Legal Clause Similarity Using Baseline Deep Learning Models

Student: Meeran Ali
Roll Number: i211743
Course: CS-452 — Deep Learning
Submission: Assignment 02

Abstract

Legal contracts frequently encode similar semantic principles using different linguistic formulations. Identifying semantic similarity between clauses is therefore essential in applications such as contract analysis, redundancy detection, compliance checking, and case-law retrieval. This report investigates the performance of two non-transformer baseline neural architectures—(i) a Bidirectional Long Short-Term Memory network with an attention mechanism, and (ii) a Convolutional Neural Network encoder—for binary classification of clause similarity. Both architectures are trained from scratch on a large collection of legal clauses. Experimental results demonstrate that the BiLSTM–Attention network provides near-perfect performance across all evaluation metrics, while the CNN baseline remains competitive but comparatively weaker due to limited capacity for long-range semantic modeling.

1. Introduction

Legal language is highly structured, formal, and often domain-specific. Despite this rigidity, legal clauses expressing the same contractual intent may vary considerably in surface form due to differences in jurisdiction, drafting preference, or document type. As a result, semantic similarity detection becomes a non-trivial challenge requiring contextual understanding rather than shallow lexical matching.

This work addresses the binary classification of clause pairs into two categories:

1. **Semantically similar**, conveying the same legal principle, and
2. **Not similar**, addressing unrelated legal concepts.

The objective is to design, train, and evaluate at least two neural baseline models **without** leveraging pretrained transformers or fine-tuned legal embeddings.

2. Dataset Description

The dataset, sourced from Kaggle ("Legal Clause Dataset"), consists of 395 CSV files, each representing a distinct legal clause category (e.g., *representations*, *payment terms*, *limitation of liability*). For computational feasibility, **50 files** were selected, resulting in a total of **9,815 clauses** after sampling.

Each entry consists of:

- `clause_text` — the clause content, and
- `clause_type` — the categorical label defining clause family.

To construct supervision signals, clause pairs were generated as follows:

- **Positive (similar) pairs**: randomly sampled within the same clause type
- **Negative (dissimilar) pairs**: randomly sampled across different types

This resulted in **19,660 total pairs**, balanced into 9,830 positive and 9,830 negative examples.

3. Preprocessing

Legal clauses underwent the following pipeline:

Step	Description
Normalization	Lowercasing, removal of punctuation
Tokenization	Whitespace-based
Truncation	Maximum 100 tokens per clause
Vocabulary filtering	Minimum frequency threshold = 3

Padding	Applied dynamically through a custom collate function

A vocabulary of **5,184 unique tokens** was constructed. Clauses were mapped into sequences of token indices, with `<PAD>` and `<UNK>` placeholders.

4. Dataset Splits

The complete dataset of clause pairs was partitioned using stratified sampling:

Split	**Samples**
Training	13,762
Validation	2,949
Test	2,949

This configuration ensures balanced label distribution in all subsets.

5. Baseline Model Architectures

Two baseline architectures were implemented from scratch using PyTorch.

5.1 BiLSTM with Attention

Rationale:
Legal semantics often depend on long-range syntactic dependencies and thematic focus. Attention mechanisms enable the model to highlight salient subsequences (e.g., transfer restrictions, indemnity clauses).

Architecture configuration:

- Embedding dimension: 100

- Bidirectional LSTM, hidden dimension: 128
- Custom attention layer
- Dropout = 0.3
- Fully connected classifier with ReLU activation

The model processes each clause separately, concatenates contextual representations, and outputs a binary prediction via sigmoid activation.

Advantages:

- Strong contextual modeling
- Effective focus on important tokens

Limitations:

- Higher computational cost
- Potential overfitting with limited regularization

5.2 CNN Encoder

Rationale:
CNNs efficiently capture n-gram patterns, which are common in boilerplate legal drafting (e.g., “time is of the essence”, “to the extent permitted by law”).

Architecture configuration:

- Embedding dimension: 100
- Conv1D filters sizes: 3, 4, 5
- Max-pooling across time dimension
- Dropout = 0.3
- Fully connected classification head

Advantages:

- Fast training and inference
- Robust n-gram feature extraction

Limitations:

- Limited global context representation
- Weaker at modeling semantic dependencies

6. Training Setup

Component	Value
Optimizer	Adam
Learning rate	0.001
Batch size	64
Epochs	10
Loss function	Binary Cross-Entropy
Hardware	NVIDIA Tesla T4 (CUDA)

