A DEEP LEARNING BASED TEXT SUMMARIZATION PROCESS

BY

BINTI BISWAS

ID: 201-15-13760

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

Supervised By

DR. MD. ISMAIL JABIULLAH
Professor
Department of Computer Science and Engineering
Daffodil International University

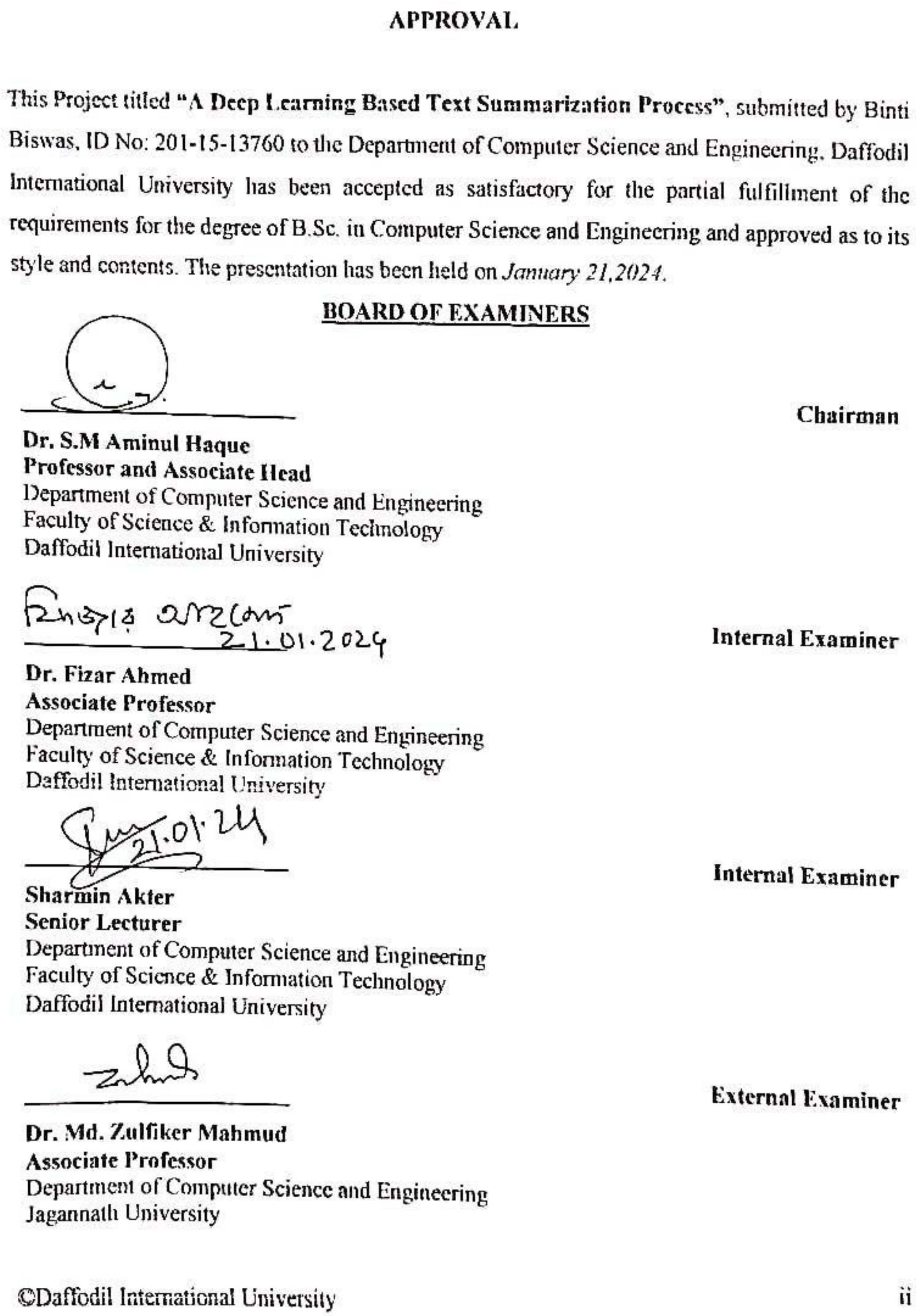
Co-Supervised By

MR. ABDUS SATTAR
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

21 JANUARY 2024



APPROVAL

This Project titled **"A Deep Learning Based Text Summarization Process"**, submitted by Binti Biswas, ID No: 201-15-13760 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 *January 21,2024.*

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

ফিজার আহমেদ
21.01.2024

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

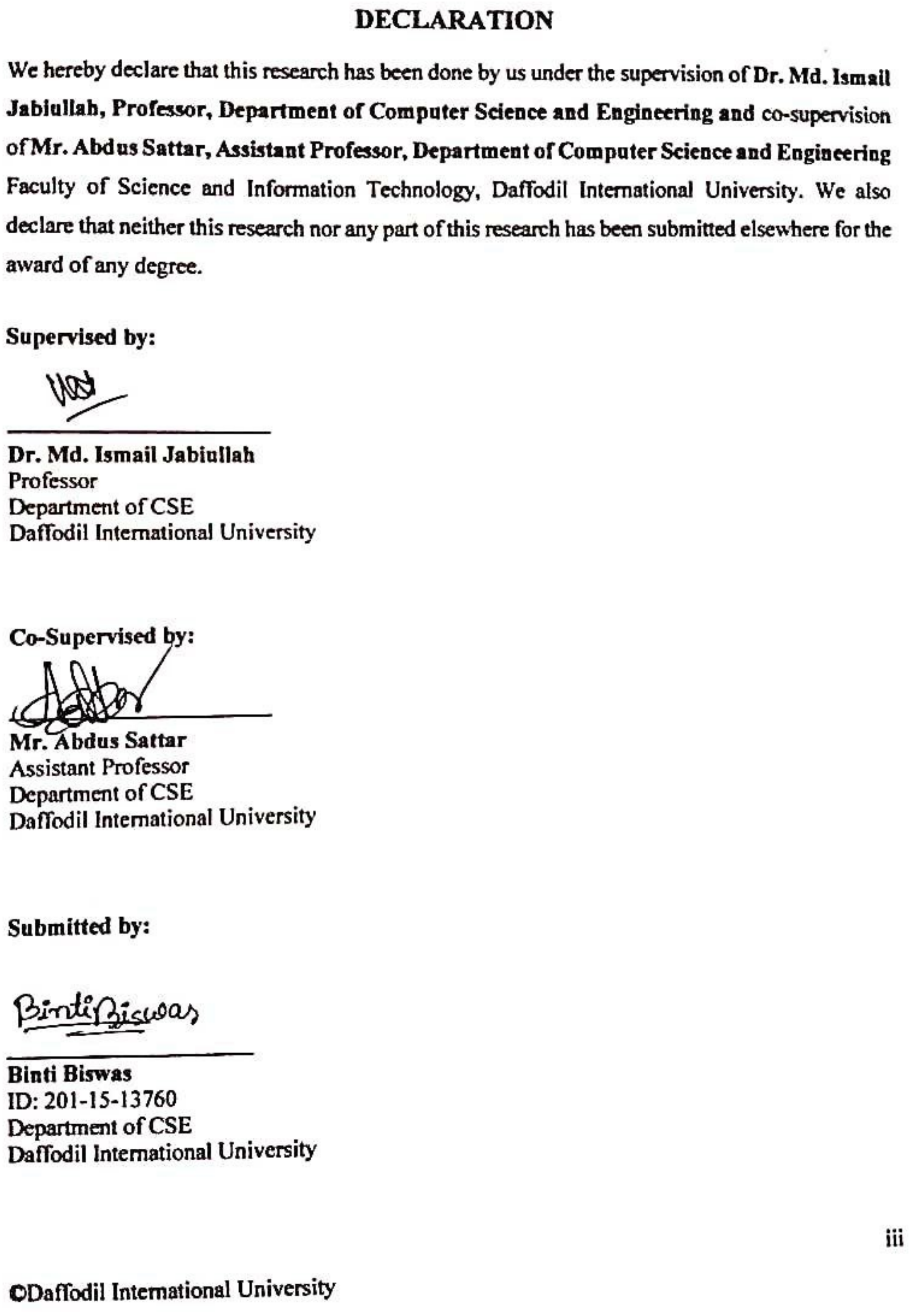
21.01.24

Internal Examiner

Sharmin Akter
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Zulfiker Mahmud
Associate Professor
Department of Computer Science and Engineering
Jagannath University



DECLARATION

We hereby declare that this research has been done by us under the supervision of **Dr. Md. Ismail Jabiullah, Professor, Department of Computer Science and Engineering and** co-supervision of **Mr. Abdus Sattar, Assistant Professor, Department of Computer Science and Engineering** Faculty of Science and Information Technology, Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

Supervised by:

Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Binti Biswas

Binti Biswas
ID: 201-15-13760
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Supervisor Dr. Md. Ismail Jabiullah, Professor**, Department of CSE, Faculty of Science and Information Technology, Daffodil International University, Dhaka. His Deep Knowledge & keen interest with supportive instructions helped us in the field of machine learning based research, finally we completed our work on "**A Deep Learning Based Text Summarization Process**". Their 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 **Dr. Md. Ismail Jabiullah, Professor, Mr. Abdus Sattar, Assistant Professor** and **Dr. Sheak Rashed Haider Noori, Professor and Head**, Department of Computer Science and Engineering, Faculty of Science and Information Technology, DIU, for his valuable support and advice to finish our project and also heartiest thanks to other faculty member and the staff of department of CSE, Daffodil International University.

At last, again we want to thank all the good wishers, friends, family, seniors for all the help and inspirations. This research is a result of hard work and all those inspirations and assistance.

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

ABSTRACT

Technological innovation has greatly improved our quality of life. But people's attention spans are getting shorter and people want to read for shorter periods of time at a rapid pace. Because of this, it's crucial to give a concise description of the most significant news item and the most logical summary that aligns with the synopsis in order to give a rapid review of the essential news or article. In this age of information, there is a vast amount of textual material at our disposal. Examples of sources include online documents, news stories, articles, and consumer reviews of various products and services. Summarizing texts is a technique for automatically summarizing any text, document, or paragraph. A summarized text is just the original material reduced to its most basic form. The primary goal of this effort is to provide a concise, easy-to-read summary that is both significant and comprehensible. since the primary barrier to communication is language. By offering a streamlined version of the material, text summary can assist in reducing the amount of time needed to read and comprehend lengthy publications. Text summaries can aid in highlighting key ideas and enhancing readers' understanding of the content as a whole. We have gathered information from the online portal Kaggle, which summarizes Amazon evaluations of products. We must apply our model in order to obtain an outline. Our model is a bi-directional RNN decoder and encoder with support for sequence-to-sequence, which uses an LSTM to provide the summary. We've encountered a number of issues with this project, including preprocessing, vocabulary counting, word embedding, and missing word counts. Our primary objectives for this project are to calculate the operational loss, create a fluid outline, and develop a more advanced technique for summarizing English texts. Our primary objective with this model was to create an abstractive summarizer based on our dataset.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

The process of condensing and smoothly summarizing a longer text while maintaining the key details and main ideas is known as summarization of text. It is a helpful tool for many uses, such as content analysis, automated essay grading, and information retrieval. Text summarization may be done in a number of ways, such as extractive and abstractive techniques. While abstractive summarizing creates fresh words and phrases that encapsulate the original text's meaning, extractive summarization chooses significant phrases and words from the original material.

These days, there are problems such as having to read a lot of material—many paragraphs—in order to grasp something. which may be newspaper articles, reviews, or essays. A newspaper would display an enormous amount of texts. It requires a lot of time to read the entire article if we want to grasp the essential news. With the speed at which technology is developing, anything can be done in an instant. Most people use technology, including computers and cellphones, to retrieve information. Furthermore, one of them is reading newspapers online, thus before reading any type for data article, a compelling synopsis is essential to entice the reader to read this particular piece of information. These days, there is a huge increase in demand for automated summary of text systems. While the abstractive strategies attempt to summarize fresh phrases and words that are already present in the material, the extractive approach seeks to extract the most significant words or paragraphs from the full document. It creates a condensed version of the initial piece that highlights the key ideas. The summaries that are created may also contain new phrases and sentences that are no longer present in the original text. Deep learning-based techniques are used in abstractive precis to generate summaries. Here, the statements that result from it may or may not be true. Taking on a document presents the seq2seq paradigm with its biggest obstacle. Furthermore, a huge document will cause the current seq-2seq paradigm to perform poorly.

