Evaluation Criteria for Summarization Project

Evaluation Metrics: ROUGE

Since our project involves comparing abstractive summaries generated by BERT and T5, we will use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. These are widely used to evaluate the quality of automatically generated summaries by comparing them with reference summaries.

Key Metrics:

- ROUGE-1:

- Measures the overlap of unigrams (individual words) between the generated summary and the reference summary.

- Purpose: This reflects the basic word-level match and gives an idea of how relevant the words in the generated summary are.

- ROUGE-2:

- Measures the overlap of bigrams (two consecutive words) between the generated and reference summaries.

- Purpose: Evaluates how well the sequence of words in the summary aligns with the reference, indicating coherence.

- ROUGE-L:

- Measures the longest common subsequence (LCS) between the generated summary and the reference summary.

- Purpose: Evaluates fluency and how closely the structure of the generated summary matches the reference. A higher score indicates better structural similarity.

- ROUGE-Lsum:

- A variant of ROUGE-L, specifically tuned for summarization tasks.

- Purpose: Focuses more on the summary-level structure (sentence ordering and coherence).

Reason for Using ROUGE Metrics:

- Standard for Summarization Tasks: ROUGE metrics are a well-accepted evaluation approach for summarizers, including for research-level tasks like your project.

- Precision and Recall: ROUGE considers both precision (how much of the generated summary is relevant) and recall (how much of the reference summary was captured).

- Quantitative Measure: It offers numeric insight into the performance of your models, helping you compare models fairly and objectively.