Ensemble Text Summarization Algorithm:

Innovative Project Report (PROJCS701)

4th Year, 7th Semester

2024

Presented By

Aritra Ghosh

B.Tech CSE

Section - A Roll No. - 74

Enrl No.: 12021002002137

Subhojit Ghosh

B.Tech CSE

Section - B Roll No. - 97

Enrl No.: 12021002002160

Under the Mentorship of **Dr. Anupam Mondal**

Abstract: A Glimpse into the Project

The exponential growth of digital content has increased the need for efficient text summarization systems that distill information while retaining core meaning. This project presents an ensemble algorithm combining BART, PEGASUS, and RoBERTa to generate high-quality summaries. Using a custom dataset of 200,000 summary-article pairs, BART and PEGASUS were incrementally fine-tuned, while RoBERTa assessed semantic similarity. A hybrid approach minimized redundancy and improved coherence by refining summaries with an abstractive model. Evaluated with ROUGE metrics and deployed via Streamlit our project offers a scalable and user-friendly solution for real-world summarization needs.

Introduction

Summarization is the process of condensing a larger piece of text into a shorter version while still capturing the essential information. It allows readers to quickly grasp the main points and key details of a document without having to read the entire text.

→ **Extractive summarization** involves selecting and combining important sentences.
→ **Abstractive summarization** generates new sentences that capture the essence of the original text.

1 ## Challenges in Summarization

Summarization research faces hurdles in dealing with dataset limitations, the computational demands of training large models, ensuring coherence and fluency in the generated summaries, and the challenges of evaluating model performance.

2 ## Real-World Applications

Summarization finds crucial applications in various domains, including news aggregation, scientific research, and social media content analysis, facilitating efficient knowledge acquisition.

Common Methods of Summarization (Traditional Approaches)

METHOD	DESCRIPTION	ADVANTAGES	LIMITATIONS
TF-IDF	Ranks terms by frequency and document importance.	Simple and fast.	Ignores semantic meaning.
Graph-based Approaches	Constructs graphs of sentences ranked by algorithms like PageRank.	Effective for extractive tasks.	Computationally expensive for large datasets.
Pointer-Generator Model	Combines extractive and abstractive approaches for flexibility.	Balances strengths of both approaches.	May struggle with highly technical texts.
Transformer Models	Leverages self-attention to capture long-range dependencies (e.g., BERT, GPT).	High-quality outputs with contextual fluency.	Demands high computational resources.

Traditional Methods vs. LLMs in Text Summarization

Method	Description	Advantages	Limitations
TF-IDF	Ranks terms by frequency and document importance.	Simple and fast.	Ignores semantic meaning.
Graph-based Approaches	Constructs graphs of sentences ranked by algorithms like PageRank.	Effective for extractive tasks.	Computationally expensive for large datasets.
Pointer-Generator Model	Combines extractive and abstractive approaches for flexibility.	Balances strengths of both approaches.	May struggle with highly technical texts.
Transformer Models	Leverages self-attention to capture long-range dependencies (e.g., BERT, GPT).	High-quality outputs with contextual fluency.	Demands high computational resources.

- **Traditional Methods**: Rule-based/statistical (e.g., TF-IDF, TextRank); extractive, domain-specific, efficient, but limited in coherence and abstraction.
- **LLMs (e.g., GPT, BART)**: Neural networks; both extractive and abstractive, domain-agnostic, fluent, semantically rich; resource-intensive and less interpretable.
- **Summary**: Traditional methods suit simple tasks; LLMs excel in handling complex texts.

Data is King: Understanding the Dataset

Dataset Characteristics

Our dataset consists of 200,000 entries, encompassing various text styles, including formal, informal, and conversational, reflecting the diverse nature of real-world text.

Challenges in Dataset Handling

Challenges in dataset handling include ambiguity in conversational data, domain-specific vocabulary, and data imbalance and redundancy, all requiring careful attention to ensure robust model training.

Improvements to the Dataset

We addressed these challenges by balancing data collection, standardizing the data format, and employing data augmentation techniques like paraphrasing and reordering, further enriching the dataset.

Methodology: Our Approach to Ensemble Summarization

Incremental Training

We fine-tuned BART and PEGASUS models incrementally, using batches of 10,000 examples, evaluating model performance after each iteration using ROUGE metrics to assess the effectiveness of each step.

