RESEARCH ARTICLE OPEN ACCESS

Aautomatic Text Summarization Using Deep Learning and NLP Model

Omprakash Choudhary*, Himanshi Nehra*, Shweta Saraswat, Kiran Ahuja****

*Department of Computer Science and Engineering, Arya Institute of Engineering & Technology Rajasthan, India

*Department of Computer Science and Engineering, Arya Institute of Engineering & Technology Rajasthan, India

****Assistant Professor, Department of Computer Science and Engineering, Arya Institute of Engineering & Technology Rajasthan, India

**Assistant Professor, Department of Computer Science and Engineering, Arya Institute of Engineering & Technology Rajasthan, India

ABSTRACT

In this new era of Technology, there is so much information on the internet. It is difficult for people to manually extract the summary of a large written document. Approximately 2.5 quintillions of data are generated every day on the internet. As a result, extracting important information from a large number of documents available is very difficult.

To solve the problem automatic text summarizing is required. Text summarizing is the process of picking the most important and meaningful information from a document or group of texts and summarizing it into a simpler form while keeping the general meanings.

Keywords- Text, Machine Learning, Neural Network.

I. INTRODUCTION

For a decade most of the report has been tired a manual manner. within the modern times,the amount of data will increase speedily in each means that over the net and from each possible supply. because the net and massive information have augmented in quality, individuals have gotte overwhelmed by the large quantity of data and documents on the market on the net. As a result, several lecturers area unit inspired to produce a technological resolution that may mechanically summarize texts. Summaries created by automatic text report embrace all crucial information from the initial material in addition as key sentences. As a result, the knowledge is delivered quickly whereas still meeting the document's original goal. Text report has been studied since the mid-twentieth century, with LUN (1958) being the primary to use a applied math approach referred to as word frequency diagrams to overtly discuss it. several different approaches are developed to this point. counting on the amount of documents, there area unit single and multi-document summary decisions. within the meantime, counting on the kind of method report follows it has two strategies theoretic and extractive text report. report systems sometimes have access to extra proof so as to search out the foremost vital document themes. When summarizing blogs, for instance, arguments or comments that follow the weblog post may well be helpful in evaluating that element of the weblog area unit as crucial and intriguing. Scientific article summaries contain a considerable quantity of data, like documented papers and conference information, that may be utilized to stress key sentences within the original study. The sections that follow check a number of the context in additional detail.

Text Summarization is classified into two types which are:

1. Abstractive Summarization

2. Extractive Summarization

1.1 Extractive Text Summarization -

The extractive text summarisation involves pull key phrases from the supply document and mixing them to form a outline. we have a tendency to establish vital words or phrases from the text and extract solely those for the outline.

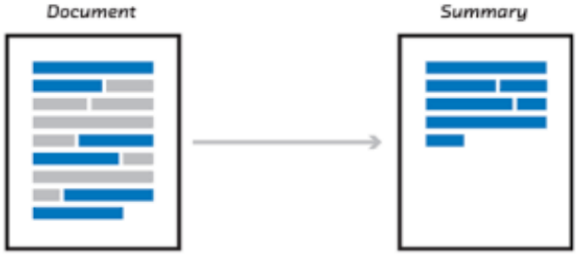


Fig -1: Extractive Text Summarization

1.2 Abstractive Text Summarization

The Abstractive text summarisation will produce new phrases and sentences that relay the foremost helpful data from the initial text. The sentences generated through this technique might not be a gift within the original document.

