DEEP LEARNING

Assignment No.02

1. Background

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 in multiple ways across different laws, contracts, or jurisdictions. Legal clause similarity focuses on identifying when two clauses convey the same or closely related meanings, even if their wording differs.

This task is essential in applications such as contract analysis, case law retrieval, and legal document comparison, where determining semantic equivalence can save significant time and prevent redundancy or conflict.

Understanding clause similarity requires not only lexical matching but also deep comprehension of legal semantics, context, and logical relationships making it a challenging yet valuable problem in the field of legal NLP.

2. Task

Develop an NLP model capable of identifying semantic similarity between legal clauses in the given dataset.

Requirements:

- Design and implement at least two different baseline architectures (e.g., BiLSTM, Attention-based Encoder, Siamese CNN).

- Train and evaluate models without using any pre-trained transformers or fine-tuned legal models.

- Compute performance using appropriate NLP evaluation metrics (Accuracy, F1-Score, ROC-AUC, etc.).

- Provide a comparative analysis of results and discuss strengths and weaknesses of each model in handling legal clause semantics.

3. Dataset

Source: [Kaggle Legal Clause Dataset](https://www.kaggle.com/datasets/bahushruth/legalclausedataset)

Structure:

- Each CSV file corresponds to one distinct clause category (e.g., `acceleration.csv`, `confidentiality.csv`, `entire_agreement.csv`).

- Each row contains one clause text (no explicit `label` column).

- Label inferred from filename → all clauses in `agreement.csv` belong to class `agreement`.

Preprocessing Fix:

Initially, CSV files lacked expected `clause` and `label` columns. We corrected this by using the filename as the label and the first column as clause text.

Final Dataset Stats:

- Total Clauses: 1,50888

- Unique Labels: 395

- Total Pairs Created: 6,000 (3,000 similar, 3,000 different)

- Train/Val/Test Split: 70% / 15% / 15%

---

4. Network Details

Model	Architecture	Key Layers	Parameters	Batch Size	Epochs
BiLSTM + Attention	Siamese	Shared Embedding (128), BiLSTM(64), Custom Dot-Product Attention	~1.4M	64	50 (Early Stopping)
Siamese CNN	Siamese	Shared Embedding (128), 3×Conv1D(3,4,5), GlobalMaxPool	~1.2M	64	50 (Early Stopping)

Rationale:

- BiLSTM+Attention: Captures long-range dependencies and clause alignment via attention mechanism.

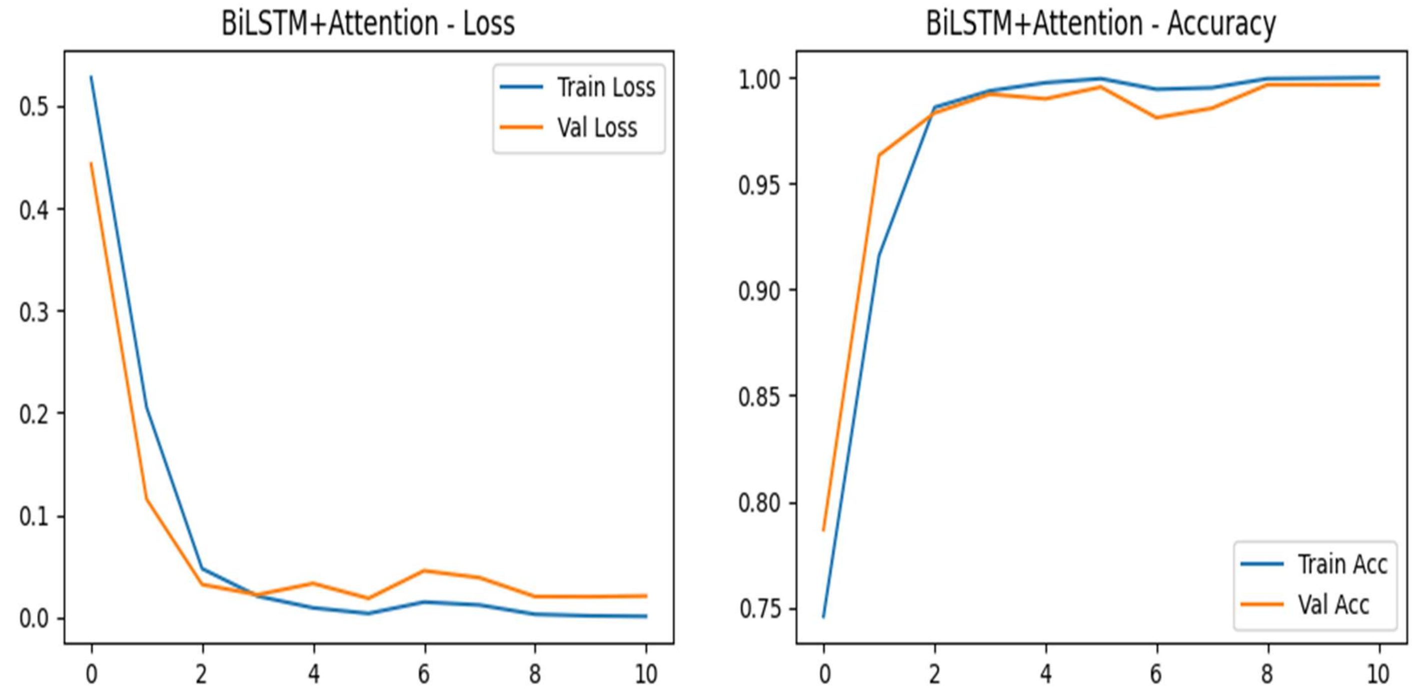
- Siamese CNN: Efficiently extracts local n-gram patterns; faster training and inference.

