Infosys Internship 4.0 Project Presentation

Title:TEXT SUMMARIZATION

NISHA

Mentor: Mr. Narendra Kumar

Table of contents

❖ Acknowledgment

- I would like to thank Infosys for giving opportunity to work as an AI/ML Intern.
- Special thanks to my mentor, Mr. Narendra Kumar, for his invaluable guidance and support.
- I am also grateful to my team and colleagues for their continuous support and collaboration throughout this project.
- Additionally, I am thankful to my family for their encouragement and assistance throughout this project.

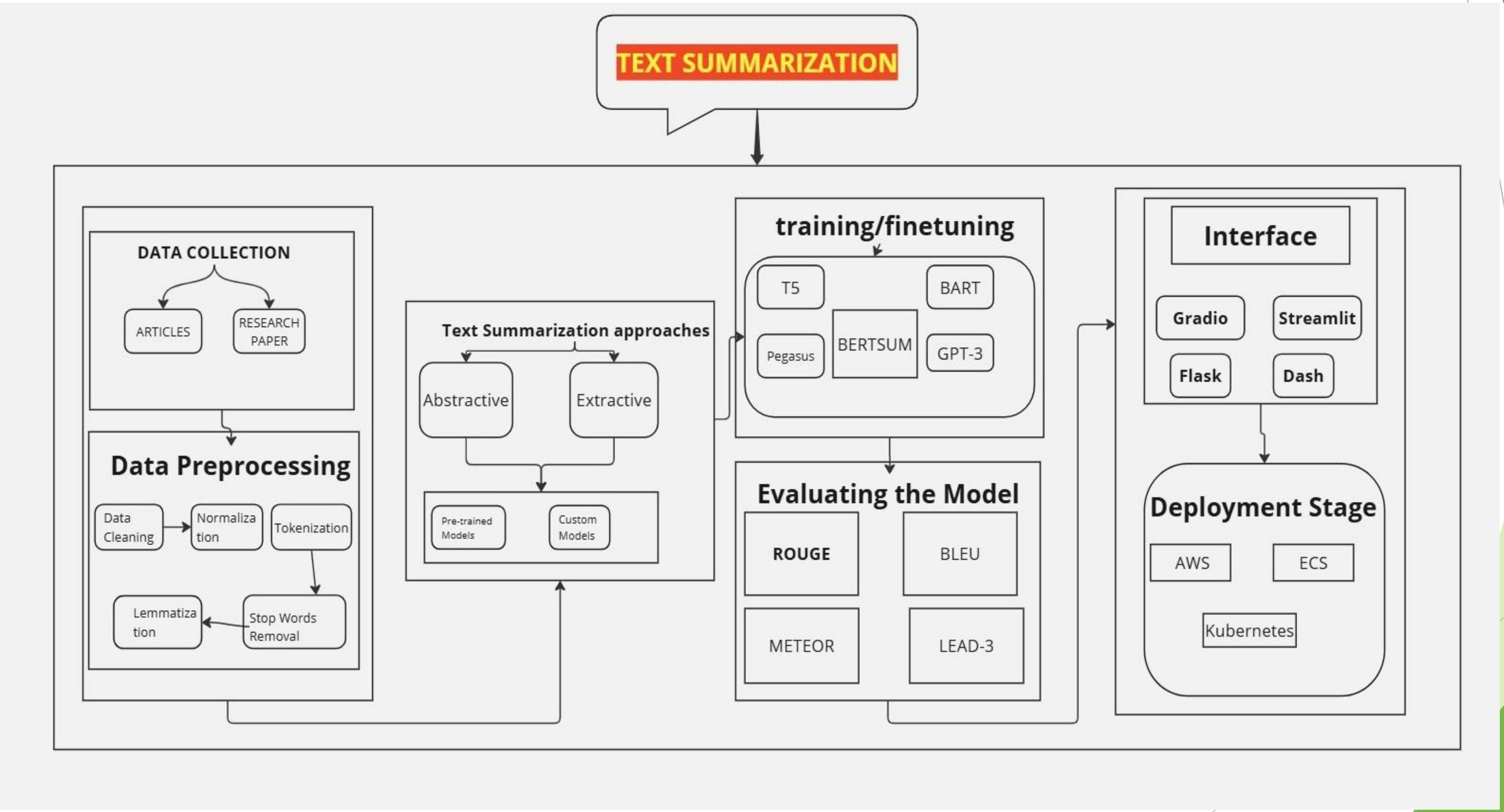
❖ Problem Statement

The project aims to develop an effective text summarization model that condenses extensive texts into concise summaries with high accuracy and relevance.

❖ Introduction

This report outlines the problem statement, workflow, data collection, preprocessing steps, summarization methodologies, user interface design, and results. It also discusses observations, conclusions, and potential future enhancements, showcasing the system's ability to improve information retrieval and readability across various domains.

