Generation of Summary from Article

A THESIS

Submitted by

Sreelakshmi TS (RCAS2021MCS207)

in partial fulfillment for the award of the degree of

MASTER OF SCIENCE
SPECIALIZATION IN
INFORMATION SECURITY AND CYBER FORENSICS

DEPARTMENT OF COMPUTER SCIENCE

RATHINAM COLLEGE OF ARTS AND SCIENCE

(AUTONOMOUS)

COIMBATORE - 641021 (INDIA)

May-2023

RATHINAM COLLEGE OF ARTS AND SCIENCE
(AUTONOMOUS)
COIMBATORE - 641021

BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **Generation of Summary from Article** submitted by **Sreelakshmi TS,**, for the award of the Degree of Master in Computer Science specialization in **"INFORMATION SECURITY AND CYBER FORENSICS"** is a bonafide record of the work carried out by her under my guidance and supervision at Rathinam College of Arts and Science, Coimbatore.

Ms Sarmila M
Supervisor

Dr.P.Sivaprakash, MTech Ph.D
Mentor

Submitted for the University Examination held on 09.05.2023

INTERNAL EXAMINER

EXTERNAL EXAMINER

RATHINAM COLLEGE OF ARTS AND SCIENCE

(AUTONOMOUS)

COIMBATORE - 641021

DECLARATION

I, **Sreelakshmi TS**, hereby declare that this thesis entitled **"Generation of Summary from Article",** is the record of the original work done by me under the guidance of **Ms Sarmila M**, Faculty Rathinam college of arts and science, Coimbatore. To the best of my knowledge this work has not formed the basis for the award of any degree or a similar award to any candidate in any University.

Place: Coimbatore **Signature of the student:**

Date: 09.05.2023 **Sreelakshmi TS**

COUNTERSIGNED

Ms Sarmila M

Supervisor

Contents

Acknowledgement

On successful completion for project look back to thank who made in possible. First and foremost, thank **"THE ALMIGHTY"** for this blessing on us without which I could have not successfully our project. I am extremely grateful to **Dr.Madan.A. Sendhil, M.S., Ph.D.,** Chairman, Rathinam Group of Institutions, Coimbatore and **Dr. R.Manickam MCA., M.Phil., Ph.D.,** Secretary, Rathinam Group of Institutions, Coimbatore for giving me opportunity to study in this college.I am extremely grateful to **Dr.R.Muralidharan, M.Sc., M.Phil., M.C.A., Ph.D.,** Principal Rathinam College of Arts and Science(Autonomous), Coimbatore.Extend deep sense of valuation to **Mr.A.Uthiramoorthy, M.C.A., M.Phil., (Ph.D),** Rathinam College of Arts and Science (Autonomous) who has permitted to undergo the project.

Unequally I thank **Dr.P.Sivaprakash, MTech,Ph.D.,** Mentor and **Dr.Mohamed Mallick, M.E., Ph.D.,** Project Coordinator, and all the Faculty members of the Department - iNurture Education Solution pvt ltd for their constructive suggestions, advice during the course of study.I convey special thanks, to the supervisor **Ms Sarmila M,** who offered their inestimable support, guidance, valuable suggestion, motivations, helps given for the completion of the project.

I dedicated sincere respect to my parents for their moral motivation in completing the project.

List of Figures

Abstract

Time is something that no one has today. Most of us have an attitude to read the heading instead of reading all the content in an article. To save time, many people use this method. The heading of the content is useful for us to not read many pages of articles and documents. Many of today's systems are used to create such a title. Such a title can be created with a technique called text summarization. Text summarization is a method of creating a summary of an article.

There are two summarization technique-extractive summarization and abstractive summarization. Extractive summarization means creating a short summary of an article by reading all the content in it. In this method, only create the summary using the lines it has. In abstractive summarization, it creates a summary of an article using new words. We build a web application here. This online application can provide an overview of any certain domain. If we conduct a search on that domain for a particular topic, it will produce a concise overview of that subject. creation of a concise summary and a suitable abstraction-based summary from online sources and papers on a certain area.

I utilised the Newsdata API to gather the news data . The user will transit via the

API and fetch the news when they search for the topic. More information can be found by searching the subject. Since we only need information from reliable websites, we rank the sites using the PageRank algorithm. I employed the TextRank algorithm for summarising. By eliminating the stopwords, we may obtain the clean sentences using the TextRank algorithm. I paraphrased the data using the GPT 2 method in order to provide abstractive summarization. Based on the sentence's score, we will create a news summary. Displaying the content in a web user interface by means of the Streamlit framework.

Chapter 1

Introduction

Nowadays, people are busy in their life. No one has the time to read. Most of us have an attitude to read the heading instead of reading all the content in an article. To save time many people use this method. The heading of the content is very useful for people to not read many pages of articles and documents. When we search for any topic many sources will be available on the internet about that topic. To understand about the topic we may need to read many contents from many articles. It will take time to read all those contents. It is a time taking process. It is a task to gather all the information about a topic. Different sites will provide different news about a topic. It is not easy to read all the news from every site. Nowadays identifying a legitimate site is a task. It is a time taking process.

To overcome this time taking process we can build a solution around the generation of a short summary and appropriate abstraction-based summary from news articles from the web on a specific domain. Any topic can be accurately summarised in this way.

By creating a web application for the summarization of an article, the users can

search for the topic and get the summary of the topic without referring to many sites. It will help the students, researchers, teachers, and readers to save time. This web application will collect the article from a different sources and combine them to give the user a summary. Initially, the web application is created for users who need authenticate information on any topic.