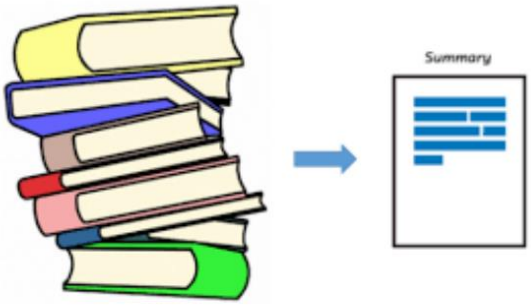


Fig -2: Abstractive Text Summarization

II. IMPLEMENTATION

We are implementing an abstractive summarization method using a deep learning technique called long short-term memory (LSTM). This is a type of recurrent neural network that can handle input sequences of any length and maintain an internal state to remember relevant information from the past. We will be using the CNN_DailyMail dataset, which contains news articles and their corresponding summaries from the CNN and Daily Mail websites, to train and evaluate our LSTM model. The model will use an encoder-decoder architecture, where the encoder processes the input article and converts it into a fixed-length representation, and the decoder generates the summary one word at a time using the encoded representation as context. To train the model, we will need to preprocess the data, extract the articles and summaries, and convert the text into numerical representations that can be input to the model. Then, we will use a supervised learning approach, where we provide the model with pairs of articles and summaries and optimize its parameters to minimize the difference between the predicted and ground-truth summaries.

2.1 Data

The data used is the CNN_dailymail dataset. it's 2 features: article and highlights. The article includes the document that's to be summarized. it's the news story. Highlights square measure the headlines of the corresponding news that square measure used as summaries.

2.2 Method

The model used is the abstractive methodology that enforced victimization deep learning techniques.

2.3 Algorithm

The rule used is that the LSTM or Long STM model that could be a variety of perennial Neural Network model.

2.4 Model

The model used is sequence to sequence model. Sequence-to-sequence learning may be a coaching model which will convert sequences of 1 input domain into the sequences of another output domain. it's usually used once the input and output of a model are often of variable lengths.

III. DATA PREPROCESSING

Performing basic pre-processing steps is extremely necessary before we have a tendency to get to the model building half. victimisation untidy and uncleaned text knowledge could be a doubtless black move. So, we are going to drop all the unwanted symbols, characters, etc. from the text that don't have an effect on the target of our downside. we are going to perform the below preprocessing tasks for our data:

- Convert everything to minuscule
- Remove (‘s)
- Remove any text within the parenthesis()
- Eliminate punctuations and special characters
- Remove stop words
- Remove short words

IV. METHODS

4.1 CleanData() it's wont to clean the info by victimisation preprocessing steps mentioned earlier.

4.2 BuildDataset() it's wont to build train and check information sets

4.3 BuildDict() it's wont to build wordbook wherever keys square measure words and values square measure random and distinctive numbers. It conjointly builds another wordbook with keys as distinctive numbers and values as words. These square measure utilized in tokenization of words so the input to the model could be a set of numbers instead of words thus on build the computation easier victimisation vectors.

4.4 Tokenize () it's wont to tokenize the info and send it to the model. Tokenizing information is vital because the networks want numerical information to figure on instead of data with characters.

V. ALGORITHMS

5.1 Graph based Algorithms:

The graph-based approach to text summarization is so Associate in Nursing unsupervised technique there in we have a tendency to use a graph to rank the mandatory sentences or words. the essential goal of the graphical technique is to extract the foremost necessary sentences from one document. during a outline, we have a tendency to verify the importance of a vertex during a graph. Text-based ranking is accomplished victimisation simplex and weighted graphs. Either the documents or the statements ar bestowed as nodes during this technique. Edges ar accustomed link any 2 nodes that share identical information. The data formatting of weightings to the nodes of the graph is employed to get sentences. alternative text summarization ways employing a graph-based approach include:

1)Page rank:

It is a Google algorithm that controls the connection of web sites with comparable content.

2)Lex rank:

This rule makes use of circular function similarity and tied models. circular function similarity may be a live of however similar 2 vectors area unit in associate degree n-dimensional house. It evaluates the circular function similarity of 2 vectors and sees whether or not they area unit heading in nearly an

equivalent general direction. It's typically utilized in text analysis to see document similarity.

3)Text rank:

It is an extension to the page rank algorithm and like the Lex rank algorithm to normalize the data.

5.2 Neural Networks algorithms:

Humans won't rethink their thoughts each second. As you browse this text, you grasp every term supported the understanding of previous statements. you do not begin over from the start and throw everything out. Your ideas area unit powerful. one thing that typical neural networks area unit incapable of, that seems to be a good weakness. contemplate the case below: you would like to classes numerous} sorts of events that happen in a very moving-picture show at various moments. It's unknown however a traditional neural network may use previous moving-picture show events to tell future ones.