We bring to you a summary that might help you save time and comprehend the content or news simply, thanks to the aid of LSTM and the Sequence to Sequence algorithm. To create the summary, we may comprehend the material using Natural Language Processing (NLP). Text summarization is a procedure that involves submitting a big quantity of text or data, compressing it using machine learning (ML), and providing us with a comprehensible condensed version of the material. In machine learning (ML), summarization of texts is not a novel concept. The fact that there are several alphabets and grammar conventions in the English language is the primary cause of this sort of problem.

Recurrent Neural Networks (RNNs) are the primary tool used by this class on the sequence-to-sequence paradigm to tackle difficult language issues like text summarization. It has been quite popular in the past several years. Because of its enhanced approach, which was presented on the sequence-to-sequence model, it offers high-quality summary. The majority of this mode's strategies fall into one of three categories: decoding, network framework, or parameter inference[1].
A kind of continuous neural network (RNN) called Long Short-Run Remembering (LSTM) may be able to recognize long-term relationships in ordered knowledge. Speech, linguistic text, and statistics are examples of process-ordered information that RNNs, a kind of neural network, are particularly well suited for. By using a unique kind of memory cell that stores data for extended periods of time and chooses to keep or remember data as needed, LSTMs may learn long-term dependencies. Because of this, LSTMs can model dependencies for far longer than older RNNs, which could only be able to learn dependencies over a limited number of time steps [2]. Speech recognition, translation of languages, language modeling, and other challenges have all been tackled with LSTMs. Additionally, they find extensive use in the processing of natural languages (NLP) applications as text categorization, machine translation, and language synthesis. LSTMs may be computationally difficult to train, especially for extremely big datasets, which is one of its drawbacks. Furthermore, because of the instability of the learning process, LSTMs might be challenging to optimize. Despite these difficulties, extended short-term dependencies (long-term dependencies) may be learned from sequential data with great strength and effectiveness using LSTMs, which makes them a valuable tool in the deep learning area [3].

There are insufficient resources in NLP frameworks for English summarizing, thus we must employ all available methods and libraries with raw code. While no model can guarantee 100% accuracy every time, it can typically produce results that are at least passable. During our study, we attempted to provide a high-quality summary for the seq2seq LSTM model summarizer.

1.2 Objectives

Making an overview of a text passage or document that captures the essential details and ideas from the original source is the aim of English Amazon items review text summarizing. Condensing the language to make it easier to comprehend and read while still delivering the most crucial information is the aim. Text summarizing has several uses, including cutting down on the time required to read and absorb a lengthy text and helping one rapidly grasp the key ideas of a document or paper. It may also be used to create summaries for reports and presentations. The following are some typical goals for summarizing an English text:

- To shorten a document to make it easier to read and understand;
- To quickly assimilate a document's main ideas;
- To decrease the amount of time required to read and absorb a large text.
- To enable readers to obtain a feel of a work without reading it in its whole; to extract important elements and ideas from a text

1.3 Motivation

A component of the processing of natural languages is text summarization. Making a clear and insightful synopsis of the book is regarded as the most fascinating and difficult task. To fully understand the paper, one must take the time to read it cover to cover. It is therefore difficult to determine the precise meaning. Automated text summarization, which counts the number of paperwork, words, as well as frequently used terms in the text while simultaneously summarizing it, is essential to solving this sort of issue. Science and technology are growing more and more entwined with our world, and they will for many years to come. These days, we read a lot of articles, books, web pages, newspapers, and other materials on the internet. Furthermore, because of the disorganized material and hazy meaning in such articles, there are moments when we become tired trying to locate the relevant information.

Many NLP techniques have been created in the contemporary period for languages like Spanish and English. We ought to step up and make a contribution in this area because of this.

Summarization is one of the many fundamental issues in NLP. We may quickly grasp the meaning of a lengthy text with the aid of an enhanced text summary. They have created sophisticated natural language processing (NLP) tools and design models for languages like English, Spanish, and others, but our available capabilities are insufficient to address the summarization challenge. The answer to this issue cannot be found without enough investigation. Consequently, we propose in our study effort a pipelined strategy to extractive and abstractive strategy integration for summarizing English materials. International languages is always favored when it comes to sharing with all communicate using these English. NLP resources in English are sufficient, therefore we should concentrate on this field and create new solutions.

1.4 Rationale of the study

Recent decades have seen a surge in research into the application of natural language processing, or NLP, approaches for text summarization, with a particular focus on applying these approaches to languages with limited resources, such as English. Still, there are a lot of obstacles to be solved before English text summarizing systems can be developed that work well. Language is an organized means of communication that allows individuals to express their ideas and comprehend problems. Vocabulary and grammar make up language. It is regarded as the fundamental form of interaction between individuals and can be expressed by written, sign, or spoken language. Many languages are developed as a result of human civilization's geographical dispersion and ongoing progress. Additionally, the goal of this research is to present a unique seq2seq based strategy utilizing cutting edge models like transformers based theories, which have demonstrated efficacy in a variety of natural language processing (NLP) applications including text summarization and machine translation.

1.5 Research Questions

This study has been completed with a great deal of enthusiasm and effort. We likewise struggled to complete this assignment. Developing a plan that is fair, realistic, and accurate faces various challenges. The following questions are intended to aid in understanding these concepts and provide an answer for researchers:

- Have I gathered real data for this research?
- What is text summarization in English?
- What is the process of summarization?
- What benefits can summarize offer?
- How should unstructured English data be preprocessed?
- Abstractive and extractive summaries: what are they?
- How do abstractive and extractive summaries operate?
- What is the process of the English text summarization model?

1.6 Expected Outcome

Our first objective was to successfully insert the necessary paperwork in the required spaces, which we did. After that, the builder makes tools for users. Recent research examines English text compression. Many analyses that were produced in the past to shorten English lectures have previously been published. In order to achieve our objective, we are trying to construct an automated method. An automatic system conditions the machine. So, you have to read the machine to grasp our suggested model. Our goal is to use our built LSTM model to produce an abstract text reduction while preserving the outstanding efficiency of the approach. Aim for precision and minimize complete loss when preparing the model.

1.7 Layout of the Report

There are six chapters in this report.

Chapter 1 contained the project's explanation, objectives, research questions, and anticipated results. The overall structure of the report is also covered in this section.

Chapter 2 covers the analysis that has previously been done in this area. The scope that arises from their restriction of this field is then demonstrated later in this second chapter. The main challenges or impediments to this research were the final issue discussed. This chapter covers the challenges faced in developing the project as well as sections on relevant studies and research summaries.

Chapter 3 provides an explanation of the conceptual evaluation of this study's endeavor. This chapter offers further information on the statistical techniques used in attempt to address the mathematical element of the investigation. With the goal to develop this project, a number of topics are covered in this section, such as the subject of the investigation and equipment, processes, collecting data process, processing of data, recommended model, approach to training, and completion requirements. It also discusses the methodologies used in this research.

The performance review, outcomes discussion, and experimental findings are all included in Chapter 4. To help in the implementation of the project, a few test images are supplied in this chapter.

A summary of the research, details on upcoming activities, and a conclusion were given in Chapters 5 and 6. This chapter provides a verifiable example to demonstrate that the project's report conforms with the requirements throughout. Impact on Sustainability, Society, and The environment: The chapter closes by highlighting the shortcomings of our attempts, which may have a lasting effect on future workers in this sector.

CHAPTER 2

Background

2.1 Introduction

Text summary is the practice of condensing a lengthy text or section into a concise version. The fundamental goal of text summarizing is to make the material easier to read, concise, and thorough. It takes longer and costs more money for the people to condense a lengthy text into a shorter one. Sometimes, even after reading the entire text, we still struggle to comprehend and identify the primary idea. However, text summary simplifies things for us and turns the work into a more effective manner. Extractive and abstractive techniques are the two main categories of methods used to extract information from unprocessed text input and apply it to a summarization model.

Extractive summarization: It is the process of picking out key passages from the text and joining them together to create a summary. To achieve this, pick the passages or statements from the source material that best capture its meaning. Because it doesn't include the creation of new words or sentences, extractive summarizing is sometimes perceived as a less difficult work than abstractive summary. But this strategy may result in a summary that lacks the coherence and fluidity of one generated by an abstractive process.

Abstractive summarization: To understand the text's semantics and provide a helpful summary, abstractive models employ more complex natural language processing (NLP), such as word embeddings. Abstractive techniques are accordingly much harder to train from beginning as they need several parameters and data.

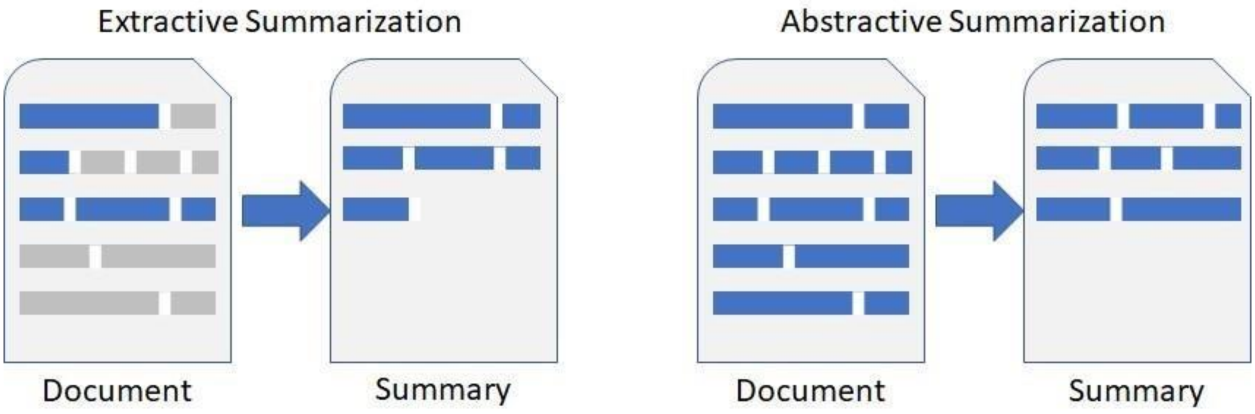


Fig 2.1: Summarizations of Abstractive & Extractive

Our data is gathered from websites such as Kaggle amazon product review. We have several steps involved in gathering and storing the information, and we also require room to store it. However, text summary provides the solution. There are two methods for handling the issue of text abbreviation: extractive and abstractive. We summarize using the abstractive method.

2.2 Related Works

The topic of data extraction is often studied using the use of everyday language. Numerous examinations in a variety of languages have been conducted in this sector. Recently, a lot of Named Entity Recognition ideas have been put out. There are very few papers for low asset dialects, and most of the approaches are language-explicit. This section will likely look at the standard operating procedures in various domains. A brief, condensed remark on a lengthy text document is the text summary by Banerjee et al. in [4]. As mentioned by Yeasmin et al., there are several types of abstractive summarization techniques [5]. According to J. Tan et al. [6], research into neural abstractive summary subsequent generations greatly enhances the accuracy of neural networks. sentence summary prototypes for the English language; however, the end-to-end method's reliance on comprehending the entire document remains a challenge, so the results cannot be conclusively demonstrated. A type of hybrid point forming network was presented by Dhar et al. [7] in their study to solve the issues of phrase repetition and factual fact replication that are difficult to correctly replicate. To enhance the attention-based sequence, they employ a hybrid pointer generating network that can generate words from vocabulary and increase accuracy in reproducing genuine information. They also utilize a coverage technique that prevents repetition. The offered method performs better in both quantitative as well as qualitative model assessments than the state-of-the-art in English.