❖ workflow



TEXT SUMMARIZATION
DATA COLLECTION
ARTICLES
RESEARCH PAPER
Data Preprocessing
Data Cleaning
Normalization
Tokenization
Lemmatization
Stop Words Removal
Text Summarization approaches
Abstractive
Extractive
Pre-trained Models
Custom Models
training/finetuning
T5
BART
Pegasus
BERTSUM
GPT-3
Evaluating the Model
ROUGE
BLEU
METEOR
LEAD-3
Interface
Gradio
Streamlit
Flask
Dash
Deployment Stage
AWS
ECS
Kubernetes

❖ Data collection

Source: I collected a dataset from the CNN/DailyMail website for this project.

Relevance: This dataset is ideal for training summarization models because of its rich and diverse content.

Dataset Description

- **Initial Dataset**: The main dataset was initially saved as dataset.csv, containing 70,000 records.

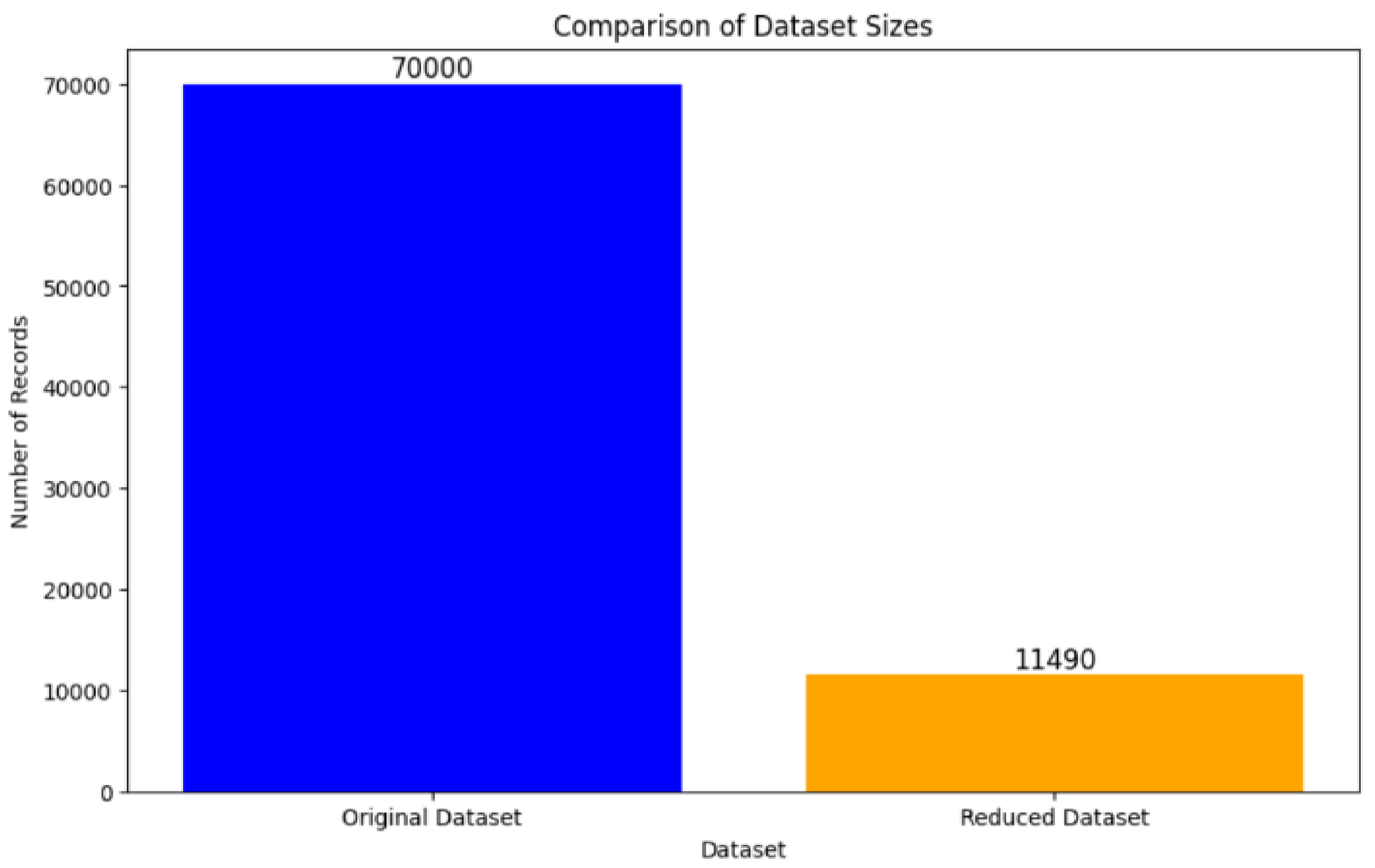
Structure:

- **id**: Unique identifier for each record.
- **article**: The body of the news article.
- **highlights**: The summary or highlights of the article.

Dataset Links

- **Master Dataset**: You can find the link to the master dataset here.

Data Reduction



- Reduced Dataset: You can find the link to the reduced dataset here.

❖ Data Preprocessing

Objective:

- The primary goal of data preprocessing was to clean and prepare the dataset for model training by removing noise, tokenizing the text, and eliminating stopwords.

