Dayananda Sagar University

SCHOOL OF ENGINEERING

HYBRID RESEARCH PAPER SUMMARIZATION WITH CITATION AWARENESS AND SECTION-WISE STRUCTURING

NATURAL LANGUAGE MODELS
(22AM3610)

Presented By:

RAKSHITHA JK ENG22AM0190
POOJYANTH M ENG23AM3003
POOJITHA R ENG22AM0160
RAHUL V ENG22AM0189

Under The Guidance of :

Prof. Pradeep Kumar K
Prof. Sahil Pocker
Dept of AIML

2024-25

1. INTRODUCTION

The rapid surge in research publications—over 2.5 million per year—makes it difficult for students and researchers to stay updated. Manually reading and summarizing papers is slow, biased, and often impractical.

This project proposes an AI-driven system to generate structured, citation-aware summaries of research papers. By combining extractive and abstractive techniques, it produces coherent, section-wise summaries that retain essential references.

The system is optimized for low-resource environments and supports text-to-speech output, enhancing accessibility and usability in education and research.

Key Focus: Accelerating access to academic knowledge through structured, accurate, and efficient summarization.

2. SDG IMPACT

SDG 4: Quality Education by leveraging AI to democratize access to complex scientific knowledge.

- By generating structured, simplified summaries of research papers, the system enables students, educators, and researchers—especially in low-resource settings—to access critical information efficiently.
- The tool condenses dense academic content, allowing users to focus on core insights without wading through exhaustive documents.
- Inclusive Learning
- With integrated Text-to-Speech (TTS), the system supports auditory learning, benefiting visually impaired users and multilingual learners.
- Low-Resource Accessibility
- Optimized to run on modest hardware, ensuring equitable access to educational tools regardless of location or economic background.

3. PROBLEM STATEMENT

Despite the availability of various summarization tools, most are not designed for academic literature. They fail to capture the structured nature of research papers, often produce generic summaries, and ignore critical citation links that provide scientific context.

Many abstractive models rely heavily on large compute resources, making them impractical for students or researchers with limited access to high-end systems. Extractive models, on the other hand, often output disjointed fragments lacking flow and coherence.

Core Challenge: Build a system that understands the scholarly format, respects references, and delivers meaningful summaries—without demanding supercomputing power.

4. OBJECTIVES

- Hybrid Summarization: Combine extractive (LexRank) and abstractive (T5-small, DistilBART) methods to ensure both informativeness and fluency.
- Citation-Aware Summarization: Retain and prioritize citation references using graph-based techniques, preserving scientific context.
- Section-Wise Structuring: Generate summaries aligned with standard research paper sections (Abstract, Introduction, Methods, etc.).
- Adaptive Summary Lengths: Support short (100 words), medium (250), and long (500) summaries based on user needs.
- Resource-Efficient Design: Ensure that the system performs well on low-power devices using lightweight models and optimization.
- Text-to-Speech Integration: Convert summaries into audio using compact TTS models to enhance accessibility for diverse learners.

5. METHODOLOGY

Our methodology integrates NLP, deep learning, and graph-based reasoning to build a citation-aware, structured summarization pipeline.

1. PDF Parsing & Section Segmentation
2. Scientific papers are parsed using PyMuPDF. Custom regex-based logic segments the content into standard sections (e.g., Introduction, Methods, Results).
3. Preprocessing
4. Text is cleaned, tokenized, and stripped of stopwords. Citations are detected using pattern matching and preserved for graph construction.
5. Citation-Aware Graph Modeling
6. A citation graph is constructed between sentences to boost relevance using overlap-based edge weighting. The graph guides LexRank to prioritize citation-rich content evaluation
7. Summaries are evaluated using ROUGE, BLEU, and BERTScore for both quality and semantic accuracy.
8. Text-to-Speech (TTS)
9. Final summaries are optionally converted to audio using Coqui TTS, enabling auditory learning and accessibility.