Every language model is dependent on automated interpretation. Machine interpreting helps create automated programs by enabling the machine to understand the text that is used. Another method for machine translation is NMT. Another NMT technique was given by Bahdanau et al. [8]. Make use of collaborative learning to modify and enhance the presentation of common encoding and inserting methods. The decoder is encoded with the content sentences vector, and the decoder produces a vector grouping effect. After some time, attention-based strategies [9] were introduced, which further enhanced NMT viability. NMT has certain limitations since it requires excessive testing and preparation costs and is unable to produce a satisfactory result for a sequence of unusual words. Reduced NMT stress was introduced to the GNMT program by Wu Y et al. [10]. One method for resolving text-related issues is RNN. A sequentially learned sequence employing RNNs was demonstrated by Nallapati et al. [11] and produced a useful visual coding effect. A predetermined number of scalar sequences of input are contained in the encoder, and the most pertinent input sequence is produced by the decoder. Enhance the functionality of an abstract phrase that doesn't employ the same DRGD as previously used [12]. enhancing the knowledge that is put to use today to summarize the unseen text. A model that summarizes the unseen text through a reinforcement process was presented by Wang et al. [13]. This time, they summarize the unseen text using a series of successive titles. The primary concept of this work is reading a text in its context. CNN is the center of the entire process [14]. The multilayer LSTM-based seq2seq learning approach was introduced by Ilya Sutskever et al. [15]. One information succession map from the target encoder for vector text. The vector shape of the decoding configuration is chosen by one. LSTM can handle lengthy successions with ease in this way. It is therefore possible to modify the inquiry in an arrangement. Using a commonly used Wikipedia text synopsis to condense lengthy content. In place of an encoder, Peter J. Liu et al. [16] presented a decoder and a decoded model that could provide fluid summaries throughout lengthy text sequences. Additionally, a summary of several texts from a sizable and comparable database may be generated using this method. Lifeng Shang et al. [17] offers a method for condensing the material. An overview of the lengthy text sequence may be found in the text replacement procedures. The common encoder and decoder offer a useful method for condensing brief texts. Text summaries can be made with the help of potential text counts.

2.3 Comparative analysis and research summary

The primary emphasis of our project endeavor is the variety of methods that society as an entire has to offer. All we have used four different approaches and added more algorithms to our dataset. Here, the main source of data has been our online-compiled dataset. In our research, we have developed a novel technique for abstractive English text summarization. To create and hone our model, we utilize our own dataset. Deep learning is the approach we employ to develop the model. Processing the approach with word embedding is ideal. With the aid of word vectors, we are able to maintain the associated lexicon in a file with a number value. We employ pre-trained word vectors for trains. Then, we carefully construct the seq2seq model. Bi-directional LSTM is employed in this technology for both encoding and decoding. The word vector learned as input in the encoder portion and in the decoder portion a portion of the output was the pertinent word vector. We discovered the anticipated outcome after training and testing our model on all the data. This are better because of their ability of sequential dependency as well as understanding of contextual information. On the other hand, some other model such as KNN, Logistic Regression or other are not suitable because they are not appropriate for handling sequential data or context.

2.4 Scope of the problem

Text summarization is a helpful tool for our modern life. We used deep learning techniques for evaluate the data and building a model. But in this model some problem can occur. It may require complex dataset, some additional sequences, imperceptible rare text set may take place in text summarization. Sometimes, this will mislead for selecting the right content or misunderstanding of context & coherence.

2.5 Challenges

The main task for this research appears to be combining and assessing all of the data sets, considering how challenging it was to assess this one massive data file. Through the use of several tools and methods, we cleaned and standardized the dataset. The size of the data sets meant that it took some time to get the desired outcome. Considering how challenging it was to analyze just single massive data file, the main problem of this research appears to be combining and evaluating all of these data sets. We used a number of instruments and methods to clean and normalize the dataset. Due to the size of the data sets, it took some time to get the desired outcome.

- Collection of Data
- Importing of data;
- Preprocessing of data
- Achieving summarizer after train model.

Additionally, additional code is required during the pre-processing phase in order to sanitize the text input. For example, removing punctuation from text requires Unicode, and raw coding is the only method that can handle it correctly. Eliminating stop words is the second step. Another difficulty is dealing with a big structured dataset. Finally, a large amount of data may provide a large vocabulary, and a large vocabulary makes it easier to create the perfect summary.

CHAPTER 3

Research Methodology

3.1 Introduction

In this section, we discuss our whole study method. Each study is unique in terms of the approaches taken to address it. Every tactic employed throughout the research endeavor is included in the methodology. A quick synopsis of every element is provided in this technique section, along with an explanation of using models. We'll talk about research methods here. Every research project has a unique approach to problem-solving. We will go over each application strategy in this part along with a brief summary. We've already spoken about how our text summarizing study uses a deep learning model. RNN algorithms were utilized to solve text-related problems. We used our gathered dataset to create an expected flawless and efficient auto system before implementing this technique. We gather and prepare the data before using the algorithm. Each component of employing methods will be covered in isolation in this section. A more thorough approach description boosts task efficiency. We will provide mathematical formulas and a graph along with a succinct explanation of the concept.

The structure is meant to seem like a casing. In this part, all of the strategy's tools are briefly discussed. The sub-section of some of the main sections aids in comprehending how the model is built up with the intention of using it. Below is a working development of all the inspection work that provides a brief overview of the entire exploratory activity. Subsections have been included in this technique section to provide a quick description of the process. The whole methodology, from data processing to model building to dataset training and testing, is provided here.

3.2 Subject of the Research and Equipment

A research theme is a field of study that is examined and investigated to clarify concepts. not only for execution but also for handling, developing models, obtaining information, carrying out tasks, and training models. We discuss the tools and procedures we employ in the instrumentation profession. We used the Python programming language, the Windows operating system, and a number of tools, such as NumPy, SkLearn, OpenCV, etc. The Google Colab system has been used for the entirety of the instruction and testing process. Python programmers are allowed to write code for data science and deep learning algorithms using Google's Colab as a platform. These algorithms make use of statistical techniques linked to summarize English evaluations about this dataset.

Libraries used:

- **Matplotlib:** One of the visualization tools offered by Matplotlib is the Pyplot graphing, calculating, plotting, charting collection of techniques. This could potentially be used, while creating forms, to indicate tale corners or to draw the limits of a narrative.
- **NumPy:** With the Python NumPy tool, vector manipulation is now easier. The topic content particularly addresses the Fourier transform, the indices transformation, and matrix computations. For dealing with matrices of various types, the NumPy packages for Python offer a multitude of tools and techniques. NumPy can help with more rational and realistic device design. NumPy is, in short, a Python package designed largely for numerical evaluation. This assertion is also known as "estimation. Py."
- **Sklearn:** Sklearn is an effective and user-friendly tool for data analysis prediction. Its design made use of three Python tools: NumPy, SciPy, and Matplotlib. These are open-source tools that each user may personalize.
- **Seaborn:** You may use matplotlib with the next version of this popular Python information visualization application. This is an easy-to-use tool for creative data visualization.
- **H5py:** Users may read binary HDF5 code using the Python programming language's h5py package. Large amounts of data, particularly integers, are managed with NumPy and subsequently stored in HDF5.
- **TensorFlow:** This collection of AI technologies appears to be free. It offers an extensive array of tools for creating and implementing diverse machine learning models, such as neural networks. TensorFlow is an excellent option for a variety of AI applications because to its adaptability, effectiveness, and user-friendliness.
- **Pandas:** This freeware toolkit for analyzing and manipulating language-specific data is offered by Pandas. There are fundamental data kinds and statistical analysis processes

available to help with systematic data administration, particularly with the management of summary information.

- **OS:** Because of the numerous features of this Python OS component programmers are able to interact with the language of programming that these people use.
- **NLTK:** A collection of Python-based libraries and applications for statistical and symbolic NLP (natural language processing) of English is known as the Natural Language Tool Kit, or simply NLTK.

3.3 Workflow

It is conceivable that a variety of approaches or strategies were applied in this investigation to obtain the results. To create a summary of this text dataset, a workflow was selected, and data was processed, collected, stop words eliminated, text cleaned up, and the output of the decoder encoder LSTM model.

Phase 1: Collection of Data: After obtaining the data from the internet, we processed it. Due to the challenges in obtaining information for the particular extraction dataset of Amazon product evaluations, which has a wealth of information for the successful completion of this study.

Phase 2: Processing of Data: Each component of the data was assessed independently once every feasible method of gathering it had been used. There are several instances of improper and stop words; tidy up the information in the surroundings. I truly oversee the last stage of the selected dataset before using it.

Phase 3: Preparing Dataset: Getting the dataset ready Preprocessing and data expansion are still being done in accordance with the summary and text. We must clean up, tokenize, arrange, remove stop words, and display the data for training purposes. Just the absolute minimum of work has been performed to prepare the data for splitting.

Phase 4: Model selection: To increase dependability, we choose a clever summarizer creation method, train it, and then evaluate it using my data. This model approach makes extensive use of filters. In the end, just one gadget was chosen to assess how well models summarized the device's design-based productivity improvements.

Phase 5: Performance Assessment: This section covers every one of the repercussions. Following the division of the dataset and the building of the models, all of the techniques' results have been detailed.

Phase 6: Conclusion and Future Work: We'll provide an overview of this area's developmental roadmap.

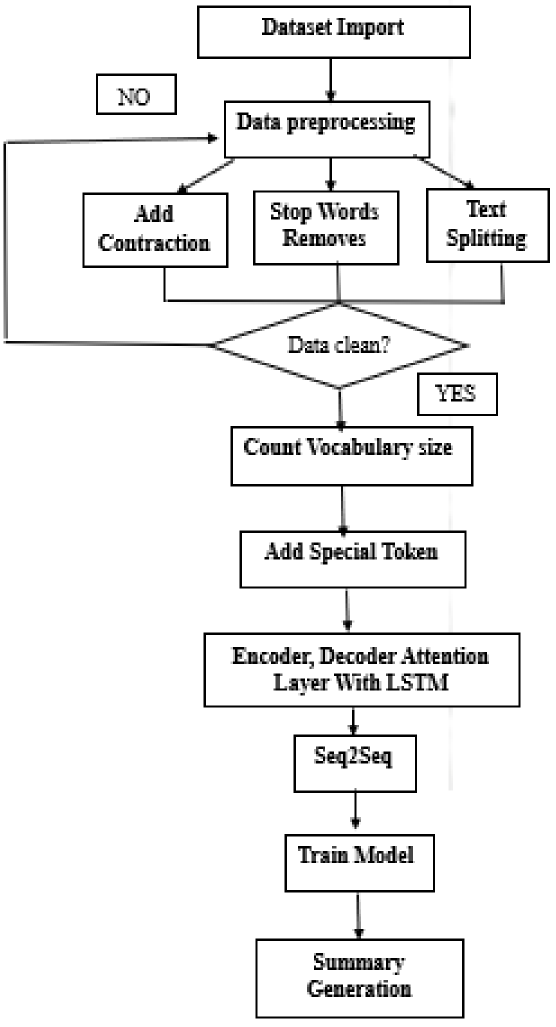
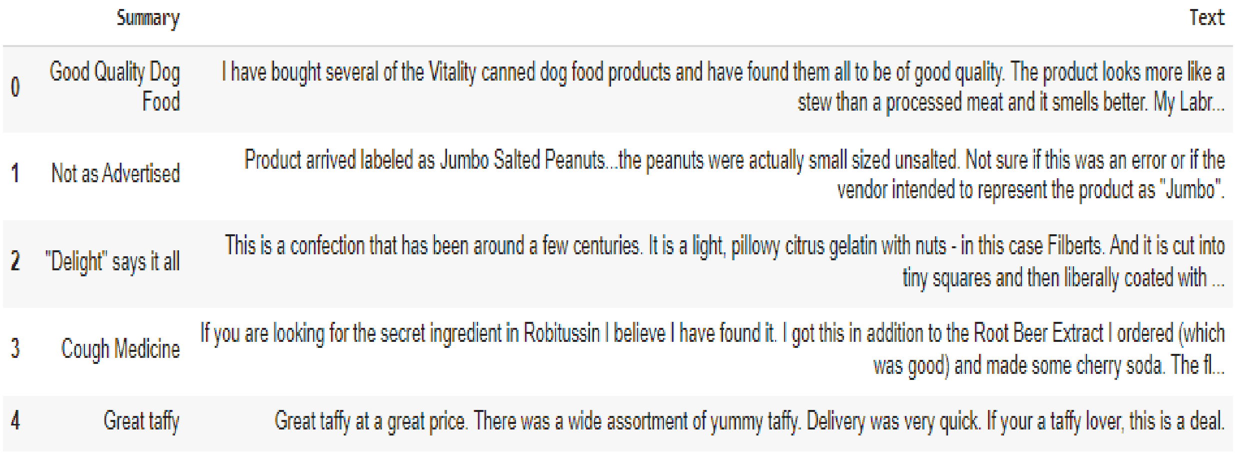


Fig 3.1: Whole research Process flow diagram of Text-summarization

The phases in our study approach that might aid in the summary of English Amazon product reviews are shown in Fig. 3.1. Our internet information was gathered from several sources. Data was gathered from a Kaggle dataset that is available to the general audience. To make sure that every data set will just include correct and pertinent information, we have gone over this data collection, eliminated any superfluous phrases, and cleaned up the phrasing. Tokenization, attention layer, and commonly used stop words should all be made easier to read. We investigate DL model techniques through model development and refinement using pre-existing data. In addition, we employ permutation techniques to address the problem of data summarizer variability to guarantee equitable representation and enhance the overall effectiveness of the model. In addition to providing a summary, our method looks for methods to use seq2seq procedures to minimize the amount of data for summarization.

