

# 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")
```

**DataSet Size:** 2,872

## 4. Dataset Link:

[https://huggingface.co/datasets/abisee/cnn\\_dailymail](https://huggingface.co/datasets/abisee/cnn_dailymail)

## 5. 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).

- T5 for basic comparison.
- Analysis of model performance , comparing fine-tuned BERT with a simpler baseline model.

6. 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..."

	Article 1: Summary: Harry Potter star Daniel Radcliffe turns 18 on Monday. He gains access to a reported £20 million (\$41.1 million) fortune. Radcliffe's earnings fr Predicted Category: Entertainment
	Article 2: Summary: Judge Steven Leifman is an advocate for justice and the mentally ill. He says about one-third of all people in Miami-Dade county jails are menta Predicted Category: Retail
	Article 3: Summary: "I probably had a 30-, 35-foot free fall," survivor Gary Babineau says. "My truck was completely face down, pointed toward the ground," survivor Predicted Category: Retail
	Article 4: Summary: Doctors remove five small polyps from President Bush's colon. All were small, less than a centimeter [half an inch] in diameter. Bush reclaimed Predicted Category: Retail
	Article 5: Summary: NEW: Atlanta Falcons owner calls Vick's actions "incomprehensible and unacceptable" Vick agrees to plead guilty to conspiracy to travel in Inter Predicted Category: Retail
	Article 6: Summary: Youssif, 5, was doused in gasoline, set on fire outside his home in Baghdad. His parents tried in vain to get help for their son, leaving "no st Predicted Category: Entertainment

## 7. 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:** T5 :
  - 
  - For each input article, the T5 model generates a summary by predicting the next word token-by-token, using its encoder-decoder architecture.
  
- **Evaluation** :
  - Evaluate models using ROUGE scores and human evaluation.
  - Compare the output quality of BERT with the T5.

## 8. Evaluation Results

### ROUGE Scores for BART and T5 Models

1. BART ROUGE Results:

- ROUGE-1: 0.3746
- ROUGE-2: 0.1773
- ROUGE-L: 0.2839
- ROUGE-Lsum: 0.3275

2. T5 ROUGE Results:

- ROUGE-1: 0.3508
- ROUGE-2: 0.1600
- ROUGE-L: 0.2673
- ROUGE-Lsum: 0.3017

### Interpretation

BART outperforms T5 across all ROUGE metrics, meaning that BART produced summaries with higher overlap in words, bigrams, and longest common sequences compared to T5.

However, the difference between the two models is moderate, indicating both perform reasonably well.

### Confusion Matrix

Predicted \ Actual	Good	Poor
Good	TP	FP
Poor	FN	TN

- **TP:** True Positives (Correctly classified good summaries)
- **FP:** False Positives (Incorrectly classified summaries as good)
- **FN:** False Negatives (Incorrectly classified good summaries as poor)
- **TN:** True Negatives (Correctly classified poor summaries)

[[2 0]

[0 0]]

### Explanation

True Positives (2): Both summaries were classified as 'Good' correctly. No False Positives or False Negatives.

The confusion matrix indicates that, for the small sample evaluated, both BART and T5 generated accurate summaries.

### Classification Report



Confusion Matrix:

```
[[2 0]
 [0 0]]
```

Classification Report:

	precision	recall	f1-score	support
Good	1.00	1.00	1.00	2
Poor	0.00	0.00	0.00	0
accuracy			1.00	2
macro avg	0.50	0.50	0.50	2
weighted avg	1.00	1.00	1.00	2

```
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/ classification.py:1531:
```

## 9. 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. BART has a slight edge over T5 in terms of ROUGE performance. The accuracy is 100%, but the limited dataset size (only 2 summaries) may not provide a comprehensive view. A larger dataset and more diverse samples are recommended for a robust evaluation



