TEXT SUMMARIZATION PROJECT

Under the Guidance of Mr. Sai Vighnesh
Presented by Apoorva Khajbage

Introduction

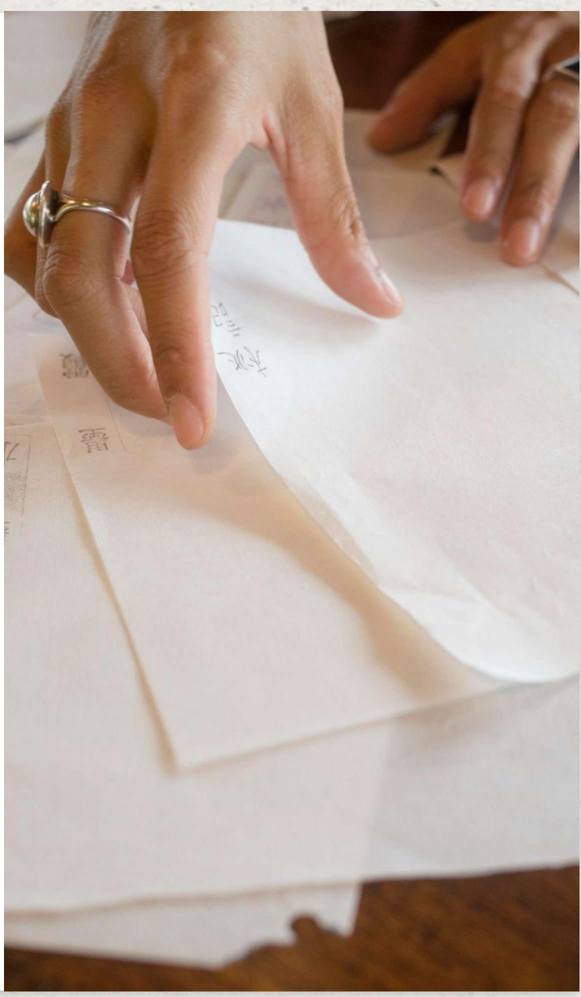
Text summarization is the process of condensing a long text into a shorter version while retaining its key information and meaning.

Importance:

- Saves time by extracting relevant information quickly.
- Enhances comprehension by focusing on key points.

Applications:

- News summarization
- Legal document summarization
- Academic paper abstracts
- Customer feedback analysis
- Chatbot and virtual assistant responses



Problem Statement

- Developing an automated text summarization system that can accurately and efficiently condense large bodies of text into concise summaries is essential for enhancing business operations.

- This project aims to deploy NLP techniques to create a robust text summarization tool capable of handling various types of documents across different domains.

- **Objective:** The system should deliver high-quality summaries that retain the core information and contextual meaning of the original text.