6. RESULTS AND EVALUATION

Evaluation Metrics

- ROUGE-1 / ROUGE-2 / ROUGE-L: Measures overlap of unigrams, bigrams, and longest common subsequences.
- BLEU: Evaluates fluency based on n-gram precision.
- BERTScore: Uses semantic similarity for a deeper quality check.

Observed Performance

- Abstractive summaries showed improved fluency over pure extractive methods.
- Citation-aware graph summaries ranked higher in factual alignment and relevance.
- Section-wise segmentation ensured better structure and readability.

Efficiency

- Inference time optimized for CPU execution.
- Memory footprint kept low via lightweight models and batch processing.

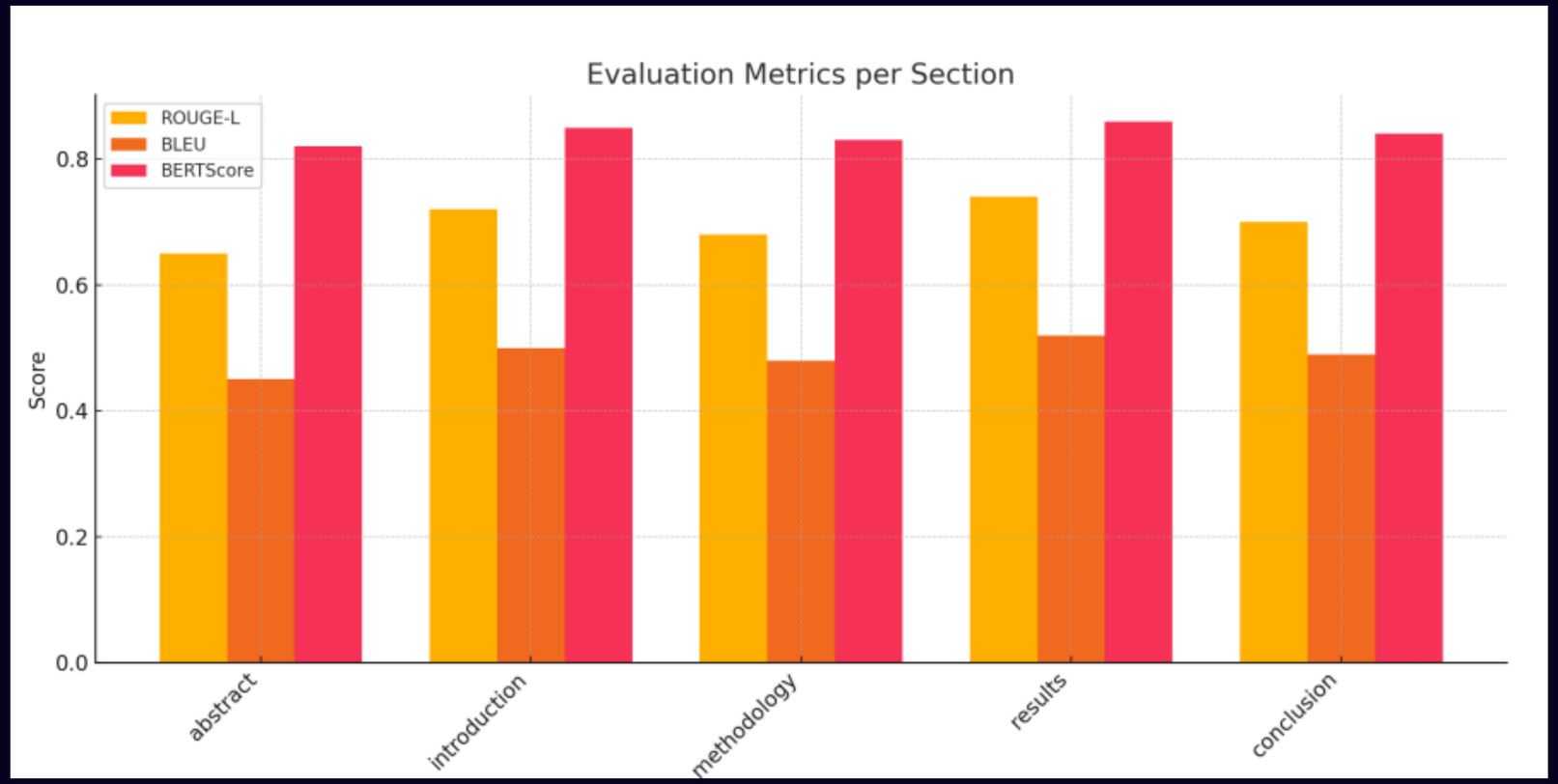

