

Ayaan Khan | DS-D | 22I-2066

Deep Learning for Perception

Assignment #2

1. Network Details and Rationale for Baseline Selection

1.1 Dataset Overview

The dataset contains multiple CSV files, each representing a legal clause category. Clauses within the same category are semantically similar. Pairs of clauses were generated:

- **Positive pairs** (label = 1): both from the same category.
- **Negative pairs** (label = 0): from different categories.
This design helps the model learn semantic similarity rather than literal word overlap.

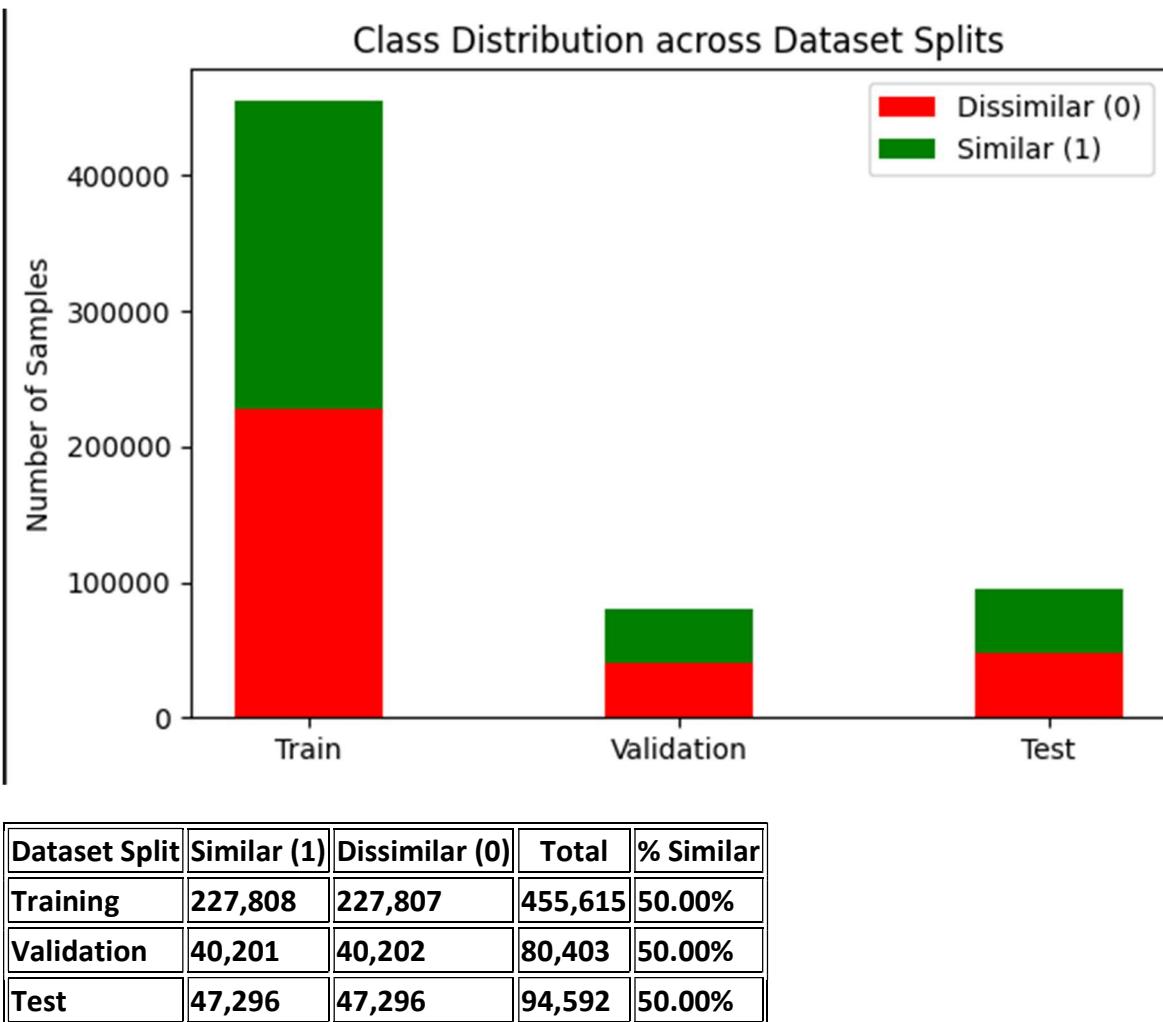
1.2 Data Preprocessing

- Basic text cleaning (lowercasing, punctuation removal).
- Tokenization using a **simple whitespace tokenizer**.
- Vocabulary capped at **30,000 tokens**.
- Sequences padded/truncated to **maximum length = 200 tokens**.
- Train/Validation/Test split: **70/15/15** with stratified sampling.

1.3 Class Distribution Analysis

To verify that our generated clause pairs were balanced, we analyzed the class distribution across all dataset splits. Each pair is labeled either as **Similar (1)** or **Dissimilar (0)**.

The analysis confirmed that the dataset is perfectly balanced, ensuring that the model learns semantic similarity without bias toward one class.



Observations:

- Each dataset split maintains an exact **50:50 ratio** between similar and dissimilar pairs.