There are many other summarization websites available on the internet to give a summary of the text. Other than that this web application will collect and combine the authenticated sites resources and provide the summary. This would enable consumers to save time by avoiding the need to visit numerous websites. This web application is different from other summarization sites.

I have implemented the Newsdata API to accurately gather news data from a variety of sources based on the user's search query. The users will be able to transit through the API and fetch relevant news corresponding to their topic of interest. More in-depth information about any subject can be obtained by running a search query. To ensure that only reliable content is being provided, we rank the various websites using PageRank algorithm which measures the importance and quality of webpages. After utilizing the TextRank algorithm for summarisation, I was able to successfully eliminate all stopwords in order to obtain clean sentences. To further enhance the summarisation process, I also employed the GPT-2 method to paraphrase the data and provide an abstractive summarization of the text. By using the Streamlit framework, we are able to generate a comprehensive news summary based on the calculated score of an individual sentence. All of the content is then presented in an interactive and user-friendly web

user interface that is accessible to everyone. This technology is revolutionizing how people access and consume news, reducing the need for manual summarization efforts while still providing accuracy and speed.

1.1 Objective of the project

There are online text summarization websites available on the internet. Those online text summarization websites will generate a summary of the text given by the users. Users can simply put the long text and then it will generate the short summary. All the online summarization websites only provide the summary by rearranging the texts.

The main objective of this project is to create an abstractive-based summarization. This means providing a summary by creating new words by combining the text from different articles.

Project Features

1. Users can search for any topic.
2. User can search by category
3. Users will get a summary of any topic.
4. Users can select the number of sentences for the summary.
5. Users will get the news from the top sites

1.2 Scope of the Project

Over the years, the amount of information on the Internet has increased tremendously due to the recent technological development. In this way, it has provided the user with ease to access information, but on the other hands, such a large amount of information provided the user with a challenge to filter over relevant information.

When searching for any particular information on the internet there will be more and more websites that contain information about the same topic. It is very difficult to understand which is an authentic website and which information is real.

Many of the existing systems of summarization have a method of just picking the sentences from the article and providing the summary. Choosing the sentences and composing the summary are not the only tasks involved in this. New words will be produced as well as the summary. It will be very useful for the students, teachers, researchers, and readers to get a summary of any topic without going through many websites. A summary will be produced by combining the content from many credible websites.

This web application will help many people to save time. As time is more precious, we will get a fast and more legitimate result.

1.3 Contributions

A lot of research has been carried out on the problem of recommendation, but the existing recommendation algorithms have the following problems:

(1) Most pre-trained model has a problem with information loss. Many text summarization methods won't give the proper text summarization. Some of the information will be lost. There will be an information loss in the generated summary. Other models have the problem of information overlapping. The generated summary may contain overlap information. This overlap will affect the summary. Some other models do not provide fluent output. The summary of any topic should be appropriate. If any information loss or any changes happened in the summary then the meaning of the summary will change. It will affect the information.

Any topic will be meaningfully summarised by this project without any information loss or overlap. Users will benefit from getting accurate summaries of any topics. The web application Will provide a summary as per the user's need. Users have the choice to select the number of sentences for the summary. It will provide a summary based on the score and will provide a true and abstractive based summary.

1.4 Module Description

Pagerank Algorithm: PageRank algorithm to determine the order of web pages in its search engine results. The algorithm calculates each page's PageRank score by examining the web's link structure and basing it on the quantity and calibre of links heading to it. A website that connects to another page is effectively endorsing that page. The PageRank algorithm considers a page's vote total as well as the calibre and relevancy of the pages that cast the vote. Pages that earn votes from other related, high-quality pages are given more weight and are assigned a better PageRank rating.

Textrank Algorithm: Text summarization and keyword extraction are two examples of natural language processing tasks that can be performed using the unsupervised, graph-based TextRank algorithm.By building a graph with the words or phrases as nodes and their co-occurrences as edges, the TextRank algorithm can analyze the relationships between words or phrases in a given text. The computer then gives each vertex a score depending on how significant the words or phrases are in the text, with more significant words or phrases receiving higher ratings. By locating the most crucial sentences and extracting them to produce a condensed version of the text, the TextRank algorithm can be used to automatically generate summaries of lengthy texts. By figuring out which words or phrases are most crucial inside the text, it can also be utilized for keyword extraction. The TextRank method is a helpful tool for jobs involving natural language processing, especially when it comes to processing vast amounts of text rapidly and effectively.

NLP Function: Stopwords are common words like "a," "an," "the," "is," "and," "in," "of," etc. that have little meaning in sentences. In NLP, eliminating stopwords can be a helpful preprocessing step to cut down on text data noise and concentrate on the more important words.

Streamlit: Streamlit is an open source app framework in python language. It helps us to create beautiful web-apps. You can quickly develop interactive web applications for activities related to data science and machine learning using the Python package Streamlit. Without needing to understand complicated web development technologies, you can make unique web apps using Streamlit. To develop a variety of interactive widgets, like sliders, checkboxes, and dropdowns, Streamlit offers a clear and straightforward syntax. With the help of built-in charting tools or personal visualisations, you can also quickly visualise data.