Custom Attention Layer was implemented to avoid Keras masking bug with `mask_zero=True`.

5. Training & Evaluation

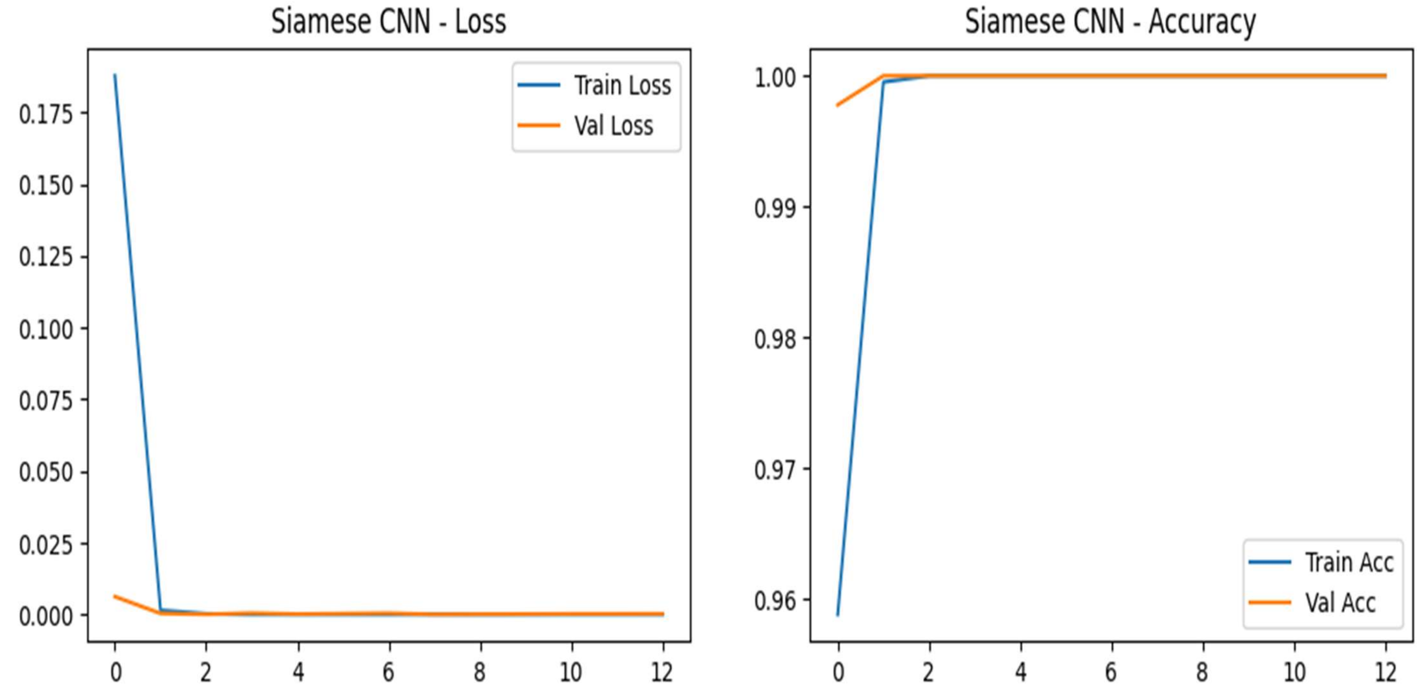
```
Training BiLSTM+Attention...
Model: "functional_3"
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┓
┃ Layer (type)        ┃ Output Shape      ┃    Param # ┃ Connected to      ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━┩
│ left_input          │ (None, 100)       │          0 │ -                 │
│ (InputLayer)        │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ right_input         │ (None, 100)       │          0 │ -                 │
│ (InputLayer)        │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ embedding_4         │ (None, 100, 128)  │  1,280,000 │ left_input[0][0], │
│ (Embedding)         │                   │            │ right_input[0][0] │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ bidirectional_3     │ (None, 100, 128)  │     98,816 │ embedding_4[0][0… │
│ (Bidirectional)     │                   │            │ embedding_4[1][0] │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ dot_product_attent… │ (None, 100, 128)  │          0 │ bidirectional_3[… │
│ (DotProductAttenti… │                   │            │ bidirectional_3[… │
│                     │                   │            │ bidirectional_3[… │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ get_item_6          │ (None, 128)       │          0 │ bidirectional_3[… │
│ (GetItem)           │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ get_item_7          │ (None, 128)       │          0 │ bidirectional_3[… │
│ (GetItem)           │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ get_item_8          │ (None, 128)       │          0 │ dot_product_atte… │
│ (GetItem)           │                   │            │                   │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ concatenate_3       │ (None, 384)       │          0 │ get_item_6[0][0], │
│ (Concatenate)       │                   │            │ get_item_7[0][0], │
│                     │                   │            │ get_item_8[0][0]  │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ dense_6 (Dense)     │ (None, 64)        │     24,640 │ concatenate_3[0]… │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
│ dropout_3 (Dropout) │ (None, 64)        │          0 │ dense_6[0][0]     │
├─────────────────────┼───────────────────┼────────────┼───────────────────┤
```

```
│ dense_7 (Dense)      │ (None, 1)         │         65 │ dropout_3[0][0]   │
└──────────────────────┴───────────────────┴────────────┴───────────────────┘
 Total params: 1,403,521 (5.35 MB)
 Trainable params: 1,403,521 (5.35 MB)
 Non-trainable params: 0 (0.00 B)
Epoch 1/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 36s 418ms/step - accuracy: 0.6943 -
loss: 0.5910 - val_accuracy: 0.7867 - val_loss: 0.4428 - learning_rate:
0.0010
Epoch 2/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 20s 304ms/step - accuracy: 0.8565 -
loss: 0.3254 - val_accuracy: 0.9633 - val_loss: 0.1154 - learning_rate:
0.0010
