractive summarization can produce a more coherent summary, but it is more difficult to achieve due to the complexity of natural language processing and the need to understand the meaning of the text. For better understanding, the summarization system underscores the needs of the developers designing. The Extractive text summarization is a lot of correct to handle massive amounts of knowledge and provide an associate degree correct outline. And it's done by looking for the foremost used words or common words and so shrinking them and presenting newly designed sentences. On the other hand, Abstractive text summarization clarifies the data and improves them, and makes a sentence. Extractive text summarization involves selecting and combining important sentences or phrases from the original text to create a summary, while abstractive text summarization involves generating new sentences and phrases that capture the meaning of the original text.

With the help of LSTM & Sequence to Sequence algorithm, we present to you a summarization that could save time & easily understand the material or news. Through Natural Language Processing (NLP) we can understand the document to generate the summary. Text summarization is like a process where a large amount of text or data has been submitted and with the help of Machine learning (ML) it compresses the data and presents us with an understandable short form of the data which is called text summarization. Text summarization is not new in Machine learning (ML). In the English language, it is used vastly. But on the other hand, in the Bangla language, it is very difficult to summarize data or form. The main reason for this kind of issue is that the Bangla language has a large number of alphabets and grammatical rules.

6.3 Further study

The model has a couple of restrictions as well. Since each study project changes persistently, we have a philosophy for how we will change from now into the foreseeable future. By using this procedure, we had the option of additionally fostering precision while likewise becoming more coordinated. We would like to add more models to improve the idea of the outcome and make this task more obvious. Though the amount of data in our assessment is limited, we really want to investigate in the future with boundless data in light of the fact that the model is good for dealing with a great deal of data and giving it exact once-overs. After we finish our survey, we really want to create a web-based, convenient application that uses man-made thinking. Moreover, this application will give Bengali text summaries. The application would likewise have the option to act as an extra layer of safety by utilizing regular language handling (NLP) methods to identify possible vindictive movements or content in the Bengali text that is inputted. As a feature of the examination, we will utilize directed and unaided learning models to create and assess our application's exhibition. This exploration will assess the presentation of both regulated and solo AI models, assisting us in understanding which one is best for our Bengali message outlines.

APPENDIX

We had to deal with a lot of challenges in order to finish the project, the first of which was deciding on our methodological strategy. Since there had not been much work done recently on this range, it was not standard work but more an investigation regarding the based venture. So, it's possible that we won't receive many offers of help from anywhere. Another problem was that gathering information was really difficult for us. Datasets for detecting tomato leaves were few. With some self-management, we were able to solve the problem.

REFERENCES

[1] 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.

[2] Rafael Ferreira et al. “Assessing Sentence Scoring Techniques for Extractive Text Summarization”, Elsevier Ltd., Expert Systems with Applications 40 (2013) 5755-5764.

[3] Nenkova, Ani, and Kathleen McKeown. "A survey of text summarization techniques." In Mining text data, pp. 43-76. Springer, Boston, MA, 2012.

[4] Shi, Tian, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. "Neural abstractive text summarization with sequence-to-sequence models." ACM Transactions on Data Science 2, no. 1 (2021): 1-37.

[5] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444.

[6] Yousefi-Azar, Mahmood, and Len Hamey. "Text summarization using unsupervised deep learning." Expert Systems with Applications 68 (2017): 93-105.

[7] Abujar, S., Masum, A.K.M., Chowdhury, S.M.H., Hasan, M. and Hossain, S.A., 2019, July. Bengali text generation using bi-directional RNN. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[8] Song, Shengli, Haitao Huang, and Tongxiao Ruan. "Abstractive text summarization using LSTM-CNN based deep learning." Multimedia Tools and Applications 78, no. 1 (2019): 857-875.

[9] Tomer, Minakshi, and Manoj Kumar. "Improving text summarization using ensembled approach based on fuzzy with LSTM." Arabian Journal for Science and Engineering 45, no. 12 (2020): 10743-10754.

[10] Masum, A.K.M., Abujar, S., Talukder, M.A.I., Rabby, A.S.A. and Hossain, S.A., 2019, July. Abstractive method of text summarization with sequence to sequence RNNs. In 2019 10th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-5). IEEE.