3.4 Data Collection and Preprocessing

Cleaning up our own dataset is best accomplished through data preparation. Because the intended outcome cannot be produced without a high-quality dataset. The data required for this investigation was sourced from the publicly accessible Kaggle database. For the summarizing, we used the summary content dataset's Amazon product review. This dataset consists of ten columns. Ten columns and 60,000 total data points make up this dataset. "Text" and "Summary" are two examples of the data columns we employ.



	Summary	Text
0	Good Quality Dog Food	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labr...
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to represent the product as "Jumbo".
2	"Delight" says it all	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and then liberally coated with ...
3	Cough Medicine	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (which was good) and made some cherry soda. The fl...
4	Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.

Fig 3.2: Sample data of review summary.

Table 3.1, which is broken down into several different categories, has a list of all the fields within each of these files:

Table 3.1: About Dataset's Column.

Column Name	**Description of the Column**
ID	Id number value.
ProductId	Id no of product
UserId	ID od User
ProfileName	Profile who review the product
HelpfulnessNumerator	how many people really discovered the review
HelpfulnessDenominator	the overall count of people that cast votes on helpfulness
Score	Review quality Score
Time	Review time
Text	**Review text**
Summary	**Summary of review**

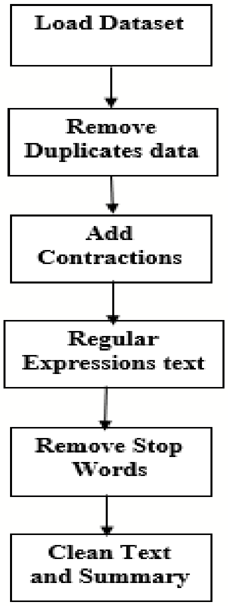
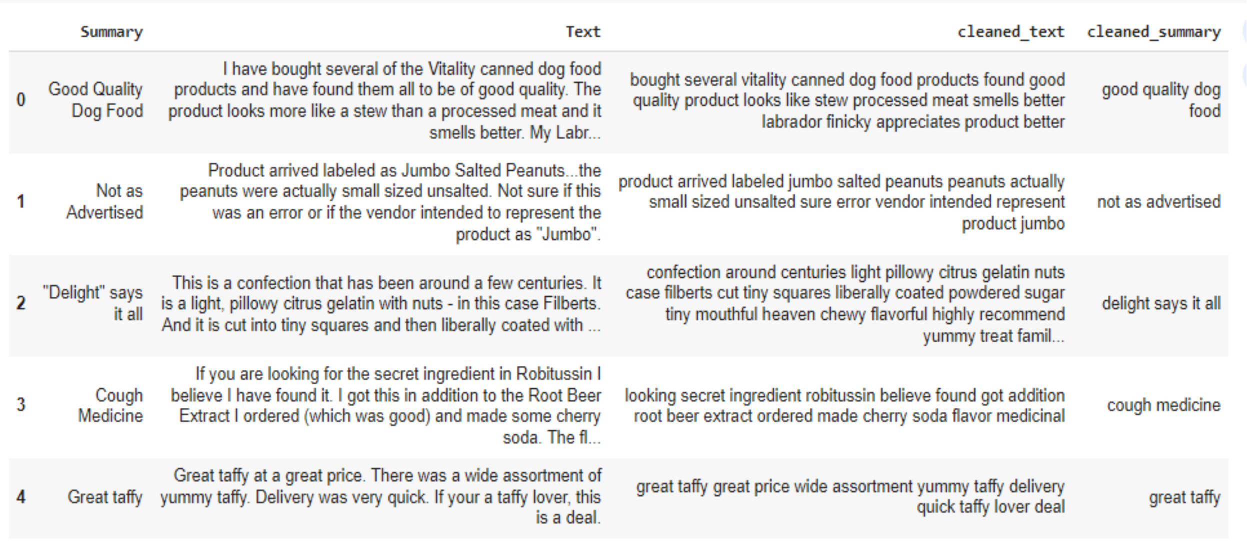


Fig 3.3: Dataset clean process

One important step before creating models is the processing of data. Numerous procedures are needed for the data preprocessing. Preparing data for English can be extremely difficult. Preprocessing starts with removing unnecessary words and spaces. This is our abbreviation for integrating it.

Table 3.2: Some English Constriction List

Short Form	Long Form
“doesn’t”	“does not”
"aren’t"	"are not"
"can’t"	"cannot"
"could’ve"	"Could have"
"couldn’t"	"could not"
"mayn’t"	"may not"
"that’d"	" that would"
"they’ll"	"they will"



	Summary	Text	cleaned_text	cleaned_summary
0	Good Quality Dog Food	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labr...	bought several vitality canned dog food products found good quality product looks like stew processed meat smells better labrador finicky appreciates product better	good quality dog food
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to represent the product as "Jumbo".	product arrived labeled jumbo salted peanuts peanuts actually small sized unsalted sure error vendor intended represent product jumbo	not as advertised
2	"Delight" says it all	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and then liberally coated with ...	confection around centuries light pillowy citrus gelatin nuts case filberts cut tiny squares liberally coated powdered sugar tiny mouthful heaven chewy flavorful highly recommend yummy treat famil...	delight says it all
3	Cough Medicine	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (which was good) and made some cherry soda. The fl...	looking secret ingredient robitussin believe found got addition root beer extract ordered made cherry soda flavor medicinal	cough medicine
4	Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.	great taffy great price wide assortment yummy taffy delivery quick taffy lover deal	great taffy

Fig 3.4: Cleaned Text & Summary

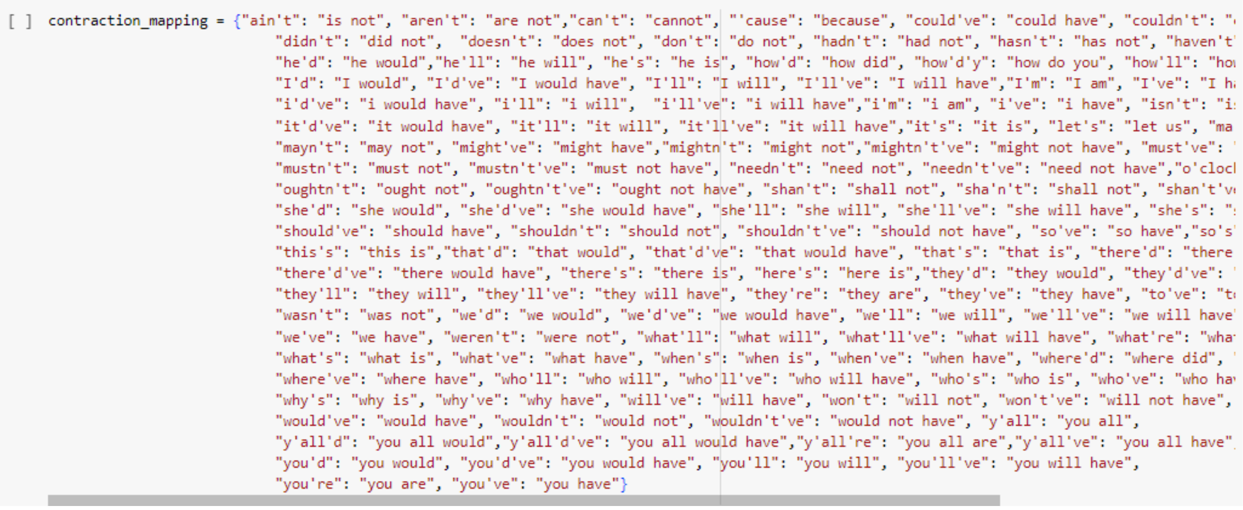
3.5 Statistical Analysis

3.5.1 Analyzing Data

Transforming the unprocessed textual input into a framework appropriate for evaluation and model training is our aim for the data pre-processing phase. The "Text" and "summary" characteristics that are required for summary generation must first be extracted from datasets. The several datasets that were collected are then combined to generate a collection of data that can be used for training and testing. Following the creation of the data database, we moved ahead and fixed errors by eliminating unnecessary punctuation and letters. Enhancing the appropriateness of word choice in the text is the aim of these exercises. The technique of converting words into readily concatenable numeric vectors is known as tokenization in models. This thorough data processing method helps our seq2seq models provide an extracted summarizer of the analysis of this dataset.

3.5.2 Contraction Mapping

Every language need diminutive for phrases that are clear. Similar to this, there are a couple of combinations in every explicit word in the English language. implies a brief form of an adjective or a brief way of spelling a word. The machine is unable to comprehend the abbreviated form of material and so cannot fully convey the meaning of the term in its entirety.



```
[ ] contraction_mapping = {"ain't": "is not", "aren't": "are not","can't": "cannot", "'cause": "because", "could've": "could have", "couldn't": "
                              "didn't": "did not",  "doesn't": "does not", "don't": "do not", "hadn't": "had not", "hasn't": "has not", "haven't
                              "he'd": "he would","he'll": "he will", "he's": "he is", "how'd": "how did", "how'd'y": "how do you", "how'll": "ho
                              "I'd": "I would", "I'd've": "I would have", "I'll": "I will", "I'll've": "I will have","I'm": "I am", "I've": "I h
                              "i'd've": "i would have", "i'll": "i will",  "i'll've": "i will have","i'm": "i am", "i've": "i have", "isn't": "i
                              "it'd've": "it would have", "it'll": "it will", "it'll've": "it will have","it's": "it is", "let's": "let us", "ma
                              "mayn't": "may not", "might've": "might have","mightn't": "might not","mightn't've": "might not have", "must've":
                              "mustn't": "must not", "mustn't've": "must not have", "needn't": "need not", "needn't've": "need not have","o'cloc
                              "oughtn't": "ought not", "oughtn't've": "ought not have", "shan't": "shall not", "sha'n't": "shall not", "shan't'v
                              "she'd": "she would", "she'd've": "she would have", "she'll": "she will", "she'll've": "she will have", "she's": "
                              "should've": "should have", "shouldn't": "should not", "shouldn't've": "should not have", "so've": "so have","so's
                              "this's": "this is","that'd": "that would", "that'd've": "that would have", "that's": "that is", "there'd": "there
                              "there'd've": "there would have", "there's": "there is", "here's": "here is","they'd": "they would", "they'd've":
                              "they'll": "they will", "they'll've": "they will have", "they're": "they are", "they've": "they have", "to've": "t
                              "wasn't": "was not", "we'd": "we would", "we'd've": "we would have", "we'll": "we will", "we'll've": "we will have
                              "we've": "we have", "weren't": "were not", "what'll": "what will", "what'll've": "what will have", "what're": "wha
                              "what's": "what is", "what've": "what have", "when's": "when is", "when've": "when have", "where'd": "where did",
                              "where've": "where have", "who'll": "who will", "who'll've": "who will have", "who's": "who is", "who've": "who ha
                              "why's": "why is", "why've": "why have", "will've": "will have", "won't": "will not", "won't've": "will not have",
                              "would've": "would have", "wouldn't": "would not", "wouldn't've": "would not have", "y'all": "you all",
                              "y'all'd": "you all would","y'all'd've": "you all would have","y'all're": "you all are","y'all've": "you all have"
                              "you'd": "you would", "you'd've": "you would have", "you'll": "you will", "you'll've": "you will have",
                              "you're": "you are", "you've": "you have"}
```

Fig 3.5: Contraction mapping using this summarizer.

3.5.3 Regular Expression

To remove an odd character or some other element that isn't necessary to remove from the text, use a standard statement. The main use of a typical statement in our investigation is to remove white space, English characters, numeric digits in the content, and composition structure text.

