Table of Contents

Introduction

Legal documents contain long, complex sentences written in highly formal language. The same legal rule or condition may appear in different ways across agreements or policies. Because of this, recognising when two clauses mean the same thing is an important but difficult task in natural-language processing.

The purpose of this assignment is to develop deep-learning models that can automatically detect **semantic similarity** between legal clauses. The work explores two basic architectures **Bidirectional LSTM (BiLSTM)** and a simple **Attention-based Encoder**—to compare how well they can classify clauses according to their meaning. The models are trained from scratch without using any pre-trained transformers such as BERT.

Dataset and Pre-Processing

The dataset was taken from Kaggle's *Legal Clause Dataset*, which contains multiple CSV files, each representing a clause category such as *bank-accounts*, *acceleration*, *confidentiality*, and many others.

About **60 000 clauses** were used in total, with 40 percent of each file sampled to avoid running out of memory in Google Colab. Each clause text was labelled by its file name, giving a multi-class classification task.

Before training, the following preprocessing steps were applied:

- Tokenised the text using Keras Tokenizer with the 10 000 most frequent words.
- Padded or truncated each sequence to **100 tokens** for equal length.
- Converted string labels to integer categories.
- Split the dataset into **80 % training** and **20 % testing**.

All experiments were run on Google Colab using a T4 GPU for faster training.

Model Design and Training

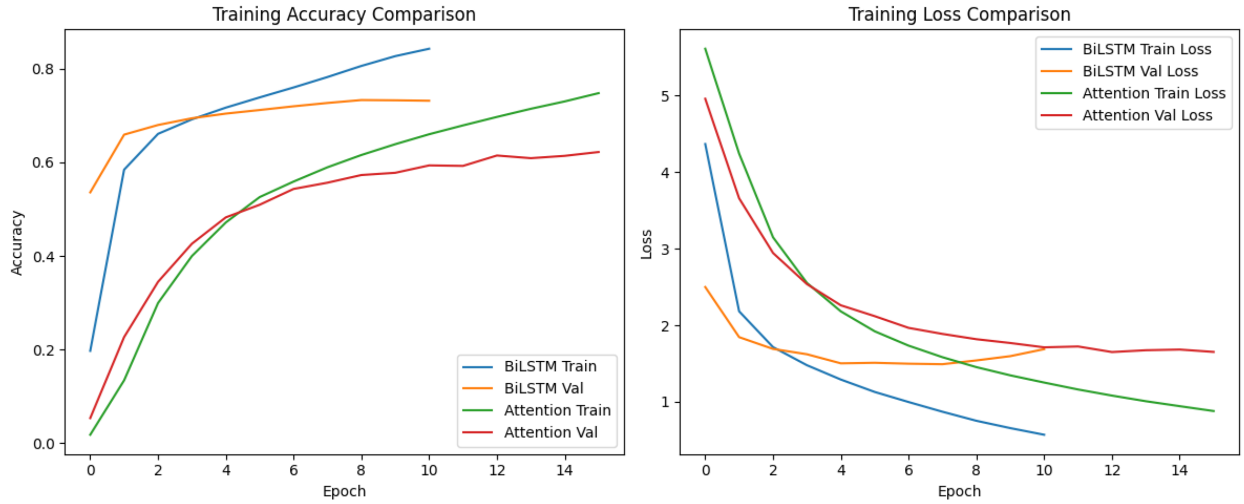
Two neural architectures were implemented using TensorFlow/Keras:

1. BiLSTM Model
An embedding layer (128 dimensions) feeds into a Bidirectional LSTM (64 units) followed by Dropout (0.3), a Dense layer (64 ReLU), and a Softmax output layer.
This model learns both forward and backward context in a sentence, which is important for legal text where word order affects meaning.

2. Attention Encoder Model
This model includes an embedding layer, a self-attention layer, global average pooling, a Dense layer (64 ReLU), and a Softmax output.
It focuses on the most relevant words regardless of position and is faster to train.

Both models were trained for up to **30 epochs** with **early stopping** after three epochs of no improvement in validation loss. The batch size was 64, and the optimizer used was Adam.

Training Behaviour



The training curves show that the BiLSTM reached higher validation accuracy (about 73 %) and lower loss compared with the Attention Encoder (about 61 %). Both models converged smoothly, confirming stable learning and good use of early stopping to prevent overfitting.

