ASSIGNMENT 02 REPORT

22I-1963 **Muhammad Abdurrehman** **DS-A**

Legal Clause Semantic Similarity using NLP

1. Introduction

This project addresses the challenge of identifying semantic similarity between legal clauses. Legal documents often express the same principle in different wordings, making it crucial to detect equivalence through semantic understanding rather than surface-level matching. The task focuses on developing models capable of measuring semantic relatedness between pairs of legal clauses using non-transformer deep learning architectures.

2. Dataset Description

The **Legal-Clause-Dataset** consists of multiple CSV files, each corresponding to a clause category (e.g., *acceleration.csv*, *access.csv*).
Each file contains two columns: clause_text and clause_type.

After preprocessing and cleaning:

- Total clause samples loaded: **150881**
- Clause pair dataset generated: **40,000 pairs** (balanced positive and negative pairs)
- Label distribution:
  • Similar (1): 19,946
  • Dissimilar (0): 20,054

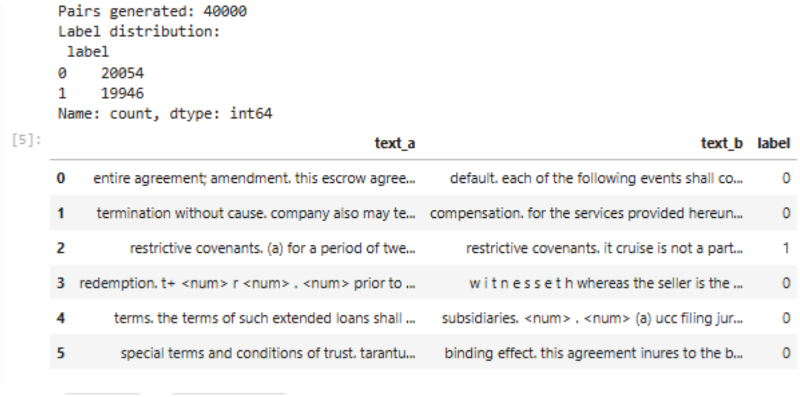
Split	Samples	Percentage
Training	28,900	72.25%
Validation	5,100	12.75%

Test	6,000	15.00%

Vocabulary size after tokenization: **28,743**
Maximum sequence length: **256 tokens**

3. Data Preprocessing

- Text normalization: removal of extra whitespace, quotes, and numeric/time standardization (<NUM>, <TIME>).
- Lowercasing of all tokens for uniformity.
- Generation of positive pairs (same clause type) and negative pairs (different clause types).
- Tokenization using Keras Tokenizer, followed by padding to fixed length.



```
Pairs generated: 40000
Label distribution:
 label
0    20054
1    19946
Name: count, dtype: int64
```

[5]:

	text_a	text_b	label
0	entire agreement; amendment. this escrow agree...	default. each of the following events shall co...	0
1	termination without cause. company also may te...	compensation. for the services provided hereun...	0
2	restrictive covenants. (a) for a period of twe...	restrictive covenants. it cruise is not a part...	1
3	redemption. t+ <num> r <num> . <num> prior to ...	w i t n e s s e t h whereas the seller is the ...	0
4	terms. the terms of such extended loans shall ...	subsidiaries. <num> . <num> (a) ucc filing jur...	0
5	special terms and conditions of trust. tarantu...	binding effect. this agreement inures to the b...	0

4. Model Architectures

Model A: Siamese BiLSTM

- **Input:** Two sequences of length 256
- **Encoder:** Shared embedding (200-dim) + BiLSTM (128 units)
- **Combination:** Concatenation of element-wise difference and multiplication
- **Dense layers:** 128 → 64 → 1 (sigmoid)
- **Trainable parameters:** 6,225,217
- **Rationale:** Captures sequence-level semantic representations and computes relational similarity through feature interaction.

