# **Project Proposal: Text Summarizer Using BERT**

### 1. Project Title:

Abstractive Text Summarization using BERT on CNN/DailyMail Dataset.

### 2. Project Overview:

The goal of this project is to develop a text summarizer capable of creating concise and informative summaries from long news articles. The summarization system will leverage model:

- Model: A BERT-based model (BART or T5) using transfer learning, fine-tuned on the CNN/DailyMail dataset.

#### 3. Dataset:

The dataset used in this project is the CNN/DailyMail dataset , which is a collection of news articles and their corresponding human-written summaries. This dataset will be accessed through Hugging Face's `datasets` library:

from datasets import load\_dataset

dataset = load\_dataset("cnn\_dailymail", "3.0.0")

### 4. Scope of the Project:

The project will focus on abstractive summarization , meaning that the summaries will be generated in natural language as opposed to just extracting keywords or sentences. The project includes:

- Data preprocessing using the CNN/DailyMail dataset.
- Implementation of two models :
- BERT-based transfer learning model (e.g., BART or T5).
- Naive model for basic comparison.
- Analysis of model performance, comparing fine-tuned BERT with a simpler baseline model.

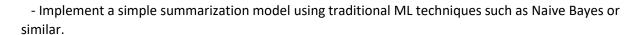
# 5. Input and Output Examples (X and y):

Example #	Input (X) — Article Text	Expected Output (y) — Summary
1	"The White House announced today that"	"The White House made a significant announcement regarding"
2	"NASA has launched its latest spacecraft"	"NASA successfully launched a new spacecraft aimed at"
3	"In a surprising move, the tech giant Apple"	"Apple made a surprise decision regarding"
4	"The latest COVID-19 vaccine trials showed"	"Recent vaccine trials have indicated that"
5	"The new law passed by Congress aims to"	"Congress passed a new law targeting"
6	"Heavy rains in the region have caused severe flooding"	"Severe floods hit the region due to heavy rains."
7	"Economists predict that the global market will"	"Economists expect global markets to experience"

# 6. Implementation Plan:

## - Data Preprocessing:

- Use `newspaper3k` and `PyPDF2` for gathering news data and converting PDF text into usable format.
- Use `datasets` to load CNN/DailyMail dataset from Hugging Face.
- **Model 1:** BERT-based Model (e.g., BART or T5) :
  - Fine-tune a pre-trained transformer model (BART or T5) using CNN/DailyMail data.
  - Use Hugging Face's `transformers` library for implementation.
- **Model 2**: Naive Bayes Model :



- Evaluation :
- Evaluate models using ROUGE scores and human evaluation.
- Compare the output quality of BERT with the Naive model.

### 7. Conclusion:

This project will explore the effectiveness of BERT-based summarization models versus simpler models on a real-world dataset. By comparing the results of these two approaches, we will gain insights into the benefits of transfer learning for NLP tasks like summarization.