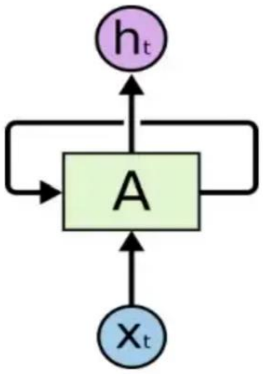


Figure 3- RNN Looping Structure

Recurrent neural networks are accustomed solve this drawback. they are networks with constitutional cycles that keep knowledge endure. within the diagram on top of, a slice of a neural network, A, evaluates a group of inputs interference and returns a price ht. A loop is accustomed communicate knowledge from one network section to subsequent. attributable to the loops, RNN appearance mysterious. Yet, after you contemplate it, they don't seem to be very all that dissimilar from a perceptron. A perennial neural network is created from many copies of identical network, every of that sends the message to a different in line. Consider what happens if you unwind the loop:

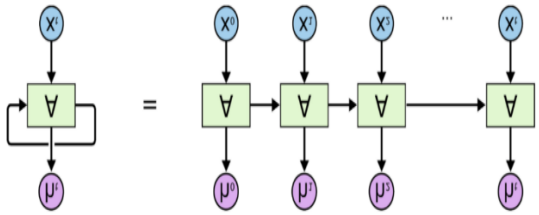


Figure 4- RNN

One of the benefits of RNNs is its ability to integrate earlier info to this task, like employing a previous video sequence to assist recognize this screen. to complete a given task, we tend to typically solely have to be compelled to look at recent information. contemplate a learning algorithmic program that tries to guess the words supported those that have precede it. we do not would like way more info to work out however the last word in "the clouds area unit already within the sky" means that - it's clear that ensuing word are sky. In things wherever the space between relevant info and also the location wherever it's required is little, RNNs will learn to leverage previous information.

However, there'll be some occasions once we need additional info. Take the text's last sentence: "I grew up in France... i'm skillful in French." consistent with current findings, ensuing word would possibly be a language name, and we'll would like additional info for France from the past to slim down that language it'll be. It's entirely attainable that the time separating relevant information and once it's needed might increase exponentially. RNNs, on the opposite hand, lose their capability to find out to attach the dots because the gap expands.

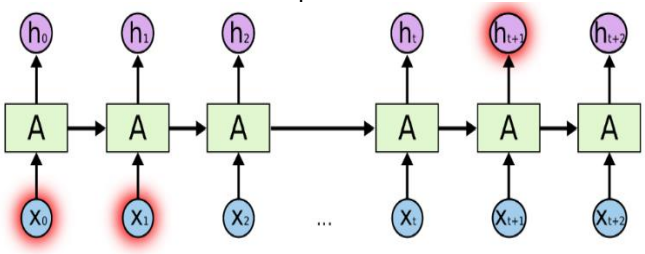
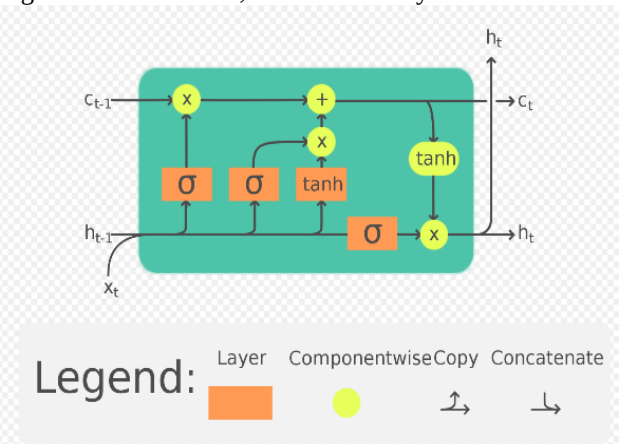


Figure 5 –LSTM Architecture

LSTMs, or Long remembering networks, ar a variety of Rnn which may learn long-run dependencies. They were 1st introduced by Hochreiter & Schmidhuber (1997), and that they have since been refined and popularized by many folks. they are being employed a lot of and a lot of oftentimes, and that they ar improbably helpful during a type of things Lstm networks were created expressly to handle the problem of long-run reliance. They should not have tried extraordinarily arduous to recall data for extended periods of time; it ought to be a state for them! A series of recurrent neural network modules helps to compensate each rnn. This continuation module in typical RNNs can have a comparatively straightforward structure, like one tanh layer.