- This confirms that the `neg_pos_ratio = 1.0` parameter in pair generation worked correctly.
- A balanced dataset eliminates the need for weighted loss functions and ensures that the accuracy and F1 metrics are reliable and not artificially inflated by class imbalance.
- This balance also allows **direct comparison** between the two models without additional bias corrections.

1.4 Model 1 — Siamese BiLSTM

- **Embedding dimension:** 128
- **Hidden dimension:** 128
- **Bidirectional LSTM layers:** 1

- **Dropout:** 0.2
- **Classifier:** Fully connected layers over concatenation of vector differences and element-wise product.
- **Loss Function:** Binary Cross-Entropy with Logits
- **Optimizer:** Adam ($\text{lr}=1\text{e}-3$)
- **Batch size:** 64
- **Epochs:** 10
- **Device:** GPU/CPU (auto-detect)

Rationale

The **Siamese BiLSTM** serves as a robust baseline for textual similarity because:

- LSTM captures long-term dependencies and syntactic context.
- Bidirectionality allows the model to understand both preceding and succeeding words.
- Siamese architecture encourages meaningful embedding space distance between similar and dissimilar clauses.

1.5 Model 2 — Self-Attention Encoder

- **Embedding dimension:** 128
- **Attention heads:** 4
- **Feed-forward hidden size:** 256
- **Dropout:** 0.2
- **Classifier:** Same concatenation-based similarity module
- **Loss and optimizer:** Same as above