1.5 Existing System

Summarize Bot: Summarize bot allows the user to read less with summarization of long text. It will compress the text and give a summary. It includes Wikipedia articles, web pages, audio, and images. This summary bot reduces the text of any article and it will not change any main points. Without changing any main points the summarize bot will provide the summary. Additionally, users can choose the summary type (generic, article, or keyword-based), the level of summarization (basic, intermediate, or advanced), and the summary length. The programme analyses the text using algorithms to extract the most crucial information, creating a summary that highlights the main ideas of the original text.

Resoomer: Generate summaries of text. Filter Documents by key topics and identify important facts and ideas. By simply downloading the extension on their browsers, or by copying and pasting the text they wish to summarize. The resoomer will provide the summary by taking the main point of the article. Users can choose the level of summarization (basic or enhanced), and the length of the summary, and enter the text from a variety of sources, including webpages, PDFs, and Word documents. Resoomer then examines the text and picks out the most crucial details, creating a summary that highlights the main ideas of the original text.

Text Compactor: Language translation tool. users need to do is paste the text and drag the slider to the percentage they would like it to be reduced. It will reduce the information to key ideas. It reduces its word content and provides a summary. Users

can choose the length of the output as well as the level of compression (low, standard, or heavy) when entering the text from a variety of sources, including webpages, documents, and social media posts. The most crucial details are then extracted from the text through analysis by Text Compactor, which creates a summary that highlights the main ideas of the original text.

Scholarcy: With Scholarcy, an online article summarizer, your content is quickly read and divided into bite-sized chunks for easier access and evaluation. Students, researchers, and other professionals use the application to rapidly grasp the main ideas of any report or paper they are working on or to write helpful summaries that they may refer to later. Additionally, you can quickly produce a summary flashcard in Word or PDF format from any report, paper, or article by extracting the important facts and citations. The program allows you to set it to extract tables, figures, and photos, and it links to open-access versions of the cited sources. When users submit PDFs of scientific papers, Scholarcy analyses the text to find the most pertinent information and creates a summary that summarises the article's main ideas. Additionally, the tool emphasizes crucial phrases and ideas to help readers better comprehend the substance of the article.

SMMRY: SMMRY is a tool designed to provide a quick way to comprehend and summarise articles and text. The program accomplishes this by condensing the text to just the most crucial sentences, rating them according to the main algorithm, and rearranging the summary to concentrate on a certain subject. Additionally, the program eliminates transitional words, pointless clauses, and overuse of instances. Users can set

the length of the summary and input material from a variety of sources, including websites, papers, and social network posts. SMMRY then examines the text and picks out the data that is most important, creating a summary that highlights the main ideas of the original text.

Split Brain Summary tool: This tool assists you in creating sentences from the article by summarising your entire content in 39 different languages. Instead of pasting text into the tool to have it summarise your information, you can instead insert a URL. The summary ratio, which you can alter by adjusting the density of the paraphrasing, can also be used to produce the variation in summaries. The ability to import files or export the outcomes to well-known formats like DOC or PDF is absent from the Split Brain Summary application.

Chapter 2

Literature Survey

MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization

PradeepikaVerma , HariOm 15 April 2019

In this research,A single document is created from numerous documents using the coverage and non-redundancy characteristics. The weighted combination of word embedding and Google-based similarity algorithms explores these properties. The text summarization task is represented as an optimization problem, where multiple text features with their optimised weights are employed to score the sentences in order to locate the relevant sentences. This is done to accommodate the relevancy feature in the system generated summaries. It employ the meta-heuristic technique known as Shark Smell Optimization for the features' weight optimization (SSO). The tests are run using six benchmark datasets (DUC04, DUC06, DUC07, TAC08, TAC11, and MultiLing13).

A hierarchical self-attentive neural extractive summarizer via reinforcement learning (HSASRL)

Farida Mohsen, Jiayang Wang Kamal Al-Sabahi 16 March 2020

The suggested model uses a hierarchical self-attention mechanism to produce sentence and document embeddings that accurately reflect the document's hierarchical structure and provide superior feature representation. While the attention method adds a second source of information to guide the summary extraction, reinforcement learning permits direct optimization with respect to assessment measures. Three well-known datasets, CNN, Daily Mail, and their merged form CNN/Daily Mail, were used to evaluate the model. According to experimental findings, the model outperformed cutting-edge methods for extractive summarization on the three datasets in terms of ROUGE scores.

Multi-task learning for abstractive text summarization with key information guide network

Weiran Xu, Chenliang Li, Minghao Lee and Chi Zhang

A multi-task learning framework-based key information guide network for abstractive text summarization. In this methodology, the key information, which is comprised of the key sentences and keywords, is encoded separately from the results of the conventional document encoder. To obtain a more complex end-to-end model, a multi-task learning framework is introduced.suggest an unique multi-view attention guidance network to get the dynamic representations of the source text and the key information in order to fuse the key information. Additionally, the abstractive module incorporates

the dynamic representations to direct the summary creation process. We test our model on the CNN/Daily Mail dataset, and the results of our experiments demonstrate that it significantly enhances performance.

Bottom-Up Abstractive Summarization

Sebastian Gehrmann, Yuntian Deng, Alexander M. Rush 31 Aug 2018

Neural network-based methods for abstractive summarization yield outputs that are more fluent than those produced by other techniques. This paper suggests a straightforward method for dealing with this problem: utilise a data-efficient content selector to identify phrases in a source document that ought to be included in the summary. This selection serves as a bottom-up attention step that limits the model to phrases that are likely to occur. We demonstrate how this method increases text compression while still producing fluid summaries. Compared to existing end-to-end content selection models, our two-step process is both easier to use and more effective, and it significantly improves ROUGE for both the CNN-DM and NYT corpus. As little as 1,000 sentences can be used to train the content selector.