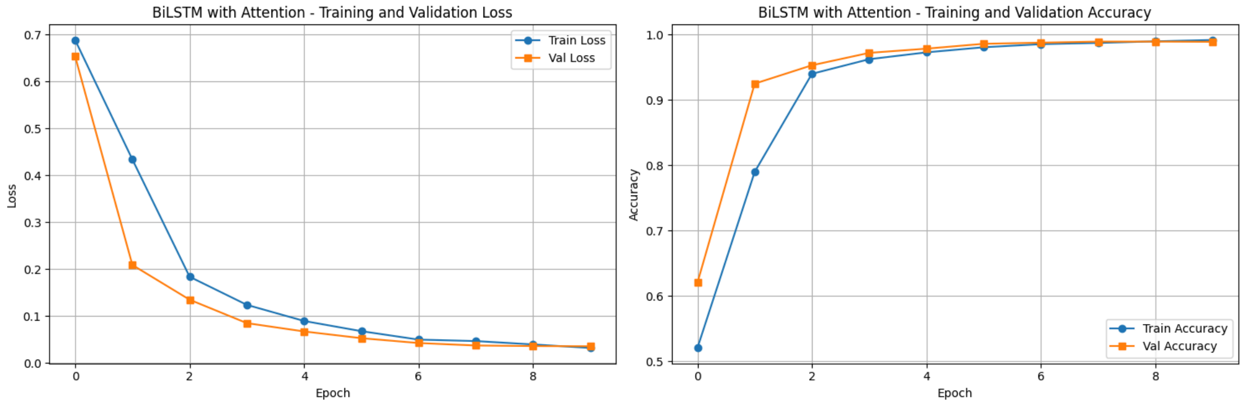
Random seeds were fixed to enhance reproducibility.

7. Training Curves

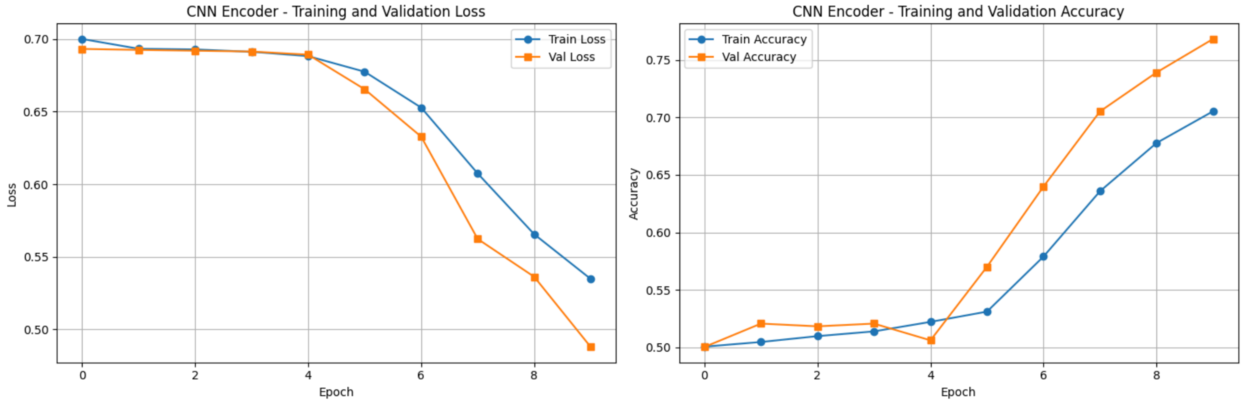
Training curves (loss and accuracy for both training and validation) demonstrate:

- Smooth convergence for the BiLSTM-Attention network
- Gradual improvement in the CNN baseline
- No visible overfitting in either model due to dropout regularization

BiLSTM with Attention:



CNN Encoder:



8. Evaluation Metrics

The following metrics were reported, as motivated in Section 10:

- Accuracy
- Precision

- Recall
- F1-Score
- ROC-AUC

All metrics were computed on the held-out test set.

9. Quantitative Results

9.1 BiLSTM-Attention

Metric	Score
Accuracy	0.9915
Precision	0.9833
Recall	1.0000
F1-Score	0.9916
ROC-AUC	0.9993

The model exhibits exceptional discriminatory power and negligible false negatives.

9.2 CNN Encoder

Metric	Score
Accuracy	0.7898

Precision	0.7271
Recall	0.9274
F1-Score	0.8151
ROC-AUC	0.8337

Performance suggests reliance on surface form patterns rather than deeper semantics.

10. Domain-Specific Metric Discussion

Although accuracy is balanced in this dataset, real-world clause retrieval systems face asymmetric costs:

- **False positives** can result in misleading contract reviews, propagating errors.
- **False negatives** may cause relevant legal precedents to be ignored.