LSTMs, on the opposite section, utilize multiplications with additions to work out little changes in information. In LSTMs, flow of knowledge through a mechanism known as cell states. Lstm networks have the flexibility of basic cognitive process or not basic cognitive process info during this manner. There seem to be 3 distinct reliance's on info at a given cell state. We'll offer an Associate in Nursing example as an example. Take, for instance, projected stock values for a selected stock. Today's stock worth is determined by:

1. the stock trend of a specific stock from the past flows is also an Associate in Nursing uptrend or downtrend within the gift state of affairs.
2. price of the stock may additionally vary from the previous days.
3.factors that have an effect on worth fall or up embody company policies and selections that ar created on the management level.

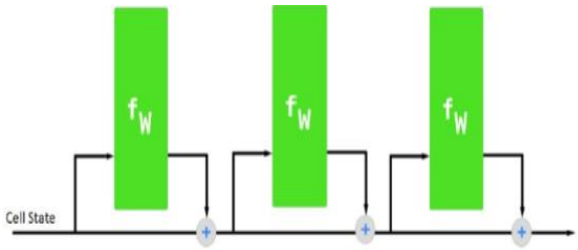


Figure 6 – LSTM convey belt

The most necessary feature of LSTM is convoy belt. Cell state machine is said to a conveyor that transmits knowledge for the duration of the cell. whereas it's not specifically a entrance, it's essential for knowledge to flow through every cell and to different cells. in line with the findings of the forget and input gates, the information passing through it's changed and updated before being passed to successive cell.

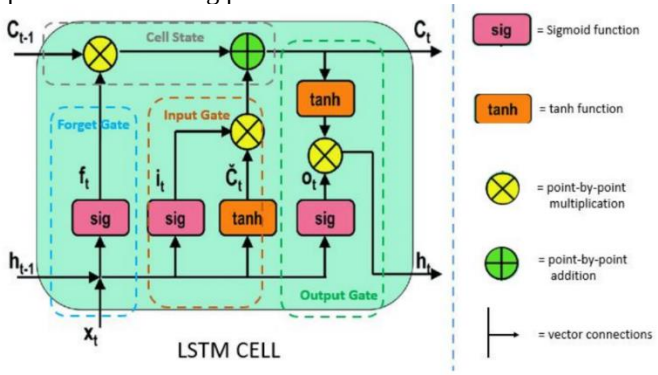


Figure 7 – LSTM Cell

Forget Gate:

Just like humans who are not to think about a number of the events and things in their everyday life activities forget gate also remove unwanted data from merging it with the cell states. It takes two inputs (xt and ht-1) and sends it to sig gate which removes the unwanted data from computer file data and sends it to the multiplication methods.

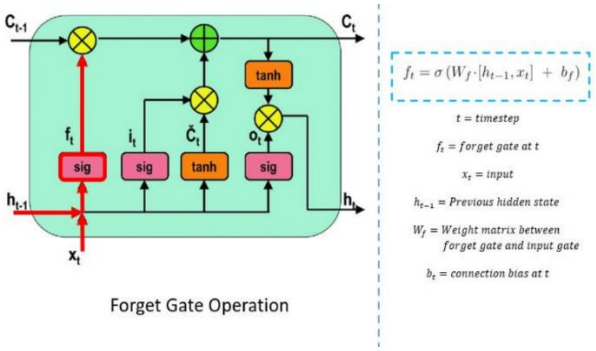


Figure 8 – Forget Gate LSTM

Output Gate:

Based on cell condition, previous cell outputs, and new knowledge, the penultimate gate selects helpful info. It achieves this employing a tanh perform to form a vector from of the cell state when the inputs and forget gates have integrated. The new input and previous cell output area unit then run through a sigmoid perform to judge that values should be outputted. The output of this cell is increased by the results of these 2 procedures.

