DOCUMENT SUMMARIZATION

By:

Mohamed Yaseen M S RA2011026040020
Kishore Ramesh RA2011026040040
Saguturu kishan sai RA2011026040031

MINOR PROJECT

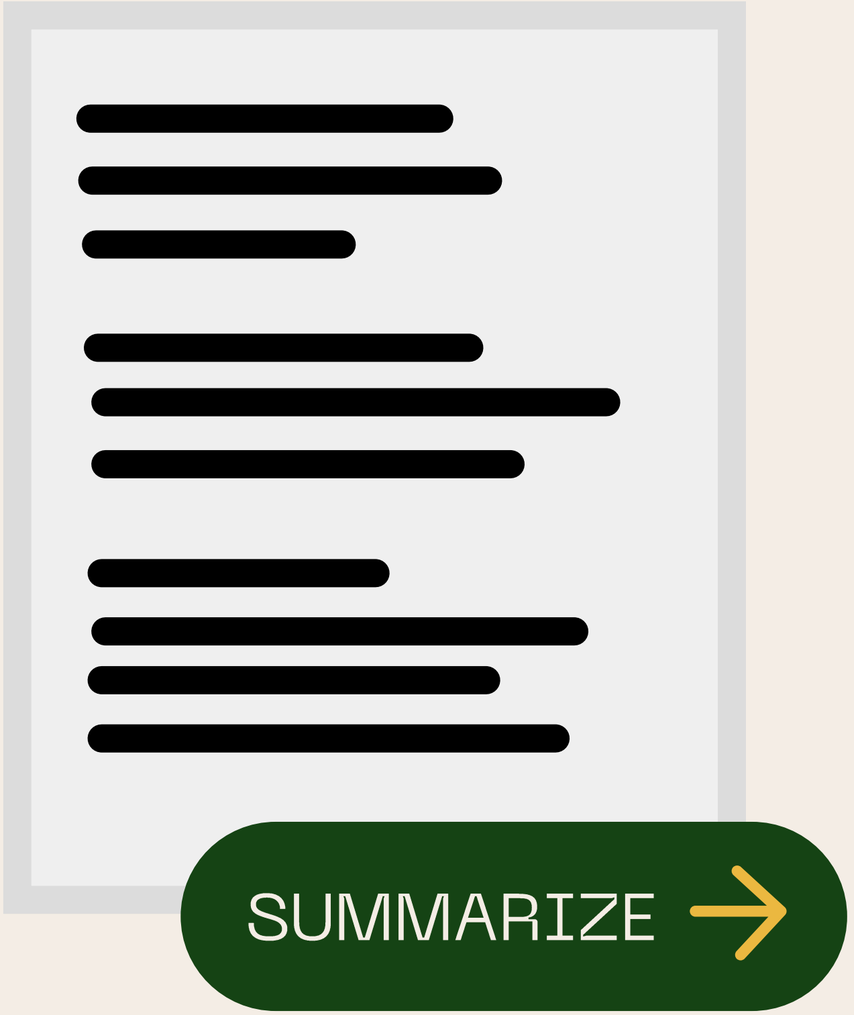
CONTENTS

Introduction

Text summarization is the process of creating a shorter version of a longer text while retaining its most important information.

It has become an important tool in natural language processing (NLP) due to the exponential growth of digital data and the need to quickly extract relevant information.

- Efficiently extract important information from long documents
- Provide a quick overview of a document's content.
- Reduce reading time and increase productivity.



Literature Survey

Document Summarization with Latent Queries by Yumo Xu and Mirella Lapata published on Transactions of the Association for Computational Linguistics, May 2022

Methodology:

- This paper presents a unified modeling framework for generic and query-focused summarization, optimizing a latent query model and a conditional language model. It also introduces a non-parametric calibration method for handling user queries at test time.

Performance Analysis:

- It demonstrates superior performance of the proposed model on various summarization benchmarks, surpassing strong baselines and existing methods on out-of-distribution queries, supported by ablation studies and qualitative analysis.

Limitation and Consideration:

- There are few limitations in the model, including potential issues with summary fidelity and capturing complex query intents. Future work suggestions include incorporating factual consistency constraints, enhancing query representation learning and exploring diverse query types.