Methodology Overview

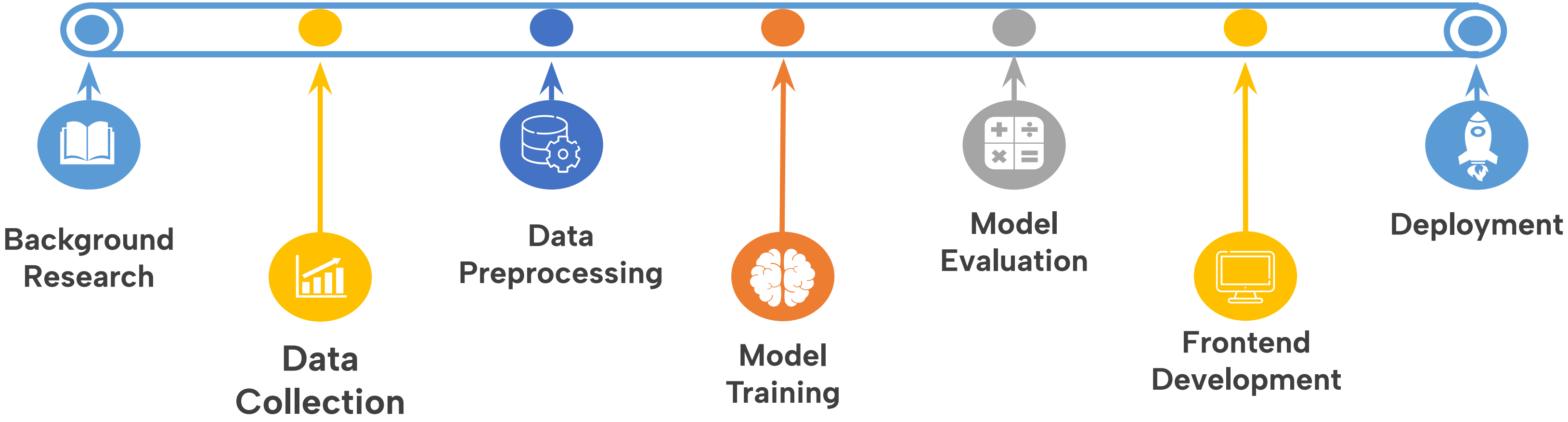
Abstractive Summarization

Involves generating a summary by paraphrasing the original text.
It uses advanced natural language processing (NLP) techniques to understand the content and create new sentences that convey the core meaning of the source material.
Generates novel sentences that may not appear in the original text.
Model Used: BART

Extractive Summarization

Involves selecting and extracting key sentences or phrases directly from the source text to create a summary.
It does not generate new sentences but instead identifies the most important content and reuses it as-is.
The summarized text retains the original wording and structure of the source.
Model Used: BERT

Workflow Of The Project



Background Research

Data Collection

Data Preprocessing

Model Training

Model Evaluation

Frontend Development

Deployment

kaggle

Dataset Description

- Dataset: CNN–DailyMail News Text Summarization Dataset

- The CNN - DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail.

