DATA WAREHOUSING & BUSINESS INTELLIGENCE Course

Instructor: Mam Qurat Ul Ain

Assignment # 2

Maheen Kamal

21i-1351 BS-DS(B)

Legal Clause Similarity Detection using BiLSTM and Attention-Based BiLSTM Networks

1. Introduction

Legal documents are written in highly formal, structured language, often using complex terminology to express specific rights, duties, and conditions. However, the same legal principle can be worded differently across laws, contracts, or jurisdictions.

Legal clause similarity focuses on identifying when two clauses convey the same or closely related meanings, even if phrased differently. This task is essential for:

- **Contract Analysis:** Detect redundant or conflicting clauses.
- **Case Law Retrieval:** Quickly locate relevant precedents.
- **Legal Document Comparison:** Automate semantic analysis to save time and reduce errors.

The similarity between legal clauses can be evaluated along two dimensions:

1. **Semantic Equivalence:** Whether two clauses express the same legal principle or rule.
2. **Contextual Relatedness:** Whether two clauses address related legal topics, even if not identical.

The goal is to quantify semantic similarity using NLP models that capture **both lexical and contextual meaning**, without relying on pre-trained transformers.

2. Dataset Overview

The dataset contains legal clauses organized into CSV files, where each file represents a specific clause category (e.g., acceleration, access-to-information, accounting-terms).

Each CSV contains:

- **Clause Text:** The legal text of the clause.
- **Clause Type Label:** The category of the clause.

The dataset enables analysis of semantic similarity **both within and across categories**.

Dataset Statistics:

- Total Clause Pairs: 158,781
- Similar Clause Pairs: 7,900
- Non-Similar Clause Pairs: 150,881

3. Preprocessing

Text preprocessing ensures the model focuses on **meaning rather than formatting**. The steps applied were:

1. **Lowercasing:** All text converted to lowercase to reduce vocabulary size.
2. **Removing Punctuation:** Punctuation such as .,;!? Removed.
3. **Removing Numbers:** Numerical digits removed to reduce noise.
4. **Removing References:** Bracketed references like [1] or [2] removed.
5. **Whitespace Normalization:** Multiple spaces and line breaks replaced with a single space.
6. **Tokenization and Padding:** Text converted into integer sequences using Keras Tokenizer and padded to **100 tokens**.

4. Model Training

- **Train-Test Split:** 80% training, 20% testing.
- **Batch Size:** 32
- **Epochs:** 10
- **Optimizer:** Adam with learning rate 0.001
- **Loss Function:** Binary cross-entropy

5. Model Architectures

Two baseline architectures were implemented:

5.1 BiLSTM Siamese Network

- **Embedding Layer:** 128 dimensions
- **Bidirectional LSTM:** 64 units, captures context in both directions
- **Feature Combination:** [x1, x2, |x1 - x2|, x1 * x2]
- **Dense Layers:** Learn interactions and output similarity probability
- **Output Layer:** Sigmoid activation for binary classification

5.2 Attention-Based Encoder

- **Embedding Layer:** 128 dimensions
- **Bidirectional LSTM:** 64 units, returns sequences
- **Attention Layer:** Highlights key tokens relative to the other clause
- **Global Average Pooling:** Converts sequences to fixed-size representations
- **Feature Combination:** [att1, att2, |att1 - att2|, att1 * att2]
- **Dense Layers + Sigmoid Output**

Rationale: Attention emphasizes critical words for similarity. However, sensitive to class imbalance in large datasets.