3.5.4 Stop word Removal

We like the language that was applied when the content was enhanced. Even though they are widely used, the aforementioned words usually have no meaning inside the categorization system or in the language context. We utilize stop words in conjunction with a bespoke module we bought from GitHub called "Stop words" to get around this problem. The vast array of English numbers in the library has been painstakingly maintained to follow the syntactic structure of our language. We might expedite data processing and reduce the number of unnecessary characters by merging these libraries. In NLP, eliminating stop words is a rather frequent procedure. For the majority of NLP research, we must adhere to this phase. Stop word remove's primary task is to eliminate useless text off the input field. However, there isn't a library for English language to accomplish this kind of job. Therefore, machined need a library for this, similar to the English language's NLTK library. Additionally, there are other terms that are uncommon, such as whitespace, extraneous space, text in another language, etc. Therefore, regular expression is used for this. As seen by the instances above, we frequently omit punctuation to improve the consistency of our endeavors at text summarization:

"it", "ve", "!","re", "me", "the", "oh", "we"etc.

3.5.5 Text Cleaning

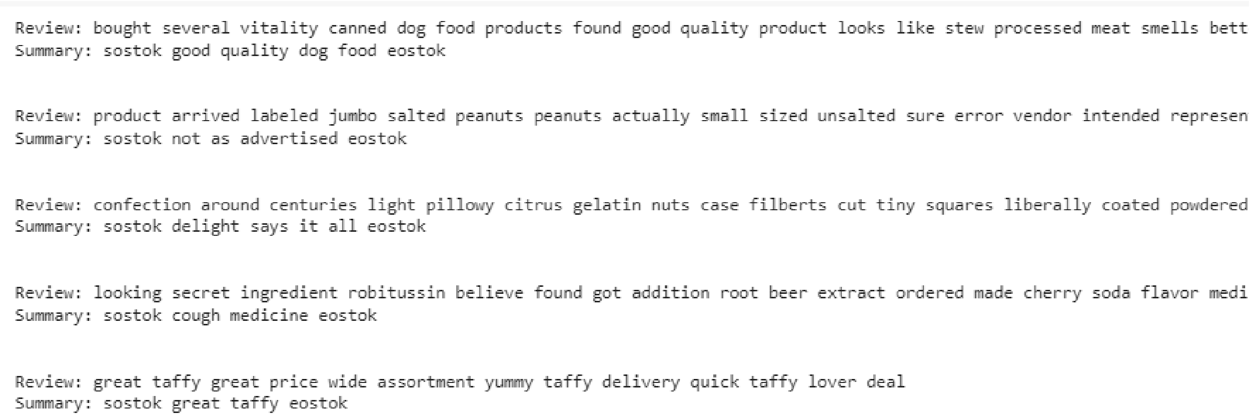
We do text change, which is a crucial step in the dataset's development. This approach consists of two primary tactics designed to improve the caliber and pertinence of written material. News is first sorted, and text are only retained for a set amount of time beyond a predefined word limit. Filtering measures safeguard our promise to provide high-quality, legally compliant, and educational content. A text correction approach is also used to remove extraneous characters like tabs, stop periods, special signs, etc. in order to systematically change the content. Standard symbols, line breaks, and certain English characters have been removed from the text. We employ a comprehensive preservation method to guarantee that the information we collect will be prepared for a detailed investigation. We have taken a lot of actions up to this point to clean up and make our dataset usable. Each word is arranged in logical sequence. Two text columns are required for the finished product. One is for text input, while the other is for text output. Ultimately, our outcomes are visible to us. Below is an illustration of preprocessed text:

Table 3.3: Example of clean up dataset.

Main Text	**Clean Text**	**Main Summary**	**Clean Summary**
I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better.	I bought several vitality canned dog food products found good quality product looks like stew processed meat smells better labrador finicky appreciates product better	Good Quality Dog Food	good quality dog food
Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to represent the product as "Jumbo".	product arrived labeled jumbo salted peanuts peanuts actually small sized unsalted sure error vendor intended represent product jumbo	Not as Advertised	not as advertised

3.5.6 Special Tokenization

Tokenization is a crucial component of contemporary data use. Tokenization is the process of turning syllables into numerical sequences so that our model can understand language. Writing down a distinct number for every word establishes the relationship between symbolic numbers and written language. By ensuring that every sequence has a consistent length during pattern training, padding raises the likelihood of agreement. Our algorithms are able to assess the various linguistic intricacies of text because they transform statements into quantifiable forms and verify that sequence length swings so frequently.Special token have been used like;: “sostok” & “eostok”



```
Review: bought several vitality canned dog food products found good quality product looks like stew processed meat smells bett
Summary: sostok good quality dog food eostok


Review: product arrived labeled jumbo salted peanuts peanuts actually small sized unsalted sure error vendor intended represen
Summary: sostok not as advertised eostok


Review: confection around centuries light pillowy citrus gelatin nuts case filberts cut tiny squares liberally coated powdered
Summary: sostok delight says it all eostok


Review: looking secret ingredient robitussin believe found got addition root beer extract ordered made cherry soda flavor medi
Summary: sostok cough medicine eostok


Review: great taffy great price wide assortment yummy taffy delivery quick taffy lover deal
Summary: sostok great taffy eostok
```

Fig 3.6: Tokenization Example.

3.5.7 Data Preparation

We did not split the data into two groups at random for model training and testing during the data preparation phase. We extract the most significant "Text" and "Summary" portions from the 60k data Amazon review summary dataset. The dataset consists of two parts: validation and training.

- Total Value of Unique 84,721.
- Total data 60k review and summary.
- Drop duplicates data then dataset contains 54,775.
- Cleaned summaries and text that include ten words or less which contains 97.94% & 91.45%.

- Add sostok and eostok special token to the summary.
- When thresh=4 , 64.8% rare words in vocabulary.
- When thresh=6, 76.75% rare words in vocabulary.
- 96% words use in model according to the text and summary.

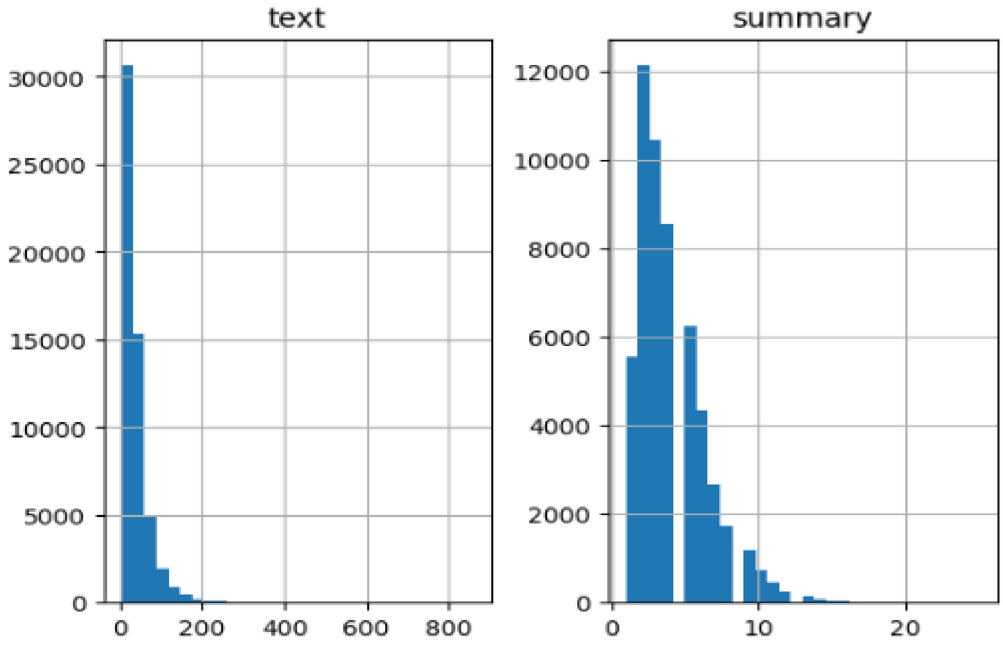


Fig 3.7: Word count for the summary and review text

3.5.8 Models

Our study uses a variety of techniques to understand text summarization. For the summarizer, we looked at the seq2seq LSTM model and the RNN encoder decoder. We leverage these models to take advantage of their capacity to characterize patterns, background, and linguistic variations. This thorough examination helps us achieve our aim of developing a dependable method for producing summaries.

1. **Encoder & Decoder:** The Random Neural Network Encoder and Decoder functions as a pair of encoders and decoders by combining two Recurring Neural Networks. It generates the encoder state and its input sequence during the encoder phase. Information is summarized into

the input sequence using the encoder state. The encoder states are used by the decoder to produce the output sequences.

Encoder: Only news structure course booklet at a time is provided by the encoder. Every word undergoes an implanting level at the beginning that transforms it into an ordered representation. A multi-facet brain networking with the resigning layers formed after entering the preceding word, or all 0's for the reading material's primary word, is also used for coordinating this described depiction. We chose to use LSTM instead of GRU, even though GRU has an advantage in preparation time since LSTM has a more robust conceptual assurance and is simpler to tweak bounds. $\hbar$.

Decoder: Our proposal splits the decoding section into two different modes: replicating mode and creation mode, based on the organization to-arrangement viewpoint. In the unlikely event if the decoder anticipated word grouping, the terms in our feedback configuration are well defined, with "I" standing for the assortment sequence, "H" for the state that is hidden, & "V" for the knowledge vector. This formula is used for evaluating the enclosed states.

$$\mathrm{h}i = (()) \mp ())............................(1)$$

The encoder's data was inserted into fixed c as =... Every t-second interval modifies the RNN.

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

&

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

where c is the invisible component and f & q are the non-linear components. In particular, we can get the stochastic translates for the decode employing the X sequence.

$$() = \prod = \mathrm{l}(\{1,....,-1\}).............(4)$$

Where $= (\, , \ldots , \quad)$. Condition statement,

$$\text{e.g.} \quad = (\, , \ldots , \mid , \ldots , \,). = (- , \, -, c)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

The probability under conditions is as

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

Context vector here.

thereafter determined as a weighted total= $\sum = 0$ ……………………………………(7)

Assume that the state that is hidden is also ($h1h$) for input sequence (h). Thus, the undetected state ($\hbar\ h1$)

h=[h;h]……………………….……………… (8)

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

RNN encoder-decoder is now presumptively lower.

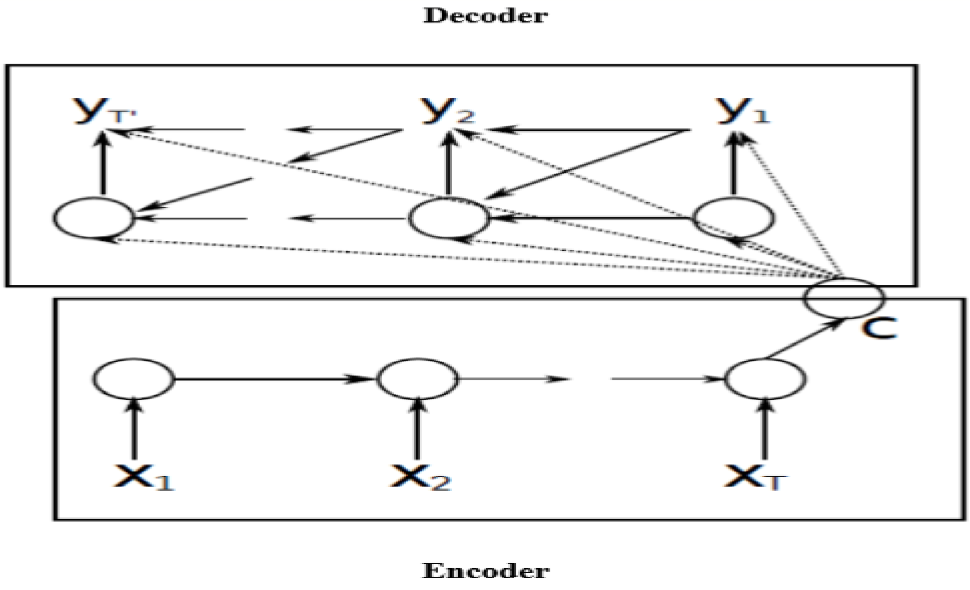


Fig 3.8: Encoder & Decoder Workflow.

2. **LSTM:** The same goal is achieved by this approach, which maximizes its cumulative memory capacity. Once the LSTM layers are connected to the embedding layers, there are density layers with periodicity dropouts. In the output layer, the sigmoid function facilitates binary classification.

