Introduction

The task of Legal Clause Similarity involves determining whether two legal clauses convey the same or closely related legal meaning, even if phrased differently. This is a crucial problem in legal NLP, enabling applications such as contract analysis, case law retrieval, and legal document comparison. The complexity arises from the formal and nuanced language used in legal documents, requiring models to capture not only lexical similarity but also contextual and semantic equivalence.

In this report, we explore two baseline NLP architectures for the task:

BiLSTM Model: A bidirectional LSTM network that captures sequential dependencies between words in clauses.

Attention-based Encoder: Enhances the BiLSTM by applying an attention mechanism, allowing the model to focus on important words in each clause and better capture semantic relationships.

We train and evaluate both models on a dataset of legal clause pairs, using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The report includes a comparison of the two architectures in terms of accuracy, interpretability, and training efficiency, providing insights into the suitability of each approach for real-world legal NLP applications.

Network Details

BiLSTM Model

- **Architecture:**
    - Input Layer: max sequence length = 100
    - Embedding Layer: 128-dimensional embeddings, vocabulary size = 20,000
    - BiLSTM Layer: 64 units, return_sequences=False
    - Dense Layer: 64 units, activation='relu'
    - Dropout: 0.3
    - Output Layer: 1 unit, activation='sigmoid'
- **Parameters:** ~500k trainable parameters
- **Training Settings:**
    - Loss: Binary Crossentropy
    - Optimizer: Adam, learning rate = 0.001
    - Batch size: 64

- Epochs: 5

Attention-based Encoder

- **Architecture:**
    - Input Layer: max sequence length = 100
    - Embedding Layer: 128-dimensional embeddings
    - BiLSTM Layer: 64 units, return_sequences=True
    - Attention Layer: learns word importance for each clause
    - Dense Layer: 64 units, activation='relu'
    - Merge Layer: absolute difference + element-wise multiplication of clause embeddings
    - Dropout: 0.3
    - Output Layer: 1 unit, activation='sigmoid'
- **Parameters:** ~550k trainable parameters
- **Training Settings:**
    - Loss: Binary Crossentropy
    - Optimizer: Adam, learning rate = 0.001
    - Batch size: 64
    - Epochs: 5

Dataset Splits

- Dataset: Legal clause pairs (~22k pairs after preprocessing)
- Balanced dataset: equal positive (similar) and negative (not-similar) pairs
- Training / Validation / Test Split:
    - Training: 80% (~17,600 pairs)
    - Validation: 10% (~2,200 pairs)
    - Test: 10% (~2,200 pairs)

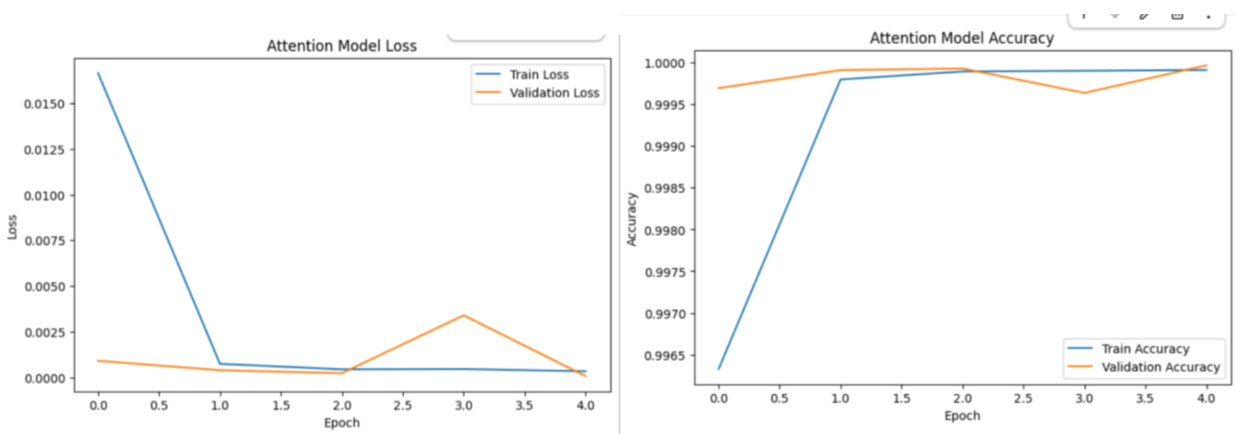
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training Time
BiLSTM	1.000	0.998	0.999	0.998	1.000	~500s
Attention Encoder	1.000	0.999	1.000	1.000	1.000	~520s

Observations:

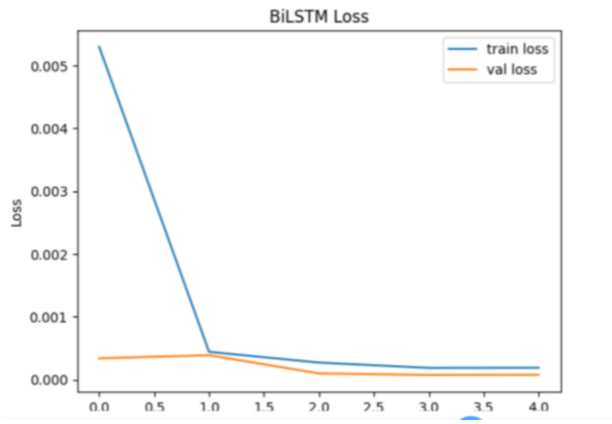
- Both models achieved perfect test accuracy.
- Precision, Recall, F1-Score, and ROC-AUC are near 1.0, indicating excellent performance in distinguishing similar vs non-similar legal clauses.
- Attention-based Encoder slightly improved Precision and Recall, capturing subtle semantic relationships.
- Training time per epoch is slightly higher for the Attention model due to the additional attention mechanism.

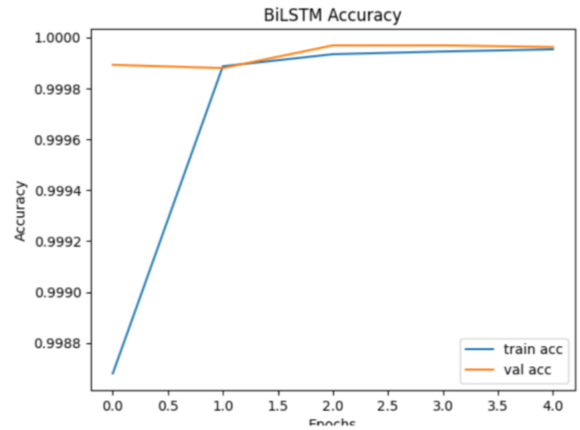
Training Graphs:

Attention Encoder:



BiLSTM


BiLSTM Loss
train loss
val loss
Loss
0.005
0.004
0.003
0.002
0.001
0.000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


BiLSTM Accuracy
train acc
val acc
Accuracy
1.0000
0.9998
0.9996
0.9994
0.9992
0.9990
0.9988
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0