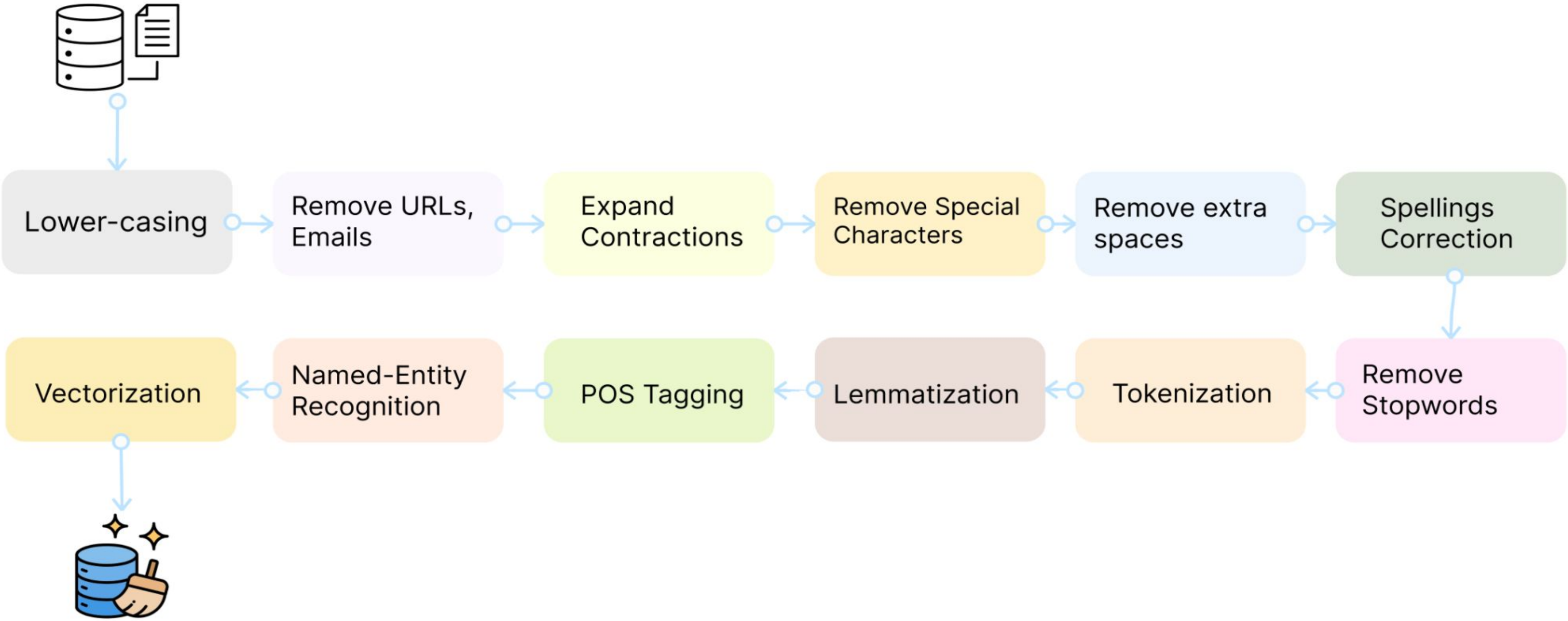
GOWRI SHANKAR PENUGONDA · UPDATED 3 YEARS AGO

▲ 172

CNN-DailyMail News Text Summarization

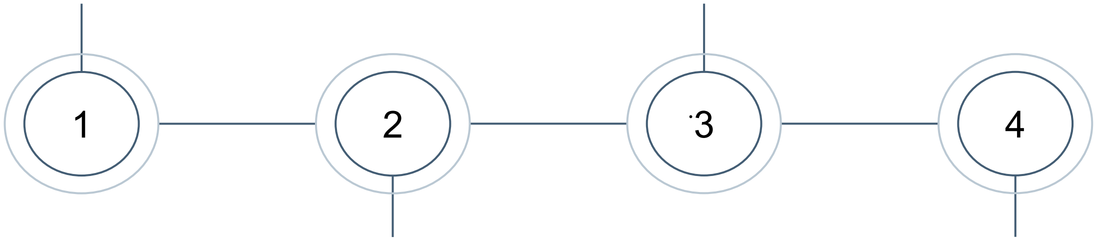
News Articles and summary from CNN-DailyMail Dataset

Preprocessing Steps:



Abstractive Summarization

- Fine-tuned BART-large-CNN for abstractive text summarization.



Data Preparation