Process:

- The dataset before preprocessing is shown in the below picture

Out[2]:

	id	article	highlights
0	8aa8d3d042356a88d25ee6fb13347184858fe770	(RollingStone.com) -- Britney Spears announce...	Britney Spears and producers still choosing so...
1	b3a6c45ccbcc6140a9fe042a385440e3a80535dc	By . Sam Adams . PUBLISHED: . 04:02 EST, 18 Ju...	Car owners would be liable even if they don't ...
2	f90015991bcec3013e502044699046581088f1a5	It is a single moment of horrifying barbarism ...	The picture was posted on a pro-government web...
3	0e029a3f67dc8df34eefc185ec5343cec72fb29d	An elderly Minnesota couple were killed after ...	Carlton and Hazel Roed of Mentor, Minnesota, w...
4	b244323ba60a10baf71a72a30ffed5162f3b2050	(CNN) -- Columbus Day often brings to mind the...	Seattle and Minneapolis will celebrate Indigen...

Outcomes:

- Below picture shows cleaned dataset after preprocessing

Out[8]:

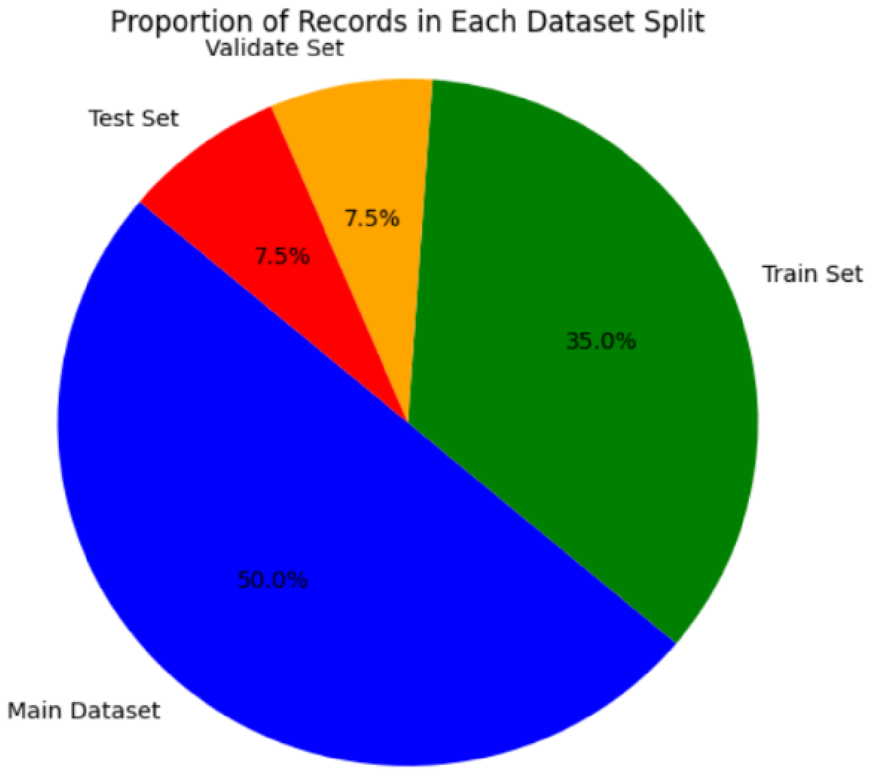
	id	article	highlight
0	8aa8d3d042356a88d25ee6fb13347184858fe770	rollingstonecom britney spears announced today...	britney spears producers still choosing songs ...
1	b3a6c45ccbcc6140a9fe042a385440e3a80535dc	sam adams published 0402 est 18 july 2012 upda...	car owners would liable even dont know dropped...
2	f90015991bcec3013e502044699046581088f1a5	single moment horrifying barbarism provides fl...	picture posted progovernment website lebanon b...
3	0e029a3f67dc8df34eefc185ec5343cec72fb29d	elderly minnesota couple killed car collided h...	carlton hazel roed mentor minnesota 2009 chevy...
4	b244323ba60a10baf71a72a30ffed5162f3b2050	cnn columbus day often brings mind nina pinta ...	seattle minneapolis celebrate indigenous peopl...

Steps Undertaken:

- **Cleaning Text Data:**
- **Tokenization:**
- **Stopword Removal:**
- **Preprocessing Articles and Highlights:**
- **Dataset Splitting**:

Dataset Splitting

- Split the preprocessed dataset into three distinct sets: training (70%), validation (15%), and test (15%) sets to facilitate model training and evaluation.
- A pie chart was generated to visually represent the proportion of records in each dataset split



❖ Abstractive Summarization

- **Definition:** Abstractive summarization creates new sentences to capture the essence of the original text.

- **Examples:** T5, BART, and GPT.

