Mandate 4
Problem Statement:

(Abstractive Summarization) Given a set of tweets pertaining to a trending topic, create an abstractive prose summary of the tweets. Do not just string the tweets together to form the summary. The summary will need to paraphrase and/or say more than what is directly said in the tweets. Propose a rubric to evaluate the accuracy of your summarization.

Evaluation Methods :-

- Rouge
- BERT score
- Human evaluation

Rouge : - The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring algorithm evaluates the **similarity between a candidate document** and a collection of reference documents. Use the ROUGE score to evaluate the quality of document translation and summarization models.

BERT Score : - BERTScore leverages the pre-trained contextual embeddings from BERT and matches words in candidate and **reference sentences by cosine similarity**. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Moreover, BERTScore computes precision, recall, and F1 measure, which can be useful for evaluating different language generation tasks.

Score generated by models : -

Model	Epoch	RougeLsum Score	Test Loss	BERT Score
Pegasus	First	44.009000	2.174100	F1 score: 0.880 Precision: 0.877 Recall : 0.884
	Last	41.143900	1.946711301	
T-5	First	30.412000	2.302800	F1 score: 0.879 Precision: 0.908 Recall : 0.852
	Last	29.693100	1.65443074	

Explanation of Scores :-

- **Rouge(Syntactic overlapping)**
  - Since Rouge calculates number of words common between two sentences like
    str1="the quick brown animal jumped over the lazy dog"
    str2="the quick brown fox jumped over the lazy dog"
    between them 8 word is common out of 9 so rouge will be → 8/9= 0.89
  - It is good for summary but **didn't help us to find the semantic of sentence** like animal in above sentence can be replace with humans and score will remain same.
  - **Since as we trained our model, Rouge is decreasing as our model is creating words by it's own(since it's abstractive summary).**

- Below images also show bit of bit of overfitting after 3rd epoch since training loss is dec. while validation loss starts increasing.

[3750/3750 2:33:01, Epoch 10/10]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	No log	1.744886	45.982000	23.539900	36.722000	42.049700	61.900000
2	2.078800	1.676235	47.831400	25.276900	38.476800	44.009000	53.722200
3	1.530100	1.688504	48.287800	24.782500	38.125300	43.800000	65.163900
4	1.203200	1.808993	47.627700	24.135300	37.527000	43.617400	61.986100
5	1.203200	1.943490	47.174000	23.387000	36.920300	43.090400	57.291700
6	0.858400	2.133166	47.849000	24.396600	37.542400	43.645100	61.194400
7	0.705400	2.215706	48.061800	23.977700	37.559900	43.766200	61.888900
8	0.650000	2.382680	46.186700	22.659700	35.591300	42.071700	66.122200
9	0.650000	2.523255	45.710400	21.940000	35.939500	41.706100	56.594400
10	0.543300	2.655459	45.185700	21.577100	35.161900	41.143900	68.733300

```
out.metrics
```

```
{'test_loss': 2.475882053375244,
 'test_rouge1': 47.0159,
 'test_rouge2': 23.5022,
 'test_rougeL': 36.772,
 'test_rougeLsum': 42.8161,
 'test_gen_len': 69.49,
 'test_runtime': 122.5222,
 'test_samples_per_second': 0.816,
 'test_steps_per_second': 0.204}
```

+ Code + Markdown

- **BERT Score**
  - Since BERT tries to find semantic similarity between two sentence so it tend to be more accurate of LLM's. So we will be focusing on this.

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

where:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

In "macro" F1 a separate F1 score is calculated for each `species` value and then averaged.

  - **Since F1 score in both models is ~ 0.88 which is consider as good.**

```
print(f"System level F1 score: {F1.mean():.3f}")
print(f"System level precision score: {P.mean():.3f}")
print(f"System level recall score: {R.mean():.3f}")
```

```
System level F1 score: 0.891
System level precision score: 0.891
System level recall score: 0.891
```

+ Code + Markdown

- **Human Evaluation** : - Looking at some summaries manually whether they are good or not. Summary generated through pegasus.
  - Summary : - "Prithvi Shaw must learn a lesson to stand up for the nation he plays for. Prithvishaw has been attacked for refusing to take selfies with social media influencer Sapna Gill, and his lawyer claims it was false."
  - Model Summary : - " Prithvi Shaw was allegedly manhandled and his car attacked with a baseball bat outside a hotel after an argument with sapna gill and her male friend shobhit after the batsman refused to click selfies with her."
  - Summary : - "Achraf hakimi's wife file for divorce and demand half of his property, but the court tells her he owns nothing. Achraf hakimi's story highlights the importance of loving one's mother, not one's spouse, as it sets a bad precedent for disinheriting children and spouses."
  - Model Summary : - "Achrafhakimi's wife filed for divorce and want a share of his wealth but the court told them that he has no property and that the bank have none either hakimi have put his entire fortune under his mother's name a long time ago."

- **TBH model summary looks more convincing then ground truth.**

```
ground_truth[45]
```

```
'This set of tweets is discussing the recent increase in the prices of LPG cylinders by the Narendra Modi government. People are expressing the
ir anger and frustration at the government for making domestic and commercial gas cylinders more expensive during the festival of Holi. They ar
e also criticizing the BJP for increasing prices and for not doing more to help the common people. They are calling for Smriti Irani to take ac
tion, and accusing the government of supporting Adani. The tweets also wish people a Happy Holi from the Modi government and express best wishe
s for the first gas cylinder price hike.'
```

```
generated_summaries[45]
```

```
'This set of tweets discusses the recent hike in the price of LPG cylinders in India during Holi, and the impact it has had on the public. Peop
le are criticizing the government for not taking action against the hike, and expressing their frustrations. They are criticizing the BJP gover
nment for not taking action against the hike, and calling for Smriti Irani to take action against the hike.'
```

Since above are the all Evaluation methods and their explanation, all methods giving good results. So we can say our model is working good.

Thanks