Figure 1: Bar chart showing comparative scores across different evaluation metrics (ROUGE, BLEU, BERTScore) for sample sections. Figure 1 clearly illustrates that BERTScore achieves higher values, particularly in semantically dense sections like Abstract and Conclusion, reflecting the model's effectiveness in capturing meaning beyond word overlap.

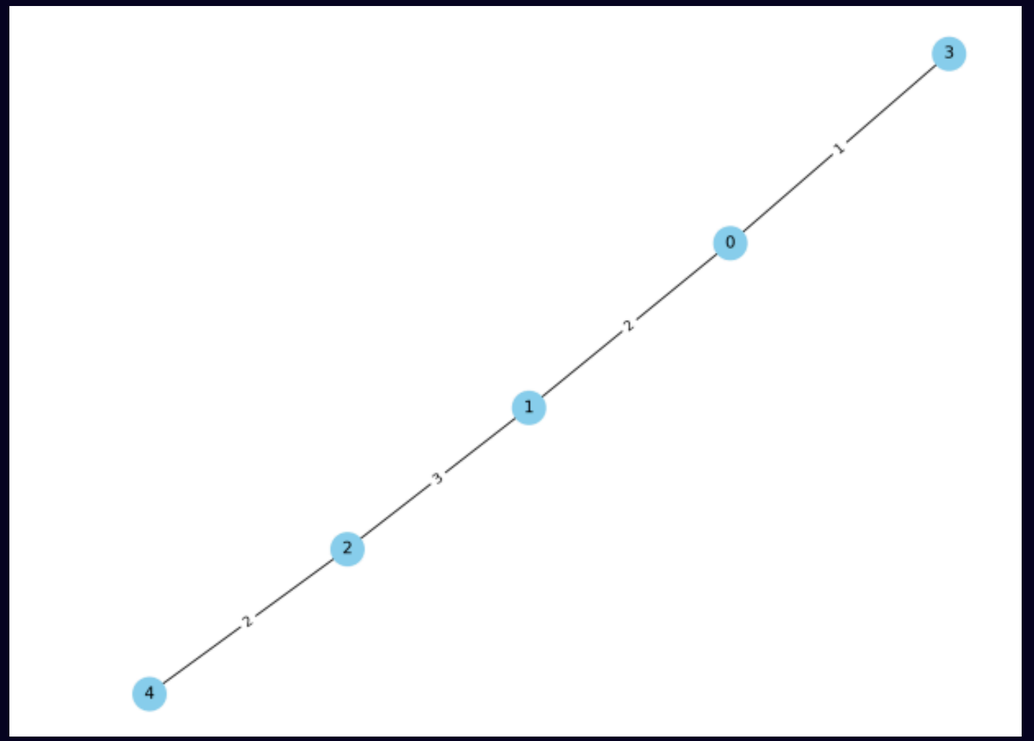

Figure 2: Citation graph of a sample academic paper showing intra- and inter-section citation relationships. Figure 2 visualizes the citation flow between different sections, showcasing how citation-aware summarization can preserve logical and referential continuity, especially in methods-heavy or literature-rich sections.

