CMPE 258 - Deep Learning
Homework - 2

Problem statement:

Perform one of the selected tasks (language modeling, Translation, Question answering, Summarization) that leverages language models. Provide two comparisons using figures or tables: 1) Choose one dataset and compare the performance of multiple models (at least two) based on evaluation metrics. 2) Test and compare the models using your own provided test samples. Add one UI for your selected application via Gradio/streamlit, or any web/mobile applications.

Task chosen: **Summarization**

Datasets: https://huggingface.co/datasets/csebuetnlp/xlsum

Models:
1. **facebook/bart-large-cnn**
2. **sshleifer/distilbart-cnn-12-6**
3. **google/bigbird-pegasus-large-bigpatent**

Metrics: **BERT Score** & **Rouge Score**

Sources:
1. Hugging-face official website,
2. Prof. Kaikai Liu's github,
3. Internet

Introduction

AI has been evolving everyday and its use cases have been reaching the horizon. Machine Learning, Language Modelling, Computer Vision, Deep Learning are some of the major areas in AI which have a major impact on our everyday lives. The sudden rise in large language models have changed the way we see the world today. Open AI, Google, Microsoft, Anthropic are some of the companies that have invested large amounts of money into the development of technology. LLMs are being used for language translation, text generation, speech recognition, summarization, chatbots etc. In this assignment, we are working on developing a summarization tool and comparing the performance of 3 language models available on hugging-face library.

Methodology

We will be using Google Colab to run our scripts. After installing all the necessary libraries and importing them, we build pipelines for each of the models being used. We call the “pipeline” function and pass the necessary arguments inside. The dataset we’re using is the “csebuetnlp/xlsum” multilingual dataset. FOr our summarization purposes, we are using the english version of it. We use the pipelines we built earlier to generate summaries given a long textual data. We do this for all the three models.

Metrics:

To evaluate the models, we are using three metrics namely Rouge score, BERT score and BLEU score. We are visualizing two of these, Rouge and BERT. The code contains functions like calculate_rouge(), calculate_bert_score() and calculate_bleu_score(). We calculate for every these for every model. If a metric has multiple values in the output, like precision, recall and f1, we consider the f1 score.

Rouge Score:

What is a Rouge Score? It measures the overlap between the generated summary and the reference summary. Higher Rouge indicates better similarity.
Rouge 1 measures overlap of unigrams. Rouge 2 measures the overlap of bigrams, Rouge L measures the LCS (Longest Common Subsequence) and Rouge-Lsum takes the average of these sentence-level LCS scores.

We use Rouge-1 for basic overlap, Rouge-2 is more sensitive to word order and phrase structure, Rouge-L considers longer sequences and is less sensitive to word order and finally, the Rouge-Lsum is for multi-sentence summaries as it considers sentence level coherence.

