Language Models Homework- 2

The aim of this report is to analyze performance of language models for summarization tasks

Dataset: cnn_dailymail version 3.0.0 from huggingface

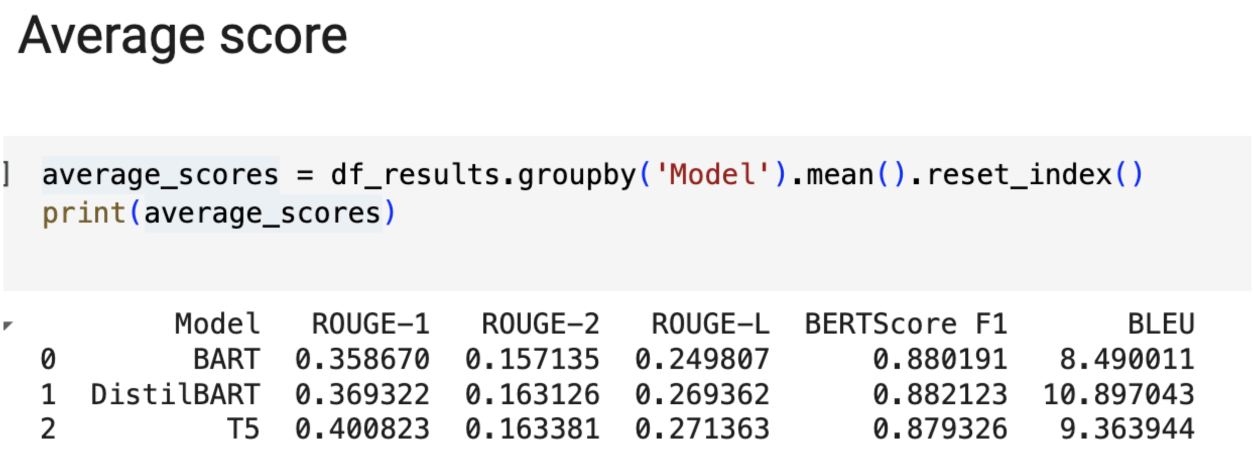
Models:

We have tested **3 huggingface** language models for this task:
1. Bart- JordiAb/BART_news_summarizer: BART combines the benefits of both autoencoder and autoregressive models. It uses a bidirectional encoder (like BERT) to understand contextual relationships between words and an autoregressive decoder (like GPT) to generate text, making it highly effective for tasks that involve both understanding and generating text.
2. T5-small: T5 simplifies the NLP landscape by framing all text-based language tasks as a text-to-text problem. Whether the task is translation, summarization, question answering, or classification, T5 treats the inputs and outputs as text strings, streamlining training and deployment pipelines.
3. DistilBart- distilbart-cnn-12-6: DistilBART is a distilled version of BART, which means it retains most of the effectiveness of BART but with fewer parameters and faster processing times. This makes it suitable for environments where computational resources or latency is a concern.

Evaluation Metrics:

1. **Rogue Score:** ROUGE measures the overlap of n-grams between the generated and reference texts. The most commonly used variants are ROUGE-N (which measures the overlap of n-grams, for example, ROUGE-1 for unigrams, ROUGE-2 for bigrams, etc.), ROUGE-L (which measures the longest common subsequence), and ROUGE-S (which measures skip-bigram co-occurrence statistics).
2. **Bert Score:** BERT Score computes the precision, recall, and F1 measure, where each token from the candidate summary is aligned with each token from the reference summary based on their BERT embeddings. This method considers the context of each token, providing a deeper semantic evaluation than n-gram overlap.
3. **Bleu Score:** BLEU compares the n-grams of the candidate translation or summary to the n-grams of the reference text, calculating a score based on the precision of n-grams. It also incorporates a brevity penalty to counteract the effect of overly short translations or summaries.

Results of 10 test samples on the following models



Average score

```
average_scores = df_results.groupby('Model').mean().reset_index()
print(average_scores)
```

```
        Model   ROUGE-1   ROUGE-2   ROUGE-L  BERTScore F1       BLEU
0        BART  0.358670  0.157135  0.249807      0.880191   8.490011
1  DistilBART  0.369322  0.163126  0.269362      0.882123  10.897043
2          T5  0.400823  0.163381  0.271363      0.879326   9.363944
```

Inference Results

ROUGE-1, ROUGE-2, ROUGE-L: DistilBART scores highest in ROUGE-L, indicating better summary structure, while T5 excels in capturing key unigrams and bigrams.
BERTScore F1: DistilBART leads, suggesting its summaries are semantically closest to the reference texts.
BLEU: DistilBART again tops the chart, indicating its summaries align closely with the reference n-grams.

Code link: https://github.com/rishabh-kr-jain/CMPE-258-HW-2