Question Answering Model Evaluation Report

1. Overview

A dataset of question-context-answer triples is used in this study to assess how well two pre-trained question answering (QA) algorithms perform. The goal is to ascertain which model, given the given context, predicts the right response the best.

2. Models Employed

Distilbert-base-uncased-distilled-squad, or DistilBERT: A lightweight transformer model that was refined using the SQuAD dataset after being extracted from BERT.
A robustly optimised BERT method that was also refined on the SQuAD2 dataset is called RoBERTa (deepset/roberta-base-squad2).

3. The dataset

A cleaned QA dataset including questions, their contexts, and matching human-annotated responses was used for the evaluation. To facilitate comparison and testing, the first 100 samples were chosen.

4. Evaluation Methodology

The identical set of 100 questions was answered by each model using the appropriate context. Next, the genuine answers and the expected responses were contrasted. The percentage of precise matches between the predicted and actual answers (case-insensitive and whitespace-trimmed) was used to calculate accuracy.

5. Results

✅ Model 1 Accuracy (DistilBERT): 83.00%

✅ Model 2 Accuracy (RoBERTa): 86.00%

According to the findings, RoBERTa performs marginally better than DistilBERT in terms of exact-match accuracy on this dataset.

6. Final thoughts

With over 80% accuracy in exact matches, both models perform well. But when it comes to correctly predicting responses, RoBERTa has a small advantage. This implies that RoBERTa is more appropriate for QA jobs that call for a high degree of precision. When speed and computational efficiency are more important than slight accuracy advantages, DistilBERT is still a good option.