- **Selected model:** T5-small Transformer model from Hugging Face

- **Training process:**

Initial Losses: Training loss started at 1.05 and validation loss at 0.96

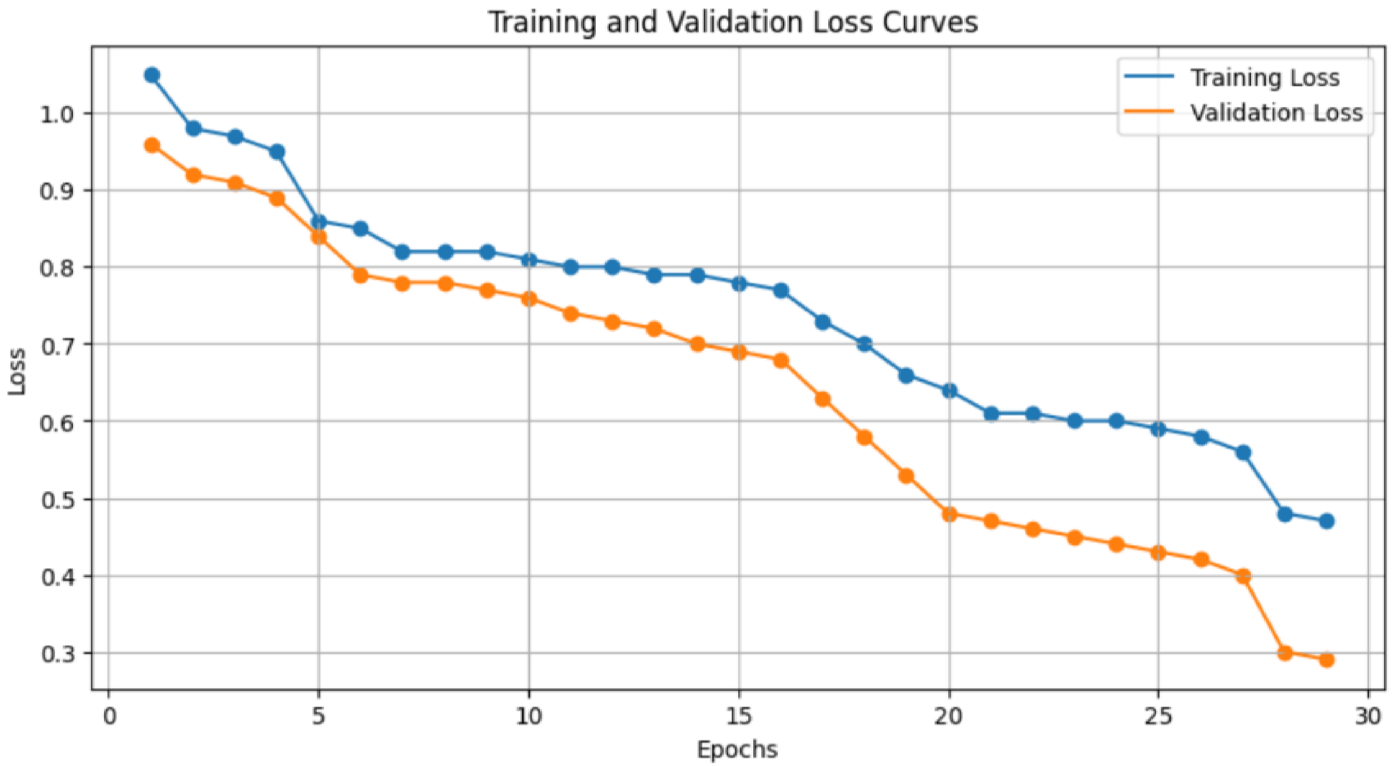
```
  warnings.warn(
Epoch 1: 100%|████████████████████████████████████████| 403/403 [10:07:02<00:00, 90.38s/it, train_loss=1.25]
Average training loss: 1.0562968504044317
Validation loss: 0.9614474930982481
Further training completed and model saved to fine_tuning.
```

Final Losses: Reduced to 0.47 for training and 0.29 for validation.

```
Average training loss: 0.47855778578759567
Validation loss: 0.2974807008135098
Model improved. Saving the model.
Training completed.
```

- # Observations:

Loss Reduction Post Hyperparameter Tuning



- The validation loss saw a significant decrease from an initial 0.96 to a final 0.29 after hyperparameter tuning

- Simultaneously, the training loss also reduced notably from 1.05 to 0.40, indicating improved model

