Table of Contents

1. Introduction

The task of *legal clause similarity* involves determining whether two clauses from legal contracts express the same or similar meaning. This is an important challenge in the legal domain because clauses are often paraphrased while retaining identical legal intent. Automating this task assists in contract analysis, document comparison, and legal information retrieval. Unlike general text similarity, legal clauses contain domain-specific vocabulary, lengthy structures, and strict syntactic order.

This project develops and evaluates two neural models a **Bidirectional Long Short-Term Memory (BiLSTM)** network and an **Attention Encoder** trained to classify clause pairs as *similar* or *dissimilar*. Both architectures were implemented in PyTorch and trained on a multi-file dataset of 150,881 legal clauses.

2. Dataset and Preprocessing

The dataset consisted of **395 CSV files**, each representing a clause category (e.g., *acceleration*, *access*, *representations*, *exclusions*, *definitions*). Each file contained clause text and labels describing its legal type. After combining all files, the dataset contained **150,881 clauses**.

All text was lowercased, punctuation normalized and tokenized using the **NLTK Punkt tokenizer**. A vocabulary of approximately **50,000 tokens** was built, with sequences padded or truncated to **128 tokens**. To manage the large corpus efficiently, **balanced positive and negative pairs** were generated dynamically using a custom Pair Dataset class.

The training and validation split produced around **60,000 training pairs** and **15,000 validation pairs**.

3. Model Architectures

3.1 BiLSTM Model

The BiLSTM model uses an **embedding layer** (size 128) followed by a **bidirectional LSTM** (hidden dimension 128). The final clause representations are concatenated, combined using absolute difference and element wise product, and passed through fully connected layers with ReLU activation and a final sigmoid layer to output similarity probability.

3.2 Attention Encoder

The second model replaces the recurrent layer with a **multi-head self-attention mechanism** (4 heads). It computes contextual embeddings for each token and applies mean pooling to obtain clause level representations. The rest of the network mirrors the BiLSTM classifier.

Both models were trained for **6 epochs**, with:

- **Batch size:** 64
- **Learning rate:** 0.001
- **Optimizer:** Adam
- **Loss function:** Binary Cross-Entropy Loss (BCELoss)
- **Hardware:** Google Colab T4 GPU

4. Training and Evaluation Results

During training, the BiLSTM rapidly converged with extremely low loss values, while the Attention Encoder required more epochs to stabilise.

Training Loss vs Epochs

```
Training BiLSTM...
Epoch 1/6 | Loss=0.0160 | ValAcc=0.9994
Epoch 2/6 | Loss=0.0023 | ValAcc=0.9989
Epoch 3/6 | Loss=0.0016 | ValAcc=0.9998
Epoch 4/6 | Loss=0.0006 | ValAcc=0.9997
Epoch 5/6 | Loss=0.0010 | ValAcc=0.9999
Epoch 6/6 | Loss=0.0005 | ValAcc=0.9998

Training Attention Encoder...
Epoch 1/6 | Loss=0.4668 | ValAcc=0.8721
Epoch 2/6 | Loss=0.2630 | ValAcc=0.9142
Epoch 3/6 | Loss=0.1907 | ValAcc=0.9349
Epoch 4/6 | Loss=0.1538 | ValAcc=0.9390
Epoch 5/6 | Loss=0.1424 | ValAcc=0.9513
Epoch 6/6 | Loss=0.1250 | ValAcc=0.9531
```