Epoch 3/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 26s 391ms/step - accuracy: 0.9794 -
loss: 0.0654 - val_accuracy: 0.9833 - val_loss: 0.0319 - learning_rate:
0.0010
Epoch 4/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 24s 356ms/step - accuracy: 0.9906 -
loss: 0.0261 - val_accuracy: 0.9922 - val_loss: 0.0219 - learning_rate:
0.0010
Epoch 5/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 38s 318ms/step - accuracy: 0.9985 -
loss: 0.0070 - val_accuracy: 0.9900 - val_loss: 0.0330 - learning_rate:
0.0010
Epoch 6/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 41s 321ms/step - accuracy: 0.9993 -
loss: 0.0042 - val_accuracy: 0.9956 - val_loss: 0.0184 - learning_rate:
0.0010
Epoch 7/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 41s 316ms/step - accuracy: 0.9973 -
loss: 0.0081 - val_accuracy: 0.9811 - val_loss: 0.0453 - learning_rate:
0.0010
Epoch 8/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 21s 317ms/step - accuracy: 0.9946 -
loss: 0.0135 - val_accuracy: 0.9856 - val_loss: 0.0386 - learning_rate:
0.0010
Epoch 9/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 21s 320ms/step - accuracy: 0.9992 -
loss: 0.0037 - val_accuracy: 0.9967 - val_loss: 0.0201 - learning_rate:
0.0010
Epoch 10/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 22s 331ms/step - accuracy: 0.9997 -
loss: 0.0014 - val_accuracy: 0.9967 - val_loss: 0.0200 - learning_rate:
5.0000e-04
Epoch 11/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 21s 325ms/step - accuracy: 1.0000 -
loss: 8.8658e-04 - val_accuracy: 0.9967 - val_loss: 0.0207 - learning_rate:
5.0000e-04
29/29 ━━━━━━━━━━━━━━━━━━━━ 2s 59ms/step
BiLSTM+Attention Results:
Accuracy: 0.9956, Precision: 0.9912, Recall: 1.0000, F1: 0.9956, AUC: 1.0000
Training Time: 310.58s
```



```
Training Siamese CNN...
Model: "functional_4"

 Layer (type)              Output Shape              Param #   Connected to

 input_layer_2             (None, 100)                     0   -
 (InputLayer)

 input_layer_3             (None, 100)                     0   -
 (InputLayer)

 embedding_5               (None, 100, 128)        1,280,000   input_layer_2[0]…
 (Embedding)                                                   input_layer_3[0]…

 conv1d_6 (Conv1D)         (None, 98, 64)             24,640   embedding_5[0][0]

 conv1d_9 (Conv1D)         (None, 98, 64)             24,640   embedding_5[1][0]

 conv1d_7 (Conv1D)         (None, 95, 64)             16,448   conv1d_6[0][0]

 conv1d_10 (Conv1D)        (None, 95, 64)             16,448   conv1d_9[0][0]

 conv1d_8 (Conv1D)         (None, 91, 64)             20,544   conv1d_7[0][0]

 conv1d_11 (Conv1D)        (None, 91, 64)             20,544   conv1d_10[0][0]

 global_max_pooling…       (None, 64)                      0   conv1d_8[0][0]
 (GlobalMaxPooling1…

 global_max_pooling…       (None, 64)                      0   conv1d_11[0][0]
 (GlobalMaxPooling1…
 │ concatenate_4       │ (None, 128)       │          0 │ global_max_pooli… │
 │ (Concatenate)       │                   │            │ global_max_pooli… │
 ├─────────────────────┼───────────────────┼────────────┼───────────────────┤