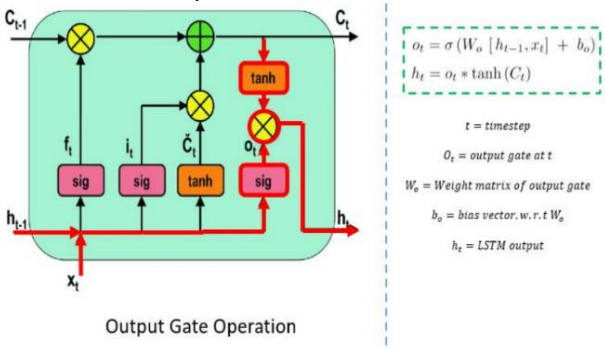


Figure 9 - Output Gate Operation

5.3 Bi-directional LSTM:

The method of permitting any neural network to store sequence knowledge in each backwards and forward directions is thought as bi- directional long-short term memory (bi-lstm). A bidirectional LSTM differs from a traditional LSTM therein its input is split into 2 channels. With a traditional LSTM, we have a tendency to could build input travel in one direction, either forwards or backwards. We can, however, have flow of data in each directions with bi-directional input, storing each the longer term likewise because the past. Let's take into account the instance for a larger Understanding. The expanse within the sentence "boys attend..." cannot get completed. Still, we are able to merely forecast the previous expanse and have our model do constant factor whereas we've got a forthcoming sentence like "boys begin of college," and bidirectional LSTM permits the neural web to figure.

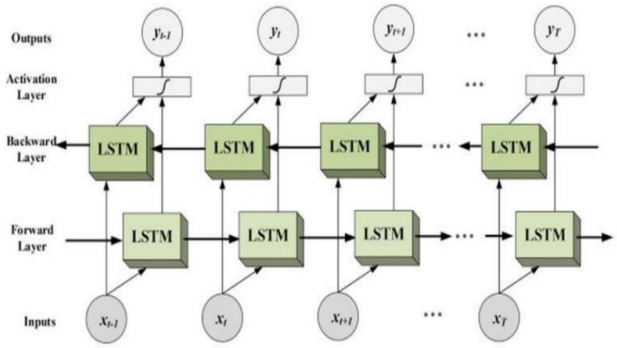


Figure 10 – Bi-Directional LSTM

VI. MODEL

Sequence-to-sequence (Seq2Seq) models are a type of machine learning model that can transform sequences of data from one domain (input) to another (output). They are often used when the input and output sequences of a model have variable lengths. Seq2Seq models consist of an encoder and a decoder, which work together to map the input sequence to the output sequence, using a special token and a focus value. The model uses long short-term memory (LSTM) units to try to predict the next state sequence based on the previous sequence.

6.1 Encoder-Decoder architecture:

The Sequence-to-Sequence model is a type of artificial intelligence that uses an encoder-decoder design to predict sequences when the length of the input and output sequences can vary. The encoder processes the entire input sequence and generates a fixed-length context vector representation that captures the entire context of the input sequence. The decoder network then uses this context vector to generate the output sequence, until it reaches the end of the sequence token.

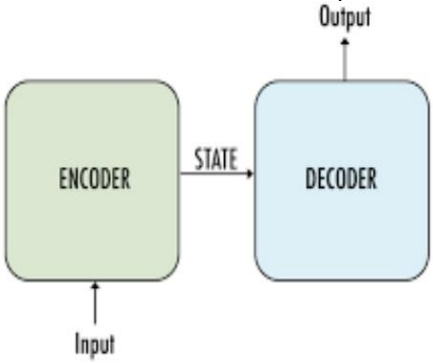


Figure 11 - Encoder Decoder

6.2 Encoder

The encoder in a Sequence-to-Sequence model is an LSTM network that reads the entire input sequence one word at a time. At each time step, it processes the input and captures the context and key information about the input sequence. It produces a hidden state output (h) and cell state (c) for each input word. The hidden state (h) and cell state (c) at the last time step are the final internal representation of the input sequence, which is used to initialize the decoder. The decoder then uses this information to generate the output sequence.