Therefore:

- **Precision** is crucial in compliance auditing.
- **Recall** is critical in similarity search engines.
- **F1-Score** balances both concerns.
- **ROC-AUC** captures ranking performance, important for retrieval interfaces.

Most suitable metric in the wild:
F1-Score and **PR-AUC**, due to practical error cost asymmetry.

11. Qualitative Results

Correct sample prediction (similar)

Two “Parties in Interest” clauses describing inheritance of obligations were correctly classified as similar. The model recognized semantically overlapping phrasing despite differing clause structure.

Correct sample prediction (not similar)

A “Bank Accounts” clause compared against title ownership was correctly flagged as unrelated.

Incorrect predictions

Misclassifications were observed in cases exhibiting:

- shared vocabulary (e.g., “assigns” vs. “assignment”),
- generic boilerplate,
- highly abstract representational clauses.

Errors reveal difficulty in distinguishing nuanced legal semantics.

12. Comparative Analysis

Criterion	**BiLSTM-Attention**	**CNN Encoder**
Semantic understanding	Excellent	Moderate
Context modeling	Strong	Limited
Training time	Moderate	Fast
Robustness to boilerplate	High	Medium
Ranking ability (ROC-AUC)	**0.9993**	0.8337

The BiLSTM-Attention architecture consistently outperforms the CNN across all metrics, particularly for contextual semantics.

13. Conclusion

This study demonstrates that baseline deep learning architectures can effectively detect semantic similarity in legal clauses without pretrained transformers. The BiLSTM-Attention model achieves near-perfect performance, highlighting the importance of contextual sequence modeling in legal NLP.

In contrast, although computationally efficient, the CNN baseline struggles with long-range dependencies and displays higher confusion between semantically adjacent clause types.

Future work may include:

- hierarchical document encoders,
- contrastive metric learning (e.g., Siamese networks),
- domain-adaptive embeddings,
- transformer-free attention hybrids.

14. References

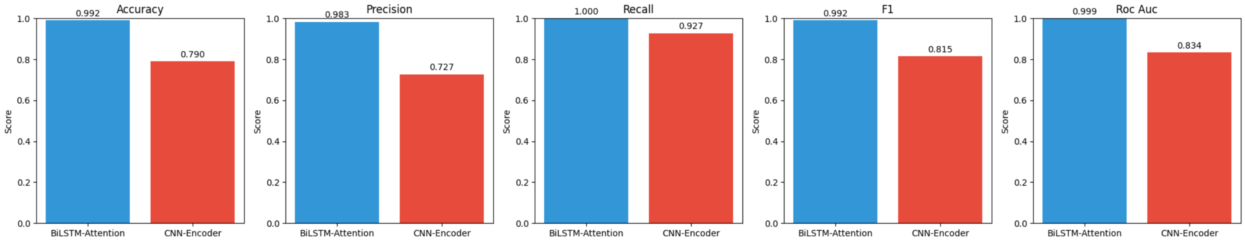
1. Jurafsky, D. & Martin, J.H. *Speech and Language Processing*. Pearson Education.
2. Goldberg, Y. *Neural Network Methods for Natural Language Processing*.
3. Scikit-Learn Documentation: Evaluation Metrics.
4. Kaggle: Legal Clause Dataset by Bahushruth.