Literature Survey

Extractive Summarization of Call Transcripts by Pratik K. Biswas and Aleksandr Iakubovich published on Transactions of the Association for Computational Linguistics, May 2022

Methodology:

- This paper proposes a novel method for extractive summarization of call transcripts, which combines topic modeling, sentence selection and punctuation restoration.

Performance Analysis:

- The performance of the method on four different use cases and compares it with another open-source summarizer. The paper reports that the proposed method achieves higher rouge-l scores and punctuation-restoration-accuracy scores than the baseline summarizer.

Limitation and Consideration:

- The paper addresses limitations and considerations, including dependency on speech-to-text conversion quality, trade-offs between summary length and information content, and the need for improved punctuation restoration.

Methodology

Keywords:
Spacy_rander, Natural Language Processing, Summarization, Streamlit, en_core_web_sm.

Step 1: Text Preprocessing Clean the text by removing stop words, punctuation, and other irrelevant information.	**Step 2: Tokenization** Break the text into individual words or phrases.
Step 3: Part-of-Speech Tagging Identify the part of speech (noun, verb, adjective, etc.) of each word in the text.	**Step 4: Named Entity Recognition** Identify and extract named entities (people, organizations, locations, etc.) from the text.
Step 5: Sentiment Analysis Determine the overall sentiment of the text (positive, negative, neutral).	**Step 6: Text Summarization** Create a concise summary of the text, highlighting the most important information.

Step 1: Text Preprocessing

Text preprocessing is the first step in our project. It is the process of bringing the text into a form that is predictable and analyzable for a specific task.

A task is the combination of approach and domain. For example, extracting top keywords with TF-IDF (approach) from Tweets (domain) is an example of a task1.

The main objective of text preprocessing is to break the text into a form that machine learning algorithms can digest. It involves cleaning and transforming unstructured text data to prepare it for analysis.

Step 2: Tokenization & Step 3: Parts-Of-Speech (POS)

Tokenization is a key step in natural language processing (NLP) where text is divided into individual words or tokens. This allows for more detailed analysis and understanding of the text's structure and meaning.

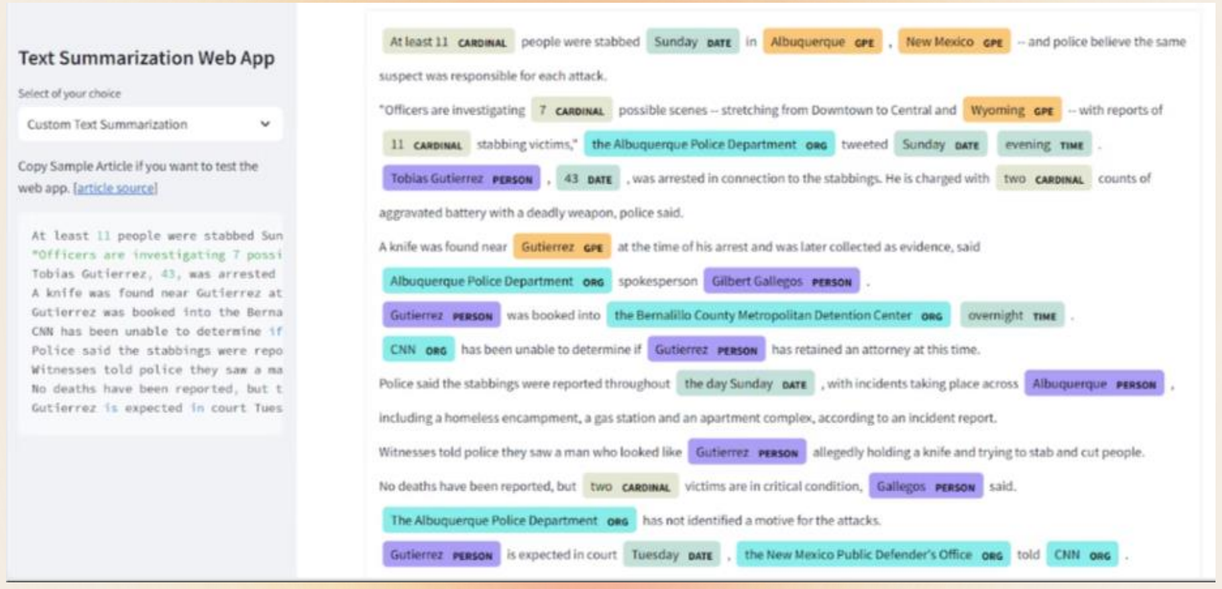
POS tagging assigns grammatical tags to each word in a sentence, indicating its syntactic category (e.g., noun, verb, adjective). This helps in analyzing the sentence's structure and extracting valuable linguistic information.

