Name: Syedd Rafay Hassan

Roll No: 22i-1955

Section: C

1. Network Details

Architecture and Rationale

Two deep learning architectures were tested to detect whether two legal clauses have similar meanings:

1. **Siamese BiLSTM (Baseline Model):**
    - Uses twin LSTM encoders sharing weights to capture sentence-level meaning.
    - Suitable for textual similarity because it learns a shared embedding space where similar clauses are close together.
    - **Rationale:** BiLSTM captures both forward and backward context, useful for complex legal language.
2. **Attention-Based Encoder (Improved Model):**
    - Adds attention layers on top of BiLSTM to focus on the most important words.
    - **Rationale:** Attention helps highlight key legal terms (e.g., *agreement*, *termination*, *warranty*).

Training Environment

- **Framework:** TensorFlow 2.19.0
- **Hardware:** GPU-enabled runtime (Physical GPU available)
- **Dataset size:** 150,881 legal clauses

- **Pairs generated:**
    - Positive (similar): 118,305
    - Negative (non-similar): 118,305
- **Train/Validation split:**
    - Train: 201,118 pairs
    - Validation: 35,492 pairs
- **Tokenizer vocab size:** ≈ 30,000
- **Sequence length:** 256 tokens per clause
- **Batch size:** 64
- **Epochs:** 12 (early stopped at 9 for BiLSTM, 7 for Attention model)
- **Optimizer:** Adam
- **Loss:** Binary Cross-Entropy

2. Baselines and Comparison

Model	Accuracy	Precision	Recall	F1	ROC-AUC	Epochs	Params	Train Time/Epoch
Siamese BiLSTM	**0.9983**	0.9967	**0.9998**	**0.9983**	**0.9999**	9	3.96M	~67s
Attention Encoder	0.9971	0.9951	0.9992	0.9971	0.9982	7	4.12M	~75s

Winner: *Siamese BiLSTM*
It slightly outperformed the Attention model in both **accuracy** and **F1 score**, and also trained faster per epoch.

3. Training Graphs

Both models showed smooth training progress:

- **Siamese BiLSTM:**
    - Training loss steadily decreased from 0.075 → 0.005.
    - Validation accuracy improved from 0.9956 → 0.9986.
    - Early stopping occurred at epoch 9 (best epoch = 6).
- **Attention Encoder:**
    - Training loss decreased from 0.13 → 0.009.
    - Validation accuracy peaked at 0.9973 (best epoch = 4).
    - Early stopping occurred at epoch 7.

Graphs showed clear **decreasing loss curves** and **rising accuracy curves**, confirming stable convergence.

4. Performance Measures and Domain Discussion

Metrics Used

- **Accuracy:** Overall correct predictions ratio.
- **Precision:** Correctly identified similar clauses out of all predicted similar.
- **Recall:** Model's ability to detect all true similar clauses.
- **F1 Score:** Balance between precision and recall — best for imbalanced or nuanced datasets.
- **ROC-AUC:** Measures overall discriminative ability across thresholds.

Rationale for Metric Choice

In legal clause matching, **false negatives** (missing a true match) are costly because similar clauses might go undetected.
Hence, **Recall** and **F1-score** are the most crucial metrics.
For real-world ("in-the-wild") systems, **F1-score** provides the best trade-off between missing matches and false alarms.

5. Correct and Incorrect Predictions

Correct Matches

1. **Label=1 Pred=1**
   Left: "complete agreement this agreement constitutes the entire agreement…"

Right: “complete agreement this agreement and the plan constitute the complete and exclusive agreement…”
Both express the same meaning — correctly classified.

2. **Label=0 Pred=0**
   Left: “notice of defaults in the event that the company receives written notice…”
   Right: “specific performance remedy at law for breach of any obligations…”
   Unrelated clauses — correctly marked as non-matching.

Incorrect Matches

1. **Label=0 Pred=1 (Prob=0.999)**
   Left: “maintenance of insurance…”
   Right: “maintenance of properties…”
   Model confused overlapping terms “maintenance of”, but the context differs.

2. **Label=0 Pred=1 (Prob=0.826)**
   Left: “investment company act none of the borrower…”
   Right: “investment company the company is not required…”
   Semantically related but legally distinct — hard for the model to separate.

Total correct: 35,430
Total incorrect: 62

6. Conclusion

The **Siamese BiLSTM** model performed slightly better overall.
It achieved:

- **Accuracy:** 99.83%

- **F1-score:** 0.9983
- **ROC-AUC:** 0.9999

Although the **Attention Encoder** introduced interpretability, its small drop in F1 and longer training time made BiLSTM the better choice.

In practical terms, this model can be effectively used to **automatically match and cluster similar legal clauses** across contracts or jurisdictions.