Preprocessed dataset .
Split into training (80%)
and evaluation (20%)

Tokenization

Used BartTokenizer to
process text.
Applied truncation for
uniform input size.

Model Training

Fine-tuned using
`Seq2SeqTrainer`.
Configured hyperparameters

Evaluation Metrics

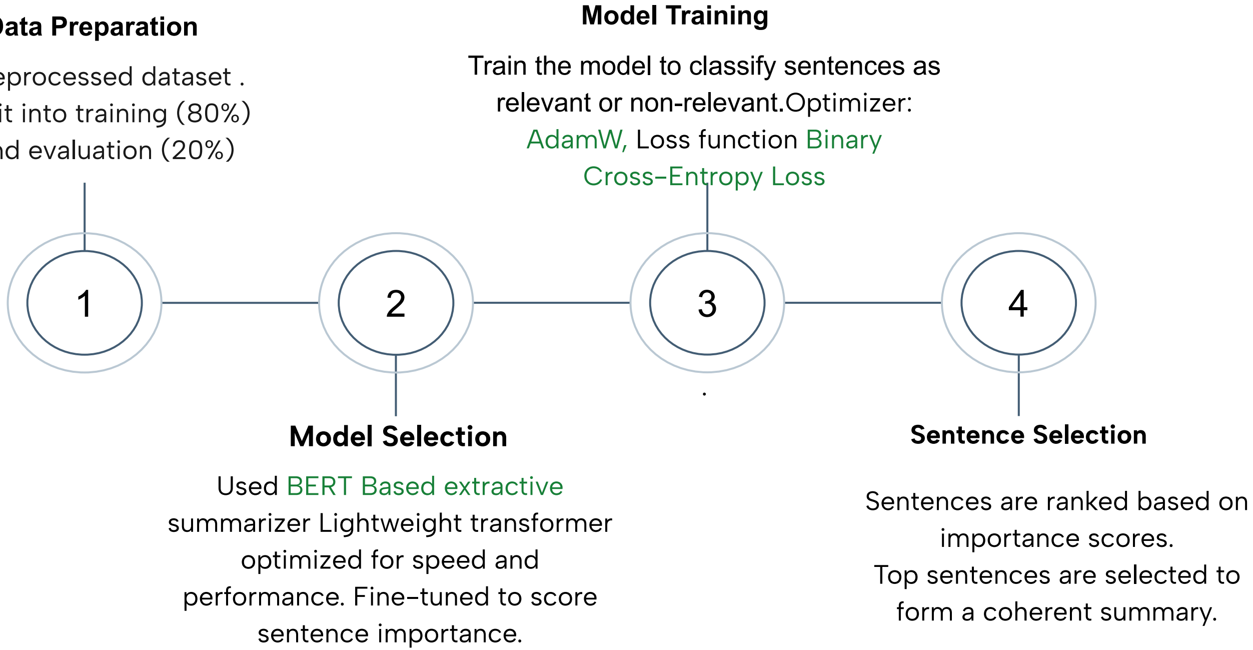
Measured performance using
ROUGE-1, ROUGE-2, and
ROUGE-L.