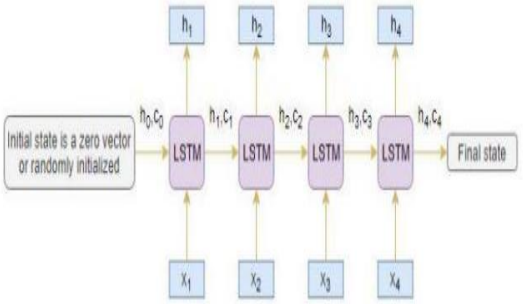


Figure 12 – Encoder

6.3 Decoder

In an abstractive summarization model using an encoder-decoder architecture, the decoder is a neural network that generates the summary one word at a time based on the encoded representation of the input document. The decoder is typically implemented using a long short-term memory (LSTM) network, which allows it to remember important information from the past as it processes the sequence. To help the decoder generate the summary, two special tokens are added at the beginning and end of the target (output) sequence. The decoder then reads the entire encoded representation, one word at a time, and uses it to predict successive words in the output sequence. The decoder is trained to predict the next word in the sequence based on the encoded representation and the previous words in the output sequence.

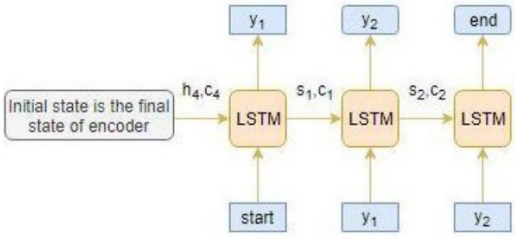


Figure 13 - Decoder

The above architecture of the model is built using the TensorFlow library which is used to build layers in neural networks. the ultimate design of the model are going to be as shown below

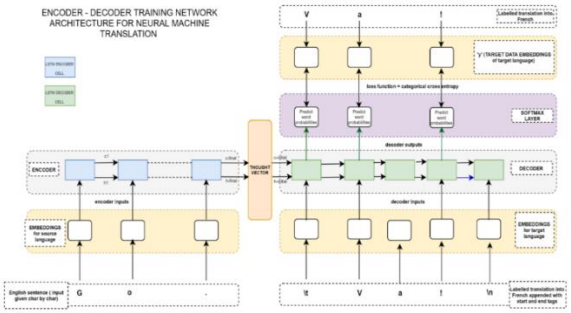


Figure 14 - LSTM Seq2Seq model architecture

VII. RESULTS AND DISCUSSION

7.1 Streamlit Application:

Streamlit is one of the most important and easiest ways to implement a website using python which usually doesn't require to integrate with the HTML/CSS part of the code. Most of the blocks that are required for building the website are predefined in the streamlit application.

7.2. Summarizing Using URL

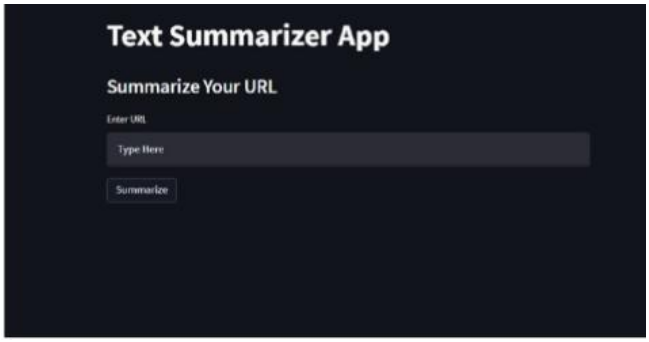


Figure15 - Home Page

When we enter any online page URL that supports markup language computer program then entire content of that page are extracted and summarized.