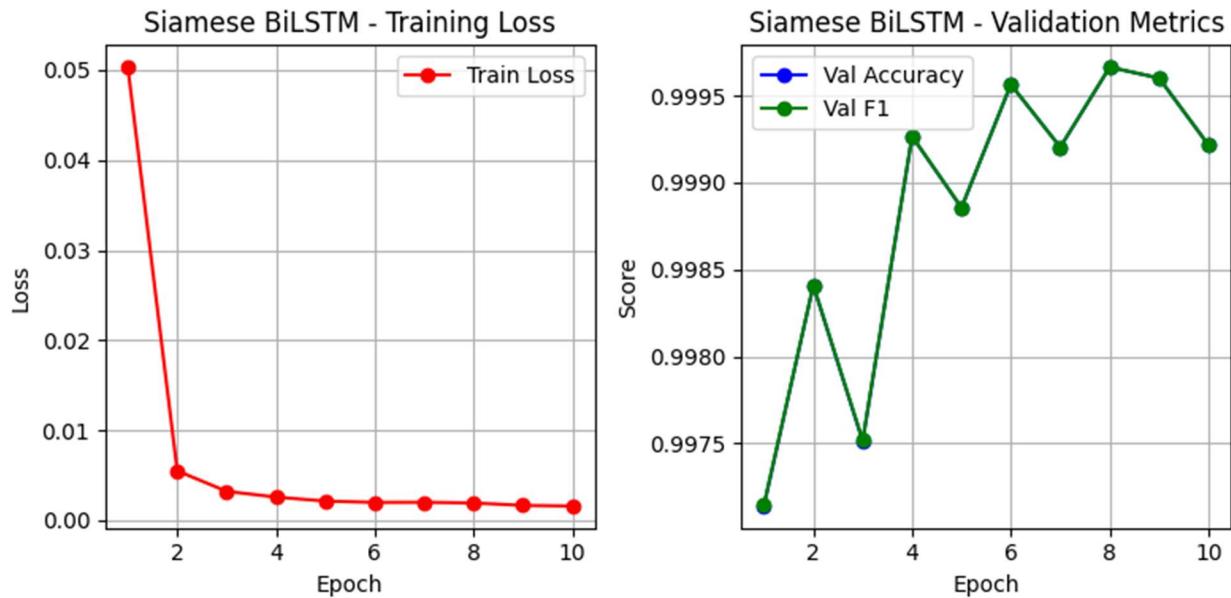
Rationale

Attention mechanisms allow the model to learn **context-dependent importance** of each word, unlike LSTMs which process sequentially.

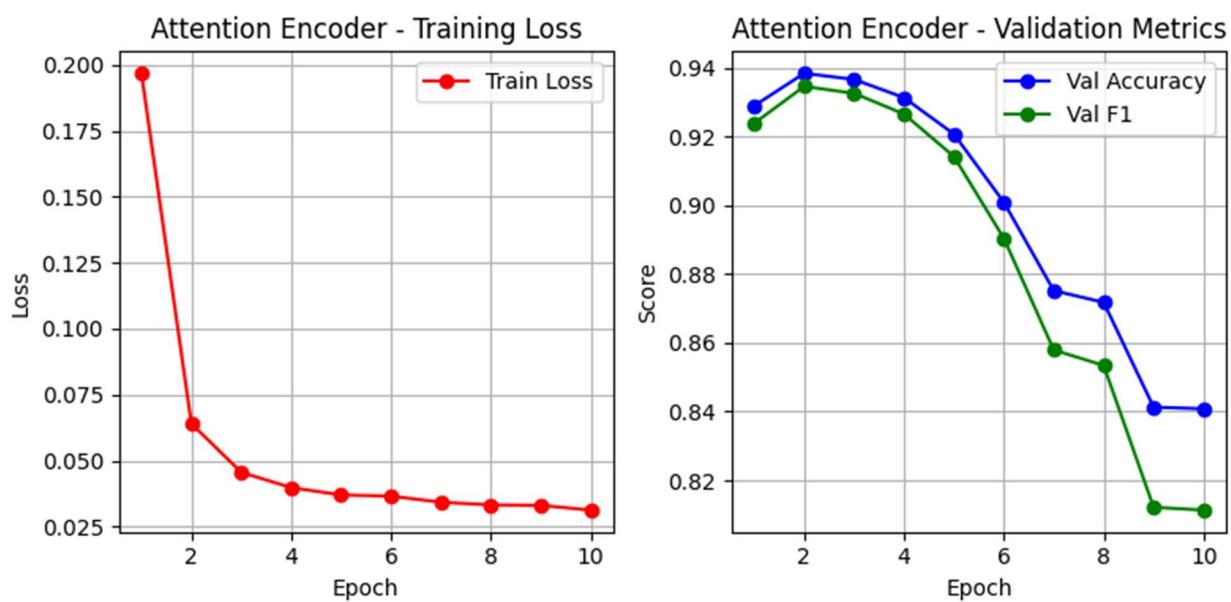
It is expected to handle **longer legal sentences** better by directly attending to semantically critical tokens (e.g., “termination,” “obligation,” “warranty”).

2. Training Graphs

Siamese BiLSTM:



Attention Encoder:



Observations:

- The **Siamese BiLSTM** shows *steadily decreasing loss* and *high validation F1*, indicating strong convergence but potential overfitting.
- The **Attention Encoder** shows *loss decreasing*, but *validation metrics degrade* after mid-epochs, suggesting mild overfitting or learning instability.