Tokenization and POS tagging are essential preprocessing steps in NLP, enabling more accurate and in-depth analysis of text data.

Step 4: Name Entity Recognition & Step 5: Sentiment Analysis

Named Entity Recognition is used to identifies and classifies named entities in text, such as people, organizations, locations, and dates. It helps extract valuable information and understand the context of the text, benefiting applications like information retrieval and text summarization.

Sentiment Analysis, or opinion mining, determines the sentiment expressed in text, whether positive, negative, or neutral. Sentiment Analysis provides valuable insights for our data-driven decision-making.



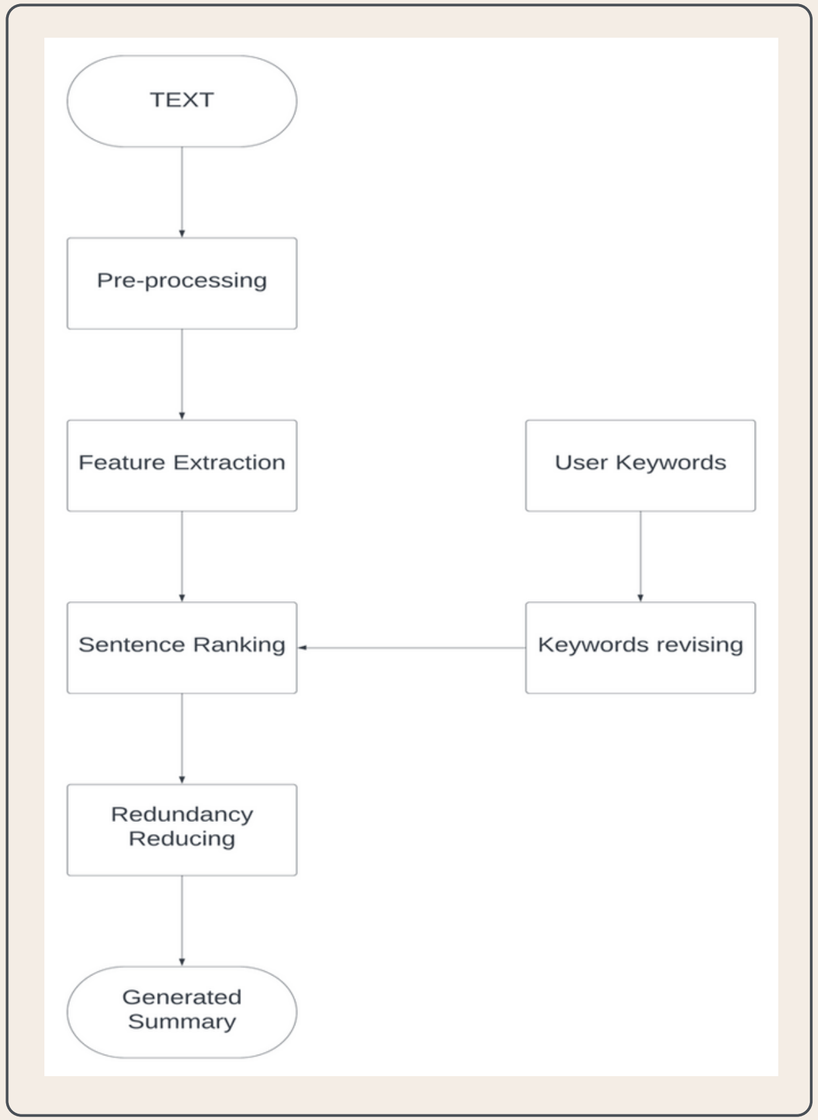
Step 6: Text Summarization & Architecture diagram

Text summarization is a vital NLP technique that condenses a given text while preserving its key information. It involves generating a concise summary that captures the main ideas.

There are two approaches:

Extractive Summarization: Selects important sentences or phrases from the original text to create a summary.

Abstractive Summarization: Generates a summary by understanding the text's meaning and expressing it in a new way.



Results and Discussion

The project implements a text summarization system using NLP techniques and a pre-trained summarizer model (spaCy) based on the Transformer architecture.

The system generates concise and informative summaries for BBC News articles and custom articles.

It leverages the Transformer's self-attention mechanism to capture important information and produce coherent summaries.

The generated summaries are generally accurate, but some details from the original text may be lost due to the limitations of automatic summarization.

The project demonstrates successful application of NLP and the Transformer architecture for text summarization.

It provides a foundation for future advancements in the field of text summarization.

Result

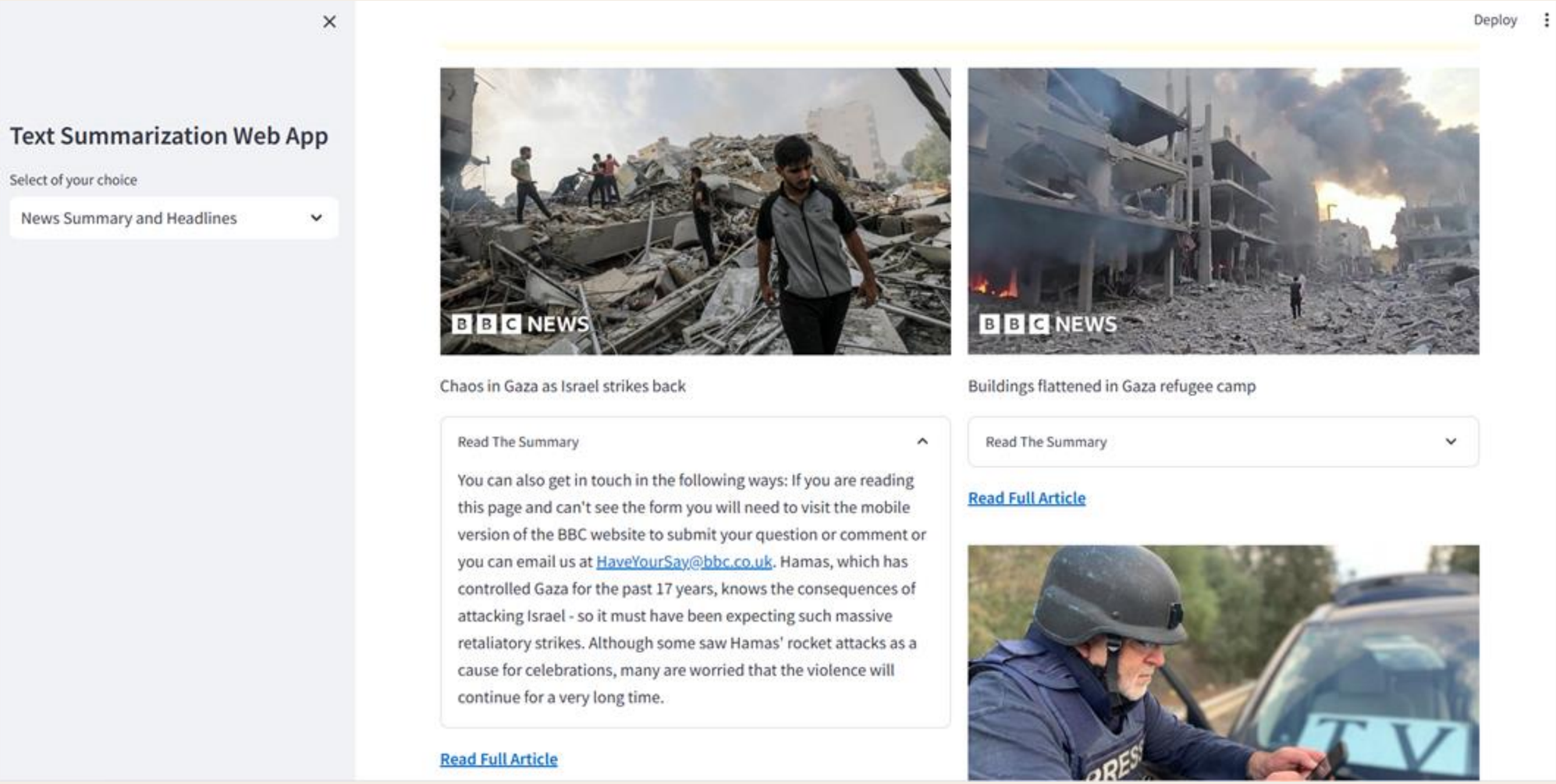


Figure : News Headlines and Summary

Result

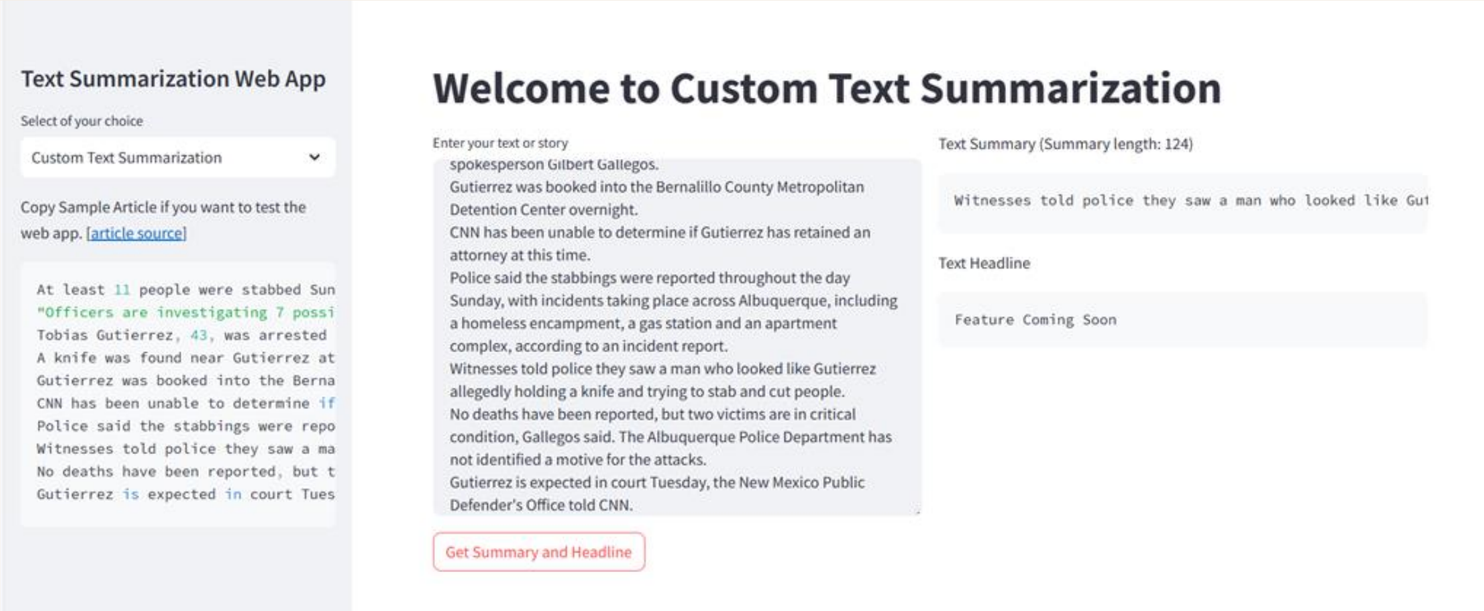


Figure : Custom Text Summarization

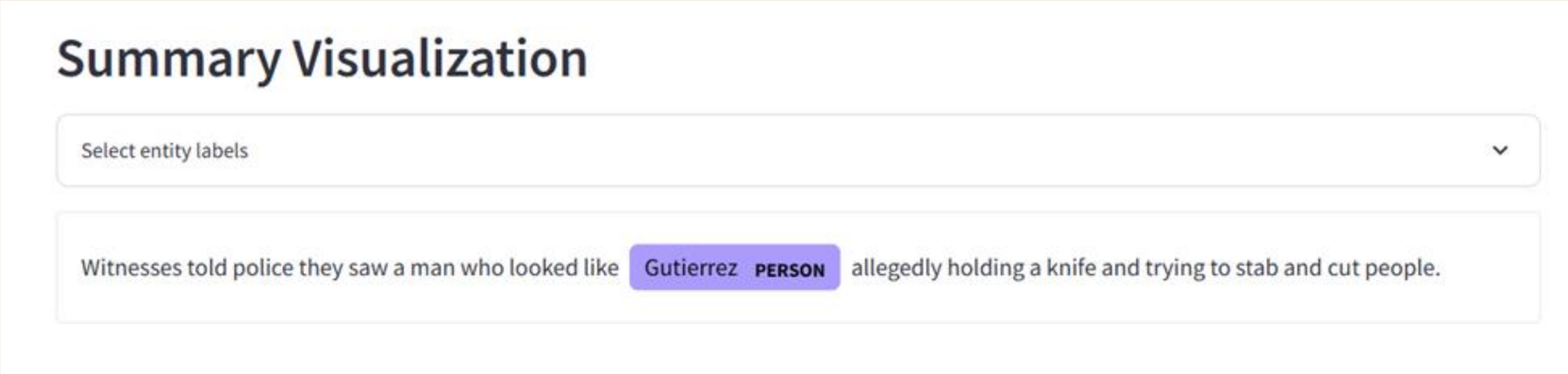


Figure: Summary Visualization

Result

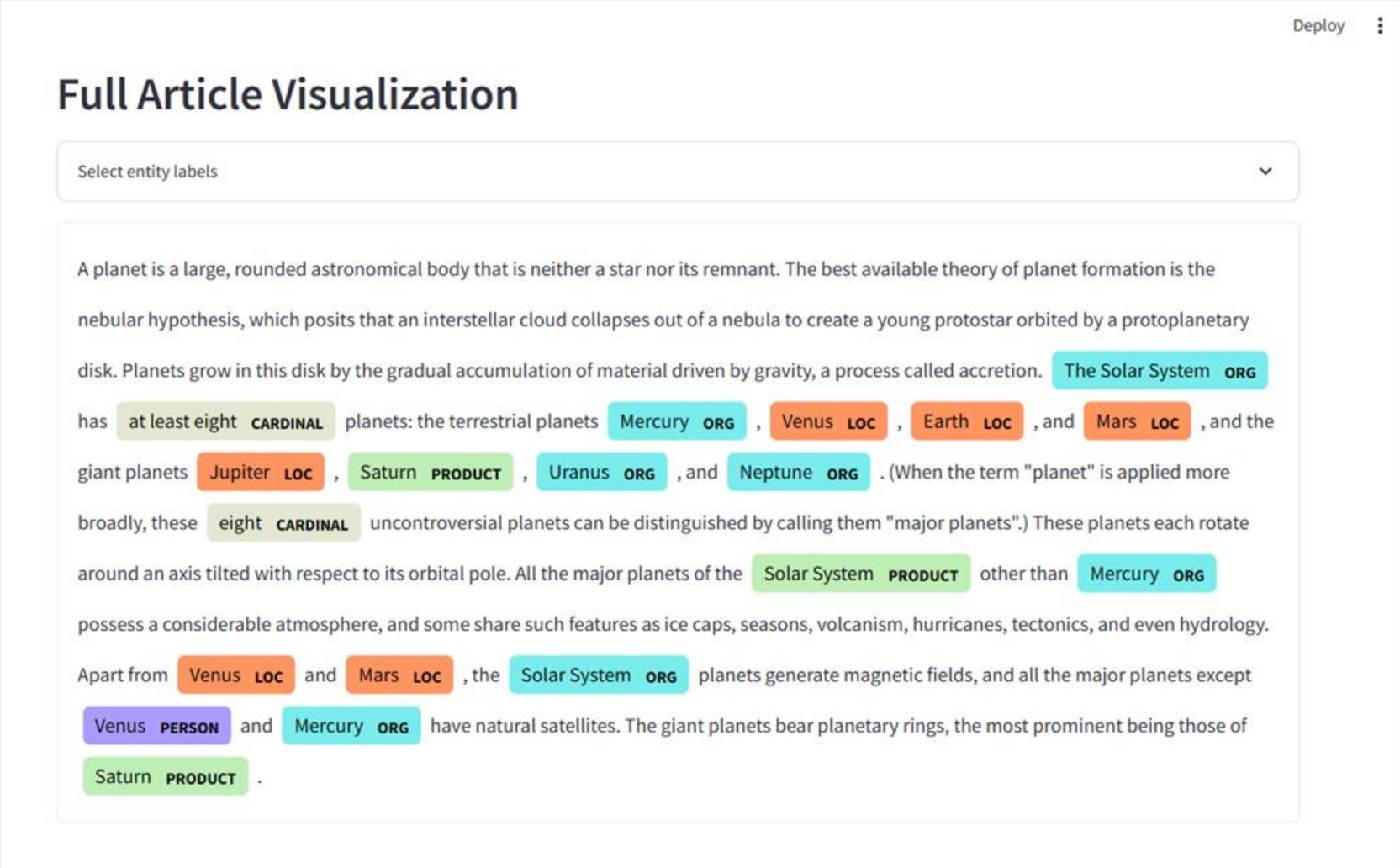


Figure: Full Article Visualization

Conclusion and Future Work

In conclusion, text summarization using NLP is a promising field with many potential applications in various industries. Our methodology has shown promising results in summarizing news articles from BBC News, and we believe that with further research and development, it can be applied to other types of text as well.

Moving forward, we plan to explore the use of deep learning models for text summarization, as well as incorporating more advanced NLP techniques such as sentiment analysis and entity recognition. We also plan to expand our dataset to include a wider range of text sources and languages.

References

[1] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Shaomeng Cao, Kechi Zhang, Zhi Jin.
"Interpretation-based Code Summarization" (2023).

[2] Aakash Bansal, Chia-Yi Su, Collin McMillan.
"Revisiting File Context for Source Code Summarization" (2023).

[3] Chia-Yi Su, Collin McMillan.
"Distilled GPT for Source Code Summarization" (2021).

[4] Yao Wan, Zhou Zhao, Min Yang, Guandong Xu, Haochao Ying, Jian Wu, Philip S. Yu.
"Improving Automatic Source Code Summarization via Deep Reinforcement Learning" (2020).

[5] Wenhua Wang, Yuqun Zhang, Yulei Sui, Yao Wan, Zhou Zhao, Jian Wu, Philip S. Yu and Guandong Xu.
"Reinforcement-Learning -Guided Source Code Summarization Using Hierarchical Attention" (2020).

[6] Yumo Xu and Mirella Lapata. "Document Summarization with Latent Queries" (2021).

[7] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, GeLI, Zhi Jin, Xiaoguang Mao and Xiangke Liao.
"An Empirical Study on Using Large Language Models for Multi-Intent Comment Generation" (2023).

[8] Partik K Biswas, Aleksandr lakubovich.
"Extractive Summarization of Call Transcripts" (2022)

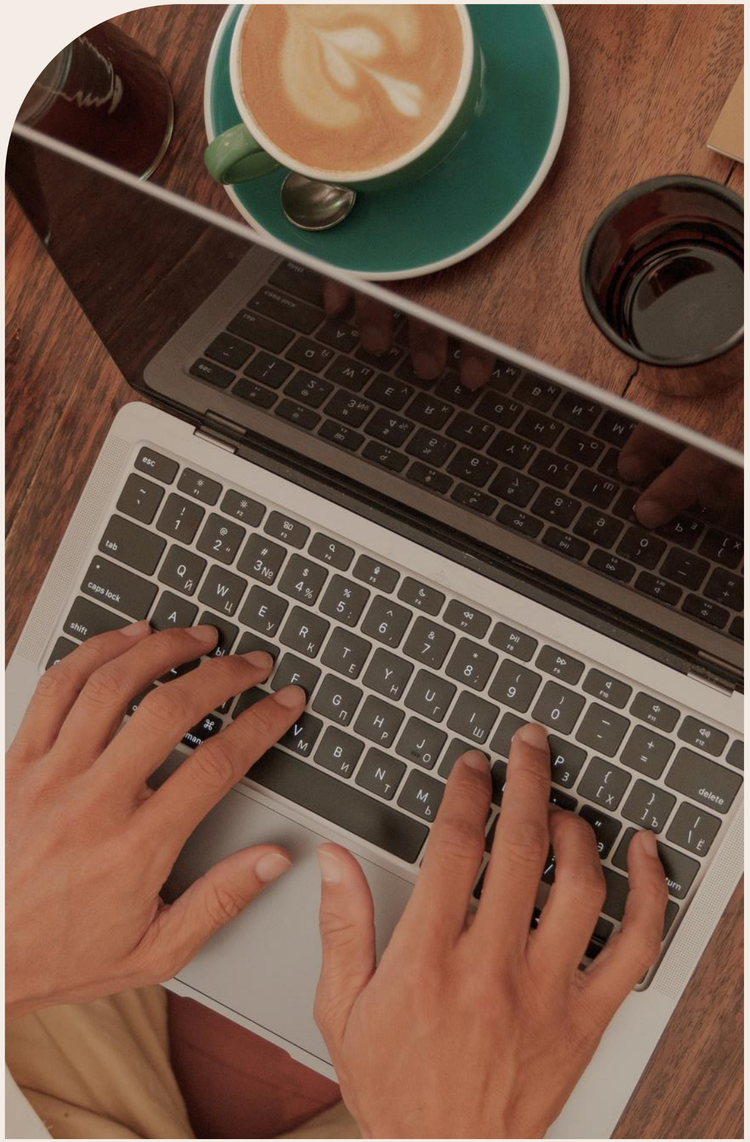
Thank you!

