AI-Powered Lecture Summarizer

Selinay Fırat
Department of Computer Engineering
selinay.firat21@ogr.atauni.edu.tr

Zeynep Odabaş
Department of Computer Engineering
zeynep.odabas211@ogr.atauni.edu.tr

Feyza Açıkgözoğlu
Department of Computer Engineering
feyza.acikgozoglu21@ogr.atauni.edu.tr

***Abstract*—This project introduces a hybrid AI-powered summarization system designed to process and summarize lecture materials efficiently. The system combines the classical TextRank extractive algorithm with an advanced Large Language Model (LLM), specifically the Mixtral-8x7B-Instruct model, to generate fluent and contextually accurate summaries. This hybrid architecture leverages TextRank's computational efficiency while enhancing coherence and readability through LLM-based abstractive generation, providing an effective and modern solution for academic content summarization.**

I. Introduction

In modern education, students are often overwhelmed by extensive lecture materials across multiple courses. Manual review is time-consuming and prone to inconsistency. Classical extractive methods, such as TextRank, can identify key sentences but often produce summaries lacking natural flow. Recent advances in Large Language Models (LLMs) enable abstractive summarization—the ability to generate fluent and human-like summaries while preserving semantic meaning.

This project aims to build a hybrid summarization system that merges TextRank with a generative LLM to produce summaries that are both computationally efficient and semantically coherent. The proposed approach not only supports students in quickly understanding lengthy materials but also enhances accessibility for learners with limited time or reading capabilities. This tool can significantly improve study efficiency and comprehension in academic environments.

II. Dataset

The dataset includes lecture notes, slides, and transcripts from open educational sources such as MIT OCW, TU Delft OCW, Khan Academy, Coursera, and edX. Additional academic texts from arXiv, DOAJ, and PubMed Central will be used to diversify writing styles.

Approximately 100 lecture documents across 5 subjects will be collected. All files (PDF, Word) will be converted to plain text using PyMuPDF and python-docx, cleaned using regex filters, and tokenized with NLTK to prepare high-quality input data.

III. Proposed Methodology

A. Text Preprocessing

Lecture materials are converted to text and cleaned by removing headers, footers, and noise. Text is tokenized into sentences and words. Stop words are removed, and all text is standardized to lowercase.

B. Hybrid Model Architecture

The summarizer consists of two stages:

1) **Extractive Stage (TextRank):** Sentences are represented as graph nodes and ranked using cosine similarity. The most central sentences are selected as key content.

2) **Abstractive Stage (LLM):** The extracted key sentences are passed to the Mixtral-8x7B-Instruct model, a transformer-based LLM capable of rephrasing and generating coherent summaries that maintain contextual meaning.

For large documents, the system performs chunk-based summarization, summarizing sections separately and merging results through an additional LLM refinement pass.

C. Summary Generation

Users can choose between extractive (TextRank-only) and abstractive (LLM-enhanced) summarization. The LLM receives top-ranked sentences and returns a fluent, concise summary. The final text is reordered based on the source structure to maintain readability.

IV. Evaluation Method

Evaluation combines quantitative and qualitative analysis:

A. Quantitative Evaluation

- ROUGE-1, ROUGE-2, and ROUGE-L metrics to measure similarity with human-written summaries.
- Average processing time and computational efficiency.

B. Qualitative Evaluation

- Human evaluation of readability, coherence, and informativeness.
- Comparison between TextRank-only and LLM-based summaries to measure improvement in fluency and accuracy.
- Feedback from 10 university students and 3 instructors will be collected to evaluate usability and satisfaction.

V. Team Contributions

- **Selinay Fırat – Data Preparation and Preprocessing:** Responsible for dataset creation, document cleaning, and tokenization.
- **Feyza Açıkgözoğlu – Model Development:** Implement the hybrid TextRank–LLM architecture and integrate the Mixtral-8x7B model.
- **Zeynep Odabaş – Evaluation and Interface:** Conduct performance evaluations and develop a user-friendly web

interface for document uploads and summary visualization.

VI. Time Plan

TABLE I
WEEKLY TASK SCHEDULE

Week	Task
1–2	Data collection, document preprocessing.
3–4	Implementation of TextRank and LLM integration.
5	Model evaluation (ROUGE and human feedback).
6	Web interface development and testing.
7	Final optimization and presentation preparation.

VII. Backup Plan

If the LLM summarization produces incoherent results:

- Use a smaller pretrained model (e.g., Flan-T5) as a fallback.
- Adjust chunk sizes or increase extractive threshold to improve clarity.
- Integrate keyword frequency-based sentence ranking to ensure minimal performance degradation.
- Conduct pilot testing on simpler datasets such as news articles to isolate errors.

VIII. Conclusion

This project presents a modern hybrid summarization framework that combines extractive and generative methods. By integrating TextRank with the Mixtral-8x7B-Instruct LLM, the system can efficiently process lecture materials and produce coherent, human-like summaries.

The approach aligns with recent advancements in AI-powered education tools and demonstrates how lightweight LLMs can enhance learning accessibility and productivity. Future work will explore multilingual summarization and domain adaptation for diverse academic materials.

Project Repository

The source code and development progress of this project are available on GitHub at: **https://github.com/zeynepodbs/AI-Powered-Lecture-Summarizer**