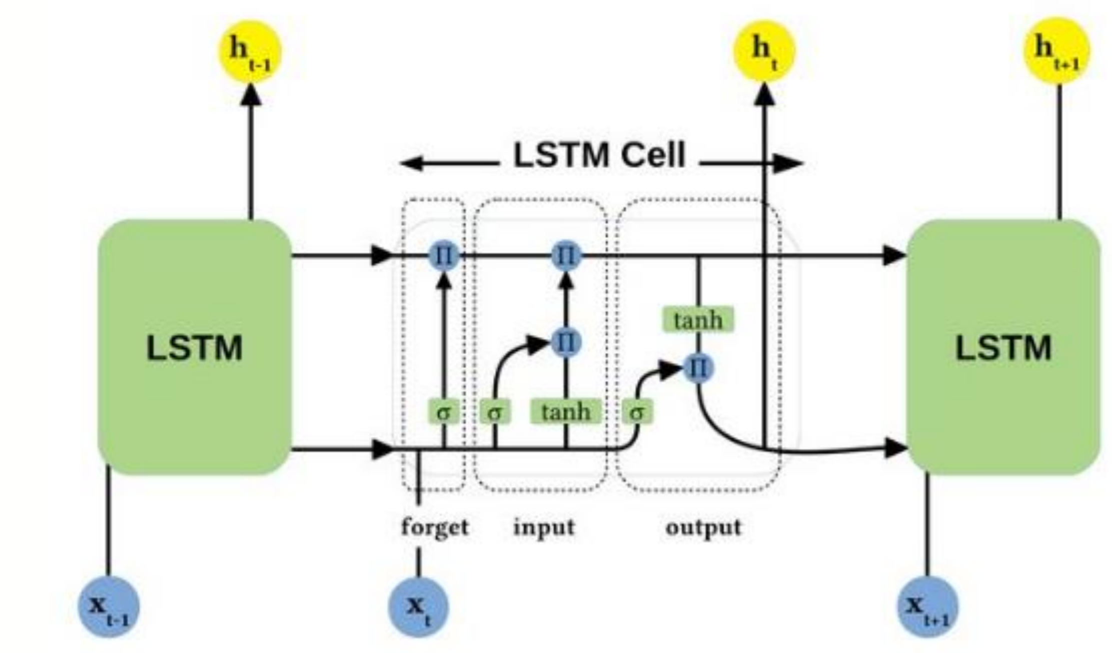


Fig 3.9: Long-Short Term Memory's overall model design.

3. **Sequence to Sequence Learning:** Sequence-to-sequence models have essentially been used for a few typical linguistic handling (NLP) programs, such as text rundown, sound acknowledgment, and title generation for machine translation. A grouping-to-succession paradigm is developed. An encoding system and a decoder are used in conjunction with a bidirectional LSTM unit in such a concept. The most logical and noteworthy RNN technique in deep learning It resolves text-related conflicts even more successfully. Transient memory is analogous to the operation of each LSTM field that constitutes an RNN LSTM cell. Both decoders and encoders are used in LSTM cells. After Encode has

delivered the information texts, the outcome is created using the important text groupings. Tokens are used to confirm the start and finish of each repetition.

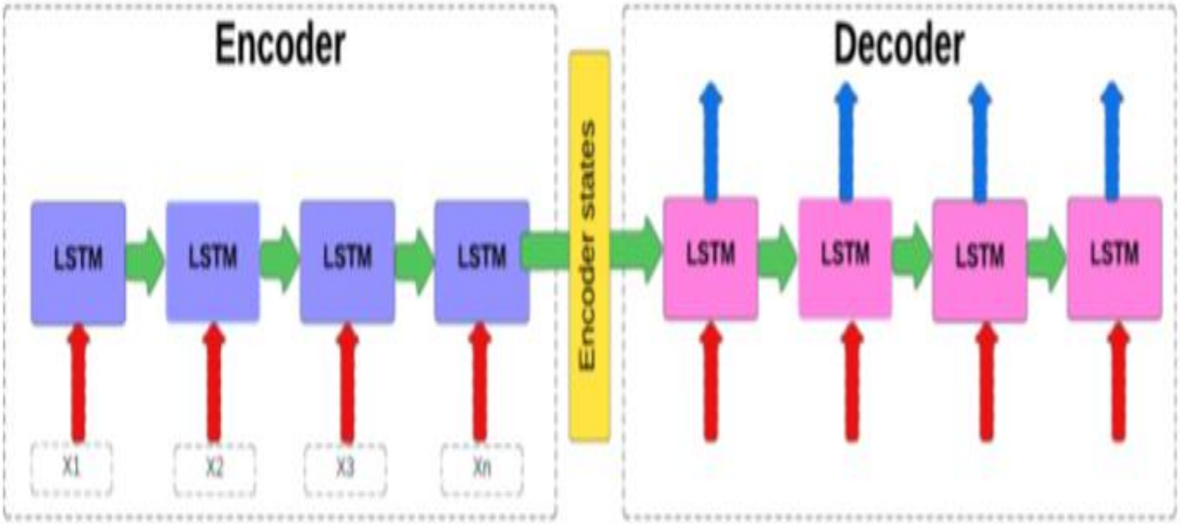


Fig 3.10: Workflow of Sequence to Sequence model.

3.5.9Proposed Model

The same goal is achieved by this approach, which maximizes its cumulative memory capacity. Once the LSTM layers are connected to the embedding layers, there are density layers with periodicity dropouts. In the output layer, the sigmoid function facilitates binary categorization.

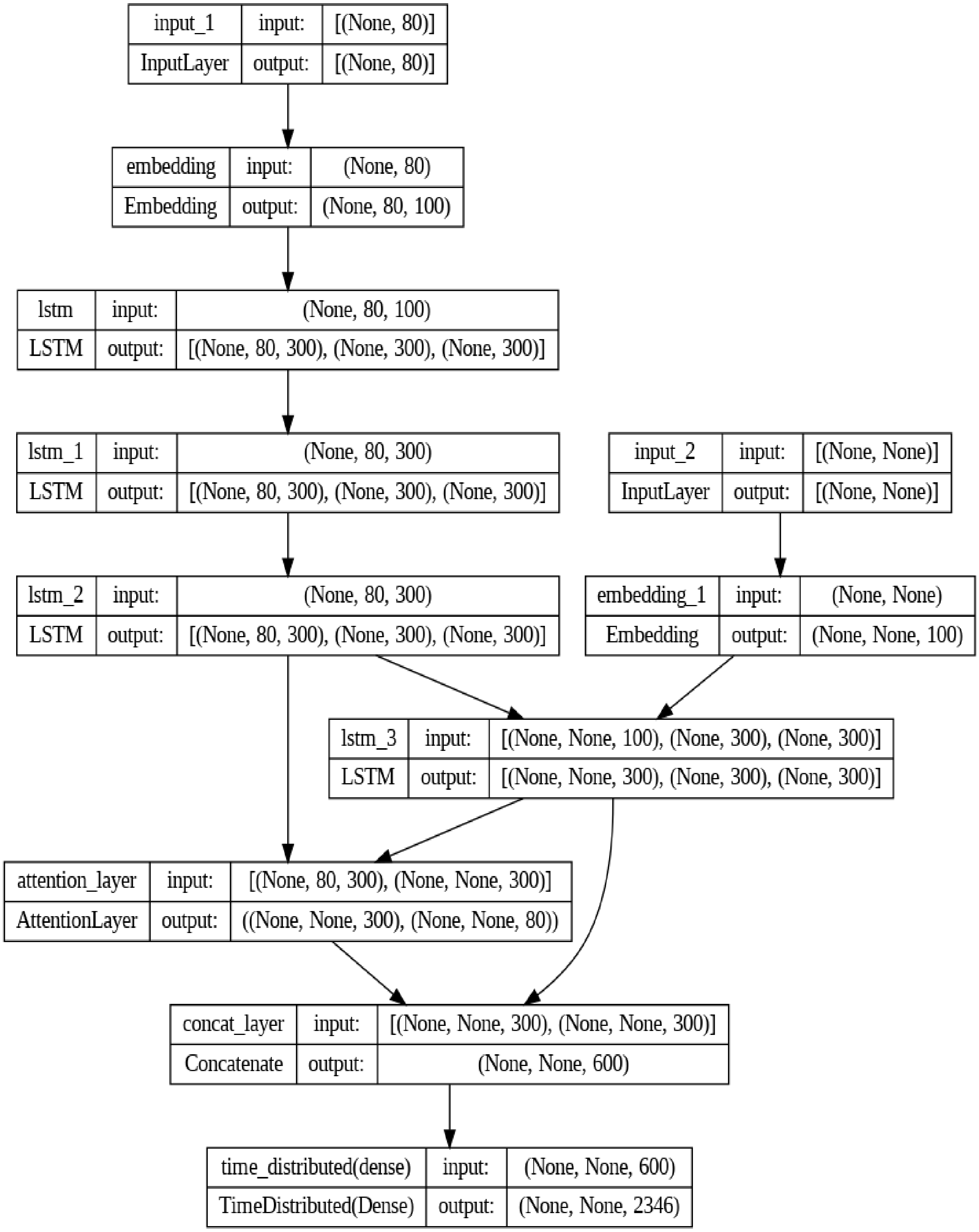

input_1 | input: | [(None, 80)]
InputLayer | output: | [(None, 80)]
embedding | input: | (None, 80)
Embedding | output: | (None, 80, 100)
lstm | input: | (None, 80, 100)
LSTM | output: | [(None, 80, 300), (None, 300), (None, 300)]
lstm_1 | input: | (None, 80, 300)
LSTM | output: | [(None, 80, 300), (None, 300), (None, 300)]
input_2 | input: | [(None, None)]
InputLayer | output: | [(None, None)]
lstm_2 | input: | (None, 80, 300)
LSTM | output: | [(None, 80, 300), (None, 300), (None, 300)]
embedding_1 | input: | (None, None)
Embedding | output: | (None, None, 100)
lstm_3 | input: | [(None, None, 100), (None, 300), (None, 300)]
LSTM | output: | [(None, None, 300), (None, 300), (None, 300)]
attention_layer | input: | [(None, 80, 300), (None, None, 300)]
AttentionLayer | output: | ((None, None, 300), (None, None, 80))
concat_layer | input: | [(None, None, 300), (None, None, 300)]
Concatenate | output: | (None, None, 600)
time_distributed(dense) | input: | (None, None, 600)
TimeDistributed(Dense) | output: | (None, None, 2346)

Fig 3.11: LSTM Proposed Model Architecture

3.6 Implementation

We must gather the data after completing the preceding procedures in order to guarantee the generation of a summarizer technology. For our job, eight elements were needed to finish the basic arrangement. Every one of those tasks must be finished if we are to meet our objective.

- Collection of Data.
- Duplicates data eliminates.
- Stop words removes.
- Data cleaning.
- Preparing Data
- Add Special Token.
- Model execution for decoder and encoder to create summary.
- Discuss about outcome.

We had to start creating the idea's code in order to put it into action. We evaluated the accuracy of LSTM seq2seq several methods. We assessed the algorithm's outputs to provide a summary after it was complete. We evaluated the summary and came to the conclusion that some summaries would be better appropriate for our purposes. A set of prerequisites that are required for any attempt at abstractive summary generation has been defined following a thorough examination of the relevant theoretical and numerical techniques and ideas. The following are potential, essential results:

1. Equipment and Software specifications

- Operating System (Windows 7 or above)
- Hard Disk (minimum 1 TB)
- Ram (Minimum 4 GB)

2. Creating Tools

- Python Environment
- PyCharm.
- Google Colab.
- Visual Studio code.

CHAPTER 4

Experiment Results and Discussion

4.1 Introduction

Information extraction, or text outline, is the most perplexing problem in the NLP domain. As a result, the computer can summarize the material without having to summarize any reactions that are included or absent from the content. Thus, it is difficult to locate an accurate summary. It is important to include chances in this summary of the material. Given that the yield provided by the machine is reliant entirely on chance. For every word read in the train, load the model and determine the possible rundown based on the weight of the associated word. An AI model should be developed with the content information following further developments. Every internal and external learning model has a backing motor during preparation. We used Tensor 1.15.0 with a backup motor for this test. Train, a particular basic boundary measure has to be described. such as layer esteem, ages, cluster size, comprehension pace, and so forth Such boundary preparation is all dependent. Reducing bad luck when traveling is essential. We used the " rmsprop " analyzer in this test to reduce mishaps and enhance the model. During testing, a well-prepared model may release the hitter. Large PC setup necessitates top-to-bottom learning model information preparation. In that regard, the GPU is helpful.