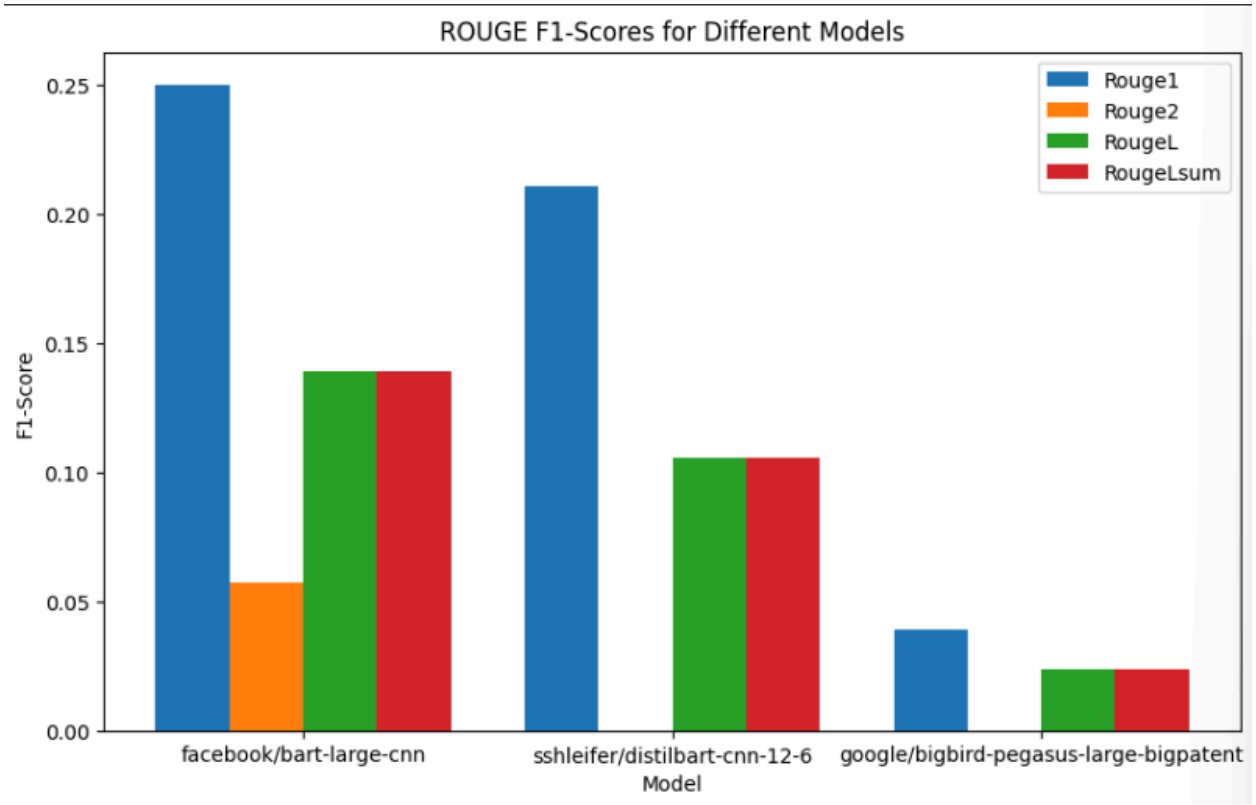

Fig. 1 This figure shows the comparison of Rouge scores between multiple models

As you can see, Rouge 1 for the facebook based model is performing the best. You may notice that the Rouge 2 scores for Distillbart and Google model are absolute zero. Zero Rouge-2 means that there are no exact bigram matches between the generated summaries and the reference summaries. This could be due to the summary length, the dataset might be such that it makes it challenging for these models to generate summaries with high Rouge score. Rouge-2 is also a stricter metric that penalizes models heavy for any deviation from the exact word order.

BERT Score:

It is used for evaluating text generation tesks such as translation and summarization. Unlike Rouge, which depends on the exact word match, BERT leverages the power of contextual embeddings from the BERT model to evaluate similarity between two sentences.

