Generation of Summary from Article

PHASE 1 REPORT

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)

DECEMBER-2022

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

BONAFIDE CERTIFICATE

This is to certify that the Phase 1 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

MR Ravisankar, M.E
Supervisor

Mr.P.Sivaprakash, MTech (Ph.D)
Mentor

Submitted for the University Examination held on 02.12.2022

INTERNAL EXAMINER

EXTERNAL EXAMINER

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

DECLARATION

I, **Sreelakshmi TS**, hereby declare that this Phase 1 entitled **”Generation of Summary from Article”,** is the record of the original work done by me under the guidance of **MR Ravisankar M.E**, 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: 02.12.2022 **Sreelakshmi TS**

COUNTERSIGNED

MR Ravisankar M.E
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 **Mr.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 **MR Ravisankar M.E.,** 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 using this method. The heading of the content is useful for us to not to 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. In this project I am going to do such a technique. 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 it only create the summary using the lines it has. In abstractive summarization , it create a summary of an article using new words . Here we create a web application. That web application has articles from specific domains. In that domain if we search for any specific topic it will generate a short summary of that topic. Generation of short summary and appropriate abstraction based summary from articles from web on a specific domain.

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 of instead of reading all the content in and article. To save time many people using this method. The heading of the content is very useful for the people to not read many pages of articles and documents. When we search for any topic many sources will available on 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 sites. Nowadays identifying the legitimate site is a task. It is a time taking process.

To overcome this time taking process we can build a solution around generation of short summary and appropriate abstraction based summary from articles from web on a specific domain . This will provide accurate summary of any topic.

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 refering many sites. This will help the students, researchers, teachers and readers to save their time . This web application will collect the article from different source and combine it to give the user a summary. Initially, the web application is created for the users who needs authenticate information of any topic.

There are many other summarization website available on internet to give the summary of the text. Other than that this web application will collect and combine the authenticate sites resources and provide the summary. This will help the users to not search many sites and it will help the users to save time. This web application is different from other summarization sites.

1.1 Objective of the project

There are online text summarization websites available on internet. Those online text summarization website will generate the 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 website only provide the summary by rearranging the texts.

The main target of this project is to create an abstractive based summarization . Which means provide a summary by creating new words by combining the text from different articles.

Project Features

1. Users can search for any topics.

2. User can search by category

3. Users will get the summary of any topic.

4. Admin can add categories and Url's.

5. Admin will get notification if data is not there while searching.

1.2 Scope of the Project

Over the years, the amount of information on the Internet has been 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 hand, such a large amount of information provided the user with a challenge to filter over relevant information.

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

Many of the existing system of summarization has a method of just picking the sentences from the article and provide the summary. This project is not just picking the sentences and creating the summary. This will create new words and generate the summary. It will be very useful for the students, teachers, researchers and readers to get the summary of a any topic without going through many websites. This project will combine the articles of many authenticate websites and create the summary.

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 of information loss. Many text summarization method won't give the proper text summarization. Some of the information will lost. There will be an information loss in the generated summary. Other models have the problem of information overlaping . The generated summary may contains overlap information. This overlap will affect the summary. Some other models 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.

This project will provide the meaningful summary of any topic without any information loss and information overlap. This will help the users to get the summary of any topic without any mistakes.

1.4 Module Description

Google News RSS API: Google News RSS Feed Create RSS feeds from any Google News webpage, search result or topic.

Beautiful soup: Beautiful Soup is a Python library for pulling data out of HTML and XML files.

Newspaper 3k: Newspaper3k package is a Python library used for Web Scraping articles.

Streamlit: Streamlit is an open source app framework in python language. It helps us to create beautiful web-apps.

1.5 Existing System

Summarize Bot: This will allow the user to read less with summarization of long text.This will compress the text and give summary. It include wikipedia articles, web pages, audio and images. This summarize 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.

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.

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 reduce its word content and provide the summary.

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 programm allows you to set it to extract tables, figures, and photos, and it links to open access versions of the cited sources.

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

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

Google news RSS API:

Google News RSS Feed Create RSS feeds from any Google News webpage, search result or topic.