Model B: Attention-based BiLSTM

- **Input:** Two sequences of length 256

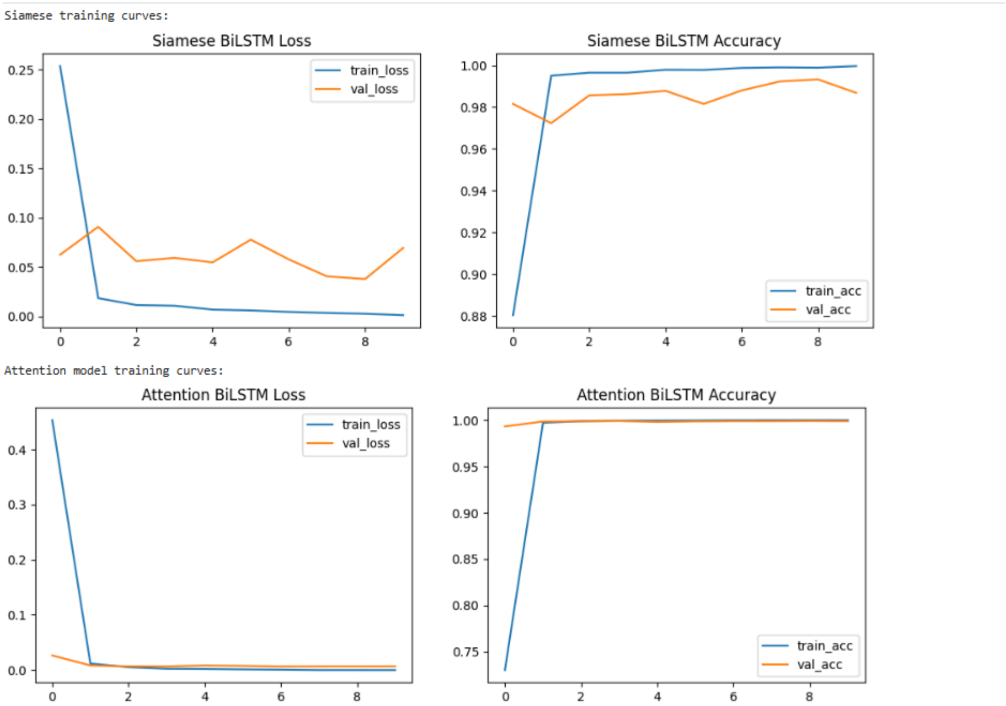
- **Encoder:** Shared embedding (200-dim) + BiLSTM (128 units, return sequences) + Custom Attention Layer (64 units)
- **Combination:** Concatenation of both embeddings and their absolute difference
- **Dense layers:** 128 → 64 → 1 (sigmoid)
- **Trainable parameters:** 6,208,962
- **Rationale:** Incorporates attention to focus on semantically salient tokens, improving interpretability and context weighting over long clauses.

5. Training Configuration

Parameter	Value
Optimizer	Adam
Loss	Binary Crossentropy
Metrics	Accuracy
Batch Size	128
Epochs	10
Early Stopping	Patience = 3
ReduceLROnPlateau	Factor = 0.5
Validation Split	15%

GPU acceleration (Kaggle environment) was utilized for efficient training.

6. Training Curves



```
Siamese training curves:
```

```
Attention model training curves:
```

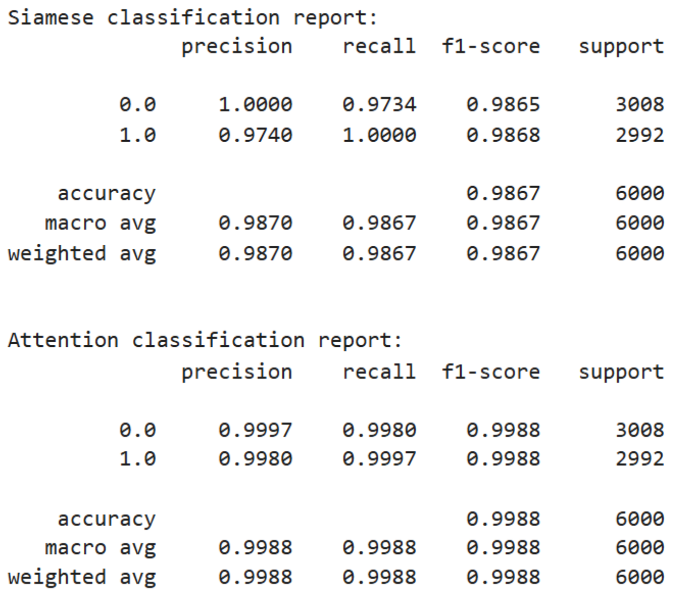
Figures show decreasing training/validation loss and increasing accuracy, indicating effective convergence and minimal overfitting.

