Institute of Engineering and Management, Kolkata

Department of Computer Science and Engineering

4th Year, 7th Semester (2021-25)

Innovative Project - Abstract and Progress Report

(PROJCS701)

Topic: Ensemble Text Summarization Algorithm

Presented by:

Sl. No.	Name	Year	Section	Roll No.	Enrollment No.
1	Aritra Ghosh	4th	A	74	12021002002137
2	Subhojit Ghosh	4th	B	97	12021002002160

Project Mentor - Dr. Anupam Mondal

Abstract

The rapid expansion of digital content has led to an increasing demand for efficient text summarization systems capable of condensing information while preserving its essence. This project presents the development of an ensemble text summarisation algorithm that leverages pre-trained transformer models—**BART, PEGASUS, and RoBERTa**—to generate high-quality summaries.

A custom dataset of **200,000 summary-article pairs** was developed for training. As illustrated in the **Training Approach diagram**, the models were fine-tuned incrementally. The dataset was split into batches of 10,000 samples each, allowing for iterative training under computational constraints. Starting with a pre-trained base model (e.g., BART or PEGASUS), models were updated sequentially, denoted as **Model M(i)**, until the entire dataset was covered. This strategy ensured the efficient use of resources while maintaining model performance.

The Process Flow was engineered as follows:

Input Text was provided to fine-tuned versions of the BART and PEGASUS models to generate individual summaries. Summaries were processed using RoBERTa to calculate **cosine similarity** at the sentence level, ensuring the removal of redundant information while preserving semantic accuracy. Finally, a **T5 model** further processed the combined summaries to generate a fluent, concise output suitable for real-world applications.