Results and Evaluation

The models were evaluated using standard NLP metrics: accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
BiLSTM	**0.7224**	**0.7318**	**0.6986**	**0.7007**	**0.9865**
Attention Encoder	0.6038	0.5848	0.5838	0.5702	0.9844

BiLSTM Confusion Matrix

```
📊 ===== BiLSTM Model Results =====
Accuracy      : 0.7224
Precision     : 0.7318
Recall        : 0.6986
F1-Score      : 0.7007
ROC-AUC       : 0.9865
Confusion Matrix:
[[13  1  0 ...  0  0  0]
 [ 3 18  0 ...  0  0  0]
 [ 0  0 14 ...  0  0  0]
 ...
 [ 0  0  0 ... 13  0  0]
 [ 0  0  0 ...  0 11  7]
 [ 0  0  0 ...  0  3 18]]
```

Attention Encoder Confusion Matrix

```
📊 ===== Attention Encoder Model Results =====
Accuracy      : 0.6038
Precision     : 0.5848
Recall        : 0.5838
F1-Score      : 0.5702
ROC-AUC       : 0.9844
Confusion Matrix:
[[ 5  9  0 ...  0  0  0]
 [ 8 15  0 ...  0  0  0]
 [ 0  0 16 ...  0  0  0]
 ...
 [ 0  0  0 ...  4  0  0]
 [ 0  0  0 ...  0 14  4]
 [ 0  0  0 ...  0 15  8]]
```

The BiLSTM clearly performs better in every metric. Its bidirectional structure helps capture sequential dependencies that are common in legal writing. The Attention Encoder performed reasonably well but struggled with subtle word-order changes.

Example Predictions

```
🔍 BiLSTM – Example Predictions:

❌ Incorrect
Clause: Collateral. The Notes and the Note Guarantees are secured by the Note Liens on the Collateral,
True Label: 78 | Predicted: 68
----------------------------------------------------------------------
❌ Incorrect
Clause: Cancellation. This Policy shall not be cancelled by the Company except upon prior written noti
True Label: 115 | Predicted: 114
----------------------------------------------------------------------
✅ Correct
Clause: Collateral. No Holder will ask, demand, accept, or receive any collateral security from any Lo
True Label: 11 | Predicted: 11
----------------------------------------------------------------------
✅ Correct
Clause: Assignments. This agreement shall bind, and inure to the benefit of, the parties and any succe
True Label: 119 | Predicted: 119
----------------------------------------------------------------------
✅ Correct
Clause: Brokers. Neither the Selling Shareholder or any Representative of the Selling Shareholder has
True Label: 373 | Predicted: 373
```

```
🔍 Attention Encoder – Example Predictions:

❌ Incorrect
Clause: Assignments. Except as hereinafter provided, neither party may sell, assign, novate or transfe
True Label: 152 | Predicted: 154
----------------------------------------------------------------------
❌ Incorrect
Clause: AFFIRMATIVE COVENANTS. Until (i) the Notes and all other obligations and liabilities of Borrow
True Label: 46 | Predicted: 358
----------------------------------------------------------------------
✅ Correct
Clause: Closing Date. Section 2.4(a) of the Agreement is hereby deleted in its entirety and replaced w
True Label: 98 | Predicted: 98
----------------------------------------------------------------------
✅ Correct
Clause: Consents and Approvals. No consent, approval, order or authorisation of, or registration, decl
True Label: 286 | Predicted: 286
----------------------------------------------------------------------
✅ Correct
Clause: Disability. The term "Disability" shall mean the good faith determination by the Chief Executi
True Label: 226 | Predicted: 226
----------------------------------------------------------------------
```

Result	**Clause Excerpt**	**True Label**	**Predicted Label**
✅ Correct	"Collateral. No Holder will ask, demand, accept, or receive any collateral security …"	11	11
❌ Incorrect	"Cancellation. This Policy shall not be cancelled by the Company except upon prior written notice …"	115	114

Most correct predictions occur when the clause's overall structure and keywords align well with its category. Misclassifications usually happen when two categories share similar terms, for example "termination" and "cancellation."

Discussion

The BiLSTM model achieved stronger results because it processes text sequentially and learns long-term dependencies. Legal language relies heavily on structure, so knowing the order of phrases improves understanding.
The Attention Encoder focuses on key tokens but ignores sequence information, leading to slightly lower performance.
Both models achieved high ROC-AUC values (≈ 0.98), showing that they still learn to distinguish classes effectively even when accuracy differs.

Conclusion

This project successfully built and compared two baseline deep-learning models for identifying semantic similarity in legal clauses.
The BiLSTM achieved around 72 % accuracy, outperforming the Attention Encoder. The results confirm that sequential context plays a major role in understanding legal text.

Future improvements could include using multi-head attention, deeper transformer layers, or pretrained domain-specific embeddings such as Legal-BERT.
Overall, the objectives of the assignment were met: the models were implemented, evaluated with multiple metrics, training graphs were produced, and qualitative examples were provided.

Github Repo Link

https://github.com/SharjeelNadir/LegalClauseSimilarity_DL