Results:

- **Training Loss:** Decreased consistently.
- **Validation Loss:** Early stopping avoided overfitting.
- **ROUGE Scores:** Improved across epochs.

Extractive Summarization

- Extract key sentences directly from the text to form summaries.



Data Preparation

Preprocessed dataset .
Split into training (80%)
and evaluation (20%)

Model Training

Train the model to classify sentences as relevant or non-relevant.Optimizer: AdamW, Loss function Binary Cross-Entropy Loss

Model Selection

Used BERT Based extractive summarizer Lightweight transformer optimized for speed and performance. Fine-tuned to score sentence importance.

Sentence Selection

Sentences are ranked based on importance scores.
Top sentences are selected to form a coherent summary.

Evaluation Metrics

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics commonly used to evaluate the quality of summaries by comparing the overlap between the generated summary and a reference summary.
- Measures the similarity of n-grams (sequence of words) between the generated and reference summaries.

1. ROUGE-1

- Measures overlap of unigrams (single words) between the generated and reference summaries.
- **Significance:** Basic word-level match, indicating content preservation.

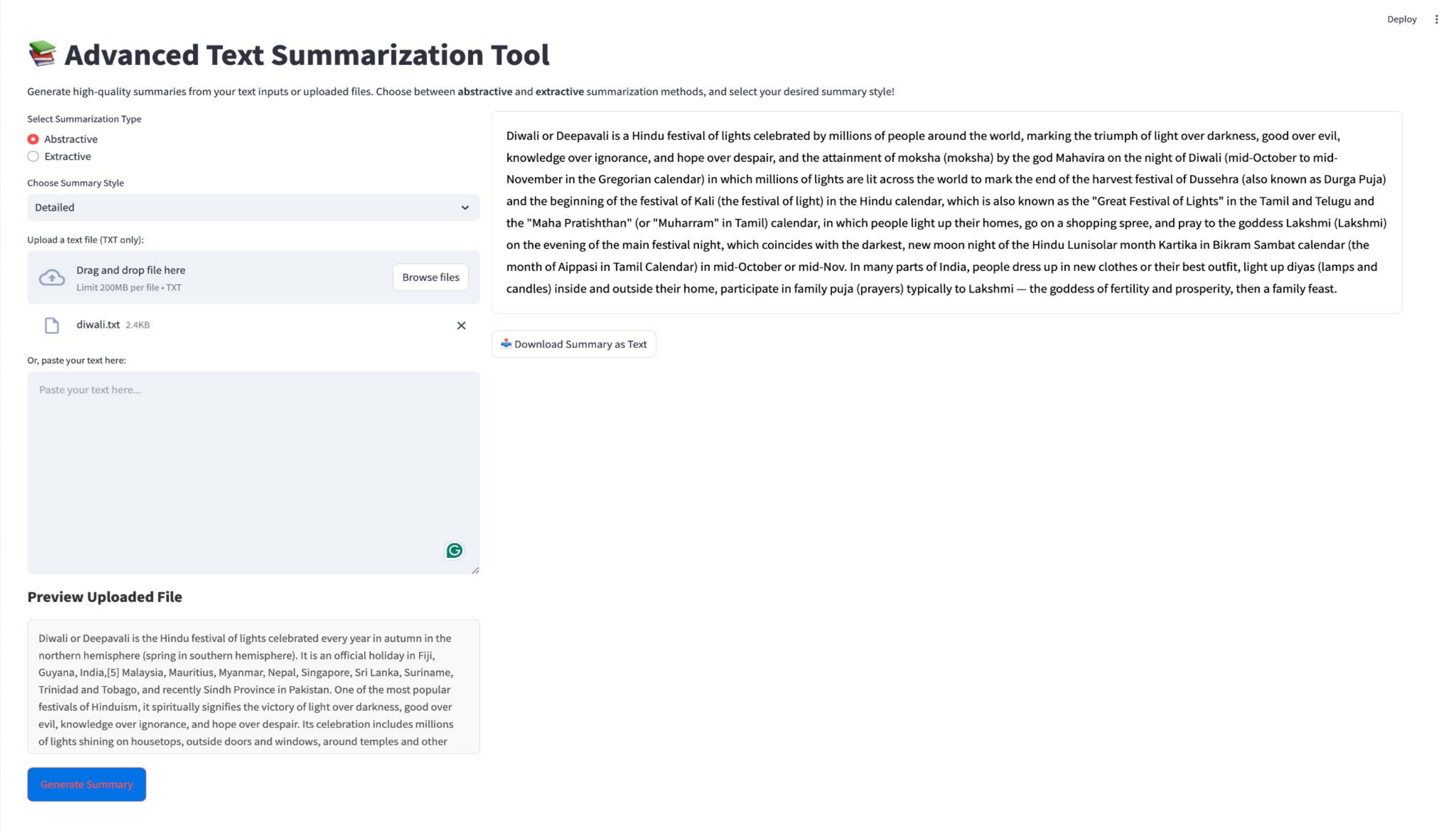
2. ROUGE-2

- Measures overlap of bigrams (pairs of consecutive words) between the summaries.
- **Significance:** Reflects fluency and contextual relevance.

3. ROUGE-L

- Measures overlap based on the Longest Common Subsequence (LCS) between the generated and reference summaries.
- **Significance:** Captures the order and sequence of words, emphasizing structure and coherence.

Abstractive Summarization Result



📚 Advanced Text Summarization Tool

Generate high-quality summaries from your text inputs or uploaded files. Choose between **abstractive** and **extractive** summarization methods, and select your desired summary style!

Select Summarization Type

● Abstractive
○ Extractive

Choose Summary Style

Detailed

Upload a text file (TXT only):