Final Refinement

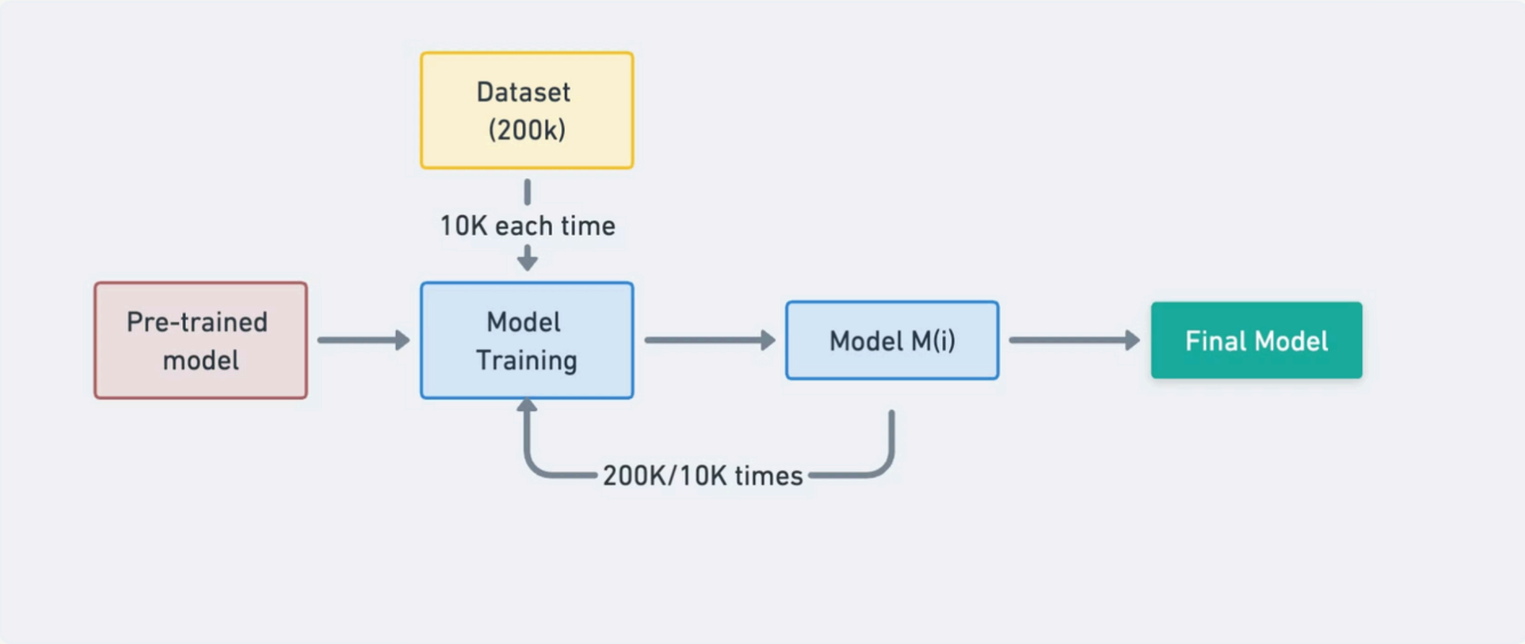
The final summary is then rephrased using T5 for enhanced coherence and fluency, ensuring a well-structured and readable output, culminating in a comprehensive and accurate summary.

1 2 3

Ensemble Approach

The ensemble approach involves generating summaries from both BART and PEGASUS models, combining them using RoBERTa to identify semantic similarities and improve overall coherence.

