Name : Maatrika P
UGID : 21WU0102056

SMA Assignment-1 Report

Dataset utilized :

1. BBC News Summary: https://www.kaggle.com/datasets/pariza/bbc-news-summary/data

2. The dataset has around 5 domains i.e, business, entertainment, politics, sports and tech, each folder running around 500 articles and their corresponding summaries. Due to delayed computing time, only the first 50 articles of business domain and their summaries were considered.

Results :

BERT-Score before fine-tune ->
- Avg BERT-Score for 50 summaries:
  - Precision : 0.898
  - Recall : 0.777
  - F1 Score : 0.833

```
Final Cumulative BERT Scores:
Precision: 0.8984292471408843
Recall: 0.7776961767673493
F1 Score: 0.8332022118568421
```

ROUGE-Score before fine-tune ->
- Avg ROUGE-Scores for 50 summaries:
  - Rouge-1: 0.35
  - Rouge-2: 0.27
  - Rouge-L: 0.29
  - Rouge-Lsum: 0.29

```
ROUGE Scores:
ROUGE-1: 0.3531693202756071
ROUGE-2: 0.27901281130255196
ROUGE-L: 0.2990974295461166
ROUGE-Lsum: 0.2978857126523605
```

BLEU-Score before fine-tuning ->
- Avg BLEU-Scores for 50 summaries:
  - BLEU-Score: 0.757

```
Average BLEU Score: 0.7575197648448169
```

METEOR-Score before fine-tuning ->
- Avg METEOR-Scores for 50 summaries:
  - METEOR-Score: 0.212

```
Average METEOR Score: 0.21299650891559285
```

More insight on the scores obtained :

BERT-Score for 50 summaries:

Precision : 0.898
Recall : 0.777
F1 Score : 0.833

BERT scores usually range from (0-1). BERT-Score is a metric for evaluating the quality of text generation, and higher values indicate better quality. An F1 Score of 0.833 suggests that the generated summaries are of high quality with a good balance of precision and recall.

Avg ROUGE-Scores for 50 summaries:

Rouge-1: 0.35
Rouge-2: 0.27
Rouge-L: 0.29
Rouge-Lsum: 0.29

A Rouge-1 score of 0.35 suggests that there is a reasonable overlap between unigram (single-word) sequences in the generated summaries and the reference summaries. Similarly, Rouge-2 measures bigram overlap, and a score of 0.27 indicates moderate overlap.

Rouge-L and Rouge-Lsum consider the longest common subsequence, which can capture longer phrases. Both having scores around 0.29 indicate a substantial matching of phrases between the generated summaries and reference summaries.

The ROUGE scores range from 0 to 1, with higher scores indicating better quality.

Avg BLEU-Scores for 50 summaries:

BLEU-Score: 0.757

A BLEU score of 0.7575 is relatively good, indicating that the generated summaries are linguistically closer to the reference summaries. The BLEU score ranges from 0 to 1, with 1 representing a perfect match, this score is strong, it suggests that there is a reasonably high level of n-gram overlap between the generated summaries and the reference summaries.

Avg METEOR-Scores for 50 summaries:

METEOR-Score: 0.212

The average METEOR score of 0.213 suggests that the generated summaries have moderate linguistic similarity to the reference summaries, it is indicative of some level of linguistic overlap between the generated and reference summaries.

Fine-Tuning the T5-small model :

Results :

BERT-Score after fine-tune ->
- Avg BERT-Score for 50 summaries:
  - Precision : 0.869
  - Recall : 0.774
  - F1 Score : 0.818

```
Final Cumulative BERT Scores:
Precision: 0.86978520154953
Recall: 0.7744840514659882
F1 Score: 0.8189856839179993
```

ROUGE-Score after fine-tune ->
- Avg ROUGE-Scores for 50 summaries:
  - Rouge-1: 0.34
  - Rouge-2: 0.23
  - Rouge-L: 0.24
  - Rouge-Lsum: 0.24

```
ROUGE Scores:
ROUGE-1: 0.340177111036897
ROUGE-2: 0.2352473718909039
ROUGE-L: 0.2493709198443118
ROUGE-Lsum: 0.24992097166199087
```

BLEU-Score after fine-tuning ->
- Avg BLEU-Scores for 50 summaries:
  - BLEU-Score: 0.664

```
Average BLEU Score: 0.6647678816485887
```

METEOR-Score after fine-tuning ->
- Avg METEOR-Scores for 50 summaries:
  - METEOR-Score: 0.208

```
Average METEOR Score: 0.20869762086067684
```

More insight on the scores obtained after fine-tuning :

Avg BERT-Score for 50 summaries:
- **Precision : 0.869**
- **Recall : 0.774**
- **F1 Score : 0.818**

Precision: Precision measures the proportion of correct n-grams in the generated summary. A higher precision (0.869) suggests that the generated summaries closely match the reference texts.

Recall: Recall indicates how many of the reference n-grams are covered by the generated summary. An acceptable recall (0.774) shows that important information is retained.

F1 Score: The F1 Score (0.818) is a balance between precision and recall. In this case, it reflects a strong overall performance in capturing essential information.

Avg ROUGE-Scores for 50 summaries:

Rouge-1: 0.34
Rouge-2: 0.23
Rouge-L: 0.24
Rouge-Lsum: 0.24

Rouge-1: Rouge-1 (0.4) measures unigram overlap between the generated summary and references. A Rouge-1 score of 0.4 suggests a substantial agreement in unigram matches.

Rouge-2: Rouge-2 (0.23) assesses bigram overlap. While it's lower than Rouge-1, this is expected as bigrams are less frequent.

Rouge-L and Rouge-Lsum: These scores (0.24 for both) evaluate the longest common subsequence between generated and reference summaries. A consistent Rouge-L and

Rouge-Lsum score reflects the quality of long phrase overlaps.

Avg BLEU-Scores for 50 summaries:

BLEU-Score: 0.664

The BLEU-Score (0.664) assesses how well the generated summary aligns with the reference text. A BLEU score around 0.664 is indicative of decent coverage and word overlap. It suggests that the generated summaries are reasonably close to the references in terms of words used.

Avg METEOR-Scores for 50 summaries:

METEOR-Score: 0.208

METEOR (0.208) evaluates the quality of generated summaries by considering precision, recall, stemming, synonyms, and stemming variants. An METEOR score of 0.208 indicates that the model performs moderately well in terms of semantic quality and diversity.