# Summarization System Project Report

## 1. Dataset Description and Preprocessing

The project utilizes the CNN/Daily Mail dataset, a widely used corpus for text summarization tasks. This dataset contains news articles and their corresponding summaries, providing a rich source of real-world text.

Preprocessing steps included:

- Tokenization
- Removal of stop words and special characters
- Sentence segmentation for extractive summarization
- Lowercasing and lemmatization using spaCy
- Preparation of input data formats suitable for both extractive and abstractive models.

## 2. Models Implemented with Rationale

Two primary summarization approaches were implemented:

1. Extractive Summarization:
   - Implemented using spaCy for sentence ranking based on similarity and relevance.
   - Chosen for its simplicity, speed, and effectiveness on informative news articles.

2. Abstractive Summarization:
   - Implemented using HuggingFace Transformers with pre-trained models like BERT and GPT.
   - Fine-tuning was performed to adapt to domain-specific language patterns.
   - Chosen to generate human-like summaries that may include rephrasing and new sentence constructions.

## 3. Key Insights and Visualizations

Key Insights:

- Extractive methods perform well on structured articles but struggle with context adaptation.

- Abstractive models provide more natural and coherent summaries after fine-tuning.

Visualizations (not included in PDF):

- Distribution of article lengths

- Comparison of summary lengths

- ROUGE scores comparison between models

## 4. Challenges Faced and Solutions

Challenges:

- Handling extremely long input texts exceeded token limits of transformer models.

- Limited labeled data for domain-specific fine-tuning.

- Ensuring the coherence and grammatical quality of abstractive summaries.

Solutions:

- Applied truncation and chunking strategies for input sequences.

- Used transfer learning and small batch sizes for efficient fine-tuning.

- Evaluated summary quality using both ROUGE metrics and human judgment.