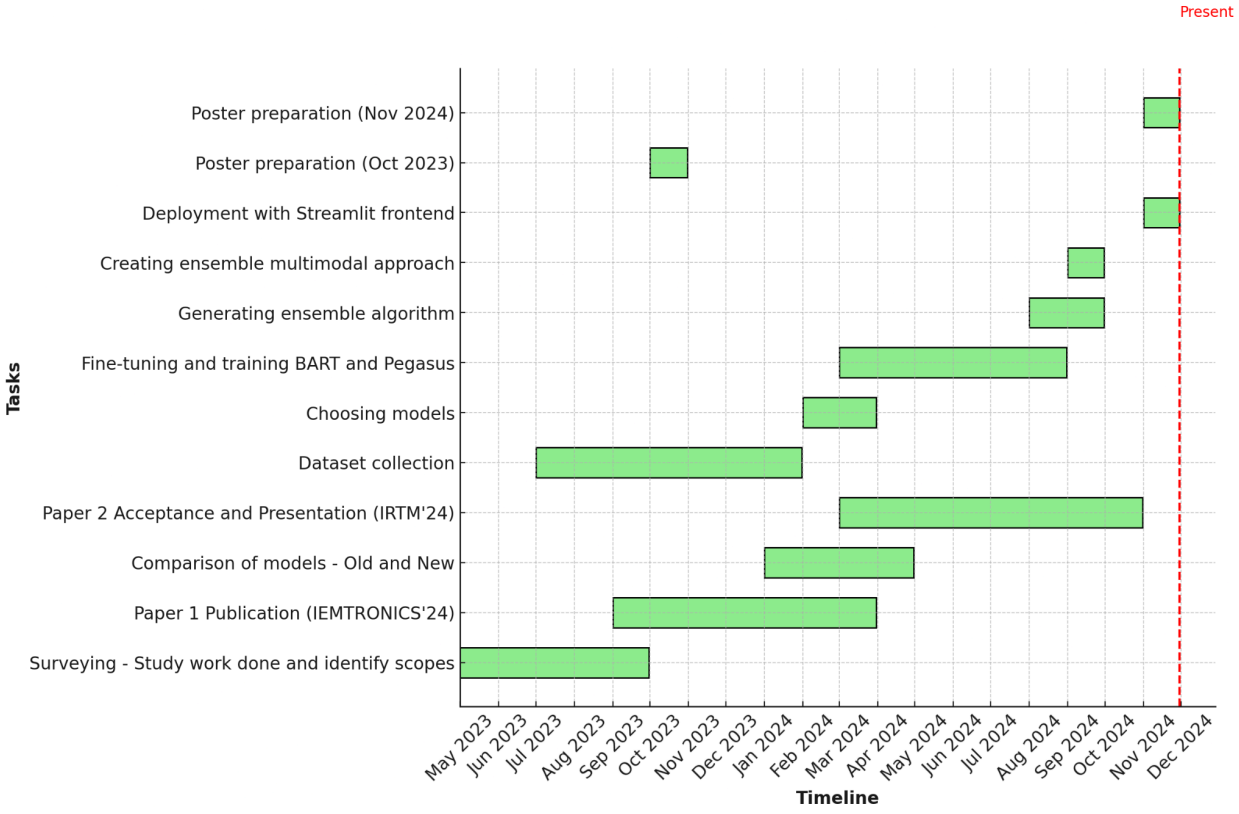
Results

The project's results are summarized in the **Performance Metrics diagram**, which highlights the effectiveness of the ensemble approach:

- **BART** achieved a **ROUGE-1 score of 67.5%**, excelling in capturing keywords and phrases.
- **PEGASUS** scored **64.3%**, demonstrating strong word and phrase matching with human-written summaries.
- The **final ensemble model** reached a **ROUGE-1 score of 71%**, showcasing the hybrid method's superiority in accuracy, coherence, and fluency.

This ensemble approach mitigates individual model biases and produces comprehensive summaries tailored for diverse textual domains. The resulting implementation, deployed via a **Streamlit interface**, offers a scalable and accessible solution for real-world summarization needs.

Progress Report



1. Timeline Overview

The project has adhered to a structured timeline:

- **Dataset Collection (May - Oct 2023):** Compiled a dataset of 200,000 summary-article pairs, covering diverse domains like news, conversational data, and transcriptions.
- **Model Training (Nov 2023 - May 2024):** Fine-tuned BART and PEGASUS models incrementally in batches of 10,000 samples to optimize training efficiency within computational limits.
- **Ensemble Method Development (Jun - Jul 2024):** Integrated model outputs using RoBERTa for redundancy reduction and semantic similarity scoring.
- **Poster Presentation (Oct 2023 & Nov 2024):** Created and presented posters highlighting project objectives, methodology, and results.
- **Project Report and Scopus Indexed Paper Submission (2024):** Submitted a detailed project report and at least one paper to Scopus-indexed journals.

2. Completed Deliverables

- **Voice-over Presentation (Nov 2024):** Delivered a comprehensive presentation detailing the project's goals, methods, results, and challenges.

- **Project Report Submission (Nov 2024):** Submitted a detailed report, including methodologies, dataset descriptions, and evaluation metrics.
- **Poster Preparation and Presentation (Oct 2023 & Nov 2024):** Highlighted the project's innovations and challenges in a visually engaging format.
- **Scopus Indexed Paper:** Progressed two Scopus-indexed publications based on performance evaluations and findings.

3. Challenges Encountered

- **Computational Resource Constraints:** Limited computational resources for training large-scale models like BART and PEGASUS.
- **Dataset Diversity and Noise:** Addressed issues such as data imbalance, conversational ambiguities, and domain-specific vocabulary challenges.
- **Redundancy in Summaries:** Difficulty in reducing repetitive information while ensuring fluency and coherence.
- **Scalability Issues:** Ensuring the system's performance remained robust under varied real-world applications.

4. Solutions Implemented

- **Incremental Fine-Tuning:** Models were trained in smaller, manageable batches to mitigate resource constraints.
- **Advanced Preprocessing:** Applied tokenization, lemmatization, stopword removal, noise reduction, and data augmentation to standardize the dataset.
- **Semantic Similarity Scoring:** Employed RoBERTa to evaluate and combine summaries based on cosine similarity thresholds, reducing redundancy.
- **Optimization Strategies:** Batch processing and parallelization were used to handle the large dataset efficiently.