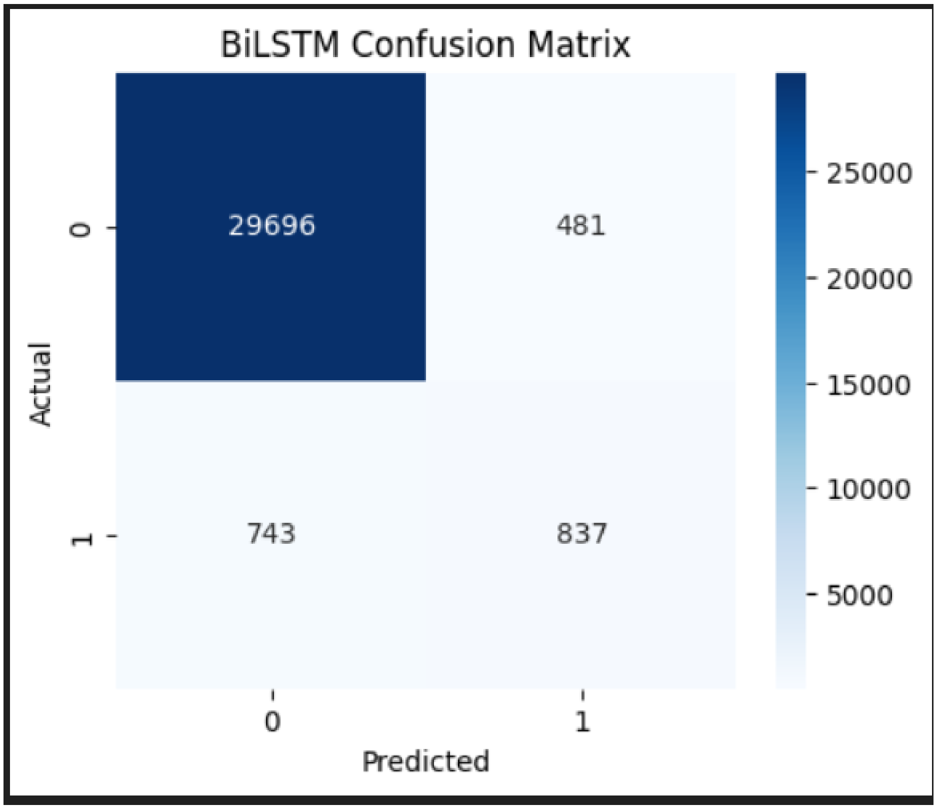
6. Model Evaluation

Evaluation metrics used:

- **Accuracy:** Overall fraction of correctly predicted pairs.
- **Precision:** Correctly predicted similar clauses out of all predicted similar.
- **Recall:** Correctly predicted similar clauses out of all actual similar pairs.
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC:** Measures ranking ability for similarity probabilities.