Shortening, which abstractive papers is a severe problem in the world of natural language processing. It is far more difficult for people to generate a region's text and synopsis from within. Given that the machine makes the most it can give its limitations. After pre-processing, the computer must be instructed to learn utilizing the data model. For every training, the model has a backend engine. For this test, we employed TensorFlow to finish the assignment. The initial values are nearly finished. Epoch, retain possibility, run measurement, batch count, number of layers, etc. are a few examples. There is now less time required for training data. The optimizer is the "rmsprop" within this particular instance for model adjustment. It is necessary to facilitate higher configurations PC data training. Lastly, we use Google Colab to train the model. It saves a lot of time and operates far faster.

The parameter's value is:

- Numbers of Epochs = 70
- Defines Batch size = 512
- Set number of neuron layer = 3
- Define model Learning rate = 0.001
- Keep the verbose value=1

4.2 Experimental Result

We are aware that no technology can produce perfect results. The yield provided by the tool is nearly accurate. Everyone is aware that no tool can provide an exact result of 100%. In essence, our prepared approach produces amazing outcomes without incurring any costs. It reacts to mistaken chemical associated with the drug sometimes. Either way, the content definition and the optimal amount of answer words are the same. After 68 epochs of training, we lower the loss, which is 0.001 then the model run early stop before 70 epochs. We save the latest version in a document called "model.json" so that we can examine the output. We then create the following JSON form to obtain the weights in the format "model.h5" so that we may save the model evaluation. Next, we establish a TensorFlow consultation in order to reload the graph that was stored in the earlier stages. Next, in order to test their summary, construct the text and summary data body at random. Next, we translate the value into the language needed for the sequence that was going to be the model's input. The logical abstractive summarizer is then created using the seq2seq model using the encoder decoder. In the past, we developed a logistic element to provide a logical response answer. A person provides the true synopsis for the associated material. Following preparation of the raw textual material, the text is converted into input phrases that are pure text. Using the gadget, the final variable response word was provided following the instruction and research.

A sentence will be chosen at random as the input from the dataset to construct the summary, and the summary length will be determined accordingly. This is the computer's satisfactory response after some brief training using the model and datasets.

Table 4.1: Some sample predicted summary

Original Review	**Predicted Review Response Summary**
awesome mayonnaise lot pepper flavor great sandwiches dips spreading chicken baking one new favorites	great product
chance try warmed still loved straight package bar soft moist great texture seem like eating prepackaged quaker bar seemed like eating something fresh bakery cafe def repurchasing trying warmed	great snack
fan dark chocolate find bars bitter newman bar cocoa strong chocolate flavor without bitter harsh one ounce bar calories flavor rich enough need chocolate fix need one square calories makes satisfying treat without diet texture smooth somewhat creamy although creamy dove dark chocolate would definitely purchase newman chocolate bars	dark chocolate dark chocolate
bar great right amount sweetness people work begging bars thanks making something wholesome tastes great	great taste
tried vita coca zico first like taste zico better also saw review coconut drink products tv zico product met claims others fell short zico	Coconut water
stopped buying grocery store cat food several experiences happened nearly learned contained learned additives food effect animals organs learned recall foods china killing pets nationwide learned national recall lists pet foods kinds foods products online everyday week made switch organics ingredients labels make smile instead foods expensive love pets buying peace mind continued health	my cat loves it

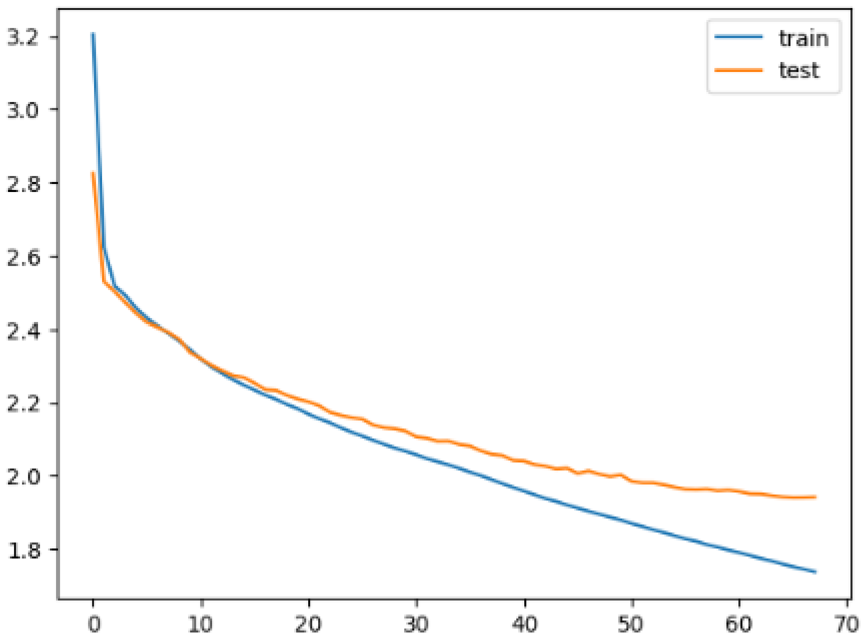


Fig 4.1: Train & Test Validation loss

4.3 Discussion

We have developed a model for abstractive text summarization in English. For different scenarios, the result from our model is superior. We made this in order to subtract the function loss. To lower the mistake rate, a learning model is employed. For chain data, the function of loss deduction is crucial. We implemented the loss mechanism for val_loss reduction to 0.001 during the data training after 68 epochs which have been . Our methodology generates a training loss earlier in the training process. Losses start to decline at a startling rate after a while. The learning value is 0.001. This section covered our model's testing. In addition, the machine responds by creating a blueprint. The horrible overview of machine reaction is quickly followed by a detailed discussion of the aforementioned.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Everything has become sophisticated and contemporary these days. where time is yet another essential component. The use of technology has made life quicker. Most individuals like reading news items, internet articles, short novels, and news periodicals in their everyday lives. Usually, individuals read those books or articles when they have enough free time. However, they never have sufficient time to devour the news, stories, or articles in their entirety. They then wish they had a brief synopsis of that subject. They will find it easier to read thanks to this text summary, and they will like how quickly they will receive the brief summary. The results of this study help people from all social classes. For example, journalists, news readers, businesspeople, and students.

Rapid access to and comprehension of information is facilitated by summarization tools. This is especially helpful in the current era of information overload, when professionals and individuals must sort through massive amounts of text in order to uncover pertinent information. Long texts are distilled into brief summaries, saving time. When reading an article, study paper, newspaper story, or another kind of writing, this may be quite helpful for those who need to rapidly understand its main elements. In order to help language learners understand and pick up new terms and grammar rules more quickly, summarization programs offer shortened versions of texts. With the use of automated summarizing, content producers and authors may produce succinct summaries of papers or articles that can be utilized as short synopses or meta descriptions.

5.2 Impact on Environment

When compared to certain other technical processes, English text summarizing is typically thought to have a negligible effect on the environment.

- **Consumption of Energy:** Text summarization may require the development and training of sophisticated NLP (natural language processing) models, which can be computationally demanding. Energy consumption for training big systems on powerful equipment might be high. The inference stage, which creates summaries from text, usually requires fewer resources after the models are trained.

- **Data Center Management:** Numerous services and methods for text summarization run on servers housed in data centers. These data centers need cooling systems and power to function. Businesses and institutions who host these kinds of activities are becoming more conscious of their environmental effect and are attempting to use renewable energy sources and increase energy efficiency.

- **Model Dimensions and Performance:** The environmental effect of models based on natural employed in text summarization might vary depending on their size. More computer resources can be needed for inference and training on larger models. Experts are hard at work creating more effective models that strike a balance between environmental factors and performance.

It's crucial to remember that text summarization's environmental effects are only a small part of the larger effects that technology has on our surroundings. Developers and companies are striving to implement sustainable practices, such as energy-efficient equipment, algorithms, and data center operations, as they become more conscious of these factors.

5.3 Ethical Aspects

Like any technology, text summarizing presents a number of ethical questions. The following are some important moral considerations when summarizing an English text:

- **Summarization Bias:** Biases in the training set of data may unintentionally be reflected and reinforced by summarization models. A biased summary may be produced by the summarizing model if the training data includes biased terminology or opinions. Bias in trained data and model results must be addressed and minimized.

- **Effect on the Quality of Information:** The initial meaning and context of the content may be changed by automatic summary. Misinformation or misunderstanding may result if the summarization model misrepresents information or misses subtleties. Developers should aim for high precision while gathering important data.

- **User Awareness and Informed Consent**: It's possible that users aren't completely aware that the material they're reading has been distilled. Respecting user autonomy necessitates informing users about the usage of summarizing and being transparent about the process of creating summaries.

The ethical ramifications of text summarization technology are becoming more widely recognized, and developers and organizations are dedicated to resolving these concerns in a responsible manner. The development and implementation of ethical summarization systems require constant communication and cooperation between members of the AI community as well as the participation of a wide range of organizations.

5.4 Sustainability Plan

Creating a sustainable strategy for English text summarizing entails implementing procedures that give ethical issues, resource conservation, and environmental effect mitigation first priority. Key elements of a sustainability strategy include the following:

- **The Efficiency of Energy:** Give top priority to the creation and use of text summarization models that use little energy. Reduce the amount of computing power needed for inference and training via optimizing hardware, software, and algorithms.
- **Computing Efficiency:** Take into account using servers and data centers that use less energy while hosting summarization services. Examine how data centers might be powered by sources of clean energy to lessen their environmental impact.
- **Openness and Responsibility:** Be open and honest about the sustainability of AI development processes. Give details on model efficiency, data center procedures, and energy usage. Create accountability systems to deal with issues related to sustainability.

CHAPTER 6

Conclusion and Future Research

6.1 Summary of the Study

The foundation of this project is English NLP. Our goal in this research is to create a model that may be used to abstractively summarize abstract literature in English. English material may be automatically summarized using this method. The entire project was completed in less than half a year. The research and project efforts are divided into many sections. Here is a detailed overview of the project's main outline.

SL. No.	Steps
1	Collecting Kaggle data: amazon product review.
2	Data Preprocessing
3	Text Cleaning
4	Add vocabulary size
5	Word coverage by model
6	Count unique size
7	Add Special Token
8	Add attention layer file
9	Encoder and Decoder with LSTM model
10	Build the Sequence to Sequence model
11	Model Train
12	Get the Review text summarizer

In the future, our proposed model will assist all scholars in creating an automatic abstractive text summarizing system in English. Let's now wrap up our report by talking about our conclusions and next steps.

6.2 Conclusion

We now receive a lot more data than ever before, and we must process it rapidly. For this reason, text summary is necessary in order to quickly comprehend the vast quantity of data. We have two options for text summarization: extractive and abstractive. Because abstractive summarizing requires a deeper understanding of the text's meaning and is more sophisticated than traditional summarization, it can yield a summary that is more cohesive. The summarizing system emphasizes the requirements of the developers designed for easier comprehension. In order to manage enormous volumes of information and produce an associate degree-correct outline, abstractive text summarization is significantly more accurate. To achieve this, one must search for the most often used or common words, reduce their size, and then display freshly created sentences. Abstractive text summarization, on the opposite hand, enhances and clarifies the facts while creating a phrase. Whereas abstractive text summarizing entails coming up with fresh words and phrases that encapsulate the original text's content, abstractive text summarization selects and combines significant phrases or words from the source material to produce a summary. We bring to you a summary that might help you save time and comprehend the content or news simply, thanks to the aid of LSTM and the Sequence to Sequence algorithm. To create the summary, we may comprehend the material using Natural Language Processing (NLP). This applied model for making a predicted summary of the review of amazon product review.

6.3 Possible impacts

English text summary has a big influence across a lot of different areas. Summarization enables readers to rapidly understand a document's main ideas while requiring to read the full text. This is especially helpful in situations where time is of the essence, such while reading the news or conducting research. It increases the audience for whom information is more accessible. Long pieces of writing or papers can nevertheless be insightful for people with short attention spans or limited time. Summarization tools are useful for professionals in a variety of disciplines, such as business and law, since they facilitate the efficient reading and understanding of papers, which in turn helps with decision-making. Language learners can benefit from summarization tools since they offer reduced versions of texts, which facilitate comprehension and help them pick up new terminology.