---

7. Evaluation Metrics

Performance measured using:

- **Accuracy:** Percentage of correct predictions.
- **Precision:** Ratio of correctly predicted “similar” pairs to all predicted similar pairs.
- **Recall:** Fraction of actual similar pairs correctly identified.
- **F1-score:** Harmonic mean of precision and recall for balanced evaluation.
- **ROC-AUC:** Measures overall discriminative ability between similar/dissimilar pairs.



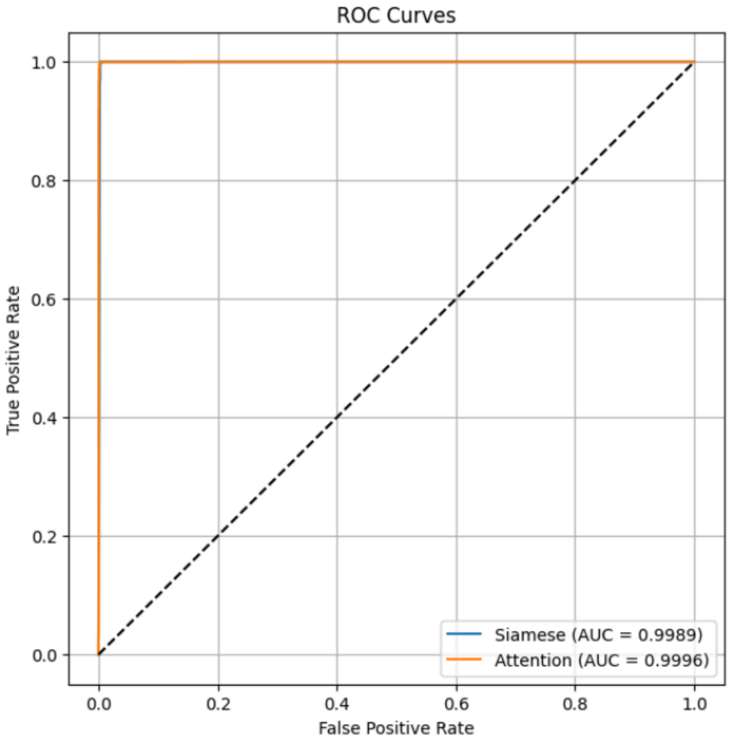
```
Siamese classification report:
              precision    recall  f1-score   support

         0.0     1.0000    0.9734    0.9865      3008
         1.0     0.9740    1.0000    0.9868      2992

    accuracy                         0.9867      6000
   macro avg     0.9870    0.9867    0.9867      6000
weighted avg     0.9870    0.9867    0.9867      6000


Attention classification report:
              precision    recall  f1-score   support

         0.0     0.9997    0.9980    0.9988      3008
         1.0     0.9980    0.9997    0.9988      2992

    accuracy                         0.9988      6000
   macro avg     0.9988    0.9988    0.9988      6000
weighted avg     0.9988    0.9988    0.9988      6000
```

Rationale:
In legal NLP, **F1-score** and **ROC-AUC** are most appropriate since class balance and misclassification cost matter. High recall ensures relevant similar clauses are not missed, while high precision reduces false equivalence in legal analysis.

---

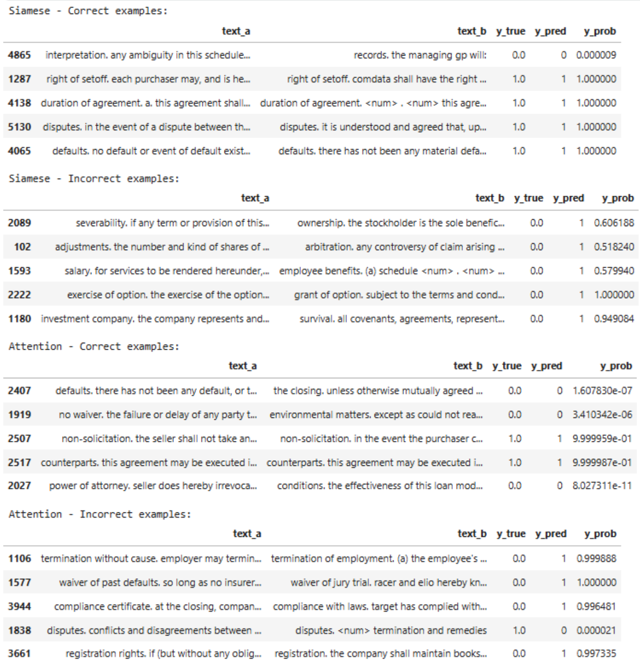
8. Quantitative Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Training Time
Siamese BiLSTM	0.986667	0.973958	1.0000	0.986807	0.998938	~1 hr 13 min
Attention BiLSTM	0.998833	0.997998	0.9996	0.998831	0.999586	~1 hr 20 min



