ENSEMBLE TEXT SUMMARIZATION

4th Year, 7th Semester Project Project Submission

Members:

- Aritra Ghosh, Section - A, Roll No - 74, Enrl. No. - 12021002002137
- Subhojit Ghosh, Section - B, Roll No - 97, Enrl. No. - 12021002002160

Project Mentor - Dr. Anupam Mondal

Abstract

Leveraging pre-trained transformer models (BART, PEGASUS, RoBERTa) to develop a hybrid text summarization system. This approach balances fluency, coherence, and accuracy to generate high-quality summaries suitable for diverse domains.

Challenges

- Dataset diversity and inconsistency.
- Computational resource constraints.
- Redundancy and scalability issues.

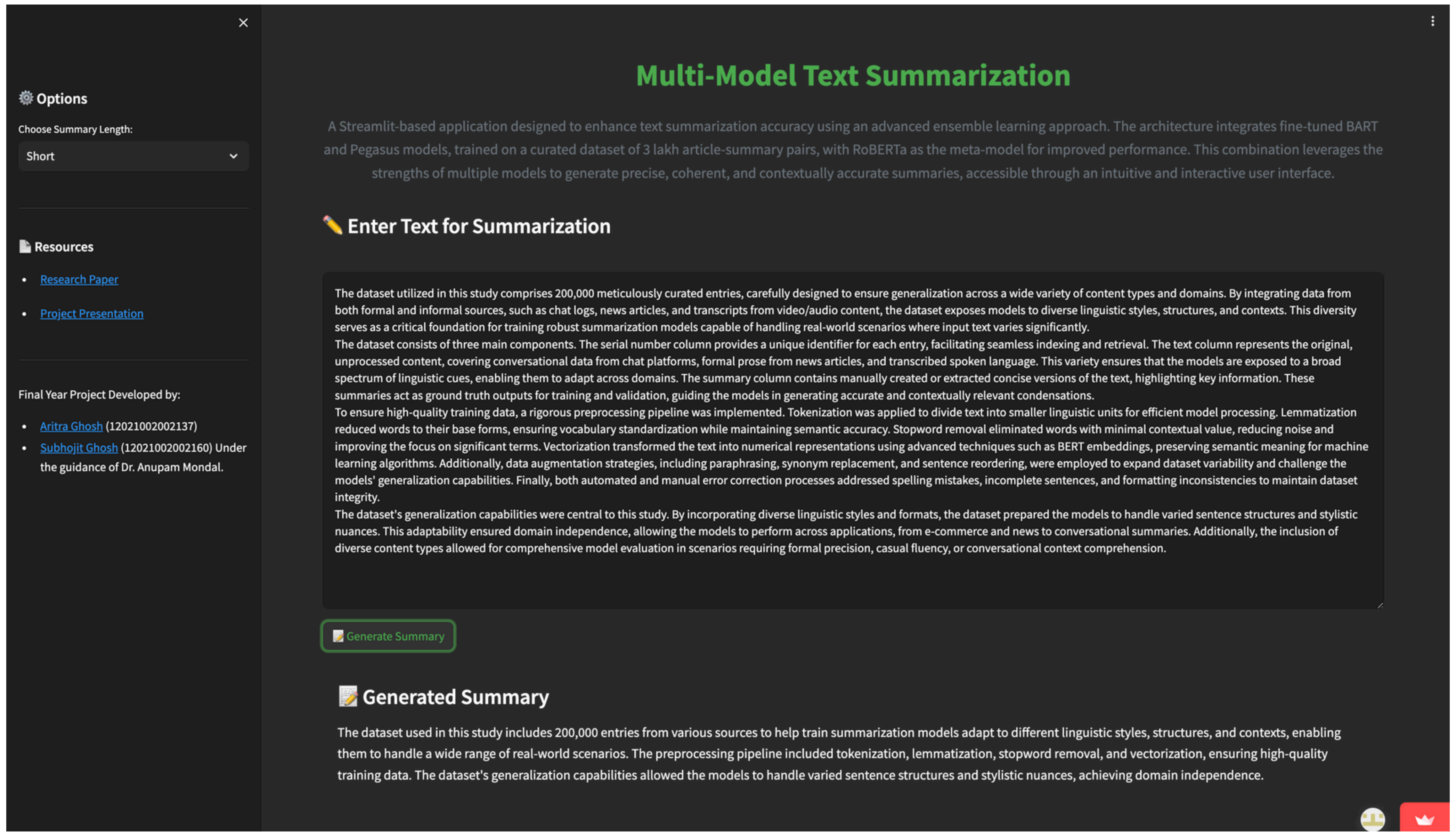
Results

Performance Metrics:

- ROUGE-1: 46.3% (PEGASUS), 44.5% (BART).
- Strengths: Enhanced semantic accuracy, fluency, and domain adaptability.

Future Scope

- Real-time and multilingual summarization.
- Enhanced evaluation metrics beyond ROUGE.
- Domain-specific adaptations for healthcare, legal, and financial sectors.




```
ROUGE Scores: {'rouge1': 0.4879544637525194, 'rouge2': 0.2996217105162541,
```

```
'rougeL': 0.38225106788784413, 'rougeLsum': 0.4691003583575095}
```

Conclusion

Transformer models like BART, RoBERTa, and PEGASUS excel in diverse summarization tasks. A stack-based ensemble approach shows promise for enhancing accuracy, coherence, and scalability in real-time applications.

Acknowledgement

We sincerely thank our mentor and college for their guidance and support throughout this study. Their insights and feedback were invaluable in shaping this research.

Dataset

- 200,000 summary-article pairs from diverse sources (news, conversational, transcriptions).
- Preprocessing steps include tokenization, lemmatization, stopword removal, and noise reduction.
- **Challenges:** Data imbalance, diversity, and domain-specific vocabulary.
- **Improvements:** Standardized summaries, data augmentation, and advanced embedding techniques (BERT, RoBERTa).

Methodology

- Fine-tuned BART and PEGASUS models using batch-wise incremental training with 10,000 examples per batch.
- Models updated iteratively from M_1 to M_{i+1} until the full dataset was covered.
- Generated separate summaries using fine-tuned models and combined summaries using RoBERTa for semantic cosine similarity assessment (>0.85).
- Merged redundant sentences retained the most relevant ones, Included unique sentences for additional insights.