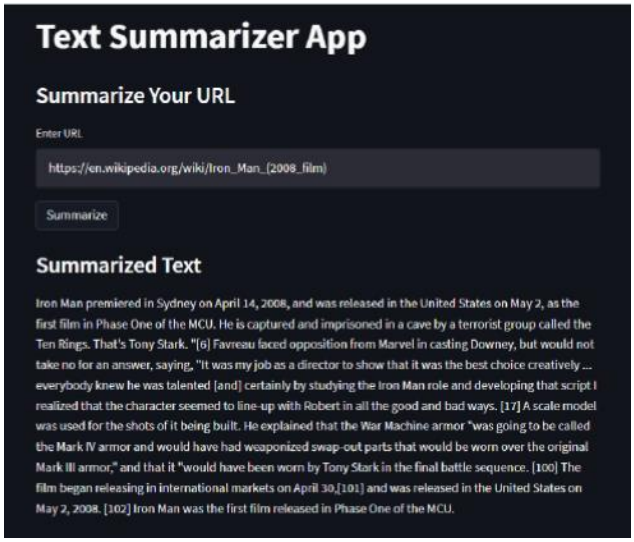


Figure 16 – Input URL

7.3 SUMMARIZING USING TEXT:

INPUT:

Figure 17-Input

OUTPUT:

SUMY LEX RANK:

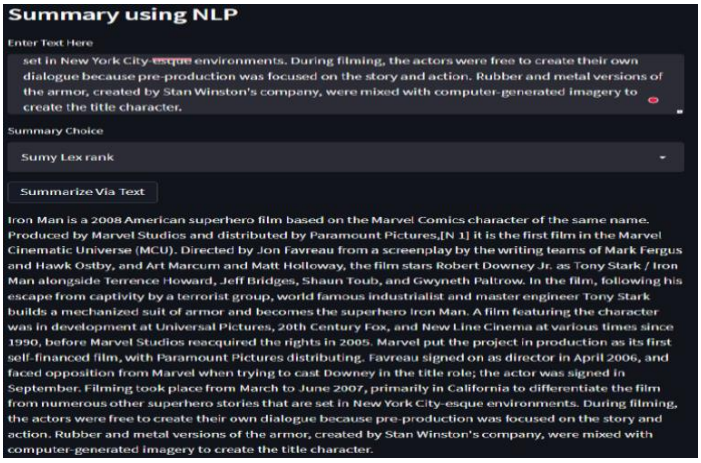


Figure 18-Output

7.4 Using LSTM

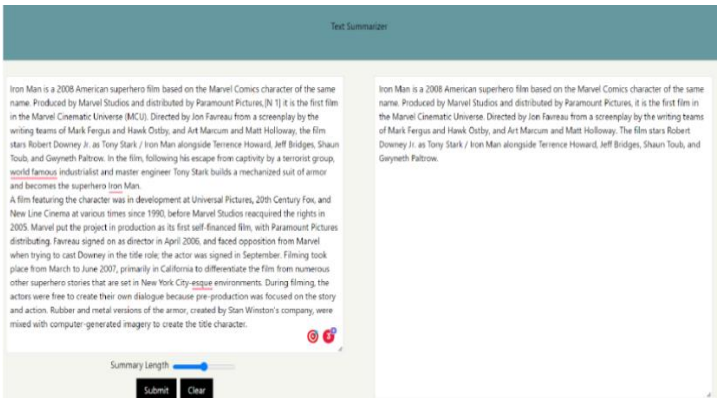


Figure 19-Input Using LSTM

VIII. CONCLUSION

The rapid growth of the internet has led to an enormous amount of data being available, making it difficult for humans to summarize large amounts of text. This has created a strong demand for automatic summarization tools in the age of information overload. According to the International Data Corporation, the total amount of digital information produced annually around the world is expected to increase from 4.4 zettabytes in 2013 to 180 zettabytes in 2025. This is a huge amount of data, and there is a pressing need for algorithms that can be used to automatically reduce the amount of information by generating accurate summaries that capture the essence of the intended messages. In addition, applying text summarization can reduce reading time and speed up the process of searching for information, playing a crucial role in the fast-paced, digital world we live in. While humans are generally good at summarizing text, automatic text summarization is essential in today's world, where there is an abundance of data and a lack of time and resources to interpret it. There are two main approaches to automatic text summarization: extractive and abstractive. Abstractive summarization, which involves generating a summary that is closer to human comprehension and may involve rephrasing or condensing the original text, is a challenging area of study that focuses on providing a summary that is more relevant, accurate, content-rich, and less repetitive. This survey aims to examine the various approaches to abstractive summarization, as well as their advantages and disadvantages, and provides a useful overview of this field of study.

