**Summary Report: Performance of Pre-trained Text Summarization Models** horizontal line

This report presents the performance evaluation of a BERT pre-trained text summarization model applied to news articles. The model was assessed using several text summarization metrics including ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (bilingual evaluation understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering) and BERTScore. ROUGE, BLEU, and METEOR all differently measure the word similarities between the generated summaries and the reference summaries. The BERTScore measures similarities across the text embeddings of the words within the generated and reference summaries. The average scores for all metrics were calculated using a dataset of 250 articles and their corresponding reference summaries.

The ROUGE score ranges from 0 and 1 with an average ROUGE F1 score above 0.3 considered reasonable, while a score above 0.5 is considered very good. BLEU scores also range from 0 to 1 with an average score above 0.3 considered reasonable, while a score above 0.5 is considered very good. METEOR scores range from 0 to 1 with an average score above 0.2 considered reasonable, while a score above 0.3 is considered very good. Lastly, BERTScore ranges from -1 to 1 with an average above 0.6 considered reasonable, while a score above 0.8 is considered very good.

**Results:**

BERT Summarizer:

1. Average ROUGE Scores:

* ROUGE-1: F1 score of 0.225 (Recall: 0.361, Precision: 0.169)
* ROUGE-2: F1 score of 0.065 (Recall: 0.109, Precision: 0.048)
* ROUGE-L: F1 score of 0.203 (Recall: 0.324, Precision: 0.152)

1. Average BLEU score: 0.035
2. Average METEOR Score: 0.304
3. BERTScore:

* Average Precision: 0.495
* Average Recall: 0.607
* Average F1: 0.544

The evaluated BERT pre-trained text summarization model achieved reasonable scores in some metrics, such as ROUGE-1, METEOR, and BERTScore recall. However, there is room for improvement in other metrics like ROUGE-2, ROUGE-L, BLEU, and BERTScore precision and F1. This indicate that further fine-tuning using training data may lead to improved performance.