There are four types of Google News RSS

1. Top headlines - Get the most recent news stories trending in your nation.Type this Url https://news.google.com/rss in your browser and it will forwarded to the main Google News feed for your country language. hl: language gl: country ceid: country: language

2. Headlines by topic - Get the topic-oriented latest news headlines for your nation These are the topics - WORLD NATION BUSINESS TECHNOLOGY ENTERTAINMENT SCIENCE SPORTS HEALTH You can get these topic oriented feeds.

3. Location headlines - Get the location-oriented latest news headlines (city, state, country, etc) This will find the news about a specific place.

4. News by search criteria - Get data by searching keywords, websites, dates, or any of these combined.

3.2 Summarize the news

Newspaper 3k

The newspaper 3k is an nlp library used for summarize the data. The Newspaper3k package is a Python library used for Web Scraping articles, It is built on top of requests and for parsing lxml. Newspaper 3k utilizes the requests library and has Beautiful Soup as a dependency while it parses for lxml.

Beautiful soup

We have huge amount of data available on internet. sometimes these data are easy to read and sometimes it may not. There is a method called web scraping . Web scraping is very useful to transform unstructured data into structured data that is easier to read analyze. Web scraping is a process of extracting copying and screening of data from various resourses. Beautiful soup is not a standard python library. Beautiful Soup is a library that makes it easy to scrape information from web pages.

3.3 Fetch the news metadata

The metadata of the data means the publishing date, News text link,

3.4 Display the content into web user interface

Streamlit Streamlit is an open source app framework in python language. It helps us to create web-apps. By using streamlit we can create apps easily.

3.5 Advantages

By using this web app, it is easy for the people to not go through many resourses for understanding the topics. The main goal of this web app is to provide the summarized information about a topic from different resourses. Users can select the range of news they want. Users can search for any topics. Users can select the category. Users will get the summary of any topic. Admin can add categories and Url's. Admin will get notification if data is not there while searching. This we app will help to get the summary of any topic by combining different resourses. So users don't want to go through many sites for the correct information.

3.6 System Design

Initially, The first step is to collect the top sites news . After that web scraping method is used to transform unstructured data into structured data, that is easier to read analyze. By using the newspaper 3k we process the data and by using the nltk function we remove the stopwords to get the output. Then using streamlit framework it will create a web application.

Chapter 4

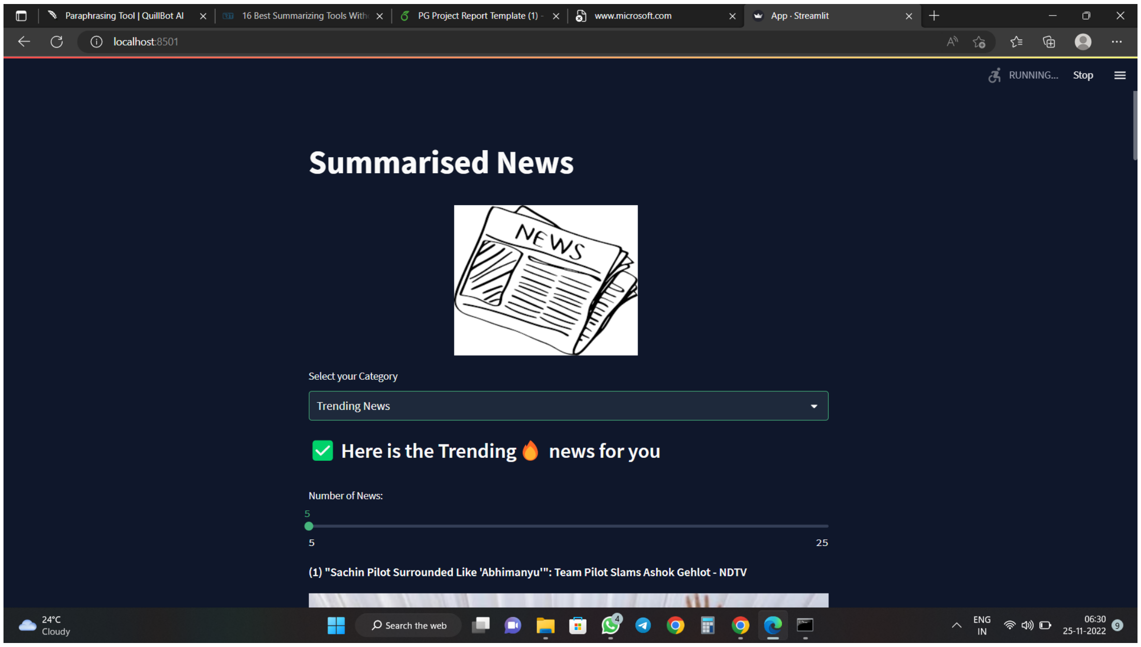
Experimental Setup

4.1 Create intent

An intent is the intention of the user interacting with the web application. where the user can select the category of news which need to be displayed which followed by Trending news, Favourite topics and search topic. This will fetch the data from the database.