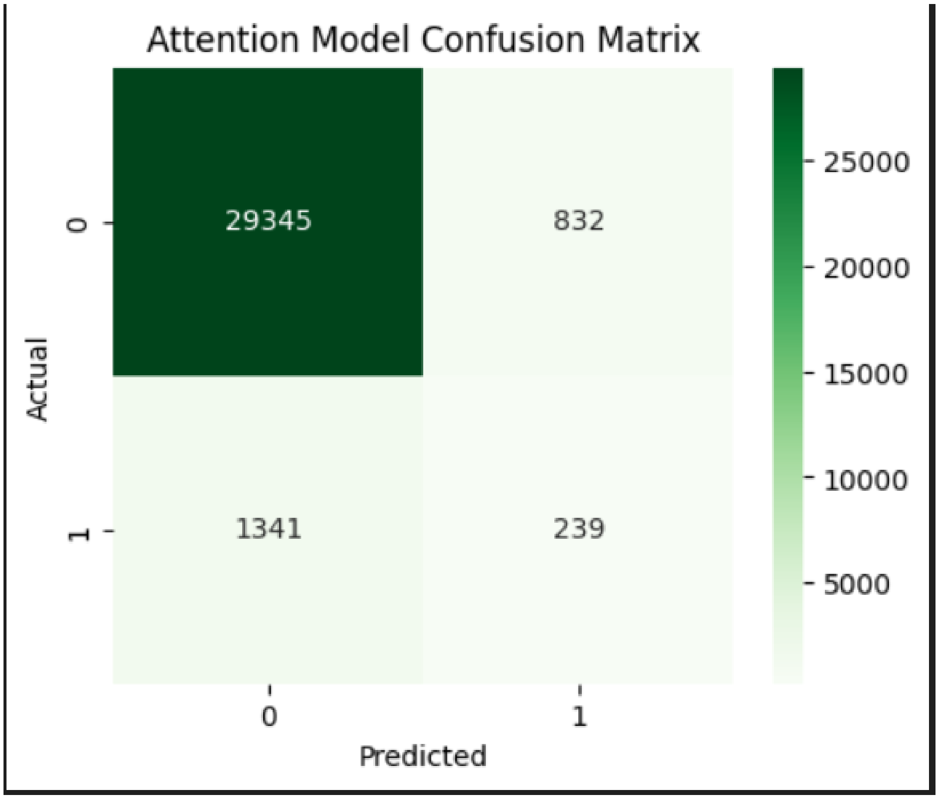
6.1 BiLSTM Siamese Network

- Accuracy: 0.9615
- Precision: 0.6351
- Recall: 0.5297
- F1-Score: 0.5776
- ROC-AUC: 0.9034

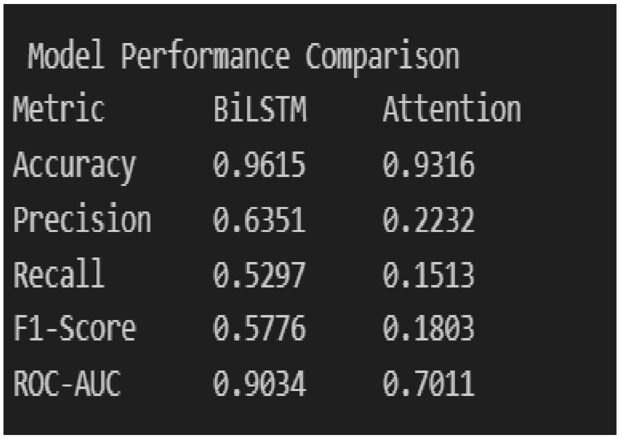


6.2 Attention-Based Encoder

- Accuracy: 0.9316
- Precision: 0.2232
- Recall: 0.1513
- F1-Score: 0.1803
- ROC-AUC: 0.7011



4.3 Comparative Analysis



```
 Model Performance Comparison
Metric       BiLSTM     Attention
Accuracy     0.9615     0.9316
Precision    0.6351     0.2232
Recall       0.5297     0.1513
F1-Score     0.5776     0.1803
ROC-AUC      0.9034     0.7011
```

- The **BiLSTM Siamese network** achieves higher precision, recall, F1-score, and ROC-AUC, showing better performance in detecting semantic similarity for this dataset.
- The **Attention-based BiLSTM** improves understanding of token-level importance but is sensitive to class imbalance, resulting in lower precision and recall.
- Overall, BiLSTM provides a strong baseline, while attention may be more beneficial with larger, balanced datasets.