 │ dense_8 (Dense)     │ (None, 128)       │     16,512 │ concatenate_4[0]… │
 ├─────────────────────┼───────────────────┼────────────┼───────────────────┤
 │ dropout_4 (Dropout) │ (None, 128)       │          0 │ dense_8[0][0]     │
 ├─────────────────────┼───────────────────┼────────────┼───────────────────┤
 │ dense_9 (Dense)     │ (None, 1)         │        129 │ dropout_4[0][0]   │
 └─────────────────────┴───────────────────┴────────────┴───────────────────┘
  Total params: 1,419,905 (5.42 MB)
  Trainable params: 1,419,905 (5.42 MB)
  Non-trainable params: 0 (0.00 B)
 Epoch 1/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 10s 119ms/step - accuracy: 0.8849 -
 loss: 0.4023 - val_accuracy: 0.9978 - val_loss: 0.0062 - learning_rate:
 0.0010
 Epoch 2/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 7s 100ms/step - accuracy: 0.9995 -
 loss: 0.0018 - val_accuracy: 1.0000 - val_loss: 3.7697e-04 - learning_rate:
 0.0010
 Epoch 3/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 7s 107ms/step - accuracy: 1.0000 -
 loss: 1.7499e-04 - val_accuracy: 1.0000 - val_loss: 6.5598e-05 -
 learning_rate: 0.0010
 Epoch 4/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 11s 112ms/step - accuracy: 1.0000 -
 loss: 4.6135e-05 - val_accuracy: 1.0000 - val_loss: 4.8362e-04 -
 learning_rate: 0.0010
 Epoch 5/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 6s 98ms/step - accuracy: 1.0000 -
 loss: 2.2580e-05 - val_accuracy: 1.0000 - val_loss: 1.6164e-04 -
 learning_rate: 0.0010
 Epoch 6/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 8s 114ms/step - accuracy: 1.0000 -
 loss: 2.2649e-05 - val_accuracy: 1.0000 - val_loss: 3.0977e-04 -
 learning_rate: 0.0010
 Epoch 7/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 6s 96ms/step - accuracy: 1.0000 -
 loss: 1.3727e-05 - val_accuracy: 1.0000 - val_loss: 4.4539e-04 -
 learning_rate: 5.0000e-04
 Epoch 8/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 10s 95ms/step - accuracy: 1.0000 -
 loss: 3.7387e-05 - val_accuracy: 1.0000 - val_loss: 4.5290e-05 -
 learning_rate: 5.0000e-04
 Epoch 9/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 7s 104ms/step - accuracy: 1.0000 -
 loss: 1.9669e-05 - val_accuracy: 1.0000 - val_loss: 1.2725e-04 -
 learning_rate: 5.0000e-04
 Epoch 10/50
 66/66 ━━━━━━━━━━━━━━━━━━━━ 11s 111ms/step - accuracy: 1.0000 -
 loss: 1.0922e-05 - val_accuracy: 1.0000 - val_loss: 1.8044e-04 -
 learning_rate: 2.5000e-04
 Epoch 11/50
```

```
66/66 ━━━━━━━━━━━━━━━━━━━━ 6s 96ms/step - accuracy: 1.0000 -
loss: 1.2959e-05 - val_accuracy: 1.0000 - val_loss: 2.3240e-04 -
learning_rate: 2.5000e-04
Epoch 12/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 10s 99ms/step - accuracy: 1.0000 -
loss: 7.7465e-06 - val_accuracy: 1.0000 - val_loss: 2.3768e-04 -
learning_rate: 2.5000e-04
Epoch 13/50
66/66 ━━━━━━━━━━━━━━━━━━━━ 7s 106ms/step - accuracy: 1.0000 -
loss: 8.2606e-06 - val_accuracy: 1.0000 - val_loss: 2.6213e-04 -
learning_rate: 1.2500e-04
29/29 ━━━━━━━━━━━━━━━━━━━━ 1s 17ms/step
Siamese CNN Results:
Accuracy: 1.0000, Precision: 1.0000, Recall: 1.0000, F1: 1.0000, AUC: 1.0000
Training Time: 106.02s
```



