NLP: A4 Do You Agree

GitHub Repository-
https://github.com/Santhosh01161/NLP_Do_You_Agree

Task 3: Sentence-BERT Evaluation & Analysis

Performance Metrics (Classification Report)
The table below summarizes the performance of the fine-tuned Sentence-BERT model on the Natural Language Inference (NLI) task. These metrics were calculated using a validation subset of 800 samples from the SNLI dataset.

Category	Precision	Recall	F1-Score	Support
Entailment	0.42	0.38	0.40	275
Neutral	0.36	0.44	0.40	260
Contradiction	0.45	0.40	0.42	265
Accuracy			**0.41**	800
Macro Avg	0.41	0.41	0.41	800
Weighted Avg	0.41	0.41	0.41	800

Discussion: Limitations, Challenges, and Improvements

Challenges Encountered

- **Hardware Constraints (VRAM):** Training Transformer architectures locally posed significant memory challenges. To prevent **Out-of-Memory (OOM)** errors, I implemented **Gradient Checkpointing** and **Mixed Precision (FP16)** training. Furthermore, a high **Gradient Accumulation** (16 steps) was used to simulate larger batch sizes without increasing memory overhead.

- **Environment Conflicts:** A critical challenge arose where the `transformers` library disabled PyTorch due to version incompatibilities with Python 3.12. This was bypassed by implementing a **manual tensor conversion** strategy for the final Web Application (Task 4) to ensure the model could still perform inference.
- **Semantic Convergence:** Training from scratch with limited data meant the model initially struggled to distinguish between 'Neutral' and 'Entailment' labels, as these categories often share high lexical overlap.

Limitations

- **Reduced Dataset Size:** Due to computational time limits, only 800 samples were used for fine-tuning. This is a small fraction of the 550k+ samples in the full SNLI corpus, which naturally limits the model's F1-score and generalizability.
- **Model Depth:** The backbone was restricted to 2 encoder layers and 4 attention heads. While this allowed for faster training, it reduced the model's capacity to capture the deep semantic dependencies required for perfect NLI classification.

Proposed Improvements

- **Advanced Loss Functions:** Implementing **Multiple Negatives Ranking Loss** or **Triplet Loss** (as detailed in the SBERT paper) would better optimize the vector space for sentence similarity compared to standard Softmax classification.
- **Transfer Learning:** Initializing the Siamese network with weights from a larger BERT model pre-trained on the full BookCorpus/WikiText datasets would significantly enhance the baseline linguistic understanding.
- **Hyperparameter Optimization:** Using automated tuning (like Optuna) for the **learning_rate** and **warmup_steps** could help the model find a more optimal global minimum during the fine-tuning stage.