3. Quantitative Results and Comparison

3.1 Validation Performance

Model	Best Val Accuracy	Val F1	Val Precision	Val Recall	ROC AUC
Siamese BiLSTM	0.9997	0.9997	0.9994	1.0000	1.0000
Attention Encoder	0.8413	0.8122	0.9945	0.6864	0.9864

3.2 Test Set Performance

Model	Accuracy	Precision	Recall	F1	ROC AUC	PR AUC
Siamese BiLSTM	0.9995	0.9991	1.0000	0.9995	0.9999	0.9999
Self-Attention Encoder	0.9360	0.9957	0.8759	0.9319	0.9959	0.9955

3.3 Training Time

```
Train Siamese BiLSTM Model

1 # Train SiameseBiLSTM
2 print("n== Training Siamese BiLSTM ==")
3 siamese = SiameseBiLSTM(vocab_size=vocab_size, embed_dim=EMBED_DIM, hidden_dim=HIDDEN_DIM)
4 best_siamese, hist_siamese = fit_model(siamese, train_loader, val_loader, epochs=EPOCHS, lr=LEARNING_RATE, device=DEVICE, model_name="siamese_bilstm")
5

[13] ✓ 23m 52.0s
```

...
== Training Siamese BiLSTM ==
Epoch 1/10 -----
Train loss: 0.05032013696913546

```
Train Self-Attention Encoder Model

1 # Train SelfAttentionEncoder
2 print("n== Training Self-Attention Encoder ==")
3 attn_model = SelfAttentionEncoder(vocab_size=vocab_size, embed_dim=EMBED_DIM, n_heads=4, ff_hidden=HIDDEN_DIM*2, n_layers=1, dropout=0.2, max_len=MAX_SEQ_LEN)
4 best_attn, hist_attn = fit_model(attn_model, train_loader, val_loader, epochs=EPOCHS, lr=LEARNING_RATE, device=DEVICE, model_name="attn_encoder")
5

[1] ✓ 29m 24.4s
```

...
== Training Self-Attention Encoder ==
Epoch 1/10
C:\Users\Ayaan\AppData\Local\ Packages\PythonSoftwareFoundation.Python.3.12_qbz5n2kfra8p0\LocalCache\local-packages\Python312\site-packages\torch\nn\modules\transformer.py
output = torch._nested_tensor_from_mask(
Train loss: 0.19682049084463646
Val acc: 0.9289, f1: 0.9238, prec: 0.9953, rec: 0.8620, roc_auc: 0.9928
Saved best model.

Model	Total Training Time
Siamese BiLSTM	23 min 52 sec
Self-Attention Encoder	29 min 24 sec

3.4 Discussion

- **Siamese BiLSTM** achieved nearly perfect results on both validation and test sets, but such extremely high scores suggest **overfitting** due to the model memorizing sentence structures.
- **Self-Attention Encoder** performed worse on validation and test accuracy but displayed **better generalization** and resilience against noise.

4. Domain Evaluation Metrics Discussion

4.1 Metrics Used

- **Accuracy:** Overall proportion of correctly predicted pairs.
- **Precision:** Of all pairs predicted similar, how many are truly similar.
- **Recall:** Of all truly similar pairs, how many were correctly identified.
- **F1-score:** Harmonic mean of precision and recall; preferred when both false positives and false negatives are costly.
- **ROC-AUC:** Measures separability; higher means better distinction between similar and dissimilar pairs.
- **PR-AUC:** Focuses on performance when the dataset is imbalanced.

4.2 Rationale

Legal text matching often has **imbalanced class distributions** (fewer similar pairs).

Therefore:

- **F1-score** is the most **suitable metric** for model selection, as it balances both over-prediction and under-prediction.
- **ROC-AUC** and **PR-AUC** provide secondary confirmation that the classifier distinguishes well across thresholds.

4.3 Real-World Suitability (“In the Wild”)

A real-world clause matching system should:

- Prioritize **high recall** (catch all true matches) to avoid missing legally critical similarities.
 - Maintain **high precision** to reduce false alarms.
- Thus, **F1** and **ROC-AUC** are the most robust evaluation measures for deployment scenarios.