diwali.txt 2.4KB

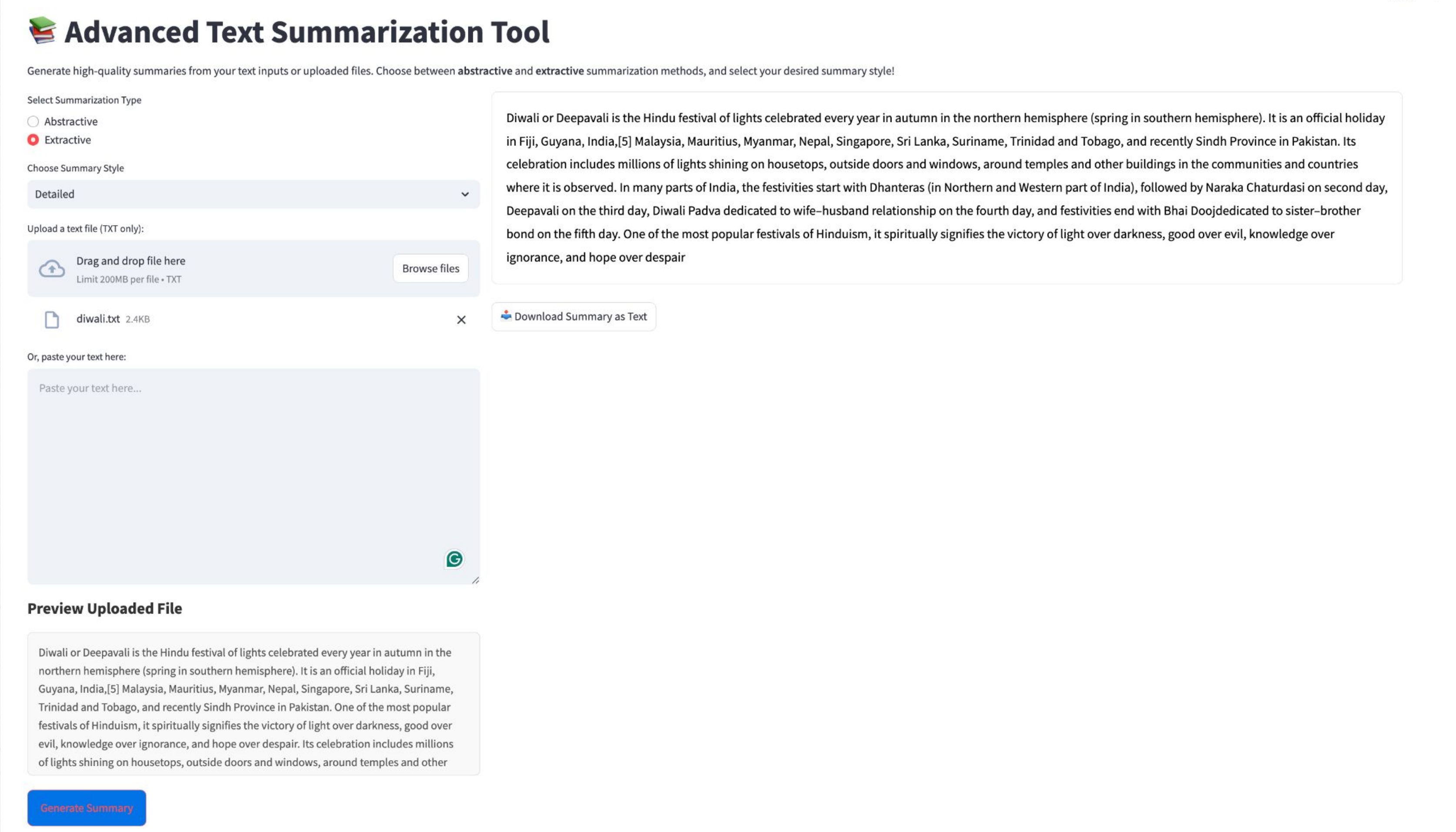
Or, paste your text here:

Preview Uploaded File

Diwali or Deepavali is the Hindu festival of lights celebrated every year in autumn in the northern hemisphere (spring in southern hemisphere). It is an official holiday in Fiji, Guyana, India,[5] Malaysia, Mauritius, Myanmar, Nepal, Singapore, Sri Lanka, Suriname, Trinidad and Tobago, and recently Sindh Province in Pakistan. One of the most popular festivals of Hinduism, it spiritually signifies the victory of light over darkness, good over evil, knowledge over ignorance, and hope over despair. Its celebration includes millions of lights shining on housetops, outside doors and windows, around temples and other

Diwali or Deepavali is a Hindu festival of lights celebrated by millions of people around the world, marking the triumph of light over darkness, good over evil, knowledge over ignorance, and hope over despair, and the attainment of moksha (moksha) by the god Mahavira on the night of Diwali (mid-October to mid-November in the Gregorian calendar) in which millions of lights are lit across the world to mark the end of the harvest festival of Dussehra (also known as Durga Puja) and the beginning of the festival of Kali (the festival of light) in the Hindu calendar, which is also known as the "Great Festival of Lights" in the Tamil and Telugu and the "Maha Pratishthan" (or "Muharram" in Tamil) calendar, in which people light up their homes, go on a shopping spree, and pray to the goddess Lakshmi (Lakshmi) on the evening of the main festival night, which coincides with the darkest, new moon night of the Hindu Lunisolar month Kartika in Bikram Sambat calendar (the month of Aippasi in Tamil Calendar) in mid-October or mid-Nov. In many parts of India, people dress up in new clothes or their best outfit, light up diyas (lamps and candles) inside and outside their home, participate in family puja (prayers) typically to Lakshmi — the goddess of fertility and prosperity, then a family feast.