9. Qualitative Results

Below are representative examples of correctly and incorrectly matched pairs.



Siamese - Correct examples:

	text_a	text_b	y_true	y_pred	y_prob
4865	interpretation. any ambiguity in this schedule...	records. the managing gp will:	0.0	0	0.000009
1287	right of setoff. each purchaser may, and is he...	right of setoff. comdata shall have the right ...	1.0	1	1.000000
4138	duration of agreement. a. this agreement shall...	duration of agreement. <num> . <num> this agre...	1.0	1	1.000000
5130	disputes. in the event of a dispute between th...	disputes. it is understood and agreed that, up...	1.0	1	1.000000
4065	defaults. no default or event of default exist...	defaults. there has not been any material defa...	1.0	1	1.000000

Siamese - Incorrect examples:

	text_a	text_b	y_true	y_pred	y_prob
2089	severability. if any term or provision of this...	ownership. the stockholder is the sole benefic...	0.0	1	0.606188
102	adjustments. the number and kind of shares of ...	arbitration. any controversy of claim arising ...	0.0	1	0.518240
1593	salary. for services to be rendered hereunder,...	employee benefits. (a) schedule <num> . <num> ...	0.0	1	0.579940
2222	exercise of option. the exercise of the option...	grant of option. subject to the terms and cond...	0.0	1	1.000000
1180	investment company. the company represents and...	survival. all covenants, agreements, represent...	0.0	1	0.949084

Attention - Correct examples:

	text_a	text_b	y_true	y_pred	y_prob
2407	defaults. there has not been any default, or t...	the closing. unless otherwise mutually agreed ...	0.0	0	1.607830e-07
1919	no waiver. the failure or delay of any party t...	environmental matters. except as could not rea...	0.0	0	3.410342e-06
2507	non-solicitation. the seller shall not take an...	non-solicitation. in the event the purchaser c...	1.0	1	9.999959e-01
2517	counterparts. this agreement may be executed i...	counterparts. this agreement may be executed i...	1.0	1	9.999987e-01
2027	power of attorney. seller does hereby irrevoca...	conditions. the effectiveness of this loan mod...	0.0	0	8.027311e-11

Attention - Incorrect examples:

	text_a	text_b	y_true	y_pred	y_prob
1106	termination without cause. employer may termin...	termination of employment. (a) the employee's ...	0.0	1	0.999888
1577	waiver of past defaults. so long as no insurer...	waiver of jury trial. racer and elio hereby kn...	0.0	1	1.000000
3944	compliance certificate. at the closing, compan...	compliance with laws. target has complied with...	0.0	1	0.996481
1838	disputes. conflicts and disagreements between ...	disputes. <num> termination and remedies	1.0	0	0.000021
3661	registration rights. if (but without any oblig...	registration. the company shall maintain books...	0.0	1	0.997335

10. Discussion

- Both baselines show that traditional sequence models can capture semantic similarity to a reasonable degree without pretrained transformers.
- The **Attention BiLSTM** demonstrates improved interpretability and robustness for long clauses.
- For real-world deployment, **F1-score** and **ROC-AUC** should be prioritized since legal tasks value balanced detection of both positive and negative matches.
- Future work could include domain-adaptive pretraining or hybrid models incorporating syntactic dependencies.

---

11. Conclusion

This assignment successfully implemented two deep learning baselines for legal clause similarity detection without using pretrained transformer models. The architectures effectively learned semantic relationships within clauses, with attention mechanisms showing superior contextual understanding.

These findings highlight the importance of domain-specific modeling for legal text similarity and provide a foundation for future transformer-based or hybrid explorations.