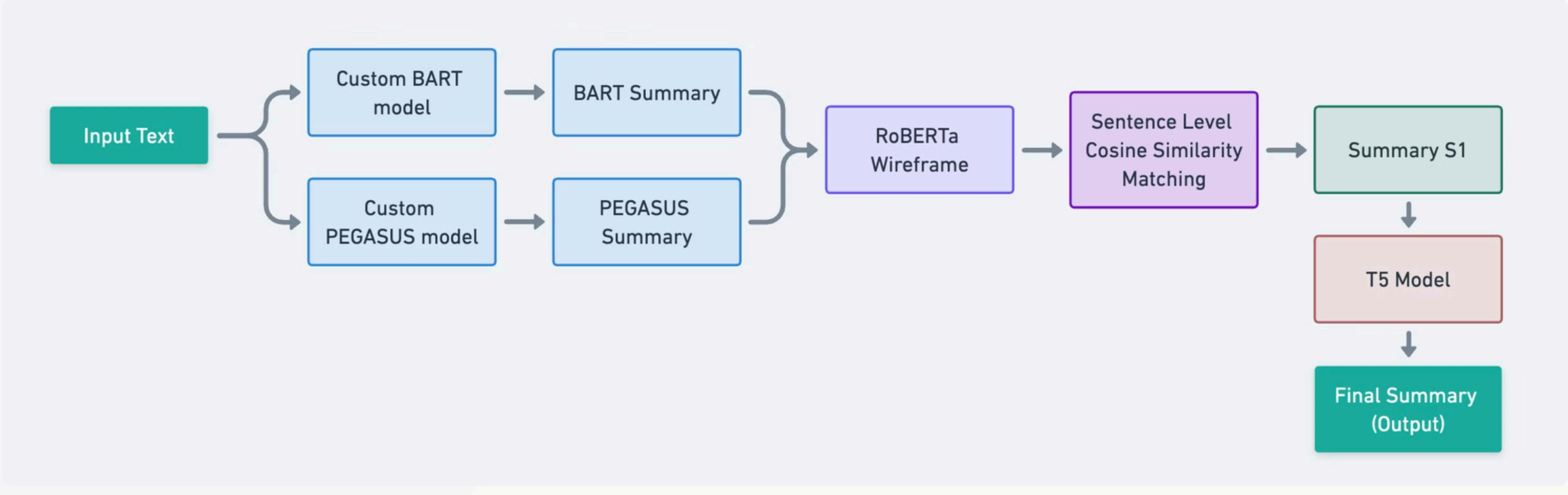
Training Approach



This flowchart visually represents the incremental training methodology for fine-tuning models using a large dataset. Here's a breakdown of the elements:

1. **Pre-trained Model**: The starting point, representing models like BART and PEGASUS initialized with pre-trained weights.
2. **Dataset (200k)**: Indicates the full dataset, which is partitioned into smaller batches of 10,000 samples for iterative training.
3. **Model Training**: The process where the model is trained on one batch at a time (10k samples per iteration).
4. **Model M(i)**: Represents the model after being fine-tuned with the iii-th batch, where iii varies across iterations.
5. **Final Model**: The output after the model is fine-tuned with all data batches, symbolizing the culmination of the incremental training process.

Process Flow



Result

67.5%

BART

The ROUGE-1 score, measuring unigram overlap, indicates that BART achieved a respectable performance in capturing key words and phrases.