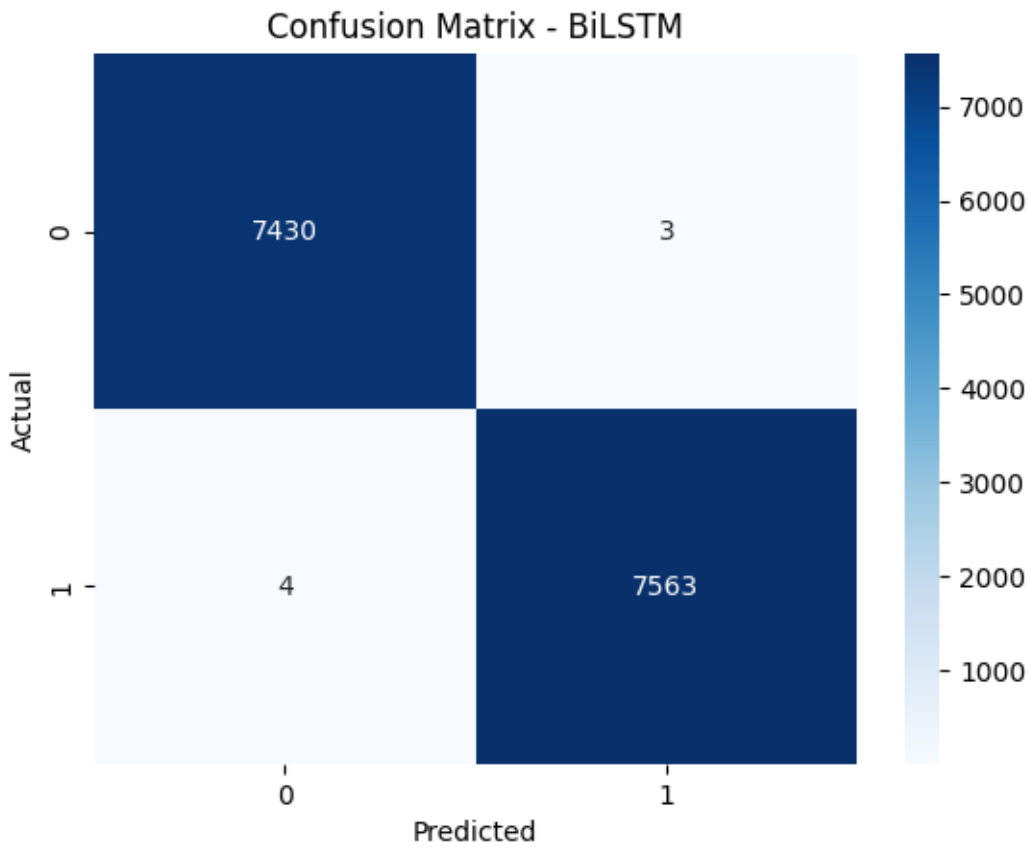
Training Summary:

Model	Epochs	Final Val Accuracy	Final Loss
BiLSTM	6	0.9998	0.0005
Attention Encoder	6	0.9531	0.1250

