

Report for DL A2

1. Introduction

Legal documents employ highly formal and structured language with complex terminology to express specific rights, duties, and conditions. The same legal principle can be articulated in multiple ways across different laws, contracts, or jurisdictions, making legal clause similarity detection a critical task in contract analysis, case law retrieval, and legal document comparison. This report presents the development and evaluation of multiple baseline NLP architectures for identifying semantic similarity between legal clauses without using pretrained transformer models.

2. Network Architecture and Design Rationale

2.1 BiLSTM-based Siamese Network

Architecture Overview: The BiLSTM-Siamese Network employs a shared encoder architecture for processing clause pairs, ensuring consistent feature extraction across both inputs. Key Components:

- Embedding Layer: 128-dimensional word embeddings (vocabulary size: 30,002)
- BiLSTM Encoder: 2-layer bidirectional LSTM with 256 hidden units per direction
- Feature Fusion: Concatenation of $[\text{encoded}_1, \text{encoded}_2, |\text{encoded}_1 - \text{encoded}_2|, \text{encoded}_1 \odot \text{encoded}_2]$
- Classification Head: Three fully connected layers ($512 \rightarrow 256 \rightarrow 128 \rightarrow 1$) with batch normalization and dropout (0.3)
- Total Parameters: 6,766,081

Rationale for Selection: BiLSTM networks excel at capturing sequential dependencies and contextual information from both directions, which is crucial for understanding legal text where word order and context significantly impact meaning. The siamese architecture with shared weights ensures that both clauses are encoded consistently, enabling robust similarity computation. The multi-scale feature fusion (concatenation, difference, and element-wise product) captures various aspects of semantic similarity.

2.2 Attention-based Encoder Network

Architecture Overview: This architecture leverages self-attention mechanisms to identify and focus on important legal terms and phrases within clauses. Key Components:

- Embedding Layer: 128-dimensional word embeddings

- Multi-Head Self-Attention: 4 attention heads with layer normalization
- Feed-Forward Network: 256-dimensional hidden layer with residual connections
- Mean Pooling: Attention-masked average pooling for sequence representation
- Classification Head: Similar structure to BiLSTM model
- Total Parameters: 4,137,857

Rationale for Selection: Attention mechanisms allow the model to selectively focus on critical legal terms and phrases while processing the entire sequence in parallel, unlike sequential RNNs. This is particularly valuable for legal text where specific keywords (e.g., "shall," "notwithstanding," "herein") carry significant semantic weight. The multi-head attention enables the model to capture different aspects of clause semantics simultaneously, while the transformer-style architecture with layer normalization and residual connections provides stable training dynamics.

Baseline Comparison Rationale: These two architectures represent fundamentally different approaches to sequence modeling: BiLSTM captures sequential dependencies through recurrent connections, while the Attention-based Encoder uses parallel attention mechanisms. This complementary design choice enables comprehensive evaluation of different inductive biases for legal clause similarity detection.

3. Training Configuration and Hyperparameters

3.1 Dataset Preparation

- Total Clauses: 150,881 legal clauses across 395 clause categories
- Clause Pairs Generated: 20,000 (10,000 similar + 10,000 dissimilar)
- Data Split: 70% training (14,000), 15% validation (3,000), 15% test (3,000)
- Vocabulary: 30,002 unique tokens with minimum frequency threshold of 2

3.2 Training Parameters

- Batch Size: 32
- Initial Learning Rate: 0.001
- Optimizer: Adam with ReduceLROnPlateau scheduler (factor=0.5, patience=3)
- Loss Function: Binary Cross-Entropy (BCE)
- Regularization: Dropout (0.3), Batch Normalization, Gradient Clipping (max_norm=1.0)
- Early Stopping: Patience of 7 epochs based on validation loss
- Maximum Epochs: 20
- Hardware: Tesla T4 GPU (15.83 GB memory) on Google Colab

3.3 Data Preprocessing

- Text normalization (lowercasing, whitespace removal)
- Special character filtering while preserving sentence structure
- Word-level tokenization
- Sequence padding with maximum length of 100 tokens

4. Training Graphs and Analysis

4.1 Loss Convergence

BiLSTM-Siamese Network:

- Initial Training Loss: 0.1581 → Final: 0.0065 (13 epochs)
- Initial Validation Loss: 0.0618 → Best: 0.0383 (epoch 6)
- Observation: Rapid convergence with minimal overfitting. Learning rate reduction at epoch 9 helped fine-tune the model. Early stopping triggered at epoch 13, indicating optimal convergence.