64.3%

PEGASUS

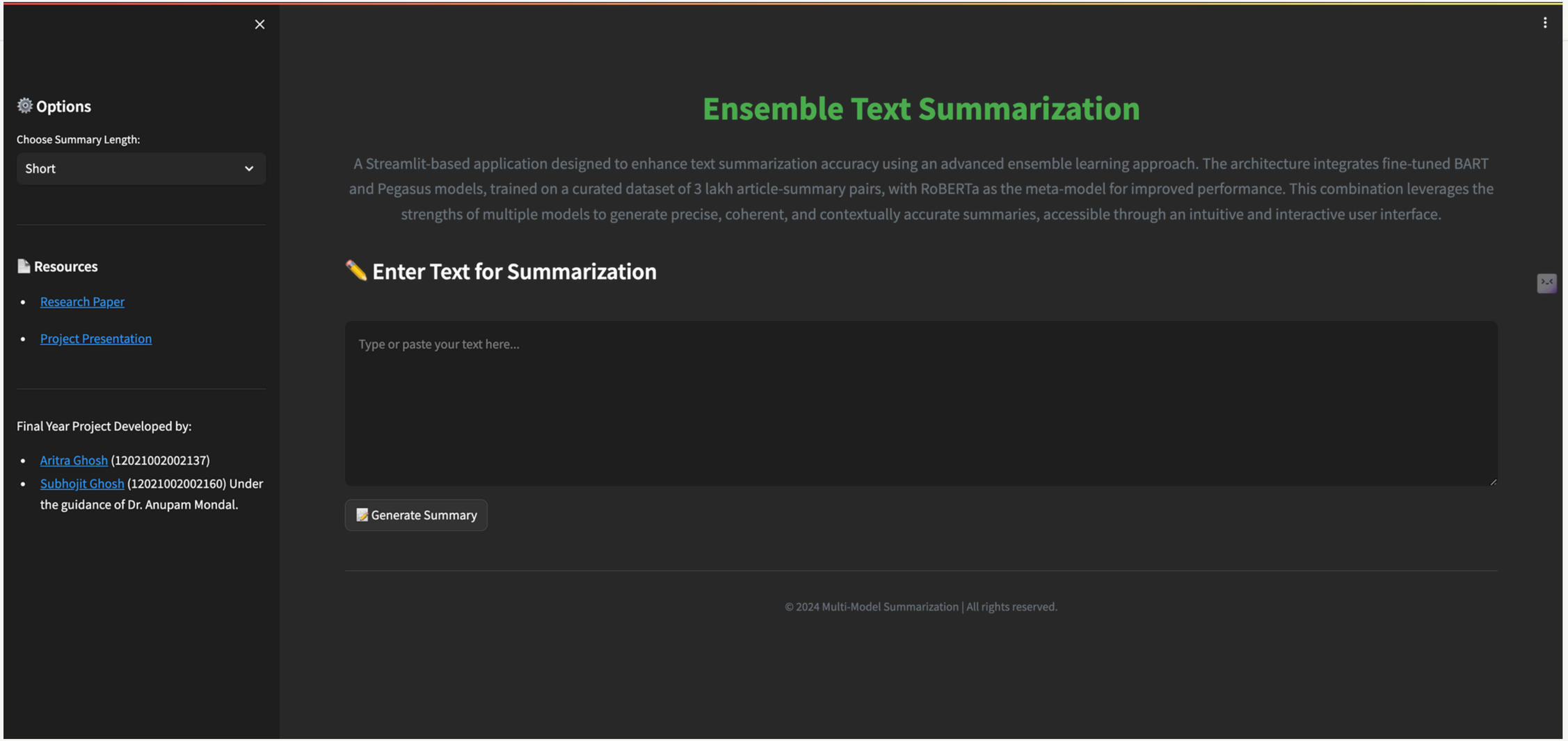
PEGASUS exhibited slightly better performance in ROUGE-1, suggesting a slightly higher level of precision in terms of word and phrase matching with human-written summaries.

71%

Final

The final summary, combining the strengths of BART and PEGASUS, achieved the highest ROUGE-1 score. This demonstrates the effectiveness of our hybrid approach in producing summaries with superior unigram overlap, indicating improved accuracy, coherence, and relevance compared to individual models.

Demo of our Application



https://summari.streamlit.app/

Resources

- **GitHub Repo link:** https://github.com/TheCleverIdiott/FYP
- **Paper1:** https://drive.google.com/file/d/1l3rKsREM1euYFlapLcoAk8YpPqd43VgY/view?usp=sharing
- **Paper2:** https://drive.google.com/file/d/1Bg0RTnMheb5lHD-xDwFDUcNVFDyvjRJF/view?usp=sharing
- **Project Report:** https://drive.google.com/file/d/12iWuXNXkGBNk5PrK6mHJO6RL5Ohwa5jS/view?usp=sharing
- **Poster:** https://drive.google.com/file/d/12iWuXNXkGBNk5PrK6mHJO6RL5Ohwa5jS/view?usp=sharing

Conclusion: Reaching New Heights in Text Summarization

Enhanced Quality

The ensemble approach utilizing BART, PEGASUS, and RoBERTa has significantly improved the quality of text summarization, striking a balance between precision, fluency, and abstraction.

Real-World Applicability

Our model demonstrates real-world adaptability and versatility across various domains, paving the way for applications in news aggregation, research paper summarization, and social media content analysis.

Acknowledgment

We would like to express our sincere gratitude to Dr. Anupam Mondal, our mentor, for his invaluable guidance and support throughout this project. We are also thankful to the Institute of Engineering and Management, Kolkata, for providing the necessary resources and creating a conducive learning environment. Lastly, we extend our gratitude to our families, peers, and faculty members for their unwavering support and encouragement.

Thank You!