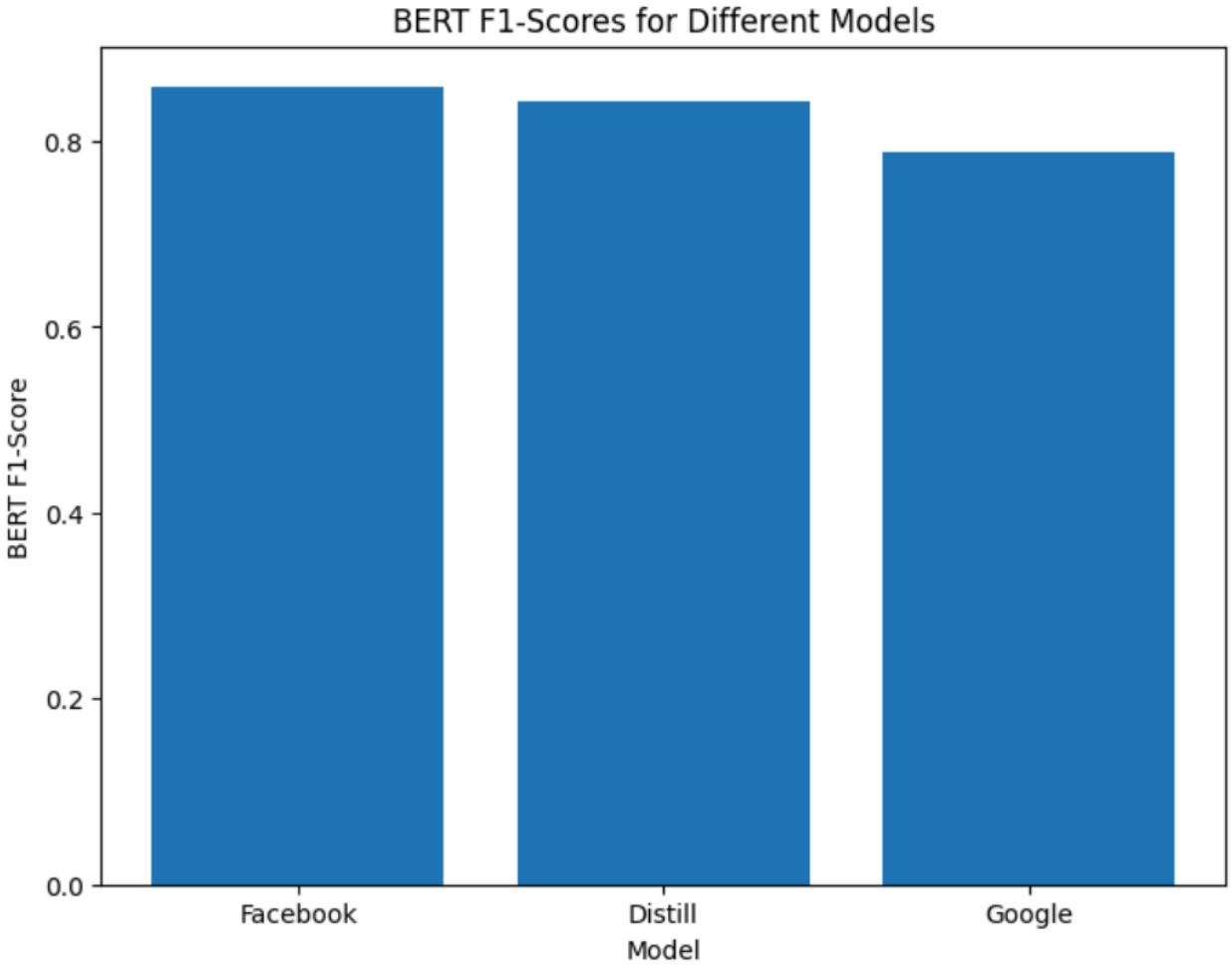

Fig. 2 Comparison of BERT scores for all 3 models based on their F1scores

BLEU Score

BLEU score is used to evaluate the quality of the machine translation system. It compares the generated translation with one or more reference translations. But BLEU can also be used for summarization although it is not recommended given some limitations. BLEU penalizes shorter summaries which is not what we want in a summarization task as we expect concise summaries. Just to test, we have used the BLEU score in this assignment.

```
facbook bleu:  1.052601808414082
Distill bleu:  1.0363935070930819
Google bleu:  0.20790245444365657
```

Fig. 3 BLEU scores for the three models

You can see the low scores here and understand why the BLEU score is not a good metrics for our requirements.

After calculating the scores, we move them into the dataframe structure to keep it nice and tidy. The dataframe consists of the metrics for each model calculated for the first 5 rows from the dataset. We consider the first 5 rows because adding more rows started causing memory related issues on Google Colab. We then calculate the average scores for each model.

Custom Testing:

We also test it with a custom article and its summary on the facebook to check the Rouge Score. The Rouge-1 had an fmeasure of 0.28169014084507044. Rouge-2 had 0, Rouge-L had 0.14084507042253522 and Rouge-Lsum had 0.14084507042253522

Deployment using Gradio:

We use gradio to give the application a nice shape and a good experience to people who want to test it out. The gradio gives 3 model options to choose from and then use the selected model to summarize.