convergence and effectiveness in generating accurate summaries

❖ Performance Matrix

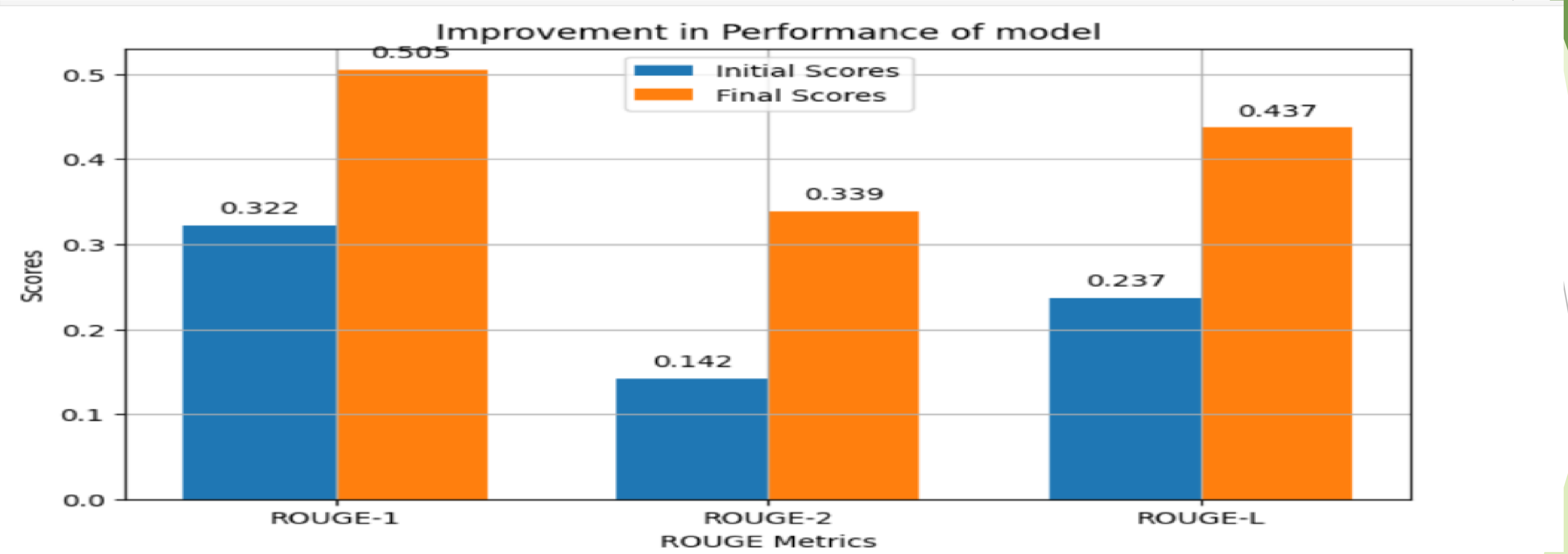
Objective:

- **Evaluate Model Performance:** Quantitatively assess how well the generated summaries match the reference summaries.
- **Available Evaluation Metrics:** BLEU, ROUGE,
- **Used Performance Matrix (ROUGE):** Recall-Oriented Understudy for Gisting Evaluation
- **ROUGE Score Analysis:**
  - **ROUGE-1:**
  - **ROUGE-2:**
  - **ROUGE-L:**

```
Special tokens have been added in the vocabulary, mak
100%|██████████| 216/216 [06:21<00:00,  1.76s/it]
ROUGE-1: 0.5051 ROUGE-2: 0.3398 ROUGE-L: 0.4377
```

❖ Observations

Performance Improvement



- Post tuning, these metrics improved to 0.505, 0.339, and 0.437, demonstrating substantial progress in summarization accuracy
- Initial ROUGE-1, ROUGE-2, and ROUGE-L scores stood at 0.322, 0.142, and 0.237 respectively

❖ Extractive text summarization

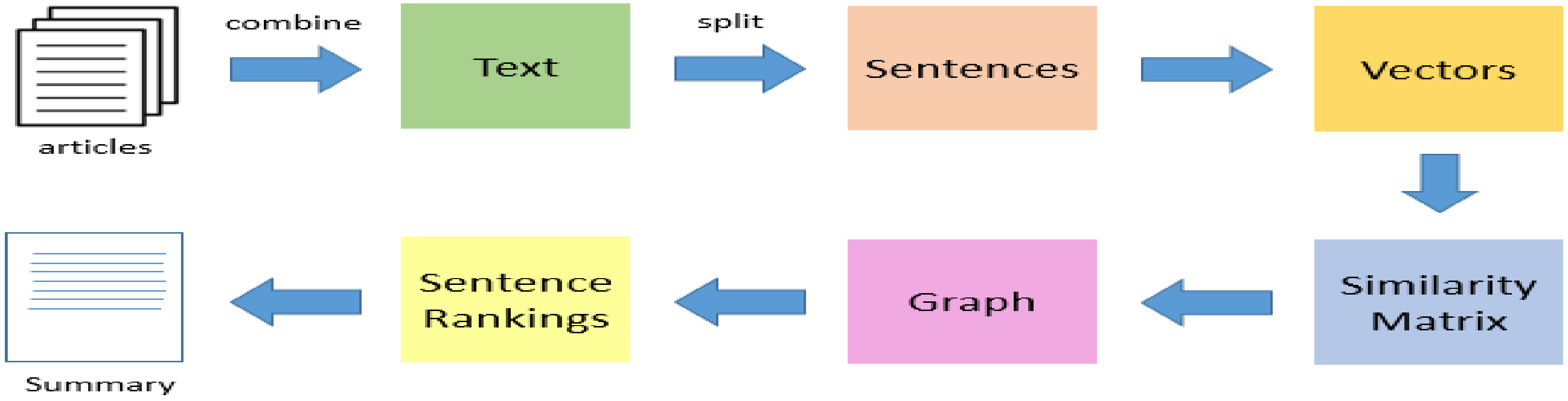
- **Objective:** The objective is to develop a system that generates concise summaries by selecting key sentences from the source text, and preserving original wording and context.