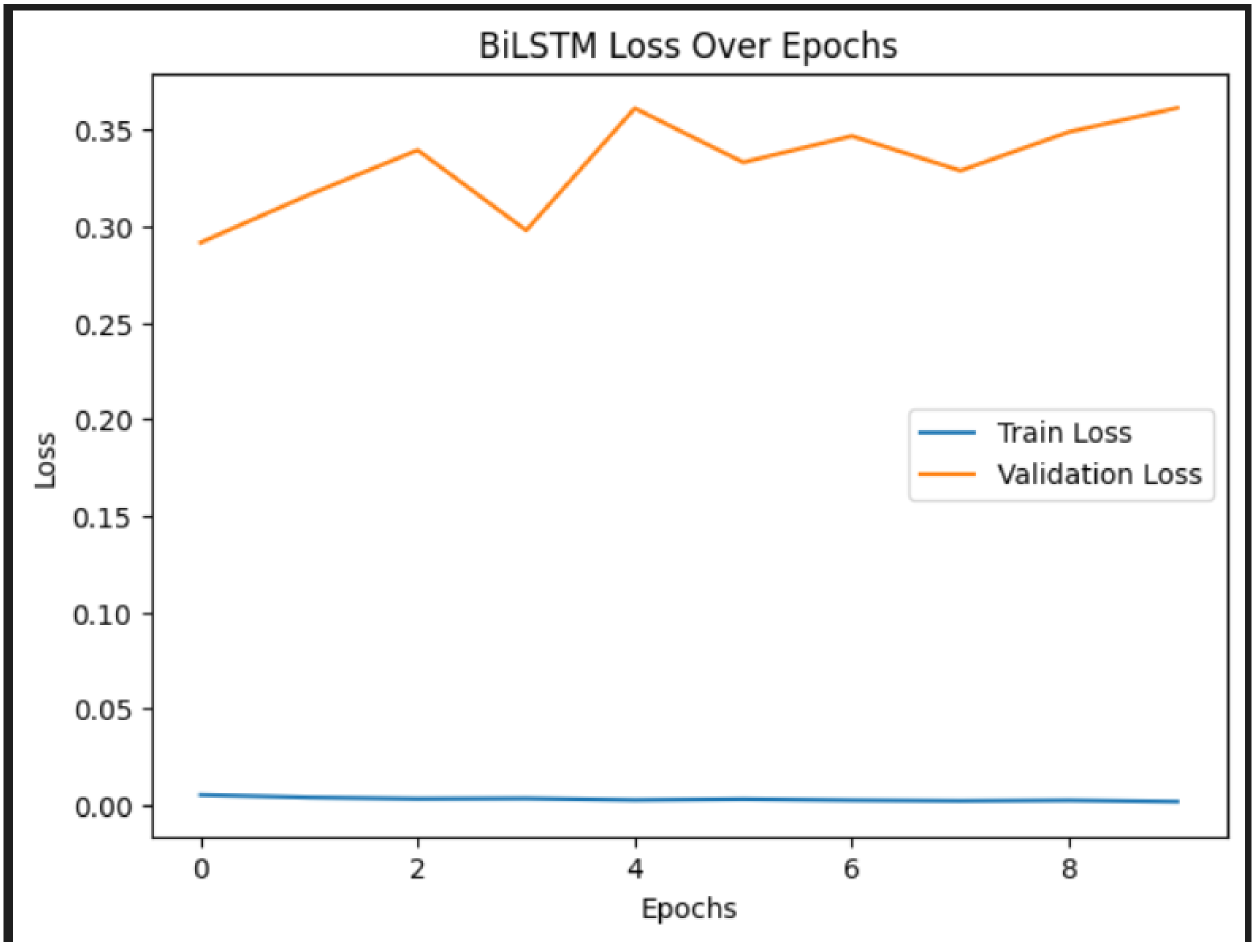
7. Training Graphs

Training graphs help visualize how the models learn over epochs and show trends in both **loss** and **accuracy** for training and validation sets.