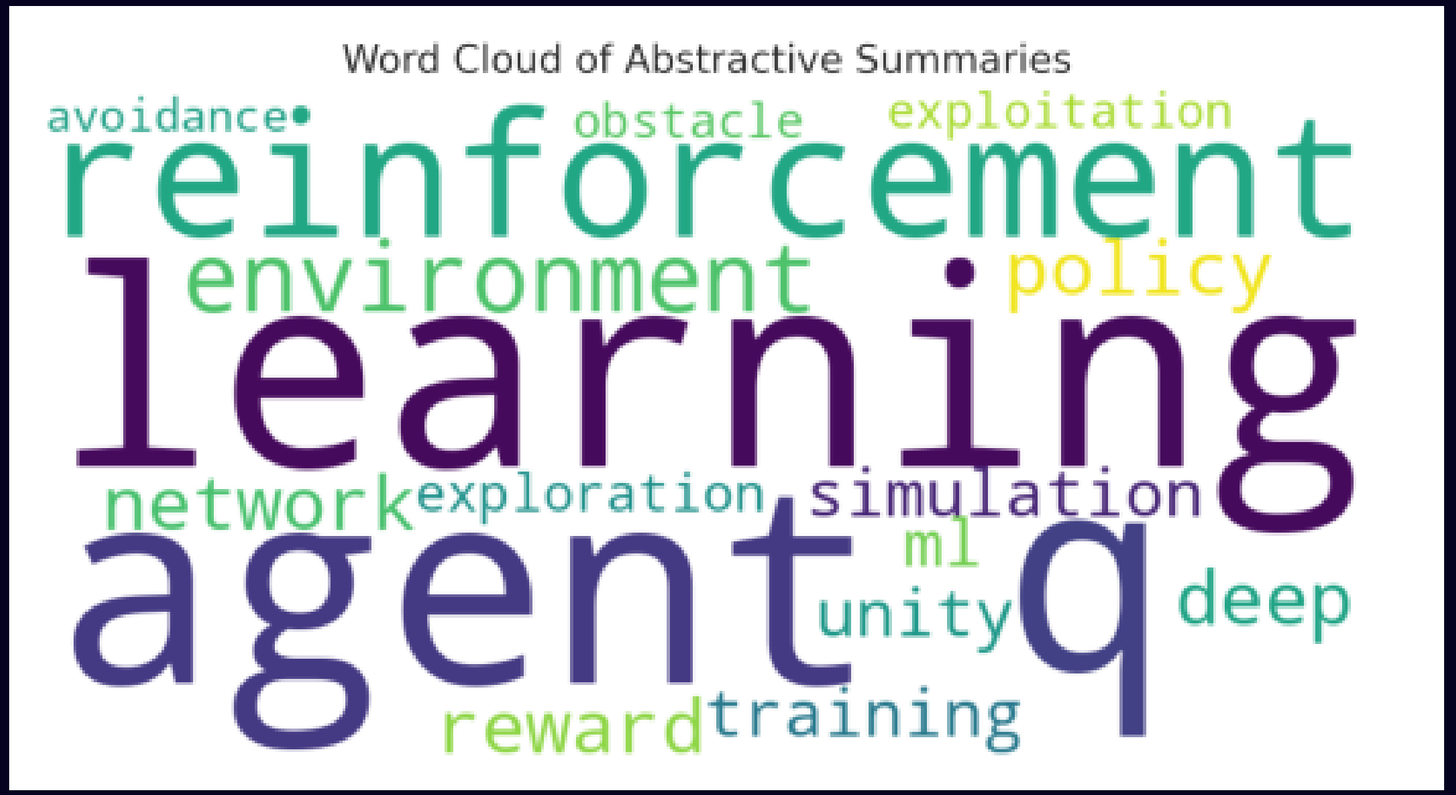

Figure 3: Word cloud of frequent terms in the abstractive summary output. As shown in Figure 3, key domain-specific terms (e.g., “reinforcement”, “policy”, “agent”) are frequent, demonstrating good coverage of central ideas in the summary output.