[11] Hanunggul, Puruso Muhammad, and Suyanto Suyanto. "The impact of local attention in lstm for abstractive text summarization." In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 54-57. IEEE, 2019.

[12] Nallapati, R., Xiang, B. and Zhou, B., 2016. Sequence-to-sequence rnns for text summarization.

[13] Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W. and Du, Q., 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *arXiv preprint arXiv:1805.03616*.

[14] Abadi, M., 2016, September. TensorFlow: learning functions at scale. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming* (pp. 1-1).

[15] Dillon, Joshua V., Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. "Tensorflow distributions." arXiv preprint arXiv:1711.10604 (2017).

[16] 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*.

[17] Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

[18] Lake, B.M., 2019. Compositional generalization through meta sequence-to-sequence learning. *Advances in neural information processing systems*, *32*.

[19] Luong, M.T., Pham, H. and Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[20] Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N., 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

[21] Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O. and Zaremba, W., 2014. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

[22] Kalchbrenner, N. and Blunsom, P., 2013, October. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1700-1709).

[23] Sennrich, R., Haddow, B. and Birch, A., 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[24] Shang, L., Lu, Z. and Li, H., 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

[25] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[26] Kalchbrenner, N. and Blunsom, P., 2013, October. Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1700-1709).

[27] Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." arXiv preprint arXiv:1908.08345 (2019).

[28] Miller, Derek. "Leveraging BERT for extractive text summarization on lectures." arXiv preprint arXiv:1906.04165 (2019).

[29] Ma, Tinghuai, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. "T-bertsum: Topic-aware text summarization based on bert." IEEE Transactions on Computational Social Systems 9, no. 3 (2021): 879-890.

[30] Kieuvongngam, Virapat, Bowen Tan, and Yiming Niu. "Automatic text summarization of covid-19 medical research articles using bert and gpt-2." arXiv preprint arXiv:2006.01997 (2020).

[31] Alexandr, Nikolich, Osliakova Irina, Kudinova Tatyana, Kappusheva Inessa, and Puchkova Arina. "Fine-tuning gpt-3 for russian text summarization." In Data Science and Intelligent Systems: Proceedings of 5th Computational Methods in Systems and Software 2021, Vol. 2, pp. 748-757. Springer International Publishing, 2021.

[32] Zhu, Qihang, Lin Li, Libing Bai, and Feng Hu. "Chinese text summarization based on fine-tuned GPT2." In Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), vol. 12167, pp. 304-309. SPIE, 2022.

[33] Hartl, Philipp, and Udo Kruschwitz. "University of Regensburg at CheckThat! 2021: Exploring Text Summarization for Fake News Detection." CLEF (Working Notes) 2936 (2021): 508-519.