Attention-Encoder Network:

- Initial Training Loss: 0.6468 → Final: 0.0832 (15 epochs)
- Initial Validation Loss: 0.5880 → Best: 0.3851 (epoch 8)
- Observation: More gradual convergence compared to BiLSTM. Learning rate reduction at epoch 11. The higher validation loss indicates this architecture requires more training data or capacity to match BiLSTM performance on this task.

4.2 Accuracy Progression

BiLSTM-Siamese Network:

- Achieved 93.61% training accuracy in epoch 1
- Reached 99.80% training accuracy by epoch 13
- Validation accuracy stabilized at 99.37% (epoch 7)
- Excellent generalization with minimal train-validation gap

Attention-Encoder Network:

- Started at 62.59% training accuracy
- Reached 97.14% training accuracy by epoch 15
- Validation accuracy peaked at 88.53% (epoch 15)
- Larger train-validation gap suggests potential for architecture refinement

5. Performance Metrics and Domain Evaluation

5.1 Quantitative Results Summary

Metric	BiLSTM-Siamese	Attention-Encoder
Accuracy	0.9930 (99.30%)	0.8463 (84.63%)
Precision	0.9920 (99.20%)	0.8411 (84.11%)
Recall	0.9940 (99.40%)	0.8540 (85.40%)
F1-Score	0.9930 (99.30%)	0.8475 (84.75%)
ROC-AUC	0.9990 (99.90%)	0.9276 (92.76%)

5.2 Metric-Specific Analysis and Rationale

Accuracy:

- Definition: Proportion of correctly classified clause pairs (similar or dissimilar)
- Appropriateness: Suitable for this task as the dataset is balanced (50% similar, 50% dissimilar pairs)
- Interpretation: BiLSTM correctly classifies 99.3% of all clause pairs, while Attention-Encoder achieves 84.6%

Precision:

- Definition: Out of all predicted similar pairs, how many are truly similar
- Domain Relevance: Critical in legal applications where false positives (incorrectly marking clauses as similar) could lead to contractual conflicts or redundancy issues
- Interpretation: BiLSTM has 99.2% precision, meaning only 0.8% of its "similar" predictions are incorrect

Recall:

- Definition: Out of all truly similar pairs, how many were identified
- Domain Relevance: Essential for comprehensive legal document analysis where missing similar clauses (false negatives) could result in overlooked precedents or redundant clauses
- Interpretation: BiLSTM identifies 99.4% of all similar clause pairs, minimizing missed similarities

F1-Score:

- Definition: Harmonic mean of precision and recall
- Appropriateness: Ideal for legal NLP as it balances both false positives and false negatives, ensuring neither type of error dominates
- System Design Consideration: For a production legal system, F1-Score should be prioritized as both precision and recall are equally critical—missing similar clauses is as problematic as incorrectly matching dissimilar ones

ROC-AUC:

- Definition: Area under the ROC curve; measures the model's ability to rank similar pairs higher than dissimilar pairs across all classification thresholds
- Appropriateness: Excellent for evaluating the quality of similarity scores rather than binary predictions, which is valuable in legal systems where confidence scores inform human review
- Interpretation: BiLSTM's AUC of 0.999 indicates near-perfect ranking ability—similar clause pairs consistently receive higher similarity scores than dissimilar pairs

5.3 Confusion Matrix Analysis

BiLSTM-Siamese Network:

		Predicted		
		Dissimilar	Similar	
Actual	Dissim.	1488	12	(99.2% correct)
	Similar	9	1491	(99.4% correct)

- False Positives (12): Only 0.8% of dissimilar pairs incorrectly marked as similar

