Task 2 Report

Title: Text Summarization using Extractive and Abstractive Models

1. Objective

To develop a system that can summarize long articles (news, blogs) into short, meaningful summaries using both extractive and abstractive approaches.

2. Dataset Description

- **Name**: CNN/DailyMail News Dataset
- **Source**: Hugging Face → ccdv/cnn_dailymail (version 3.0.0)
- **Fields**:
    - article: Full article text
    - highlights: Human-written summary

3. Preprocessing

- Articles were cleaned by removing extra spaces.
- For extractive summarization, stopwords and punctuations were excluded.
- For abstractive summarization, the BART transformer model was used directly via Hugging Face pipeline.

4. Techniques Used

a. Extractive Summarization

- Based on frequency of important words using spaCy.
- Selected top 3 scored sentences from the original text.

b. Abstractive Summarization

- Used Hugging Face's facebook/bart-large-cnn model.
- Summary generated was grammatically fluent and paraphrased.

5. Evaluation

- Comparison done between:
    - Extractive output
    - Abstractive output
    - Original reference summary (highlights)
- (Optional) ROUGE evaluation used to compare quality.

6. Key Insights

- Extractive method is fast but less human-like.
- Abstractive summary is fluent, close to real summaries.
- BART works well out-of-the-box without fine-tuning.

7. Challenges

- Loading full dataset requires good internet.
- Summarizer models need GPU or take time on CPU.
- Abstractive models have max token limits (e.g., 1024).

8. Conclusion

This task successfully demonstrated both types of summarization approaches using real-world news data. Abstractive models like BART provide high-quality results with minimal setup.