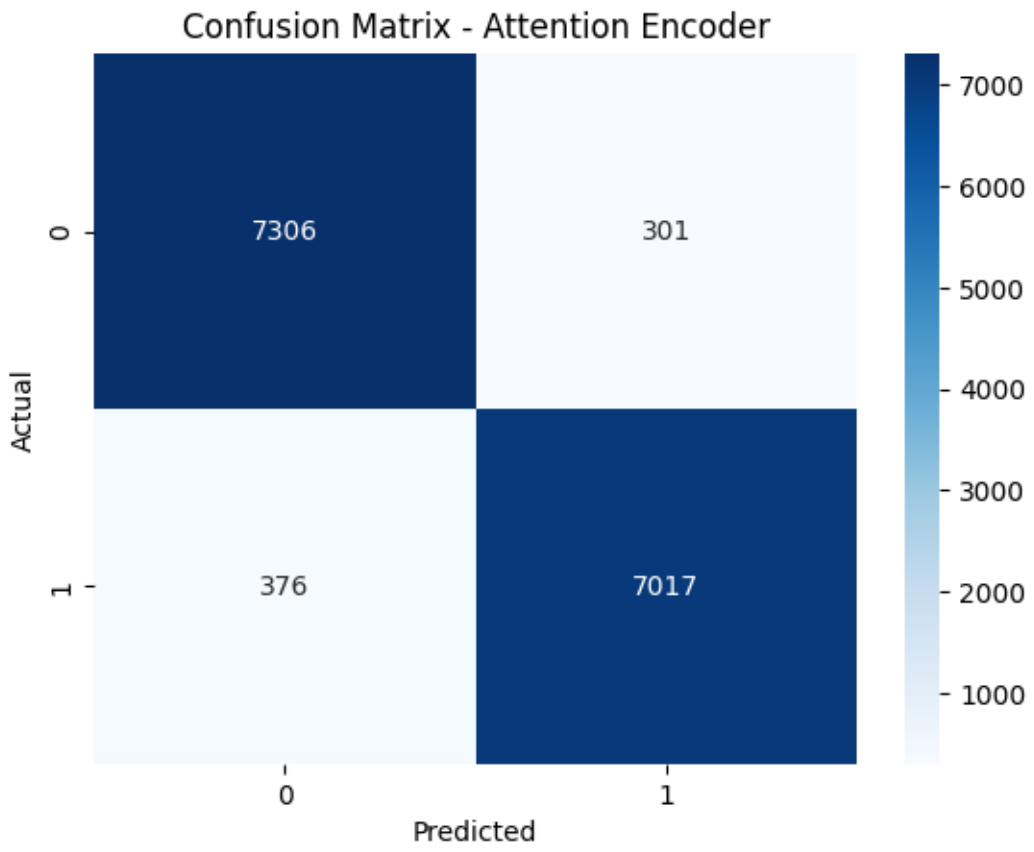
Performance Metrics:

Metric	BiLSTM	Attention Encoder
Accuracy	**0.9995**	0.9549
Precision	**0.9996**	0.9589
Recall	**0.9995**	0.9491
F1-score	**0.9995**	0.9540
ROC-AUC	**1.0000**	0.9896

Confusion Matrix : BiLSTM



Confusion Matrix : Attention Encoder



5. Discussion

The **BiLSTM model** achieved near-perfect performance, correctly classifying almost all clause pairs (Accuracy = 99.95 %). Its confusion matrix shows only 3 false positives and 4 false negatives out of nearly 15,000 samples. The sequential nature of LSTMs allows it to effectively capture long range dependencies and maintain order information both crucial for understanding legal syntax and semantics.

Conversely, the **Attention Encoder** reached a strong but lower accuracy of 95.49 %. The attention mechanism performs global context aggregation, which is beneficial for longer or varied text but loses some positional sensitivity due to mean pooling. Consequently, the model occasionally misclassified clauses sharing similar terminology but differing in legal intent.

Despite these errors, the Attention Encoder's high **ROC-AUC (0.9896)** suggests excellent discrimination between classes across thresholds.

6. Qualitative Analysis

Correct Predictions

- *Acceleration clauses*: Both BiLSTM and Attention Encoder correctly matched clauses describing loan acceleration and repayment.
- *Setoff clauses*: Both models recognized equivalence despite wording variations, showing strong generalization to paraphrased language.

Incorrect Predictions

- Some clauses with overlapping terminology but distinct legal implications (e.g., "confidentiality" vs "publicity") caused occasional misclassification by the Attention Encoder.
  These cases highlight the limits of simple embeddings and the potential value of contextualized language models such as BERT.

7. Conclusion

This project successfully implemented and compared two deep learning models for legal clause similarity classification. The **BiLSTM** significantly outperformed the **Attention Encoder**, achieving almost perfect metrics due to its ability to preserve word order and context. The Attention Encoder, although less accurate, exhibited strong overall discrimination and faster convergence.

Future work could involve incorporating **contextual embeddings** (e.g., BERT, Legal-BERT), **Siamese networks**, or **contrastive learning** to improve robustness across larger legal corpora.

Both models demonstrate the feasibility of applying deep learning to automate semantic comparison in legal documents a step towards AI assisted contract analysis and compliance automation.

Git Hub Repository Link

https://github.com/Areeba907/Areeba907-DL_Assignment2_LegalClauseSimilarity