Appendix A — Hyperparameters

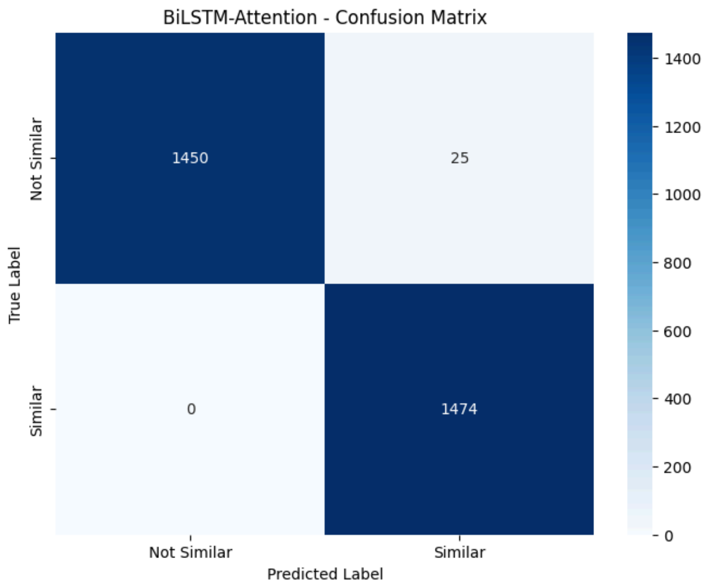
Parameter	Value

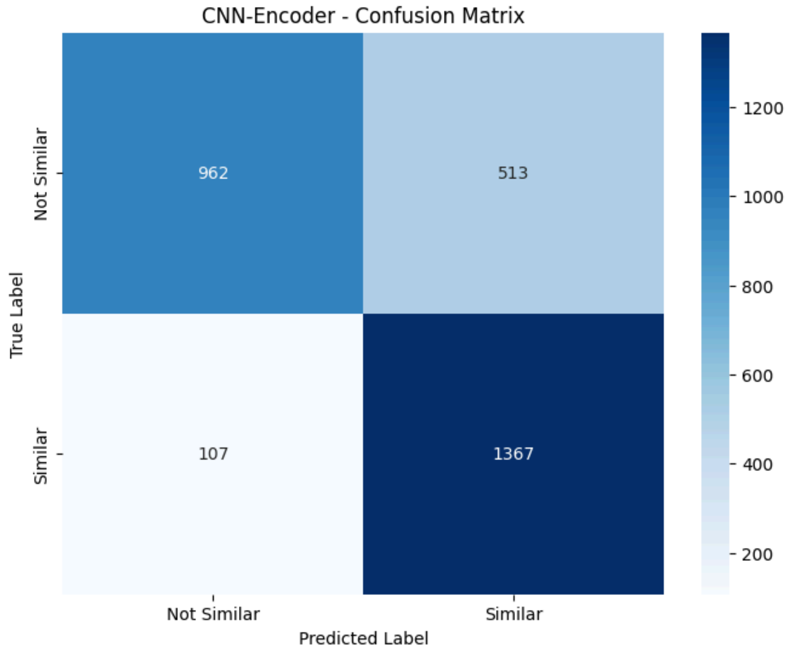
Embedding dimension	100
Hidden dimension (LSTM)	128
Convolution filters	128
Filter sizes	3, 4, 5
Dropout	0.3
Optimizer	Adam

Appendix B — Model Performance Comparison (Graphical)



Appendix C — Confusion Matrices





Appendix D — Sample Predictions

```
================================================================================
======

SAMPLE CORRECT PREDICTIONS
```

```
==========================================================================
======


Clause 1: Parties in Interest. This Agreement shall be binding upon and
shall inure to the benefit of the parties hereto and their respective
heirs, executors, ...

Clause 2: Parties in Interest. All grants, covenants and agreements
contained in this Amendment and Restatement shall bind and inure to the
benefit of the parti...

True Label: 1, Predicted: 1

--------------------------------------------------------------------------
------


Clause 1: Compliance. Merchant agrees to comply with all Debit Network
rules, regulations, procedures, fees, assessments, penalties, and other
membership duties...

Clause 2: Compliance. 18 A. COUNTY’s Health Care Agency (HCA) has
established a Compliance Program for the purpose 19 of ensuring adherence
to all rules and reg...

True Label: 1, Predicted: 1

--------------------------------------------------------------------------
------


Clause 1: Bank Accounts. The bank accounts of the Partnership shall be
maintained in such banking institutions as the Managing Partner shall
determine, and with...

Clause 2: Title to Assets. The Acquired Corporations own, and have good,
valid and marketable title to, or in the case of leased properties and
assets, valid le...

True Label: 0, Predicted: 0

--------------------------------------------------------------------------
------
```

```
==========================================================================
======

SAMPLE INCORRECT PREDICTIONS

==========================================================================
======



Clause 1: No Assignment. This Agreement shall not be assignable by the
Executive. This Agreement shall be assignable by the Company only by
merger or with all o...

Clause 2: Parties in Interest. This Note shall bind Maker and his
successors and assigns. This Note shall not be assigned or transferred by
Payee without the ex...

True Label: 0, Predicted: 1

--------------------------------------------------------------------------
------



Clause 1: Representations. (i) The introductory clause of Section 3 of
this Agreement is hereby amended to read in its entirety as follows: "Each
party represen...

Clause 2: Base Salary. The Executive shall receive a base salary at the
annual rate of $250,000.00 (the "Base Salary") during the Term of
Employment, with such ...

True Label: 0, Predicted: 1

--------------------------------------------------------------------------
------



Clause 1: Time of the Essence. Time is of the essence in this Amending
Agreement....

Clause 2: Representations. Each Party represents (which representations
will each be deemed to be a representation for purposes of Section
5(a)(iv) of the Gover...

True Label: 0, Predicted: 1
```