Chapter 3

Methodology

3.1 Fetch the news

News Data API:

Newsdata.io is a powerful news aggregator that uses sophisticated natural language processing and machine learning algorithms to efficiently gather news articles from multiple sources into one unified platform. With this innovative platform, users can easily find the latest news stories from various categories, topics, and sources in an organized and efficient manner. Furthermore, this platform also enables users to have more control over their news consumption by offering options for customizing their experience for maximum personalization.

Users of this platform are able to easily search for news articles based on keywords, browse through content by topic, source, or date, and filter the results based on sentiment and relevance. It provides access to an impressive amount of news articles from over 50,000 sources around the world - including major news outlets, blogs, and even social media accounts.

Newsdata.io is a powerful platform that offers an extensive range of data analytics tools, such as trend analysis, data visualization, and topic modeling. These features allow users to quickly identify the main topics and trends discussed in the news articles they read, making it easier to gain insights from data and develop effective strategies. The comprehensive platform offered by this service provides users with access to a vast historical news archive, allowing them to gain insight into news articles from the past and track changes in the news narrative over the course of time. This powerful tool can prove invaluable for organizational research, providing data that can be used to inform decisions and develop strategic initiatives.

Newsdata.io offers convenient and powerful APIs that enable users to quickly and easily integrate the platform's comprehensive data and insightful analytics into their applications and workflows. It is a versatile platform that can be utilized for a number of different purposes, including research, media monitoring, market analysis, and content creation.

3.2 Ranking the sites

Page Rank Algorithm

The PageRank algorithm is widely used by search engines to evaluate and rank websites in terms of their authority and relevance. This algorithm works by analyzing the internal links between web pages, as well as external links from other websites. Pages with a higher PageRank score are considered more important and relevant to users, and thus these pages are ranked higher in search engine results pages.

Here's how the algorithm works:

1. A typical algorithm for web crawling begins with a set of seed pages, which are usually well-known, authoritative websites such as top news portals or official government pages. These seed pages are used as the starting points for the crawling process, from which the crawler will explore related links and discover new web pages.

2. The algorithm then proceeds to evaluate the connections between web pages by calculating the number and quality of links pointing to them. Pages with a higher amount of authoritative, high-quality links are seen as more significant and are assigned a higher PageRank score accordingly.

3. Link quality is an important factor in determining the value of a link and is taken into consideration. Links coming from high-authority and well-respected pages are considered more valuable than links coming from low-authority sites, as these links add more value to the overall content.

4. The algorithm begins by recursively calculating the PageRank scores of all the

web pages that are connected to the seed pages. This process then continues until all webpages have been comprehensively evaluated, and a PageRank score has been assigned to each one. This process is repeated until the calculated scores converge, thereby providing an accurate representation of the relative importance of each page on the web.

5. When an individual enters a search query into a search engine, the PageRank scores of the respective web pages are analysed in order to determine their relevance and authority. Pages with higher PageRank scores are considered more reliable and trustworthy, thus they are ranked higher in the search results for better visibility and engagement with users.

3.3 Summarize the Text

Text Rank

TextRank is a graph-based ranking algorithm that is commonly used in natural language processing for text summarization. TextRank is a powerful algorithm that works by representing the text as a graph, where each sentence is represented as a node. The edges between the nodes capture the relationships between the sentences, and these edges are then weighted based on how similar each sentence is to one another. By doing this, it allows for quick and accurate extraction of important information from any given text.

Here's how the algorithm works for text summarization:

1. Before the text can be used for further analysis, it needs to be preprocessed in order to remove all unnecessary elements such as stop words, punctuation marks, and other non-essential elements. This allows for a clean dataset that is ready for deeper analysis and more advanced tasks.

2. Sentences in texts are broken down into individual components and then represented as nodes in a graph structure, which allows for an easier analysis and understanding of the text. This data-driven approach to content analysis has become increasingly popular due to its ability to quickly provide meaningful insights from the text.

3. The edges between the nodes of a network are generated based on the similarity between the sentences, which can be accurately calculated using advanced techniques

like cosine similarity or Jaccard similarity. These sophisticated algorithms provide a reliable method for measuring the similarity between two pieces of text and are essential for creating meaningful relationships between nodes in a graph.

4. The graph is examined and then scored using a sophisticated iterative algorithm, where the score of each sentence is determined and calculated based on the scores of the sentences that are linked to it. This scoring mechanism helps to ensure that each sentence is assessed and evaluated accurately, providing more accurate results.

5. The algorithm continues to iterate, considering every sentence in the original text and running through each option until the scores for all sentences become consistent and converge. At that point, the highest-scoring sentences are chosen as the final summary, producing a concise yet comprehensive output that incorporates the most important details.

3.4 Text paraphrasing

GPT-2 (Generative Pretrained Transformer 2)

GPT-2 (Generative Pretrained Transformer 2) is a powerful deep learning language model that can be utilized to generate natural sounding, human-like text without any manual input. It also allows for paraphrasing, which is the process of expressing the same meaning of a sentence or text using different words or phrases, making it incredibly useful for businesses looking for efficient ways to produce content. This AI writing assistant has revolutionized the industry as it can be used to quickly and accurately generate large amounts of content in a short period of time.