REFERENCES

[1] Kiran Ahuja, Harsh Sekhawat, Shilpi Mishra, Pradeep Jha (2021). Machine Learning in Artificial Intelligence: Towards a Common Understanding. Turkish Online Journal of Qualitative Inquiry, 12(8):1143-1152.

[2] H. Arora, M. Kumar, T. Rasool and P. Panchal (2022) Facial and Emotional Identification using Artificial Intelligence. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 1025-1030.

[3] Dr. Himanshu Aora, Kiran Ahuja, Himanshu Sharma, Kartik Goyal, Gyanendra Kumar (2021). Artificial Intelligence and Machine Learning in Game Development. Turkish Online Journal of Qualitative Inquiry (TOJQI), 12(8):1153-1158.

[4] Abhinav Agarwal, Himanshu Arora, Shilpi Mishra, Gayatri Rawat, Rishika Gupta, Nomisha Rajawat, Khushbu Agarwal (2023). Security and Privacy in Social Network. Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and Computing 1432:569577.

[5] S. Mishra, M. Kumar, N. Singh and S. Dwivedi (2022). A Survey on AWS Cloud Computing Security Challenges & Solutions. IEEE 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 614-617.

[6] Shweta Pachauri, Deeksha Sharma, Dr. Rahul Misra (2022). Role of Computer Education in Indian Schools. International Journal of Recent Research and Review, XV(3), 15-18.

[7] A. Dhoka, S. Pachauri, C. Nigam and S. Chouhan (2023). Machine Learning and Speech Analysis Framework for Protecting Children against Harmful Online Content. IEEE 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), 1420-1424.

[8] Rahul Misra, Dr. Ramkrishan Sahay (2018). Evaluation of Student Performance Prediction Models with TwoClass Using Data Mining Approach. International Journal of Recent Research and Review, XI(1): 71-79.

[9] Rahul Misra, Dr. Ramkrishan Sahay (2018). Evaluation of Five-Class Student Model based on Hybrid Feature Subsets. International Journal of Recent Research and Review, XI(1):80-86.

[10] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani (2023). House Price Prediction using Machine Learning Algorithm. IEEE 7th International Conference on Computing Methodologies and Communication (ICCMC), 982-986.

[11] Dr. Himanshu Arora, Gaurav Kumar soni, Deepti Arora (2018). Analysis and Performance Overview of RSA Algorithm. International Journal of Emerging Technology and Advanced Engineering. 8(4), 10-12.

[12] Vipin Singh, Manish Choubisa, Gaurav Kumar Soni (2020). Enhanced Image Steganography Technique for Hiding Multiple Images in an Image Using LSB Technique. TEST Engineering & Management 83, 30561 -30565.

[13] Shilpi Mishra, Divyapratap Singh, Divyansh Pant, Akash Rawat (2022). Secure Data Communication Using Information Hiding and Encryption Algorithms. IEEE 2022 Second International

Conference on Artificial Intelligence and Smart Energy (ICAIS), 1448-1452.

[14] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1334-1342.

[15] Akash Agarwal, Himanshu Arora, Monika Mehra, Debosmit Das (2021). Comparative Analysis of Image Security Using DCT, LSB and XOR Techniques. IEEE 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 1131-1136.

[16] P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 569-574, 2023.

[17] Himanshu Arora, Pramod Kumar Sharma, Km Mitanshi, Aayush Choursia (2022). Enhanced Security of Digital Picture and Text Information with Hybride Model of Hiding and Encryption Techniques. IEEE 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 1238-1241.

[18] Jha, P., Dembla, D. & Dubey, W. Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model. Multimed Tools Appl (2023).