Extractive Summarization Result



📚 Advanced Text Summarization Tool

Generate high-quality summaries from your text inputs or uploaded files. Choose between **abstractive** and **extractive** summarization methods, and select your desired summary style!

Select Summarization Type

○ Abstractive

● Extractive

Choose Summary Style

Detailed

Upload a text file (TXT only):

diwali.txt 2.4KB

Or, paste your text here:

Preview Uploaded File

Diwali or Deepavali is the Hindu festival of lights celebrated every year in autumn in the northern hemisphere (spring in southern hemisphere). It is an official holiday in Fiji, Guyana, India,[5] Malaysia, Mauritius, Myanmar, Nepal, Singapore, Sri Lanka, Suriname, Trinidad and Tobago, and recently Sindh Province in Pakistan. One of the most popular festivals of Hinduism, it spiritually signifies the victory of light over darkness, good over evil, knowledge over ignorance, and hope over despair. Its celebration includes millions of lights shining on housetops, outside doors and windows, around temples and other

Diwali or Deepavali is the Hindu festival of lights celebrated every year in autumn in the northern hemisphere (spring in southern hemisphere). It is an official holiday in Fiji, Guyana, India,[5] Malaysia, Mauritius, Myanmar, Nepal, Singapore, Sri Lanka, Suriname, Trinidad and Tobago, and recently Sindh Province in Pakistan. Its celebration includes millions of lights shining on housetops, outside doors and windows, around temples and other buildings in the communities and countries where it is observed. In many parts of India, the festivities start with Dhanteras (in Northern and Western part of India), followed by Naraka Chaturdasi on second day, Deepavali on the third day, Diwali Padva dedicated to wife–husband relationship on the fourth day, and festivities end with Bhai Doojdedicated to sister–brother bond on the fifth day. One of the most popular festivals of Hinduism, it spiritually signifies the victory of light over darkness, good over evil, knowledge over ignorance, and hope over despair

Challenges

1. Data Challenges

- **Data Diversity:** Ensuring the dataset is representative of various domains and text styles.
- **Preprocessing Issues:** Handling noisy and inconsistent data formats.

2. Model Challenges

- **Long Texts:** Summarizing lengthy documents without losing context or coherence.
- **Balancing Trade-Offs:** Striking a balance between informativeness and brevity in summaries.
- **Abstractive Complexity:** Generating human-like summaries while maintaining factual accuracy.

3. Computational Challenges

- **Resource Constraints:** High memory and computational requirements for training transformer models.
- **Fine-Tuning:** Optimizing hyperparameters for better performance across diverse datasets.

Key Learnings

1. Technical Insights

- **Understanding Summarization:** Gained a deep understanding of extractive and abstractive methods.
- **Model Familiarity:** Learned about the advanced transformer architectures like BART and BERT.
- **Importance of Preprocessing:** Learned how clean and well-prepared data directly impacts model performance.

2. Practical Takeaways

- **Evaluation Metrics:** Realized the importance of using ROUGE for consistent performance evaluation.
- **Fine-Tuning Skills:** Improved ability to fine-tune large models effectively for specific tasks.
- **Challenges in Summarization:** Recognized the complexities of generating context-aware, accurate summaries.

Future Scope

1. Multilingual Summarization

- Extend support for multiple languages to serve a global audience.

2. Domain-Specific Models

- Build customized summarization tools for specialized domains like legal, medical, or technical content.

3. Real-Time Summarization

- Enable summarization of streaming data, such as live events or social media content.

Conclusion

- Successfully developed a summarization tool supporting both extractive and abstractive methods.
- Evaluated performance using ROUGE metrics, demonstrating high-quality summaries.
- Addressed challenges in preprocessing, model training, and evaluation.
- Demonstrated potential applications across domains like media, education, and customer service.

THANK YOU