GPT-2 uses a combination of techniques to paraphrase text, including:

1. Synonym Replacement is a powerful feature of GPT-2 that can effectively replace words in the original text with appropriate synonyms that convey the same meaning, all while maintaining the overall structure and flow of the sentence. This can be incredibly useful for writers who are looking to diversify their writing style or to simply make their content more engaging.

2. With the Sentence Restructuring capabilities of GPT-2, users are now able to quickly and easily change the order of clauses or phrases in a sentence to create a new sentence with the same meaning, as well as to switch the tense or voice in a sentence. Such abilities can help streamline workflows while still ensuring that the original intent is met and conveyed accurately.

3. Semantic Interpretation: GPT-2, the powerful deep learning technique, is capable of extracting the underlying meaning of any given text and rephrasing it using more natural sounding words and phrases while still retaining its original message. This invaluable capability allows for complex ideas to be expressed much more concisely, making it a highly sought-after technology in many industries.

4. Contextualization is a key feature of GPT-2, as it can take into account the context of the text to generate an accurate and relevant paraphrased version that fits perfectly in that specific context. This technology has revolutionized the way we create content, allowing us to produce more targeted and efficient material than ever before.

3.5 Fetch the news metadata

The metadata of the data means the publishing date, News text link, author, language and description of the news article.

3.6 Display the content into web user interface

Streamlit

Streamlit is an increasingly popular open-source app framework that has become a go-to choice for Machine Learning and Data Science teams. It enables developers and data scientists to quickly and effortlessly create interactive web applications with their models and data, requiring no prior web development knowledge. Its powerful features make it easier for teams to share their insights with others, giving them the ability to showcase their work in a visually engaging way.

Streamlit is a user-friendly tool that allows people to quickly create and customize powerful web applications using Python. It offers an array of built-in tools and components that make it much easier to streamline the development process and create sophisticated, dynamic apps with minimal effort. The powerful and convenient framework offers a simple yet intuitive Application Programming Interface (API) for quickly creating user interfaces, as well as advanced data visualization and debugging tools that can help identify problems more accurately. This makes it easier to develop applications in shorter periods of time and with fewer resources.

3.7 Advantages

This web app provides an invaluable service to the people by making it easy for them to quickly understand a topic without having to search through multiple resources. The main purpose of this application is to provide users with a comprehensive summary about any given topic derived from various sources. Users of a news-reading platform can select the desired range of news articles, search for any topics they want, and choose the desired category that best suits their interests. Additionally, users will get an easy to understand summary of any given topic, along with the ability to select the number of sentences they want in their summary.

This web application is designed to help users quickly and efficiently get a summary of any given topic by leveraging the power of AI to intelligently combine different resources. Instead of spending hours going through multiple sites, users can now get all the correct information they need from one single source, saving them both time and energy.

3.8 System Design

The first and foremost step in the process of collecting news is to do an initial gathering of relevant sources. This requires searching for reliable and authoritative sources, reading up on the latest developments, and checking out what other people are saying about it. Once you have identified the right sources, you can start to compile your news collection by organizing it into categories, noting down important points or facts related to a particular event or story, and then finally putting together an effective summary of the news. After that, we can utilize our analysis to accurately rank the sites according to their relevance and quality of content, ensuring that the most relevant and useful information is presented first. After ranking the websites, we can effectively summarize their content by removing stopwords, measuring Cosine similarity, and scoring sentences for their relevance. This enables us to quickly extract the key information from each website, allowing for more efficient processing and improved accuracy. The GPT-2 model is a tool for paraphrasing tasks thanks to its ability to accurately reproduce text in a natural language that is close to that of a human. This AI system can quickly and easily generate new versions of existing text without compromising on the quality, making it an invaluable asset for any organization or individual. By leveraging the power of the streamlit framework, it is now possible to quickly create a custom web application. This streamlit based web application provides an easy-to-use interface for users to interact with their AI writing assistant to pick up tasks, view progress, and access quality content in an efficient manner.

Chapter 4

Experimental Setup

4.1 Create intent

An intent is defined as the intention of the user when they are interacting with a web application. By specifying an intent, the user can select a specific category of news which they would like to be displayed, followed by options to choose the number of sentences they want to see. This input will then be used to fetch relevant data from the web application.

4.2 Get User Input

Users are now able to easily choose a specific topic from the available categories, such as World, Nation, Business, Technology, Entertainment, Sports, Science and Health. The user can then select the number of sentences that should be displayed for each given topic; allowing them to quickly find the content they are looking for. Additionally, there is also an option to search all topics simultaneously and efficiently.