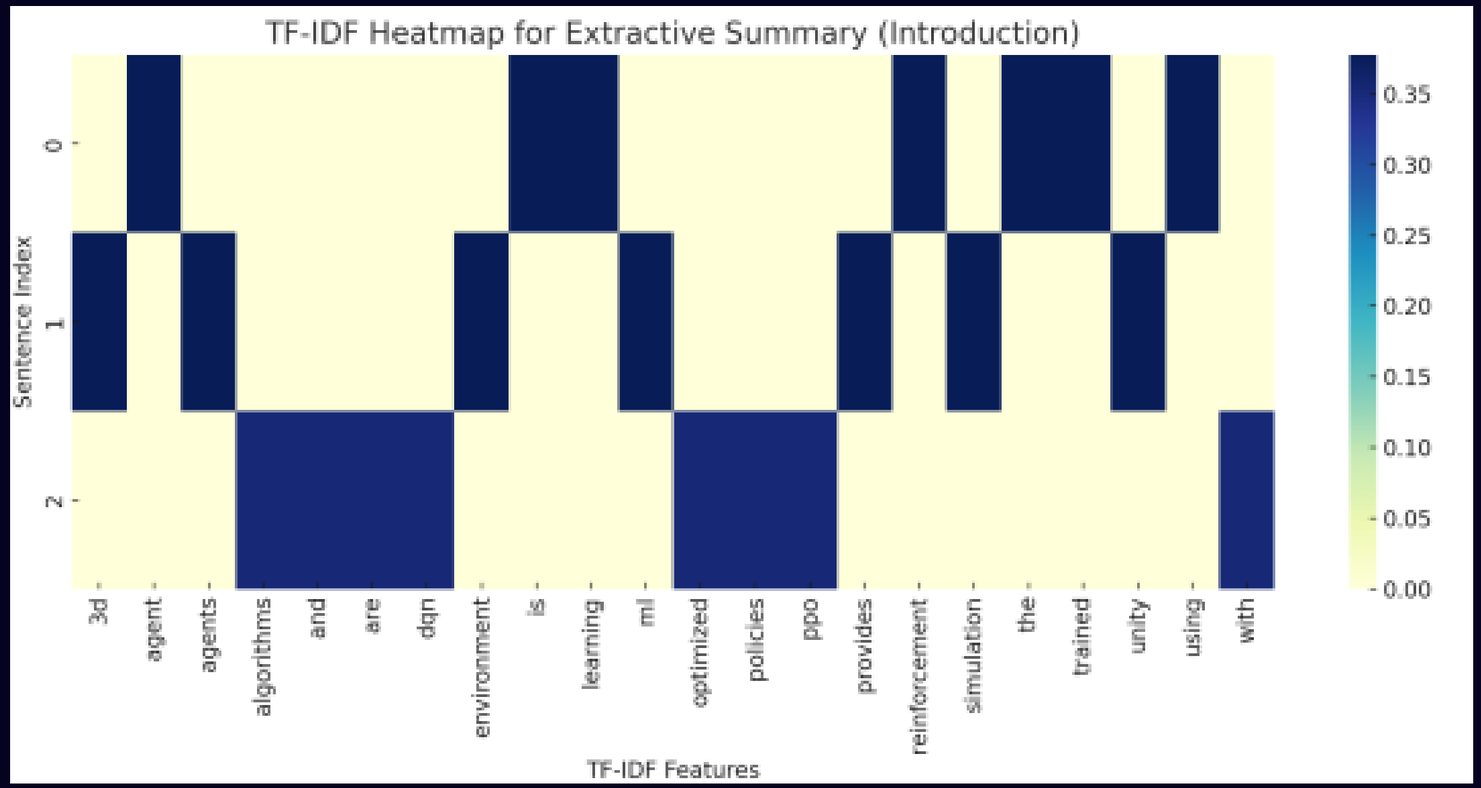

Figure 4: TF-IDF heatmap of term importance across selected sections. The TF-IDF heatmap in Figure 4 highlights the relative importance of terms. Notably, technical terms in the Methodology section and summarizing verbs in the Conclusion appear as dominant, confirming the relevance-focused nature of the summarization process.

7. CONCLUSION

In this project, we proposed and implemented a lightweight, hybrid research paper summarization system that combines extractive techniques (LexRank) with transformer-based abstractive models (T5-small/DistilBART), augmented by citation graph analysis and section-wise structuring. The goal was to enhance readability, coherence, and contextual accuracy while ensuring computational efficiency suitable for low-resource environments.

The final output structure aligns with the expectations of academic readers by maintaining a logical flow and preserving referential integrity, which is critical in scientific communication.

«« THANK YOU »»