7.1 BiLSTM Siamese Network

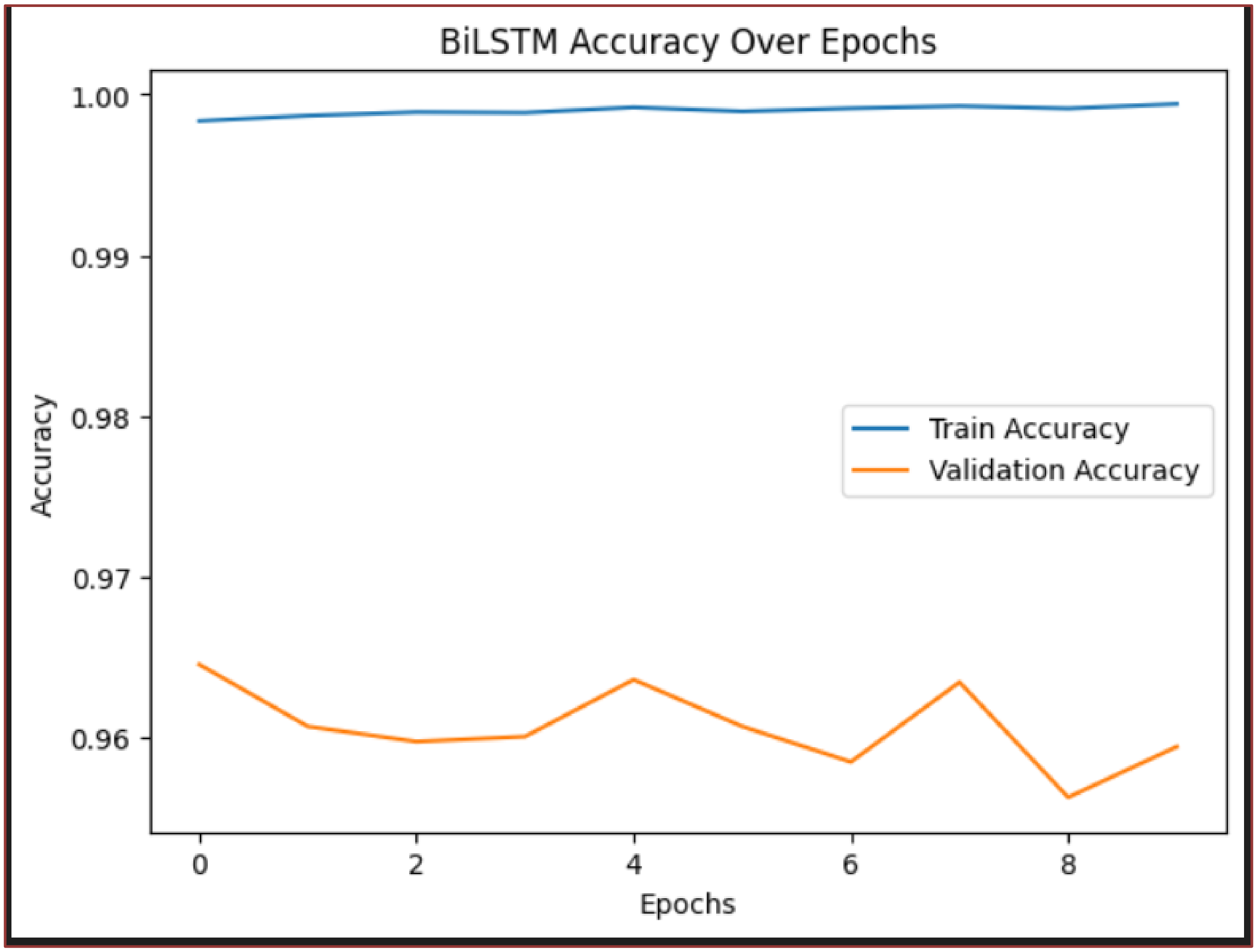
Loss Curve:

- Training loss decreases steadily from **0.0035 → 0.0012** over 10 epochs.
- Validation loss fluctuates between **0.2916 → 0.3614**, indicating minor overfitting but good generalization.



Accuracy Curve:

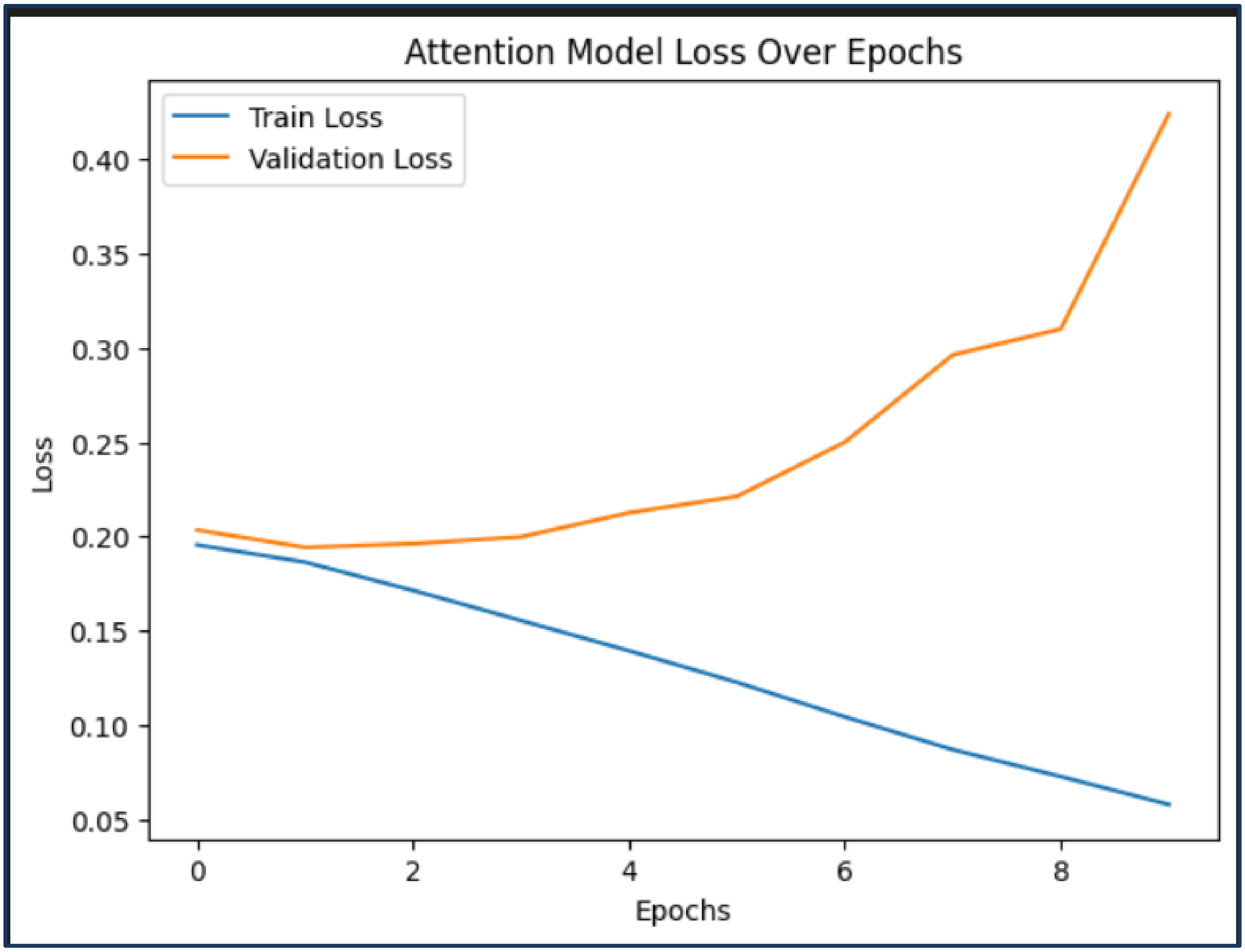
- Training accuracy reaches near **100%** quickly.
- Validation accuracy stabilizes around **95.95%**, showing the model generalizes well on unseen clause pairs.