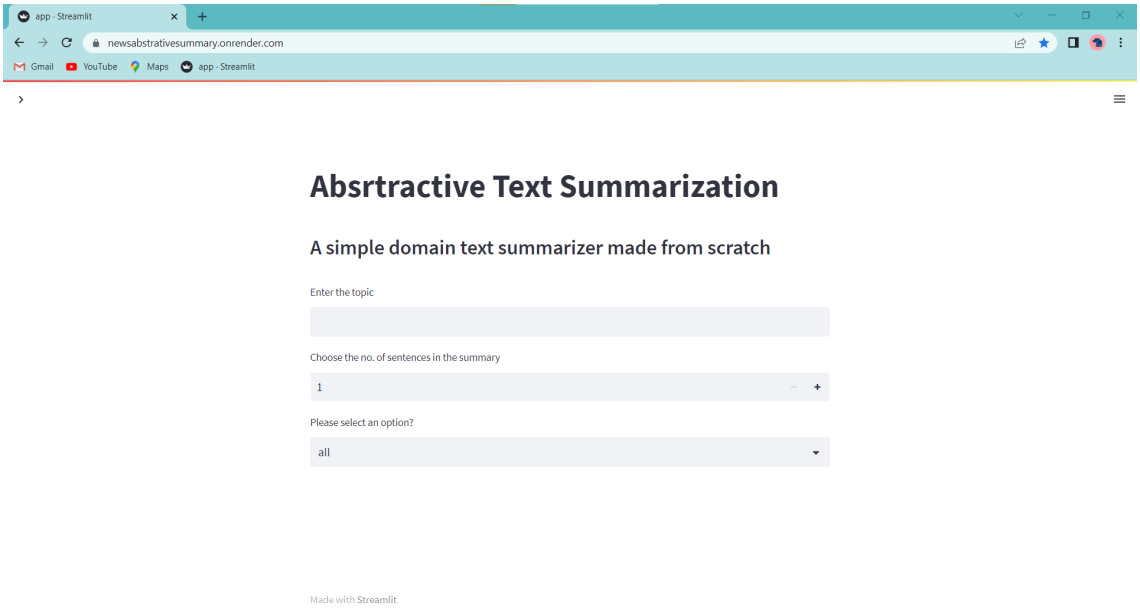

Figure 4.1: Home Page

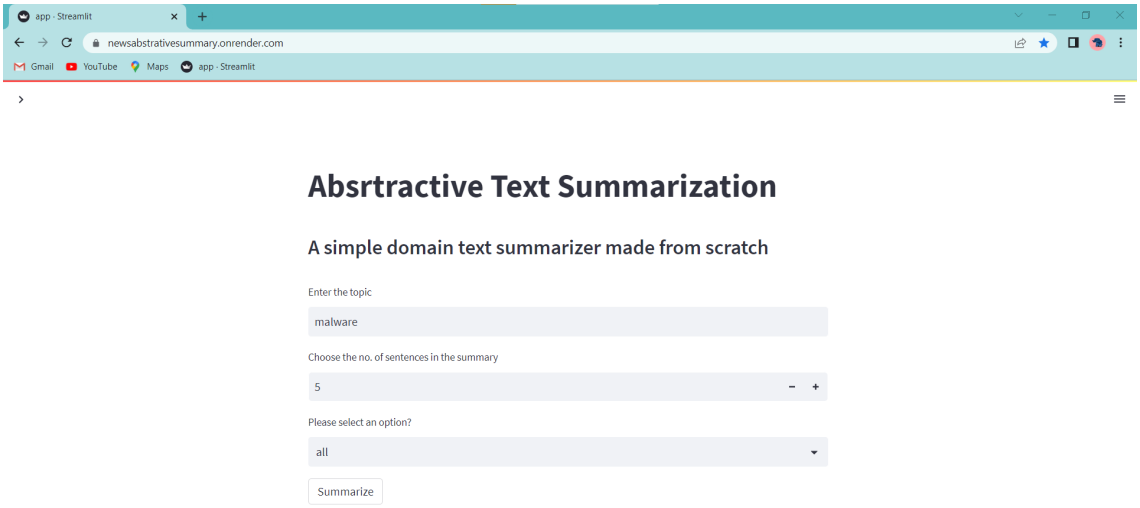

Figure 4.2: Input

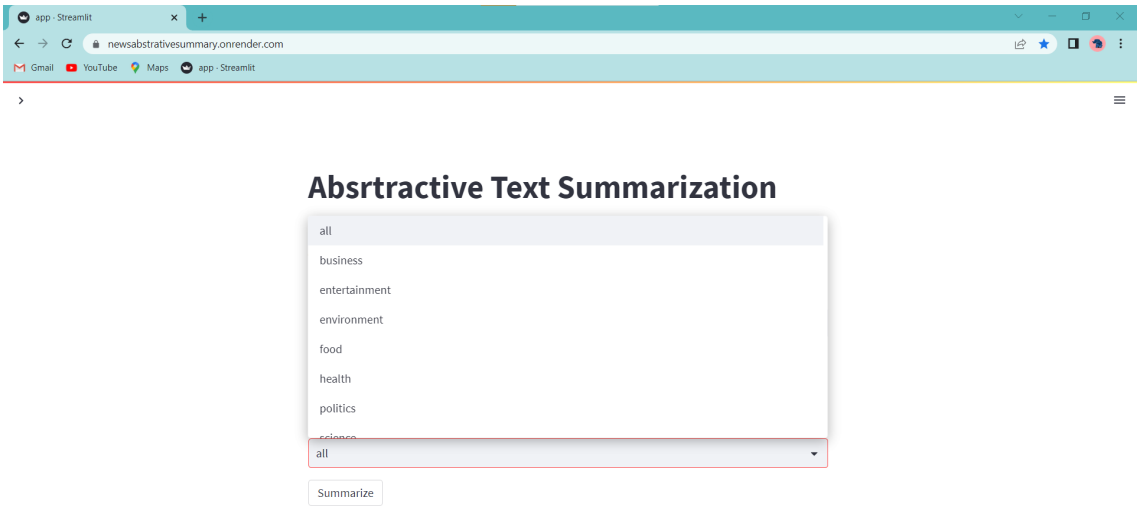

Figure 4.3: Category

4.3 Get Output

When a user searches for a specific topic, It quickly generates the most relevant news articles about that topic. This feature allows users to select the number of sentences they would like as a summary and also offers them the ability to narrow their search even further by selecting from a variety of categories and then searching for topics related to that category.