- False Negatives (9): Only 0.6% of similar pairs missed
- Balanced Performance: Nearly perfect classification in both categories

Attention-Encoder Network:

		Predicted		
		Dissimilar	Similar	
Actual	Dissim.	1258	242	(83.9% correct)
	Similar	219	1281	(85.4% correct)

- False Positives (242): 16.1% of dissimilar pairs incorrectly marked as similar
- False Negatives (219): 14.6% of similar pairs missed
- Performance Gap: Significantly higher error rate compared to BiLSTM

5.4 ROC Curve Analysis

The ROC curves demonstrate exceptional separation capability:

- BiLSTM-Siamese (AUC=0.9990): Near-vertical rise indicating almost perfect discrimination with minimal false positive rate even at maximum sensitivity
- Attention-Encoder (AUC=0.9276): Strong performance with good separation, though requiring higher false positive rates to achieve maximum true positive rate

6. Qualitative Results: Matched Legal Clauses

6.1 Correctly Matched Similar Clauses (BiLSTM-Siamese)

Example 1 - Termination Without Cause (Similarity: 1.0000):

- Both clauses address employer's right to terminate employment without cause
- Semantic equivalence despite different phrasing ("written notice" vs "Good Reason")
- Analysis: Model correctly identifies core legal concept of at-will termination rights

Example 2 - Insurance Requirements (Similarity: 1.0000):

- Both mandate insurance maintenance obligations
- Different contexts (contractor agreement vs grantor obligations) but same legal principle
- Analysis: Demonstrates model's ability to generalize beyond surface-level text matching

Example 3 - Fees and Expenses (Similarity: 1.0000):

- Both specify fee and expense allocation in agreements
- Different specific contexts (arbitration vs agreement preparation)
- Analysis: Model captures the underlying concept of cost allocation despite contextual differences

6.2 Correctly Matched Dissimilar Clauses (BiLSTM-Siamese)

Example 1 - Financial Statements vs Adjustments (Similarity: 0.0000):

- Completely different legal concepts with no semantic overlap
- Analysis: Model correctly assigns near-zero similarity to unrelated clause types

Example 2 - Additional Documents vs Investment Company Act (Similarity: 0.0002):

- Distinct legal areas (document requirements vs regulatory compliance)
- Analysis: Minimal similarity score reflects accurate semantic distance measurement

Example 3 - Miscellaneous vs Execution (Similarity: 0.1869):

- Both relate to agreement formalities but serve different functions
- Slightly higher score reflects some procedural similarity
- Analysis: Model nuances capture partial overlap in agreement execution context

6.3 Incorrectly Matched Cases (BiLSTM-Siamese)

False Negative Example - "Person" Definitions (Similarity: 0.1342):

- Both define "Person" in legal terms with nearly identical meaning
- Error Analysis: Model likely confused by different enumeration styles and word ordering
- Implication: Suggests need for better handling of definitional clause variations

False Positive Example - Modification vs Reinstatement (Similarity: 0.8921):

- Semantically different concepts incorrectly marked as similar
- Error Analysis: Possible confusion due to shared legal terminology and clause structure patterns
- Implication: Indicates model may over-rely on lexical similarity in certain cases

7. Comparative Performance Discussion

7.1 BiLSTM-Siamese Network Strengths

1. Sequential Dependency Modeling: Excellent at capturing word order and contextual relationships critical in legal language
1. Robust Encoding: Bidirectional processing ensures comprehensive context understanding from both directions
1. Consistent Similarity Computation: Shared siamese weights ensure symmetric and reliable similarity scoring
1. Superior Generalization: Minimal overfitting with 99.3% test accuracy

7.2 BiLSTM-Siamese Network Weaknesses

1. Long Sequence Handling: May struggle with very lengthy legal clauses due to vanishing gradient issues
1. Sequential Processing Bottleneck: Cannot parallelize within sequences, leading to slower inference
1. Fixed Context Window: Limited ability to capture very long-range dependencies beyond LSTM memory