- **Examples:** Techniques like TextRank, TF-IDF, and pre-trained models such as BERT and GPT.

Model training methodology

- **Used textrank algorithm**

Methodology



- **Performance Metrics and Scores:**

  - **Used Performance Matrix (ROUGE):**

    - **ROUGE Score (**Recall-Oriented Understudy for Gisting Evaluation).

- **RESULT:**

  - Precision: The average precision score was approximately 0.119.

  - Recall: The average recall score was approximately 0.813.

  - F1 Score: The average F1 score was approximately 0.203

```
Summarization complete. Summarized data saved to 'summarized_train_new1.csv'.
Average Precision: 0.11943901651820747
Average Recall: 0.8127170619817642
Average F1 Score: 0.20283625093778002
```

- **Observations**

  - **High Recall:** High recall indicates that the model captures most of the relevant information from the original text.
  - **Moderate Precision and F1 Scores:** The model shows moderate precision (0.119) and a balanced F1 score (0.203), indicating the potential for improving sentence relevance selection in summaries.
  - **Resource Efficiency:** The model is resource-efficient and suitable for environments with limited computational power.

❖ User Interface

- **Overview**

  - The interface is developed using the Gradio library, providing an interactive platform for text summarization.
  - Users can input text or upload PDF files and choose between abstractive and extractive summarization methods.

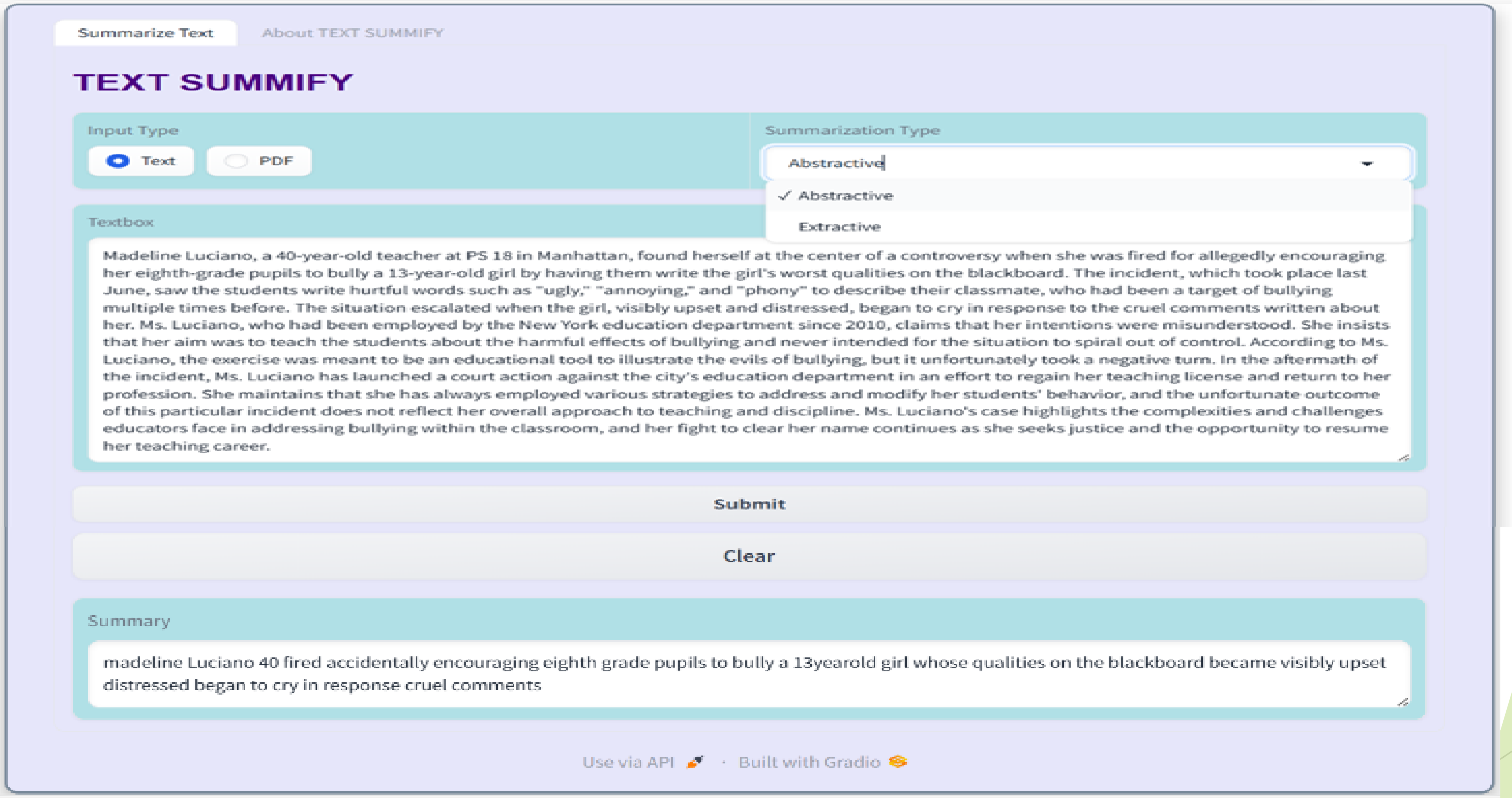
- **Implementation Overview**

Uses the T5 model to generate abstractive summarization

Implements the TextRank algorithm using CountVectorizer, cosine similarity, and PageRank to rank and select sentences for the extractive summarization