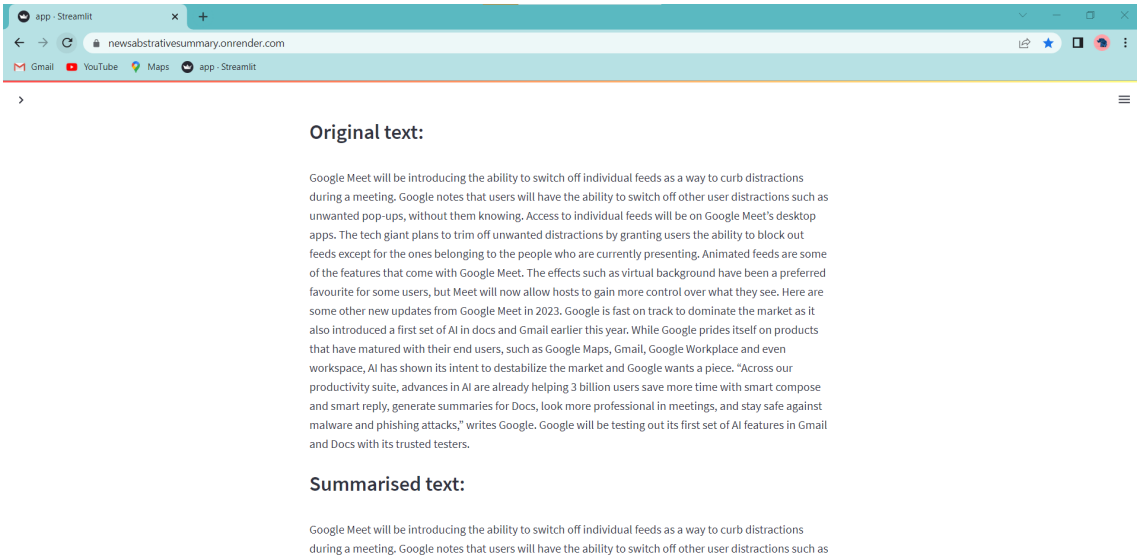

Figure 4.4: Original text

Figure 4.5: Summarized Text

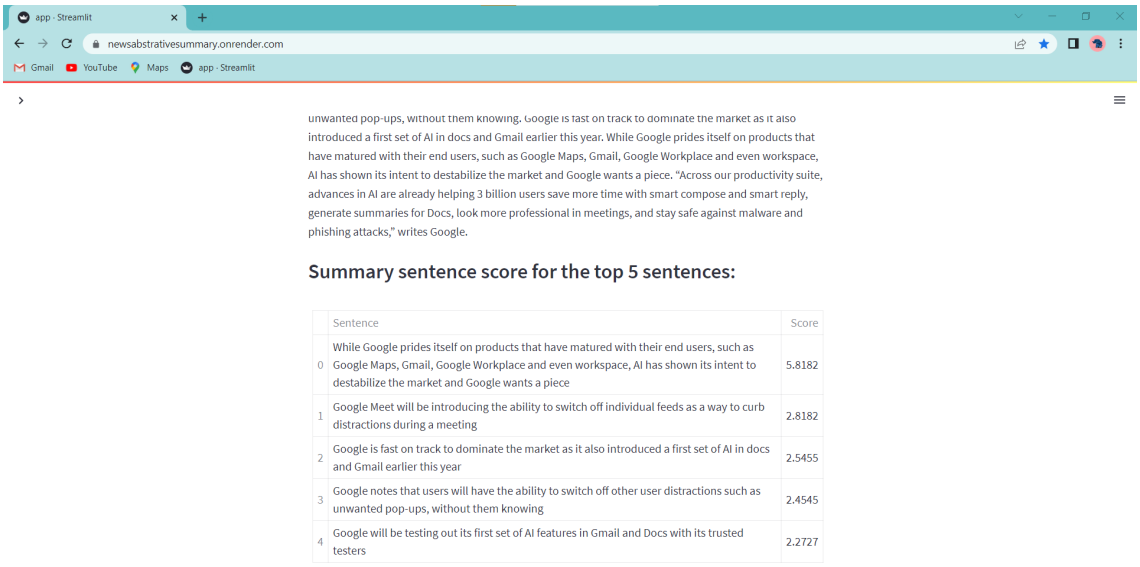

Figure 4.6: Scoring the sentence

4.4 Model Training

Step 1: API creation

It will generate the request by concatenating the parameters coming from the front end with the URL of the API.

Step 2: API Response

We will get the response of the request of the parameters along with the url that we give.

Step 3: Newsdata API

Newsdata.io is a News API that provides access to news articles from all over the world. making requests to the API to retrieve news articles.

Step 4: NLP function

Remove the stopwords

Stop words are frequently used words that are omitted from searches to speed up the indexing and parsing of online pages. Even though stop words are used by the majority of Internet search engines and NLP (natural language processing). Examples of common stop words are a, an, and, but, how, in, on, or, the, what, will.

Step 5: GPT-2 Paraphrasing

GPT-2 that can effectively replace words in the original text with appropriate synonyms that convey the same meaning, all while maintaining the overall structure and flow of the sentence.

