**Text Summarization Project Report**

# Objective

This project focused on developing a text summarization system using both extractive and abstractive techniques. The dataset used for training and testing was the CNN/DailyMail dataset, sourced from Kaggle. The objective was to explore how well extractive (spaCy) and abstractive (T5-small Transformer) summarization methods perform on real-world news articles.

# Problem Faced

- Compatibility issues with Hugging Face's `TrainingArguments-` parameters due to version mismatches.  
- Frequent runtime crashes in Google Colab, especially during model fine-tuning and `pyarrow` import errors.  
- Tokenizer deprecation warnings such as `as\_target\_tokenizer` needing updated syntax.

# Solutions Implemented

- Switched to a simpler pipeline using the T5-small model for inference-only summarization.  
- Avoided full fine-tuning to prevent crashing by limiting processing to single-sample evaluation.  
- Used `spacy` for extractive summarization to give a quick baseline for comparison.  
- ROUGE evaluation metrics were used to compare the quality of summaries with ground truth highlights.

# Summary of Findings

- Extractive summarization captures key sentences based on word frequency but lacks coherence.  
- Abstractive summarization via T5-small generates fluent and semantically rich summaries, better matching the reference.  
- ROUGE scores validate the advantage of Transformer-based abstractive summarization over simple extraction.

# Conclusion

This experiment demonstrated the effectiveness of modern NLP techniques for summarizing lengthy news articles. Although resource limitations restricted model fine-tuning, inference using pre-trained models proved highly effective. Future work can explore full fine-tuning and deployment for real-world summarization applications.