Team:

Mohamed Yaseen M S	RA2011026040020
Kishore Ramesh	RA2011026040040
Saguturu Kishan Sai	RA2011026040031

Guide:

Mr. Muthurasu N



Conference Certificate

INSTITUTION'S INNOVATION COUNCIL (Ministry of HRD Initiative) ICTACADEMY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
VADAPALANI CAMPUS

SRM VADAPALANI

NATIONAL CONFERENCE ON TECHNOLOGY FOR SOCIETY (NCTS'23)

TRANSFORMING SOCIETY WITH TECH: OUR PATH TO SDG SUCCESS

Certificate of Participation

This is to certifiy that

MOHAMED YASEEN

of SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, VADAPALANI CAMPUS College has participated and presented a paper titled TEXT SUMMARIZATION USING NLP in the National Conference on Technology for Society (NCTS'23) organized by department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, held on 18th and 19th October, 2023.

Dr.P.Durgadevi	Dr.B.Prabha	Dr.S.Prasanna Devi	Dr.C.Gomathy	Dr.C.V.Jayakumar
Organizing Secretary	Organizing Secretary	HOD (CSE) & Convener	VP (Academics & Placements)	Dean (FET)

INSTITUTION'S INNOVATION COUNCIL (Ministry of HRD Initiative) ICTACADEMY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
VADAPALANI CAMPUS

SRM VADAPALANI

NATIONAL CONFERENCE ON TECHNOLOGY FOR SOCIETY (NCTS'23)

TRANSFORMING SOCIETY WITH TECH: OUR PATH TO SDG SUCCESS

Certificate of Participation

This is to certifiy that

KISHORE RAMESH

of SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, VADAPALANI CAMPUS College has participated and presented a paper titled TEXT SUMMARIZATION USING NLP in the National Conference on Technology for Society (NCTS'23) organized by department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, held on 18th and 19th October, 2023.

Dr.P.Durgadevi	Dr.B.Prabha	Dr.S.Prasanna Devi	Dr.C.Gomathy	Dr.C.V.Jayakumar
Organizing Secretary	Organizing Secretary	HOD (CSE) & Convener	VP (Academics & Placements)	Dean (FET)

INSTITUTION'S INNOVATION COUNCIL (Ministry of HRD Initiative) ICTACADEMY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
VADAPALANI CAMPUS

SRM VADAPALANI

NATIONAL CONFERENCE ON TECHNOLOGY FOR SOCIETY (NCTS'23)

TRANSFORMING SOCIETY WITH TECH: OUR PATH TO SDG SUCCESS

Certificate of Participation

This is to certifiy that

SAGUTURU KISHAN SAI

of SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, VADAPALANI CAMPUS College has participated and presented a paper titled TEXT SUMMARIZATION USING NLP in the National Conference on Technology for Society (NCTS'23) organized by department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, held on 18th and 19th October, 2023.

Dr.P.Durgadevi	Dr.B.Prabha	Dr.S.Prasanna Devi	Dr.C.Gomathy	Dr.C.V.Jayakumar
Organizing Secretary	Organizing Secretary	HOD (CSE) & Convener	VP (Academics & Placements)	Dean (FET)