Chapter 5

Results and Discussions

This web application is providing the most up-to-date and accurate news on any topic which the user selects, giving them access to detailed, timely information about the topic at hand. It is incredibly reliable, highly efficient, and incredibly convenient for users who need quick access to trustworthy news sources. Users have the incredible advantage of being able to customize their experience by selecting what type of news they would like to receive and even the number of sentences they would prefer for each summary. This level of customization allows users to tailor the content they receive to their exact needs, which is incredibly powerful and convenient. There are many categories like sports, entertainment, business, environment, food, health, politics, science etc. In search topics, users can search any topic the user need. The topic that the user provides will show on the page. This application provides a summary of any news the user wants.

Chapter 6

Conclusion

The primary goal of this project was to develop a web application that would deliver up-to-date, accurate news to users. I have implemented the Newsdata API to accurately gather news data from a variety of sources based on the user's search query. The users will be able to transit through the API and fetch relevant news corresponding to their topic of interest. More in-depth information about any subject can be obtained by running a search query. To ensure that only reliable content is being provided, we rank the various websites using PageRank algorithm which measures the importance and quality of webpages. After utilizing the TextRank algorithm for summarisation, I was able to successfully eliminate all stopwords in order to obtain clean sentences. To further enhance the summarisation process, I also employed the GPT-2 method to paraphrase the data and provide an abstractive summarization of the text. By using the Streamlit framework, we are able to generate a comprehensive news summary based on the calculated score of an individual sentence. All of the content is then presented in an interactive and user-friendly web user interface that is accessible to everyone. This technology is revolutionizing how people access and consume news, reducing the

need for manual summarization efforts while still providing accuracy and speed. This web application is designed to be easy to use, reliable and provide the most up-to-date information from trusted sources. It is our aim to make sure that all users have access to the latest news, as quickly and accurately as possible. Thus, by developing a web application, the user can quickly get accurate summaries of news from multiple sources in a convenient manner. This saves time and offers a great convenience to users who want to stay updated with the current happenings in the world. Not only that, but it also helps users to have access to news from around the globe with minimal effort on their part. When the user select or search any topic the application provides the summary of the news. The user can select the category of the news. In this, the topic summary is not getting by combining different article. Ranking the sites and provide each sites abstractive summary. This highly useful web application gives users the ability to quickly get an overview of a certain topic without having to read through lengthy articles. In the future, it promises to provide even more condensed information as well as summaries of any type of content, not just news topics.

6.1 Future Works

The article summarising online applications may be improved and enhanced by adding more categories and subsections. With advancing technology, the administrator of this web application may decide to upgrade the software to improve its functionality and keep it up to date with modern trends. Such improvements are necessary to ensure that users get an optimal experience when using this web application.

References

1. Fang, Y., Zhu, H., Muszynska, E., Kuhnle, A., Teufel, S.H.: A proposition-based abstractive summarizer. In: COLING, Osaka, pp. 567–578 (2016)

2. S. Modi and R. Oza, "Review on Abstractive TextSummarization Techniques (ATST) for single and multidocuments," 2018 International Conference onComputing, Power and Communication Technologies(GUCON), Greater Noida, Uttar Pradesh, India, 2018,pp. 1173-1176

3. J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," International Conference on Data Science Communication, pp. 1–3, 2019.

4. Sreejith C, Sruthimol M P and P C Reghuraj, "Box Item Generation from News Articles Based Paragraph Ranking using Vector Space Model", International Journal of Scientific Research in Computer Science Applications and Management Studies, Vol. 3,2014.

5. M. Moradi, G. Dorffnerand M. Samwald,"Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Comput.

Methods Programs Biomed.,pp. 105117, Vol. 184, 2020.

6. B. Mutlu, E. A. Sezerand M. A. Akcayol,"Multi-document extractive text summarization: A comparative assessment on features," Knowledge-Based Syst., pp. 104848, Vol. 183, 2019.

7. M. Afsharizadeh, H. Ebrahimpour-Komlehand A. Bagheri, "Query-oriented text summarization using sentence extraction technique," Fourth Int. Conf. Web Res., pp. 128–132, 2018.

8. M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroniand R. Leonardi,"A freeWeb API for single and multi-document summarization," ACM Int. Conf. Proceeding Ser., Vol. Part F1301, 2017.

9. T. Jo,"K nearest neighbor for text summarization using feature similarity," Proceedings of Int. Conf. Commun. Control. Comput. Electron. Eng., pp. 1–5, 2017

10. D. Bartakke, S. D. Sawarkar, and A. Gulati, A Semantic Based Approach for Abstractive multi-Document Text Summarization, International Journal of Innovative Research in Computer and Communication Engineering, 4(7), India, 2016.

11. Moratanch, N., Chitrakala, S.: A survey on abstractive text summarization. In: 2016 International Conference on Circuit, power and computing technologies (ICCPCT) (pp. 1–7). IEEE (2016)

12. Nallapati, R., Zhou, B., Ma, M.: Classify or select: neural architectures for extractive document summarization. arXiv preprint arXiv:1611.04244 (2016)

13. Narayan, S., Papasarantopoulos, N., Lapata, M., Cohen, S.B.: Neural extractive summarization with side information. arXiv preprint arXiv:1704.04530 (2017)

14. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252 (2016)

15. Ramesh Nallapati, Bowen Zhou, et al "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond".The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 26 Aug 2016.

16. K.Cho, B .van Merrienboer, D.Bahdanau, Y.Bengio " On the Properties of Neural Machine translation: Encoder Decoder Approaches". Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8),7 Oct 2014.

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

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

19. Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing

20. Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

21. Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics

22. Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization.

23. Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction

24. Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

25. Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies