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:

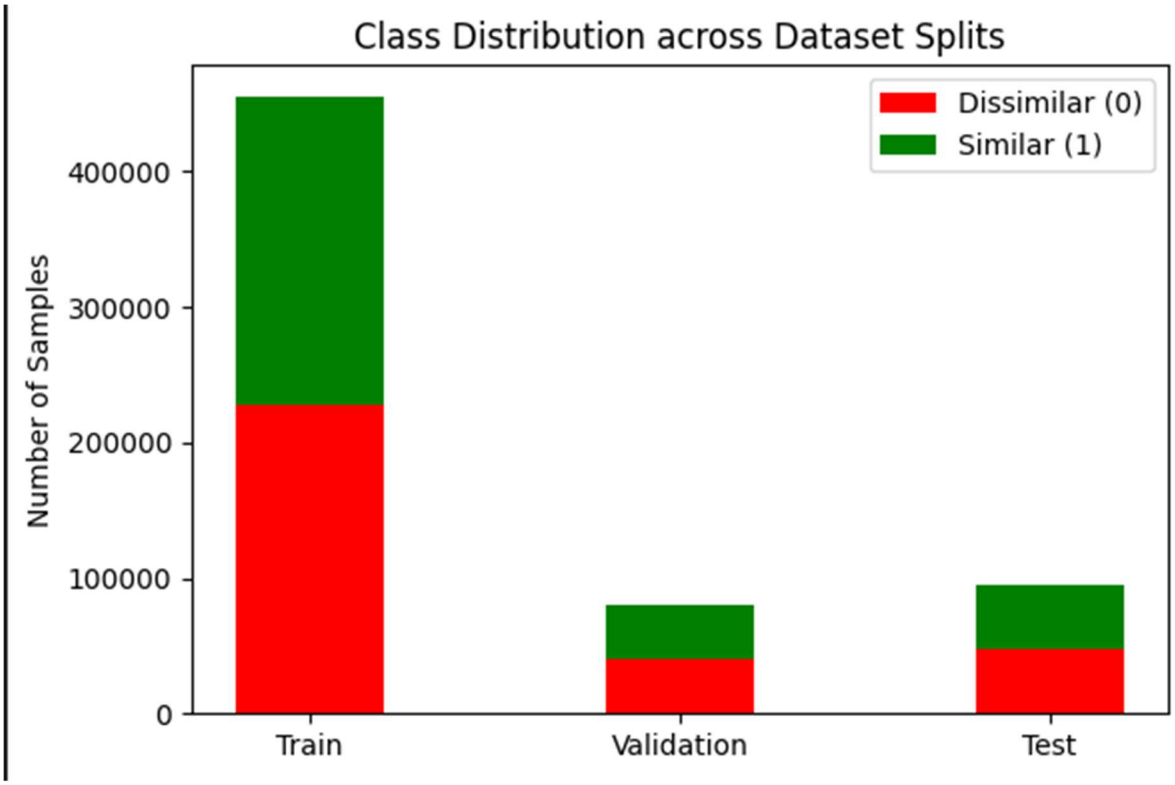
- **Positive pairs** (label = 1): both from the same category.
- **Negative pairs** (label = 0): from different categories.
  This design helps the model learn semantic similarity rather than literal word overlap.

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.



Dataset Split	Similar (1)	Dissimilar (0)	Total	% Similar
Training	227,808	227,807	455,615	50.00%
Validation	40,201	40,202	80,403	50.00%
Test	47,296	47,296	94,592	50.00%

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 (lr=1e-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

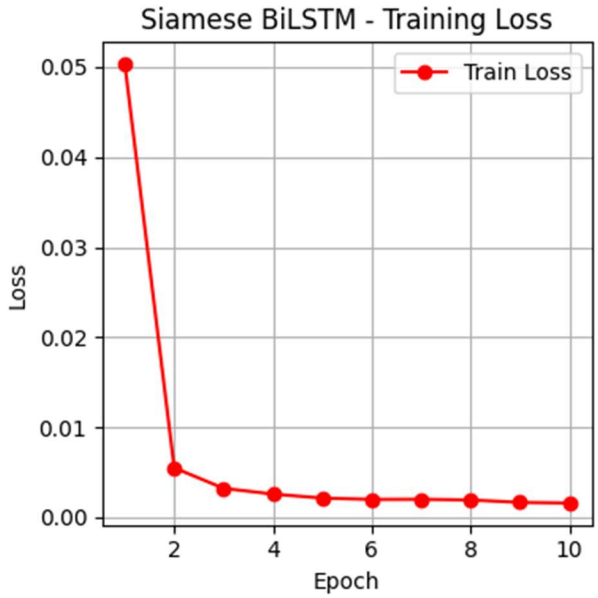
- **Embedding dimension:** 128
- **Attention heads:** 4
- **Feed-forward hidden size:** 256
- **Dropout:** 0.2
- **Classifier:** Same concatenation-based similarity module
- **Loss and optimizer:** Same as above