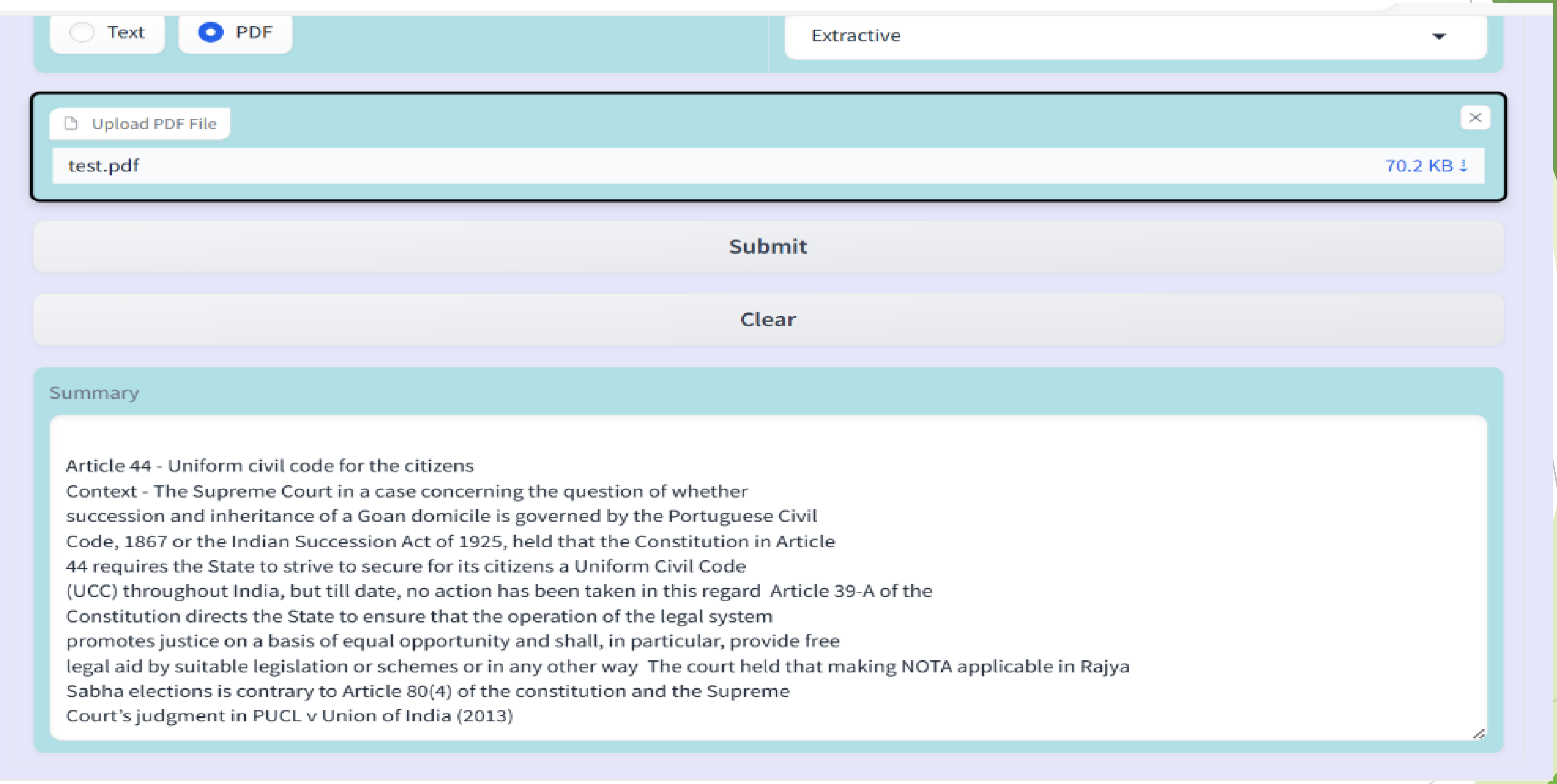
PDF Text Extraction: Uses PyMuPDF (fitz) to extract text from PDF files.

❖ Results



- The above figure shows the interface that summarizes long text into a summary using abstractive summarization with the T5-small model

❖ Results



The above figure shows the interface that summarizes by extracting the text from the pdf using textrank algorithm.

- **Observations**

- Efficient performance in summarizing both text and PDF inputs.
- It produced accurate summaries using the T5 model for abstractive summarization and TextRank for extractive summarization.
- Summarized outputs were concise an preserved essential information from the original texts or documents.