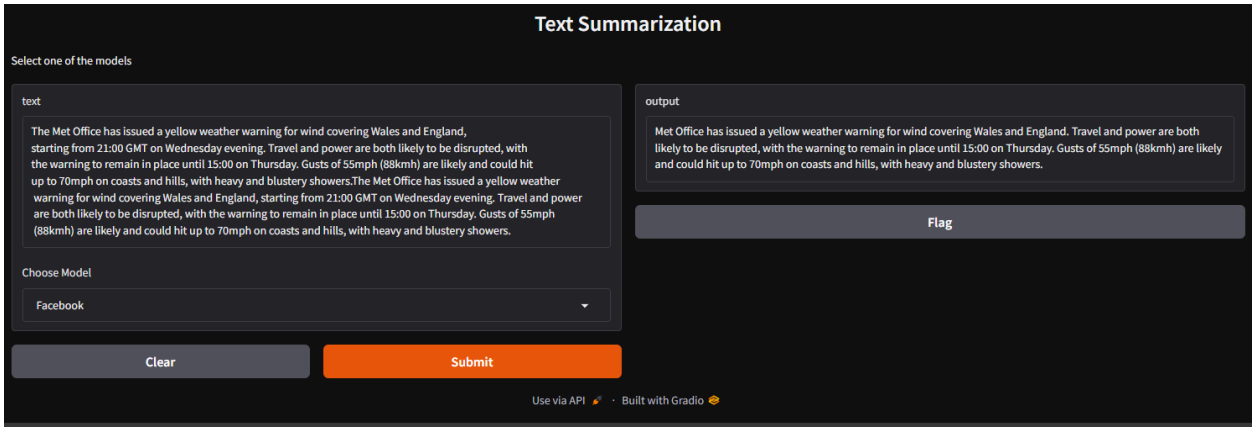

Fig. 3 Summary generated by the Facebook model

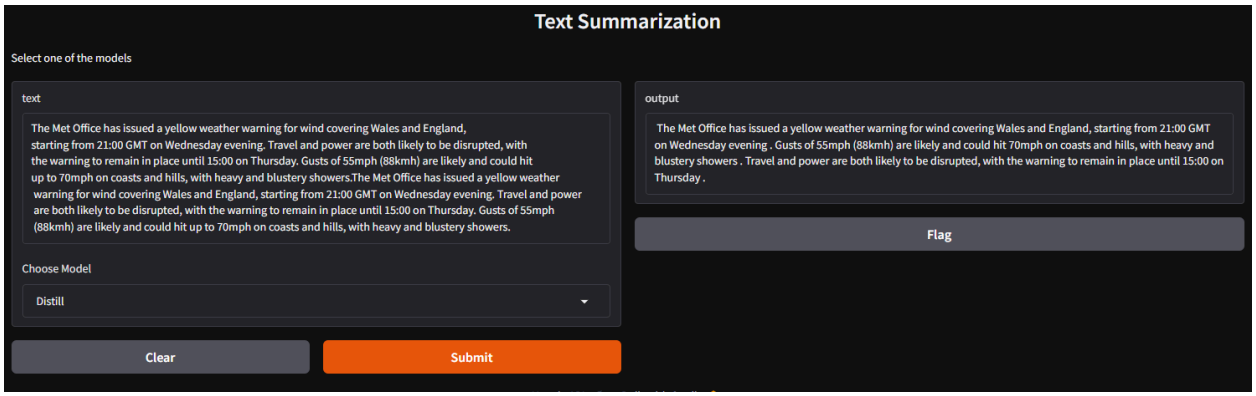

Fig. 4 Summary generated by the Distillbart model

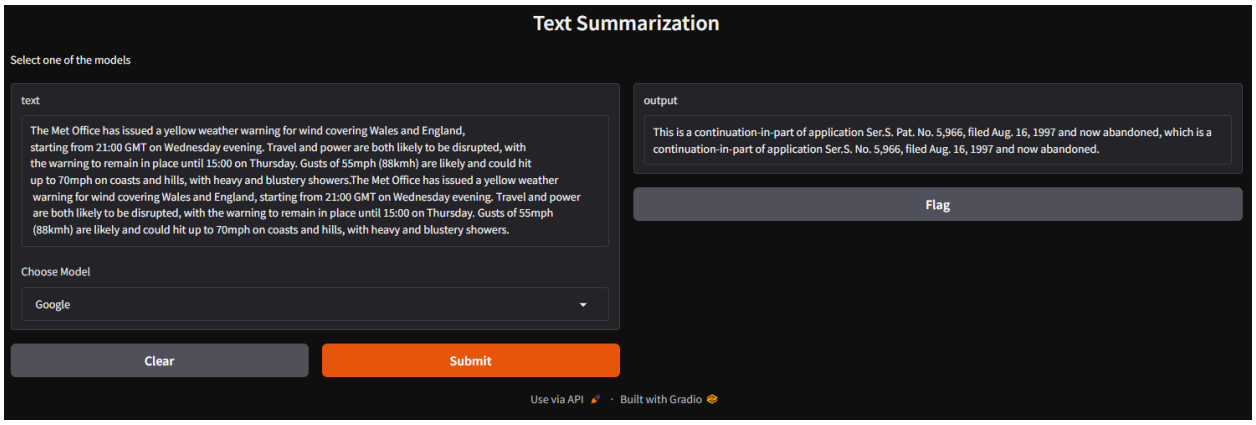

Fig. 5 Summary generated by the Google model

Conclusion

To meet the requirements of the assignment, we have compared the performance of multiple models based on BERT and Rouge scores and noticed that the facebook model performed well according to Rouge-1 and BERT scores.