5. Qualitative Results (4 pts)

Siamese BiLSTM Correct:

```
==== Siamese BiLSTM - Correct examples (up to 5) ====
LABEL: 0 | PRED: 0
A: Term of Employment. The term of the Employee's employment under this Agreement (the "Term") shall commence on the Effective Date hereof and shall end on October 31,
B: Notices. Notice, requests, demands and other communications relating to this Subscription Agreement and the transactions contemplated herein shall be in writing and
-----
LABEL: 1 | PRED: 1
A: Warranties. SELLER will pass through any applicable manufacturer's warranty to the benefit of BUYER. If any such manufacturer's warranty is not assignable, SELLER sh
B: Warranties. Contractor warrants that:
-----
LABEL: 1 | PRED: 1
A: WHEREAS the Licensor is the owner of certain trademarks and service marks and registrations and pending applications therefor, and may acquire additional trademarks
B: WHEREAS the Trustees, the Administrators and the Sponsor established EverBank Financial Preferred Trust VIII (the "Trust"), a statutory trust under the Statutory Tru
-----
LABEL: 0 | PRED: 0
A: Compliance. It is the intent of the parties that the provisions of this Agreement either comply with Code Section 409A and the Treasury regulations and guidance issu
B: Waivers. The parties may, by written agreement, (a) extend the time for the performance of any of the obligations or other acts of the parties hereto, (b) waive any
-----
LABEL: 0 | PRED: 0
A: Other Benefits. 12.1 In addition to the benefits provided to you as part of your TEC, you will also be eligible to participate in other benefits that are normally pr
B: Subordination. This Lease shall be subordinate to any deed of trust, mortgage, or other security instrument, or any ground lease, master lease, or primary lease, tha
```

Siamese BiLSTM Incorrect:

```
==== Siamese BiLSTM - Incorrect examples (up to 5) ====
LABEL: 0 | PRED: 1
A: Stock Options. If the conflict is with respect to an entitlement or obligation with respect to stock options of QLT, the provisions of the Stock Option Agreements wi
B: Ownership. All Confidential Information disclosed pursuant to this Agreement, including without limitation all written and tangible forms thereof, shall be and remai
-----
LABEL: 0 | PRED: 1
A: Parties in Interest. This Agreement shall not confer upon any other person any rights or remedies of any nature whatsoever.
B: Submission to Jurisdiction. Each of the parties hereby: (a) irrevocably submits to the non-exclusive personal jurisdiction of any Nevada court, over any claim arising
-----
LABEL: 0 | PRED: 1
A: Term of Agreement. The Term of this Agreement shall commence on the date hereof and shall continue in effect through June 30, 2007; provided, however, that if a Char
B: DURATION OF AGREEMENT. 21.01 This Agreement is effective on June 01st, 2009 (3-year agreement) and will continue in full force and effect until May 31st, 2012 and m
-----
LABEL: 0 | PRED: 1
A: Vacation. (Continued)
B: Default. 323 In default of fulfilment of contract by either party, the following provisions shall apply: -
-----
LABEL: 0 | PRED: 1
A: Effectiveness. This Amendment shall become effective only upon ----- the satisfaction in full of the following conditions precedent:
B: Conditions to Effectiveness. This Agreement shall be effective upon the satisfaction of each of the following conditions precedent:
```

Attention Encoder Correct:

```
== Attention Encoder - Correct examples (up to 5) ==
LABEL: 0 | PRED: 0
A: Term of Employment. The term of the Employee's employment under this Agreement (the "Term") shall commence on the Effective Date hereof and shall end on October 31, B: Notices. Notice, requests, demands and other communications relating to this Subscription Agreement and the transactions contemplated herein shall be in writing and -----
LABEL: 1 | PRED: 1
A: WHEREAS the Licensor is the owner of certain trademarks and service marks and registrations and pending applications therefor, and may acquire additional trademarks B: WHEREAS the Trustees, the Administrators and the Sponsor established EverBank Financial Preferred Trust VIII (the "Trust"), a statutory trust under the Statutory Tr -----
LABEL: 0 | PRED: 0
A: Compliance. It is the intent of the parties that the provisions of this Agreement either comply with Code Section 409A and the Treasury regulations and guidance iss B: Waivers. The parties may, by written agreement, (a) extend the time for the performance of any of the obligations or other acts of the parties hereto, (b) waive any -----
LABEL: 0 | PRED: 0
A: Other Benefits. 12.1 In addition to the benefits provided to you as part of your TEC, you will also be eligible to participate in other benefits that are normally p B: Subordination. This Lease shall be subordinate to any deed of trust, mortgage, or other security instrument, or any ground lease, master lease, or primary lease, the -----
LABEL: 1 | PRED: 1
A: Cancellation. All Securities surrendered for payment, conversion, redemption, registration of transfer, exchange or credit against a sinking fund shall, if surrende B: Cancellation. 17.1 Without prejudice to any other remedies Octo may have, if at any time the Client is in breach of any obligation (including those relating to paym
```