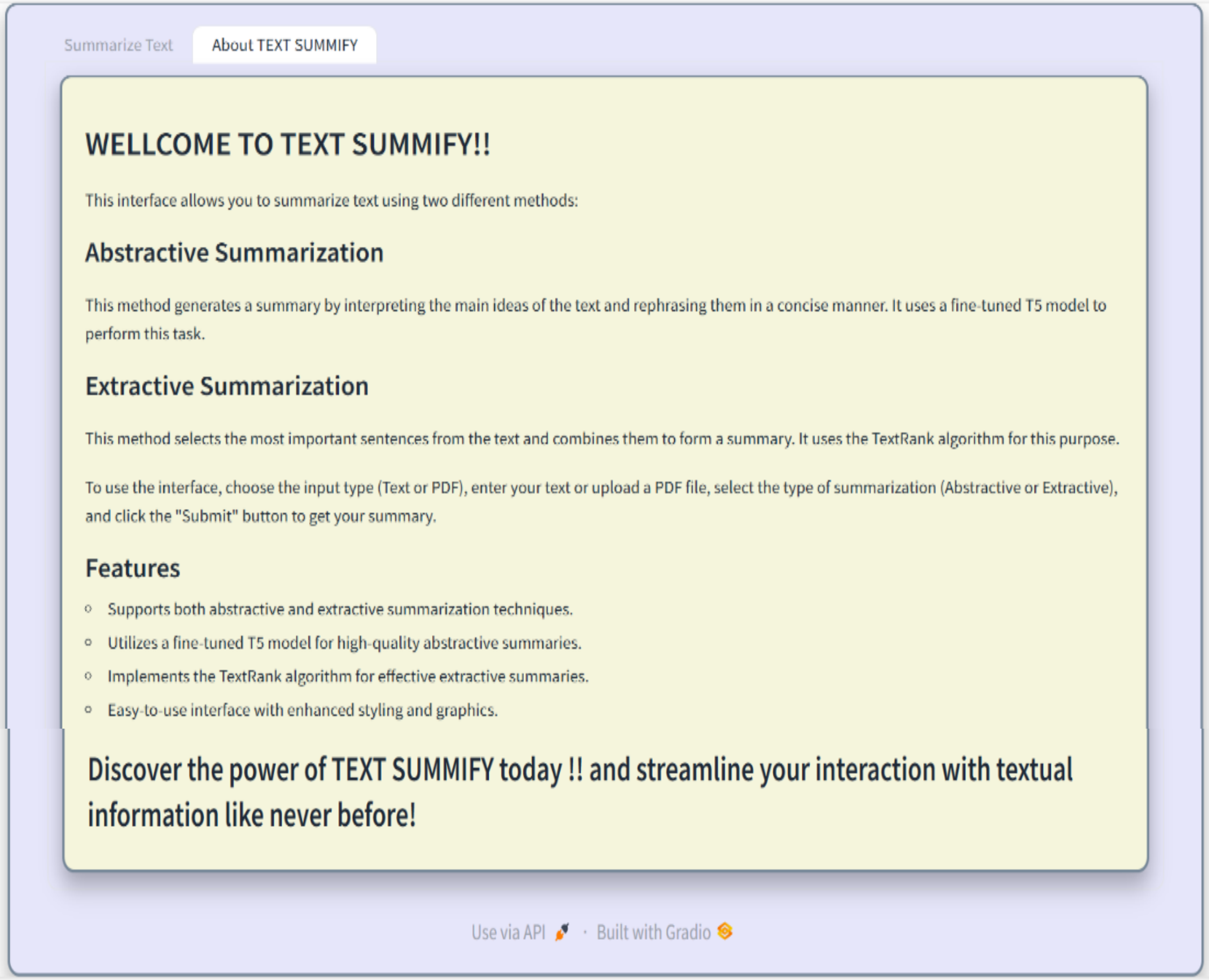


Fig:Description page

❖ Challenges Faced and Solutions

Challenge: Initially faced difficulties in improving model performance due to resource limitations and increased training times on local systems.

Solution: Overcame these limitations by migrating computations to Google Colab, utilizing powerful GPUs to expedite training and inference processes. Additionally, downsized the dataset through strategic sampling and preprocessing techniques to optimize efficiency without compromising quality.

❖ Conclusion

- The development of an automated text summarization system using advanced NLP techniques has proven effective and efficient.
- The Gradio interface provided a user-friendly platform for summarizing long texts and PDF documents.
- This project highlighted the potential of NLP technologies to enhance information retrieval and readability, benefiting various business applications and beyond.

❖ Future scope

- **Model Improvement**: Implementing more advanced pre-trained models such as T5-base or T5-large could further improve the accuracy and coherence of the summaries.
- **Multi-language Support**: Expanding the system to support multiple languages to cater to a more diverse user base.
- **Additional Summarization Techniques:** Exploring and incorporating other summarization methods, such as neural network-based extractive summarization and hybrid models, to offer users more choices.
- **Integration with Other Platforms:** Integrating the summarization tool with popular content management systems and collaboration platforms to broaden its applicability.

Thank you 🌟