6.4 Implications of Further Study

Additionally, there are a few limitations with the model. Every research project undergoes constant change; thus, we have a philosophy for how we want to evolve going forward. We have the opportunity to improve coordination and also encourage accuracy by employing this process. To strengthen the concept of the result and to render this task more evident, we would want to include more models. The framework is good at handling a lot of data and providing it precise once-overs, so even if the quantity of information available for our evaluation is restricted, we definitely want to examine further down the road with endless data. Once our survey is complete, we would really want to develop a handy, web-based application that makes use of artificial intelligence. Additionally, this app will provide text summaries in English. Additionally, the program would have the capability to function as an additional security measure by employing natural language processing (NLP) techniques to detect potentially retaliatory actions or phrases within the entered English text. We will design and evaluate our application's display as part of the test using both guided and unassisted learning approaches. This investigation will evaluate the formatting of both controlled and independent AI models, helping us determine which is most appropriate for the English we use for message outlines. There are several restrictions on the model. Since research projects often undergo constant change, we will eventually need to include these modifications into our model updates.

- Big amount of dataset;
- Additional sequences added;
- Research scope increased.
- There are no word count limitations.

We intend to create a mobile or online application that will be accessible from any location. We also have other plans, including developing a website and a mobile app, to make the study easier to access when it is finished.

Reference

[1] 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.

[2] 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.

[3] 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.

[4] Yeasmin, S., Tumpa, P.B., Nitu, A.M., Uddin, M.P., Ali, E., Afjal, M.I, "Study of abstractive text summarization techniques," American Journal of Engineering 6(8), 253–260 (2017)

[5] Tan, Jiwei et al. "From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach." International Joint Conference on Artificial Intelligence (2017).

[6] N. Dhar, G. Saha, P. Bhattacharjee, A Mallick, and M.S. Islam et al., "Pointer over Attention: An Improved Bangla Text Summarization Approach Using Hybrid Pointer Generator Network" 2021 24th International Conference on Computer and Information Technology (ICCIT), 202

[7] S. Banerjee, S.K. Naskar and S. Bandyopadhyay, Bengali named entity recognition using margin infused relaxed algorithm, in Text, Speech, and Dialogue (P. Sojka, A. Hora´k, I. Kopecˇcek, and K. Pala, eds.), (Cham), pp. 125–132, Springer International Publishing, 2014.

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

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

[10] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. 2016 Sep 26

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

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

[13] Wang, Li, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization." arXiv preprint arXiv:1805.03616 (2018).

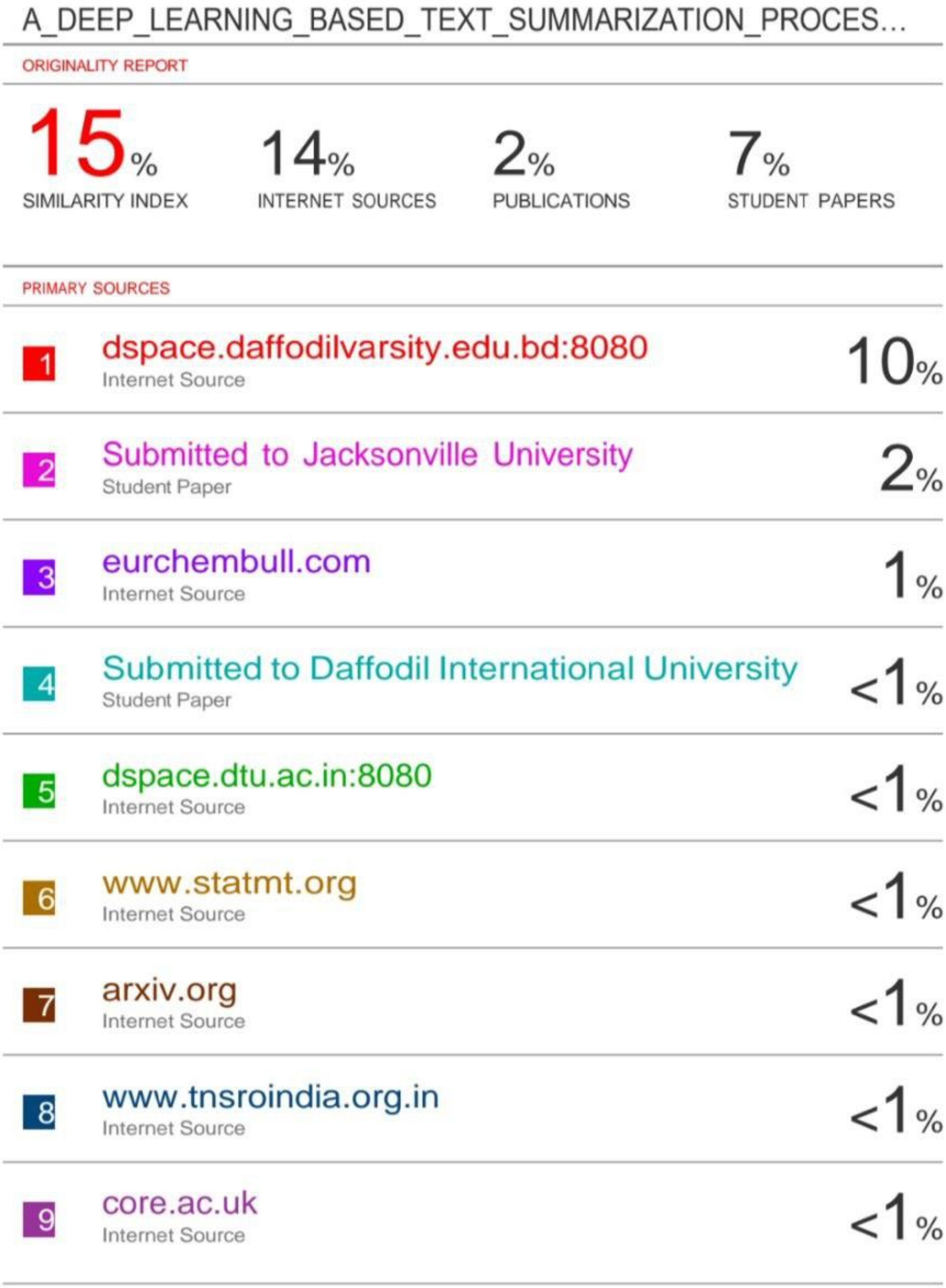
[14] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

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

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

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

Plagiarism Report:



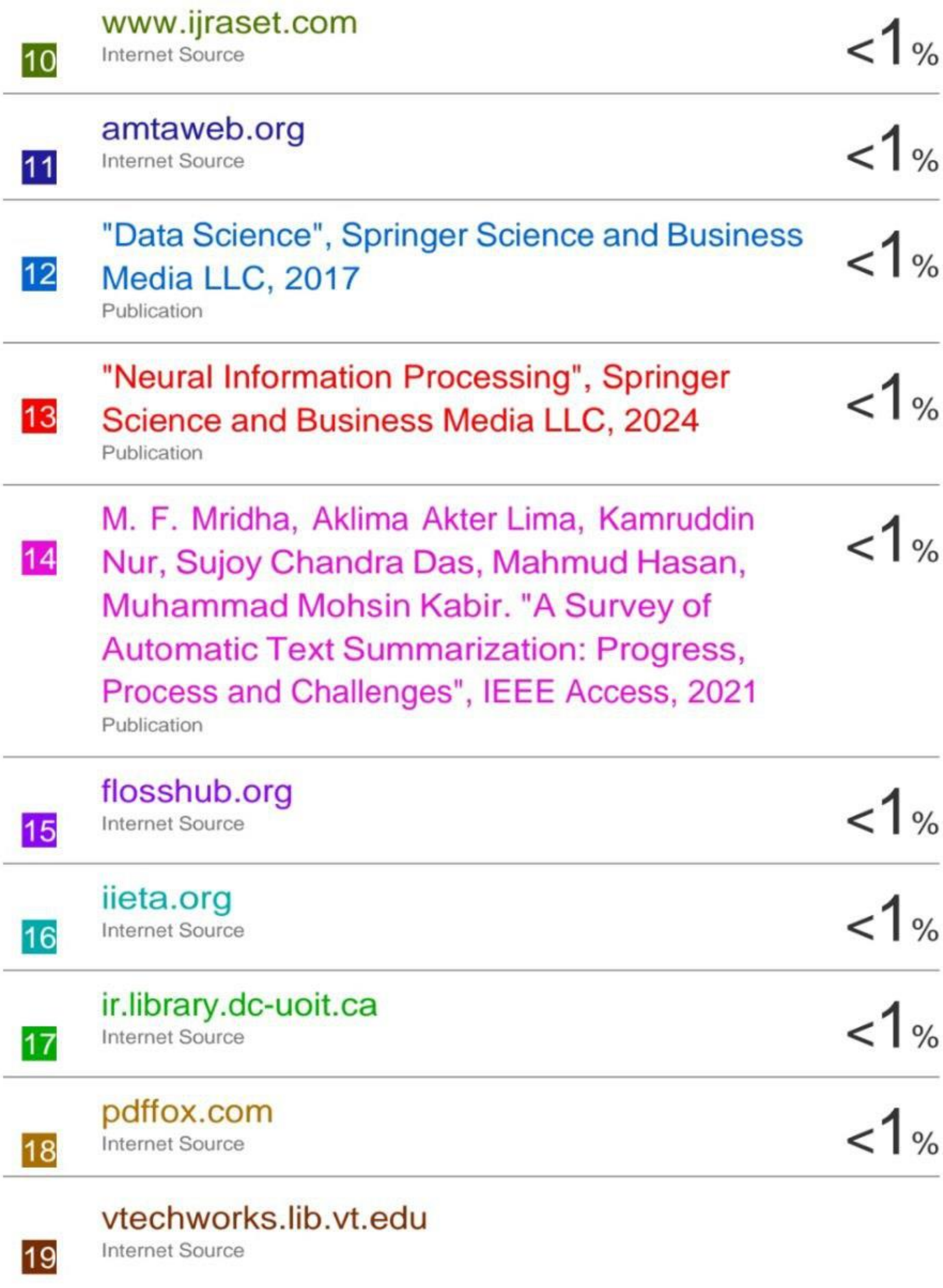
A_DEEP_LEARNING_BASED_TEXT_SUMMARIZATION_PROCES...

ORIGINALITY REPORT

15%	14%	2%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	10%
2	Submitted to Jacksonville University Student Paper	2%
3	eurchembull.com Internet Source	1%
4	Submitted to Daffodil International University Student Paper	<1%
5	dspace.dtu.ac.in:8080 Internet Source	<1%
6	www.statmt.org Internet Source	<1%
7	arxiv.org Internet Source	<1%
8	www.tnsroindia.org.in Internet Source	<1%
9	core.ac.uk Internet Source	<1%



10	www.ijraset.com Internet Source	<1%
11	amtaweb.org Internet Source	<1%
12	"Data Science", Springer Science and Business Media LLC, 2017 Publication	<1%
13	"Neural Information Processing", Springer Science and Business Media LLC, 2024 Publication	<1%
14	M. F. Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, Muhammad Mohsin Kabir. "A Survey of Automatic Text Summarization: Progress, Process and Challenges", IEEE Access, 2021 Publication	<1%
15	flosshub.org Internet Source	<1%
16	iieta.org Internet Source	<1%
17	ir.library.dc-uoit.ca Internet Source	<1%
18	pdffox.com Internet Source	<1%
19	vtechworks.lib.vt.edu Internet Source	

		<1%
20	www.arxiv-vanity.com Internet Source	<1%
21	www.pet-info.org Internet Source	<1%
22	Nobel Dhar, Gaurob Saha, Prithwiraj Bhattacharjee, Avi Mallick, Md Saiful Islam. "Pointer over Attention: An Improved Bangla Text Summarization Approach Using Hybrid Pointer Generator Network", 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021 Publication	<1%
23	Xingxing Ding, Ruobing Wang, Zhong Zheng, Xuan Liu, Quan Zhu, Ruiqun Li, Wanru Du, Siyuan Shen. "DoS: Abstractive text summarization based on pretrained model with document sharing", 2022 4th International Conference on Intelligent Information Processing (IIP), 2022 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		