4.2 Get User Input

User can select a topic according to the category and user can also view the trending news and also can select the number of news which need to be displayed from 5 to 15. The user can select the favourite news . In favourite news user can select the topics like World, Nation, Business, Technology, Entertainment, Sports, Science and health. One more option is there to search a topic.



Summarised News

Select your Category

Trending News

✅ Here is the Trending 🔥 news for you

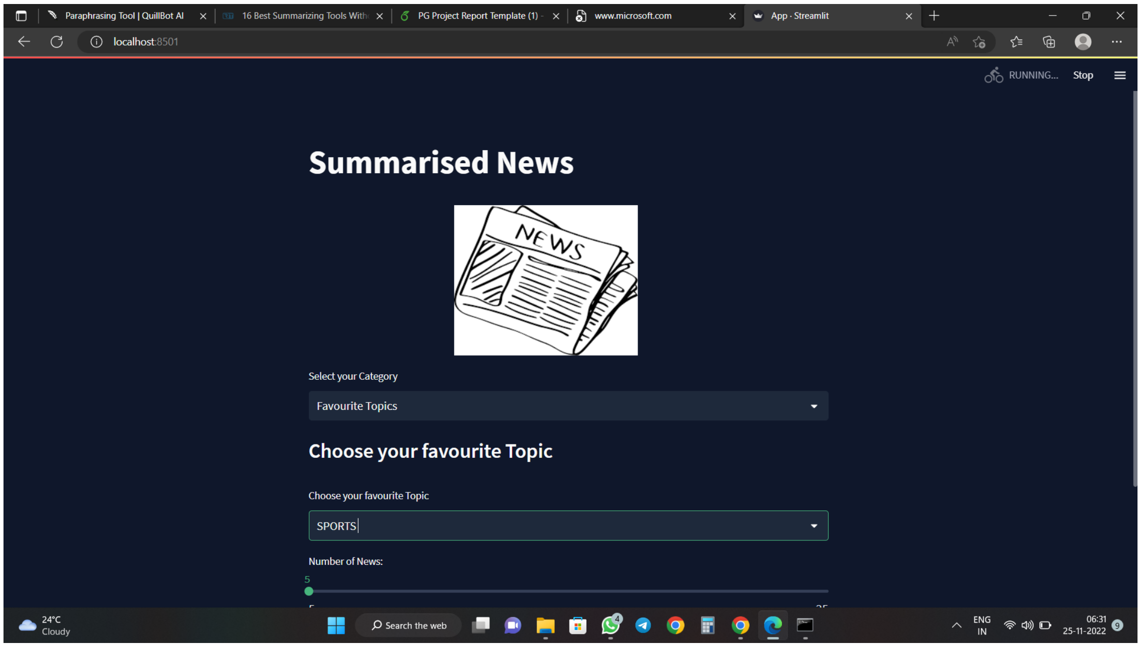
Number of News:

5

5 25

(1) "Sachin Pilot Surrounded Like 'Abhimanyu'": Team Pilot Slams Ashok Gehlot - NDTV

Figure 4.1: Trending News



Summarised News

Select your Category

Favourite Topics

Choose your favourite Topic

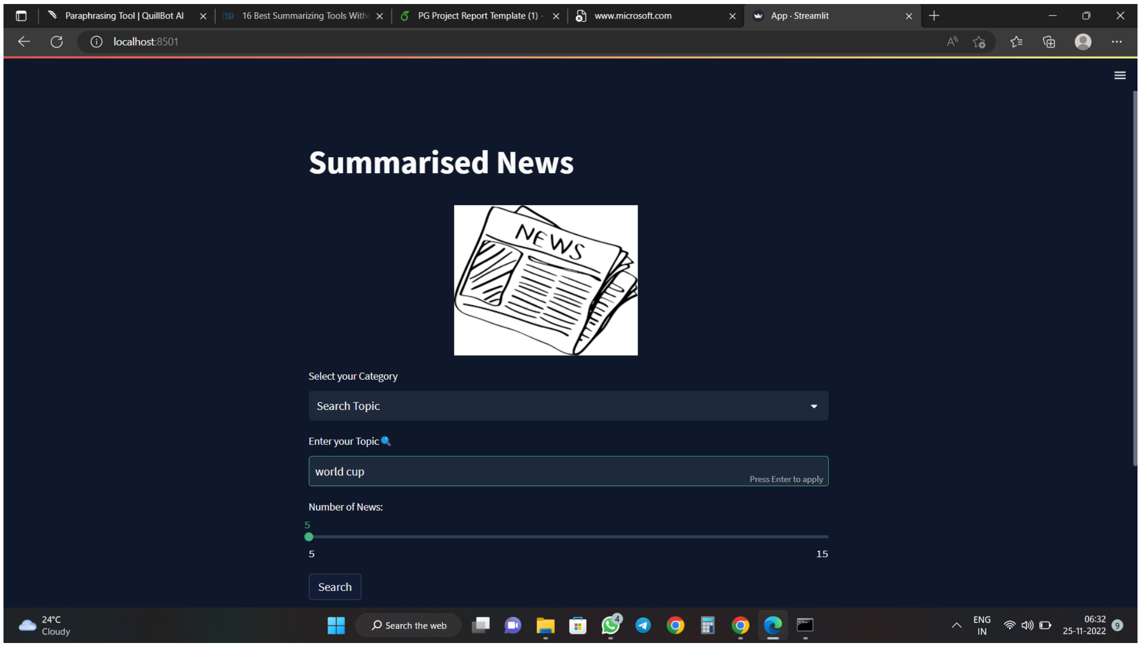
Choose your favourite Topic

SPORTS

Number of News:

5

Figure 4.2: Favourite Topic



Summarised News

Select your Category

Search Topic

Enter your Topic 🔍

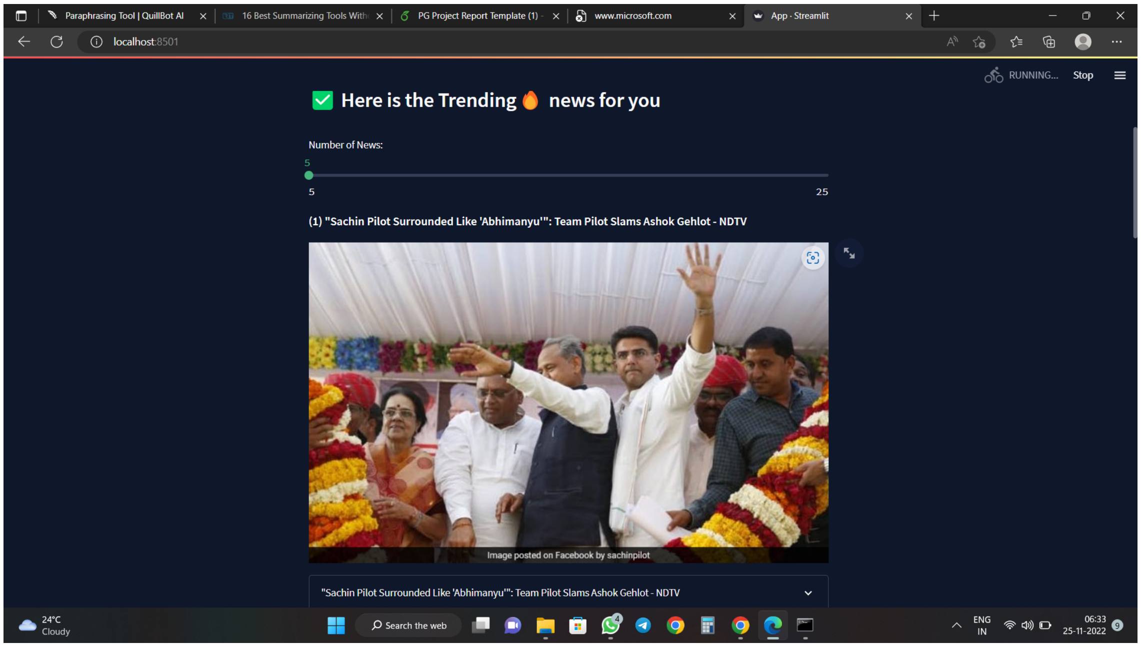
world cup

Number of News: 5

Figure 4.3: Search topic

4.3 Get Output

when the user select the trending news category , the trending news and the summary of the news will get. The user have an option to select the number of news they want. When the user select the favourite topic category there is options for World, nation, business, technology, entertainment, sports, science, health. From these topics user can select any topic and user will get the news. Another category is search topic. In search topic user can search any topic they want. It will provide the output of any topic. In every category the user search, there is an option for the user to select the number of news.



✅ Here is the Trending 🔥 news for you

Number of News:

5

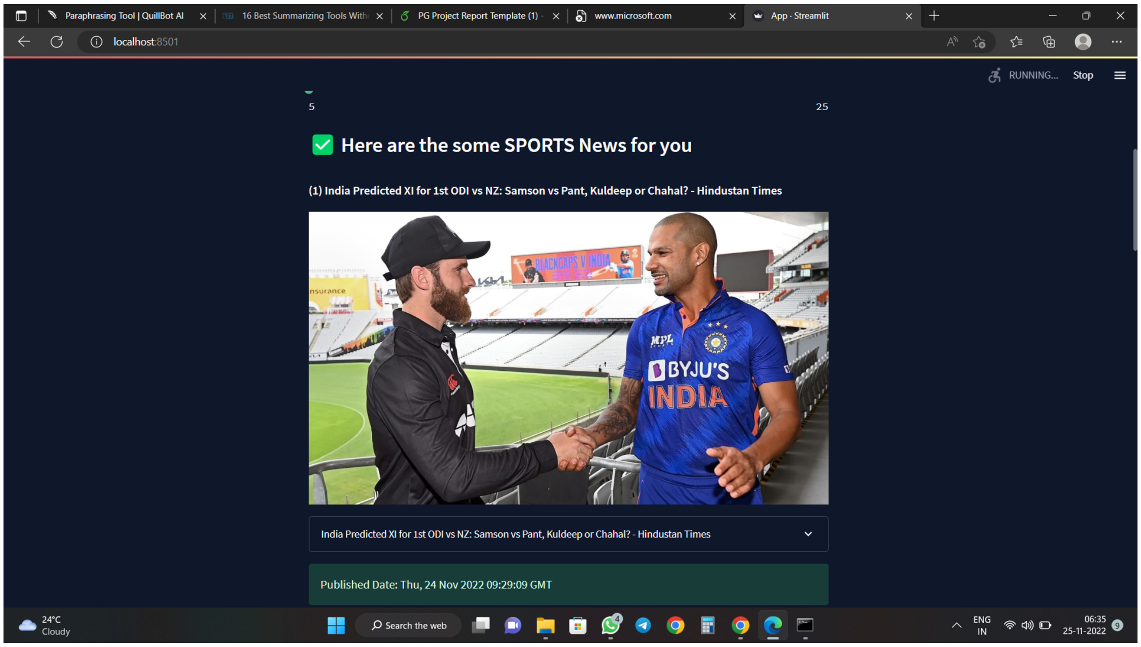
5 25

(1) "Sachin Pilot Surrounded Like 'Abhimanyu'": Team Pilot Slams Ashok Gehlot - NDTV