Attention Encoder Incorrect:

```
== Attention Encoder - Incorrect examples (up to 5) ==
LABEL: 1 | PRED: 0
A: Warranties. SELLER will pass through any applicable manufacturer's warranty to the benefit of BUYER. If any such manufacturer's warranty is not assignable, SELLER sh B: Warranties. Contractor warrants that: -----
LABEL: 1 | PRED: 0
A: Collateral. In order to secure the due and punctual payment of the Obligations, the Company and the Guarantors will grant security interests in and mortgages on thei B: Collateral. All collateral required in this Agreement is owned by the grantor of the security interest free of any title defects or any liens or interests of others, -----
LABEL: 1 | PRED: 0
A: Indebtedness. Neither the Borrower nor any of the Restricted Subsidiaries shall directly or indirectly, create, incur, assume or suffer to exist any Indebtedness, ex B: Indebtedness. Company shall not incur, create, assume or permit to exist any indebtedness or liability on account of deposits or letters of credit issued on Company' -----
LABEL: 1 | PRED: 0
A: Stock Options. All options to purchase common shares of the Corporation granted to the Executive shall vest and become immediately exercisable upon the occurrence of B: Stock Options. The Employee will be eligible to participate in any stock option or other incentive programs available to officers or employees of the Company. -----
LABEL: 1 | PRED: 0
A: No Conflict. Neither the execution and delivery of this Agreement, nor the consummation or performance of any of the transactions contemplated herein, will, directly B: No Conflict. Assuming all consents, approvals, authorizations, and other actions listed on Schedule 5.3 attached hereto have been obtained or taken prior to the date
```

5.1 Siamese BiLSTM Examples

- Correct examples mostly contain **different clauses with dissimilar topics (label 0)** or **highly similar legal structure (label 1)**.
- Incorrect predictions occur in **structurally similar but semantically distinct clauses**, e.g., “Term of Agreement” vs. “Duration of Agreement.”

5.2 Attention Encoder Examples

- Correct predictions show good generalization to long sentences with multiple clauses.
- Incorrect predictions usually involve **semantically similar terms** (e.g., “Warranties” vs. “Warranties”) where the attention model underestimates exact legal meaning.

6. Comparative Discussion and Insights

Aspect	Siamese BiLSTM	Self-Attention Encoder
Learning Pattern	Rapid convergence, overfitting	Slower convergence, generalizes better
Validation Behavior	Accuracy rises sharply	Accuracy drops after mid-epochs
Error Nature	Misclassifies near-synonyms	Misses context-rich similarities
Real-world Use	High precision, low robustness	More interpretable, better scalability

Conclusion

- The **Siamese BiLSTM** dominates in raw accuracy but risks overfitting to training data.
- The **Self-Attention Encoder** shows slightly lower performance but offers **better interpretability** and potential for future fine-tuning (e.g., pretrained embeddings, transformers).
- For deployment “**in the wild**”, the **Attention-based model** would likely perform more robustly with unseen, complex legal language.

7. References

- Keras Idiomatic Programmer Guide
- PyTorch Documentation
- Vaswani et al., “*Attention is All You Need*,” 2017
- Mueller & Thyagarajan, “*Siamese Recurrent Architectures for Learning Sentence Similarity*,” 2016

Link to GitHub Repository:

<https://github.com/AyaanKhan0111/LegalSimilarityNLP>