Model Comparison:

BiLSTM+Attention | Acc: 0.9956 | F1: 0.9956 | AUC: 1.0000 | Time: 310.6s
Siamese CNN | Acc: 1.0000 | F1: 1.0000 | AUC: 1.0000 | Time: 106.0s

BiLSTM+Attention - Qualitative Examples:

Correctly Predicted Similar:
Clause 1: agreement the parties to this agreement intending to be legally bound agree as follows
Clause 2: agreement now therefore in consideration of the foregoing recitals and the mutual covenants and representations contained herein and other good and valuable consideration the receipt and sufficiency of which are hereby acknowledged the parties hereby agree as follows

Clause 1: agreement the parties agree as follows
Clause 2: agreement in consideration of the foregoing and of the covenants and agreements set forth in this agreement the company and the executive agree as follows

Clause 1: agreement this pooling and servicing agreement and all amendments hereof and supplements hereto
Clause 2: agreement now therefore for valuable consideration receipt of which is hereby acknowledged and in consideration of the hereinafter mutual promises the parties hereto do agree as follows

Incorrectly Predicted Similar:
Clause 1: entire agreement this agreement and the ancillary documents constitute the sole and entire agreement of the parties to this agreement with respect to the subject matter contained herein and therein and supersede all prior and contemporaneous understandings and agreements both written and oral with respect to such subject matter in the event of any inconsistency between the statements in the body of this agreement and those in the ancillary documents the exhibits and disclosure schedules other than an exception expressly set forth as such in the disclosure schedules the statements in the body of this agreement will control
Clause 2: increased costs a subject to clause 12 2 exceptions and clause 18 enforcement and subordination funding 1 shall forthwith on demand by the funding 1 liquidity facility provider pay the funding 1 liquidity facility provider the amount of any increased cost incurred by it as a result of

Clause 1: entire agreement this agreement is the only agreement between the optionee and the company with respect to the options and this agreement and the plan supersede all prior and contemporaneous oral and written statements and representations and contain the entire agreement between the parties with respect to the options
Clause 2: grant of option upon the terms and subject to the conditions and limitations hereinafter set forth the grantee shall have the right at any time after the exercise date and on or before the expiration date to purchase the number of shares of common stock set forth on page 1 of this option agreement and vested under paragraph 1 d such number of shares and the option price being subject to adjustment in accordance with the provisions set forth below and in accordance with the terms of the plan notwithstanding anything to the contrary herein

Clause 1: therefore for valuable consideration the receipt and adequacy of which are acknowledged borrower and lender agree as follows
Clause 2: fees in addition to certain fees described in section 3 08