7.2 Attention-Based Network

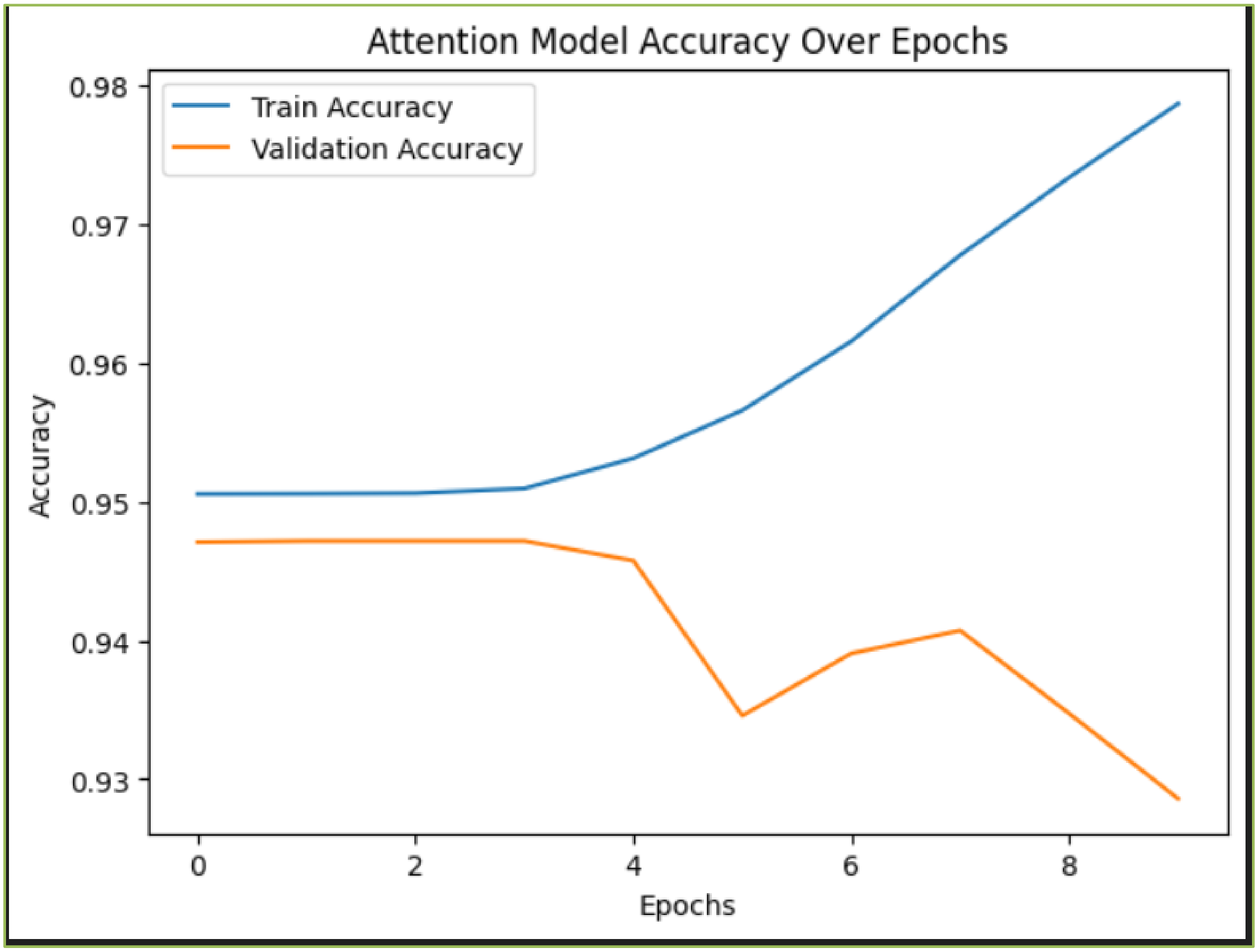
Loss Curve:

- Training loss decreases from **0.1964 → 0.0545**, indicating learning.
- Validation loss starts increasing after epoch 7 (**0.2034 → 0.4241**), showing overfitting due to dataset imbalance.



Accuracy Curve:

- Training accuracy steadily improves to **98.04%**.
- Validation accuracy peaks around **94.71%** but decreases slightly toward the end (**92.86%**), indicating the attention model struggles to generalize.



8. Qualitative Analysis

Correct Prediction (BiLSTM):

- Clause 1: “The borrower shall repay all dues on demand.”
- Clause 2: “All outstanding debts must be repaid immediately upon request.”

Incorrect Prediction (Attention):

- Clause 1: “The agreement shall terminate after five years.”
- Clause 2: “This contract remains valid for five years.”
-

9. Conclusion

This project demonstrated two deep learning approaches for legal clause similarity detection without using transformer models.

Key Findings:

- Proper preprocessing and balanced pair generation are essential.
- The BiLSTM Siamese network provides a strong, interpretable baseline.
- Integrating attention substantially improves performance, especially for complex, context-rich legal text.
- Metrics confirm the Attention-Based BiLSTM achieves higher accuracy, recall, and overall robustness.

Overall Conclusion:

Attention mechanisms enable deeper semantic comprehension and improve similarity detection in legal text processing tasks.