7.3 Attention-Encoder Strengths

1. Selective Focus: Self-attention mechanism identifies and emphasizes critical legal terms
1. Parallel Processing: Faster training and inference compared to sequential RNNs
1. Long-Range Dependencies: Better theoretical capability for capturing distant word relationships
1. Interpretability: Attention weights can reveal which terms influence similarity decisions

7.4 Attention-Encoder Weaknesses

1. Data Requirements: Requires substantially more training data to achieve competitive performance (84.6% vs 99.3%)
1. Quadratic Complexity: Attention mechanism's $O(n^2)$ complexity becomes problematic for very long sequences
1. Training Instability: More gradual convergence and larger validation gap suggest optimization challenges
1. Underfitting on Current Dataset: 15.37% error rate indicates architecture may need refinement

7.5 Overall Comparison

Winner: BiLSTM-Siamese Network
The BiLSTM-Siamese Network significantly outperforms the Attention-Encoder across all metrics:

- 15% absolute improvement in accuracy (99.3% vs 84.6%)
- 7% higher ROC-AUC (0.999 vs 0.928)
- More balanced performance with minimal false positives and negatives
- Better suited for legal domain where sequential understanding and contextual relationships are paramount

The Attention-Encoder, while showing promise with 92.76% ROC-AUC, would require:

- Larger training dataset
- Architecture refinements (more layers, different attention mechanisms)
- Extended training with more sophisticated optimization strategies

8. System Design Recommendations for Real-World Deployment

8.1 Optimal Metric Selection

Primary Metric: F1-Score
For a production legal clause similarity system, F1-Score should be the primary optimization target because:

1. Balanced Error Consideration: Legal applications cannot tolerate high rates of either false positives (incorrect similarity) or false negatives (missed similarity)
1. False Positive Consequences: Incorrectly matching dissimilar clauses could lead to:
 - Contractual conflicts from combining incompatible terms
 - Regulatory violations from applying incorrect precedents
 - Wasted legal review time investigating non-matches
1. False Negative Consequences: Missing similar clauses results in:
 - Overlooked precedents in case law research
 - Redundant or conflicting clauses in contract drafting
 - Incomplete contract analysis missing key similarities
1. Practical Workflow Integration: F1-Score optimization ensures the system provides reliable recommendations that legal professionals can trust, reducing the need for extensive manual verification

8.2 Secondary Metrics

- ROC-AUC: Monitor for ranking quality when presenting multiple candidate matches
- Precision@K: For systems returning top-K similar clauses, optimize precision at relevant K values
- Recall@Threshold: Ensure minimum recall at acceptable confidence thresholds for high-stakes applications

8.3 Deployment Recommendation

Deploy BiLSTM-Siamese Network with the following configuration:

- Confidence threshold tuning based on use case (e.g., 0.95 for contract automation, 0.80 for research assistance)
- Human-in-the-loop validation for borderline cases (0.70-0.90 similarity range)
- Continuous monitoring of false positive/negative rates in production
- Periodic retraining with domain-specific legal clause pairs

9. Conclusion

This study successfully developed and evaluated two baseline NLP architectures for legal clause similarity detection without relying on pretrained transformers. The BiLSTM-Siamese Network achieved exceptional performance with 99.3% accuracy and 0.999 ROC-AUC, demonstrating that carefully designed baseline architectures can achieve near-perfect performance on legal text similarity tasks. The Attention-Encoder, while underperforming on this dataset (84.6% accuracy), shows theoretical promise and warrants further investigation with larger datasets and architectural refinements. The comprehensive evaluation framework encompassing quantitative metrics, confusion matrix analysis, ROC curves, and qualitative examples provides a robust foundation for understanding model behavior. The BiLSTM-Siamese Network's superior performance makes it the recommended choice for deployment in real-world legal clause similarity systems, where both precision and recall are critical for maintaining legal accuracy and compliance. Future work should explore:

1. Hybrid architectures combining BiLSTM sequential modeling with attention mechanisms
1. Incorporating legal domain knowledge through specialized embeddings
1. Multi-task learning with related legal NLP tasks
1. Evaluation on cross-jurisdictional and multilingual legal texts