Lab Assignment – 2
Fine-tune Pre-trained Language Models for Classification Tasks

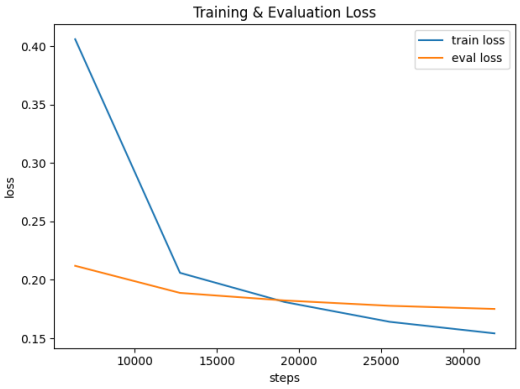
Dataset
- Source: `SetFit/ag_news` (Hugging Face). We recombined the provided splits and created an 80/10/10 stratified split. Final sizes: train 102,080, val 12,760, test 12,760.
- Task: 4-class news topic classification (World, Sports, Business, Sci/Tech).
- Preprocessing: Lowercasing is implicit with bert-base-uncased; tokenization via WordPiece using `AutoTokenizer` with `max_length=128`, truncation on, dynamic padding at batch time.

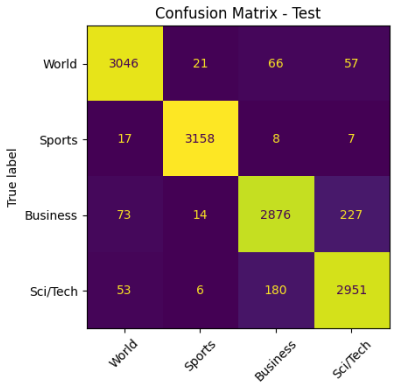
Tokenizer
- Name: `bert-base-uncased` (`AutoTokenizer`)
- Vocab size: 30,522 tokens.

Model Configuration
- Backbone: BERT-base (`AutoModelForSequenceClassification`) with 12 layers, hidden size 768, 12 attention heads, ≈110M parameters; num_labels=4.
- Fine-tuning method: LoRA adapters (r=8, α=16, dropout=0.1) targeting attention/feed-forward projections; ~1.34M trainable params (~1.21%).

Training Details
- Hardware: Apple Silicon (MPS).
- Epochs: 5 (within the required 3–5). Batch size: 16.
- Optimizer / LR schedule: AdamW with lr=5e-5, weight_decay=0.01, warmup_ratio=0.1.
- Loss: Cross-entropy.

Results (Test set)
- Accuracy: 0.9429
- Precision (macro): 0.9429
- Recall (macro): 0.9429
- F1 (macro): 0.9428
- Loss: 0.1824
- Artifacts: `loss_plot.png and confusion_matrix.png`





Observations & Improvements
- Performance is in the expected 94–95% AG News range for BERT-base; macro metrics ≈ accuracy → balanced class performance.
- Common errors: Business vs. Sci/Tech headline overlap.
- Next steps: try lr=3e-5, small prompt length sweep (e.g., 96/128/160), or modest epochs/early stopping; try RoBERTa-base for a +0.5–1% gain in some setups.