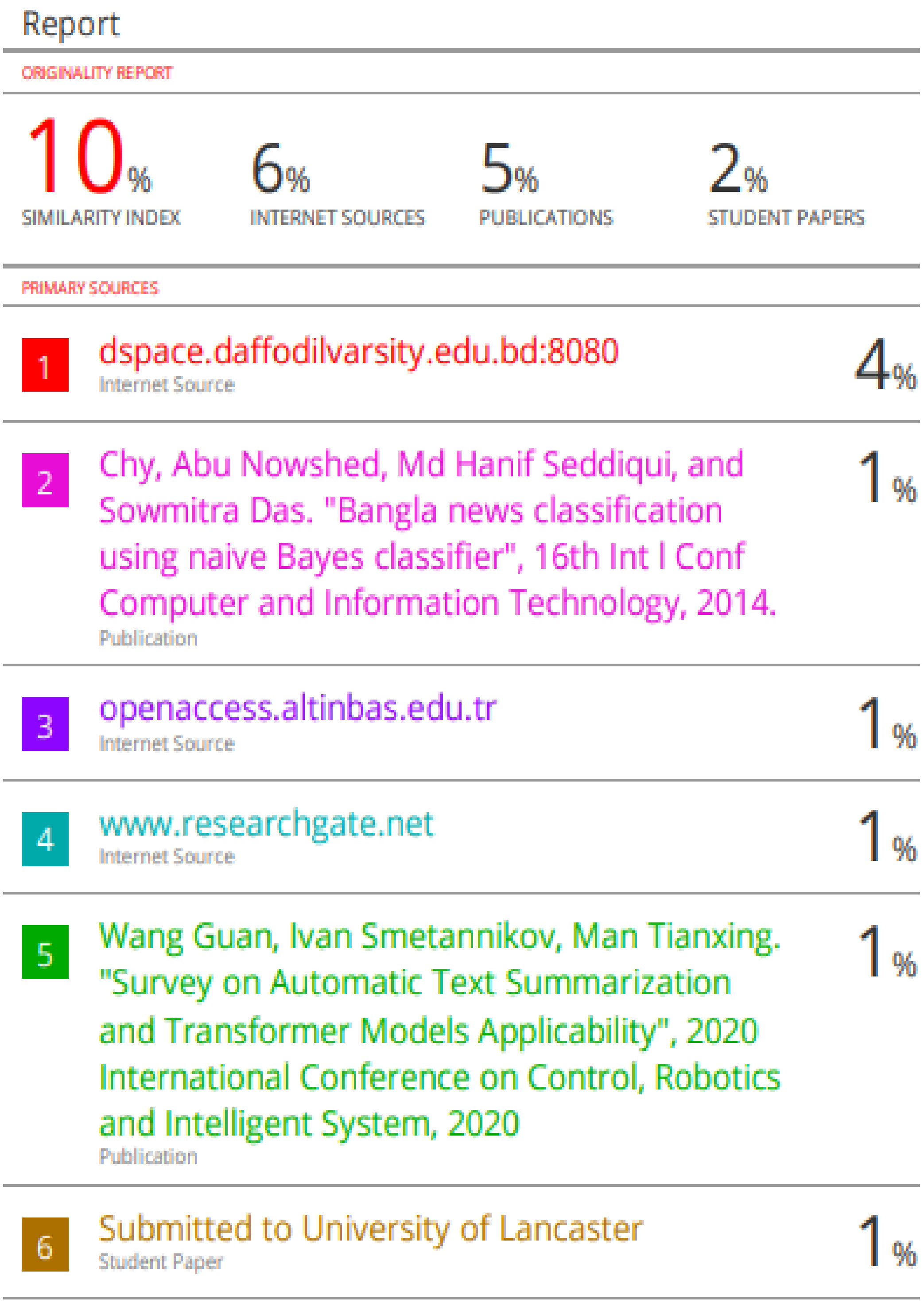
[34] Guan, Wang, Ivan Smetannikov, and Man Tianxing. "Survey on automatic text summarization and transformer models applicability." In Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System, pp. 176-184. 2020.

[35] Wang, Guan, Weihua Li, Edmund Lai, and Jianhua Jiang. "KATSum: Knowledge-aware Abstractive Text Summarization." arXiv preprint arXiv:2212.03371 (2022).

[36] Yang, Zhenmin, et al. "Research on Automatic News Text Summarization Technology Based on GPT2 Model." 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture. 2021.

[37] Kieuvongngam, Virapat, Bowen Tan, and Yiming Niu. "Automatic text summarization of covid-19 medical research articles using bert and gpt-2." arXiv preprint arXiv:2006.01997 (2020).

[38] Zhu, Qihang, et al. "Chinese text summarization based on fine-tuned GPT2." Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021). Vol. 12167. SPIE, 2022.



Report

ORIGINALITY REPORT

10%	6%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Chy, Abu Nowshed, Md Hanif Seddiqui, and Sowmitra Das. "Bangla news classification using naive Bayes classifier", 16th Int l Conf Computer and Information Technology, 2014. Publication	1%
3	openaccess.altinbas.edu.tr Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	Wang Guan, Ivan Smetannikov, Man Tianxing. "Survey on Automatic Text Summarization and Transformer Models Applicability", 2020 International Conference on Control, Robotics and Intelligent System, 2020 Publication	1%
6	Submitted to University of Lancaster Student Paper	1%