Image posted on Facebook by sachinpilot

"Sachin Pilot Surrounded Like 'Abhimanyu'": Team Pilot Slams Ashok Gehlot - NDTV

Figure 4.4: Trending News



5 25

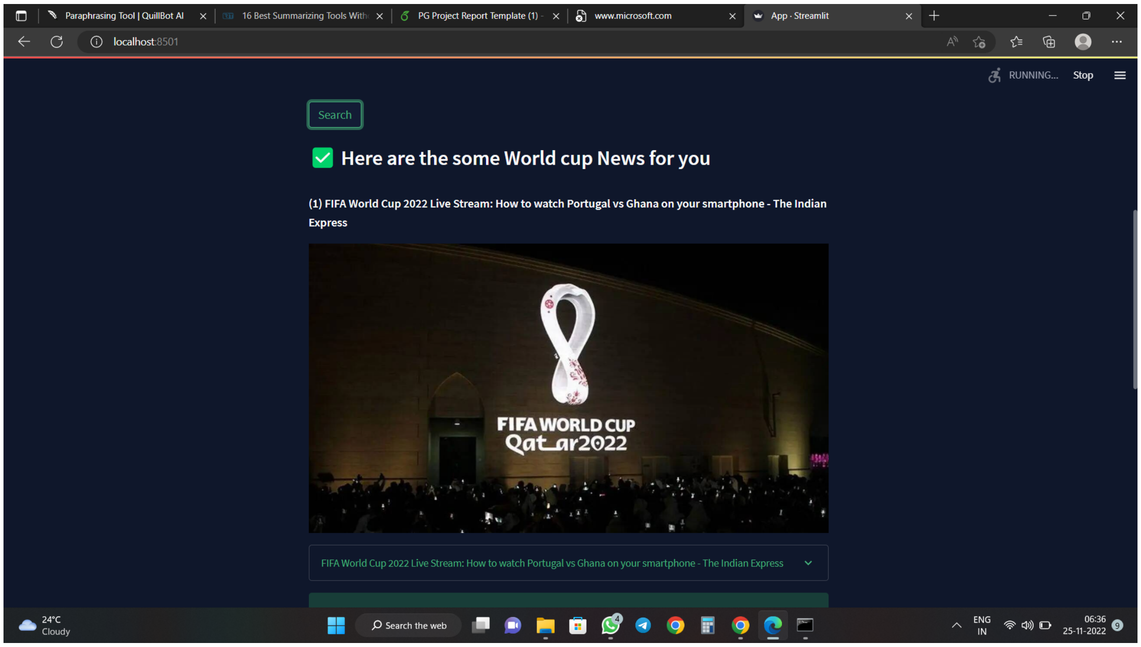
✅ Here are the some SPORTS News for you

(1) India Predicted XI for 1st ODI vs NZ: Samson vs Pant, Kuldeep or Chahal? - Hindustan Times

India Predicted XI for 1st ODI vs NZ: Samson vs Pant, Kuldeep or Chahal? - Hindustan Times

Published Date: Thu, 24 Nov 2022 09:29:09 GMT

Figure 4.5: Favourite Topic



☑ Here are the some World cup News for you

(1) FIFA World Cup 2022 Live Stream: How to watch Portugal vs Ghana on your smartphone - The Indian Express

FIFA World Cup 2022 Live Stream: How to watch Portugal vs Ghana on your smartphone - The Indian Express

Figure 4.6: Search topic

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: Beautiful soup

With the use of Beautiful soup package we will extracting copying and screening of data from various resourses (web scraping)

The beautiful soup parses the unwanted data and helps to organize and format the messy web data by fixing bad HTML and present to us.

Step 4: Newspaper 3k

The data will be passed to the article module of the newspaper 3k library. This is for generating the summary.The newspaper3k is a Python library used for scraping web articles After that it will 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.

Chapter 5

Results and Discussions

This web application is providing the accurate news of the topic which the user selects. Users have the option to select what category of news they want. If the user want to know about the trending news, In this web application it provide the option to select the trending news option. If the user want to know about sports, there is option for favourite topic. From favourite topic user can select the sports. Then the sports news will provide by the application. Not only sports news there are many other topics in the favourite topic. Other topics in favourite news are World, News, Business, Technology, Entertainment, Sports, Science and Health. There is one more category called search topic. In search topic user can search any topic the user need. The topic that user provide will show in the page. This application provides the summary of any news the user wants.

Chapter 6

Conclusion

This project goal was to create a web application which provide the news. The goal is to provide the accurate news to the users. Thus by creating the web application the user get the summary of the news from different sites. When the user select or search any topic the application provides the summary of the news. The user can select the range of the news. In this, the topic summary is not getting by combining different article. Only the single site provided news summary is now showing by this application. This web application is very useful for the users to not read too much of information about a topic. If the user wants to read the content the link of the news is available under the summary. The future work of this application will provide much more summarized data.

6.1 Future Works

This article summarizing web application will get updated according to the system updates and it can automatically add the categories of the news and articles. In future the admin can enhance the performance of this web application on according to the future technical developments. This web application can also be develop as an API which can be support in both android and iOS in future.

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