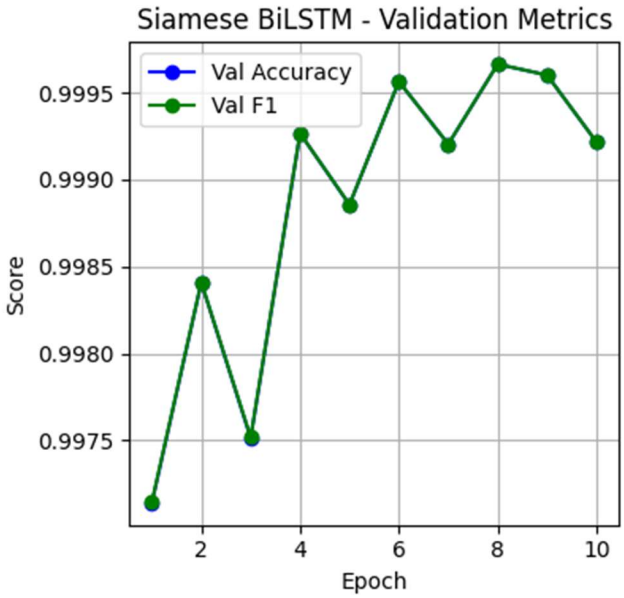
Rationale

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”).

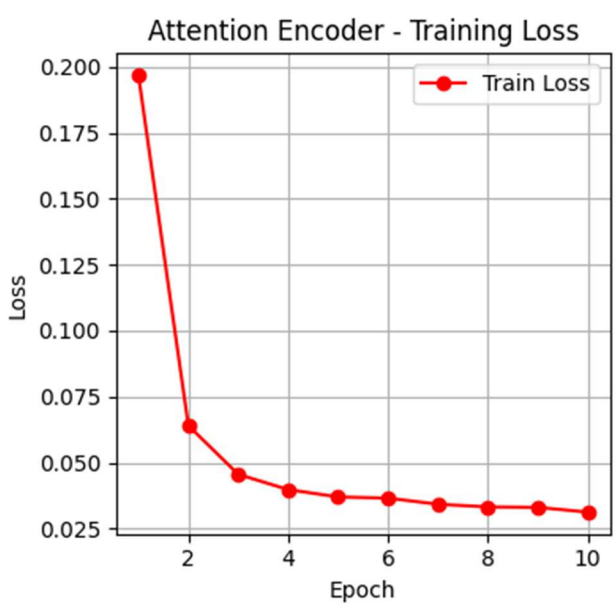
2. Training Graphs

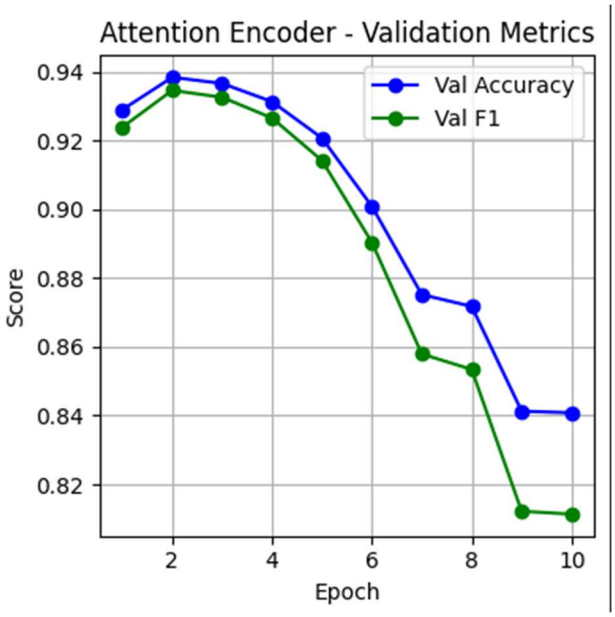
Siamese BiLSTM:





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

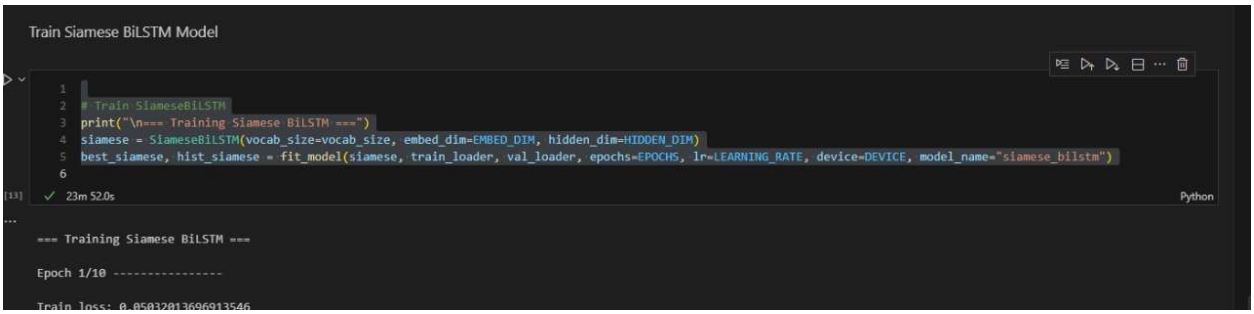
3.1 Validation Performance

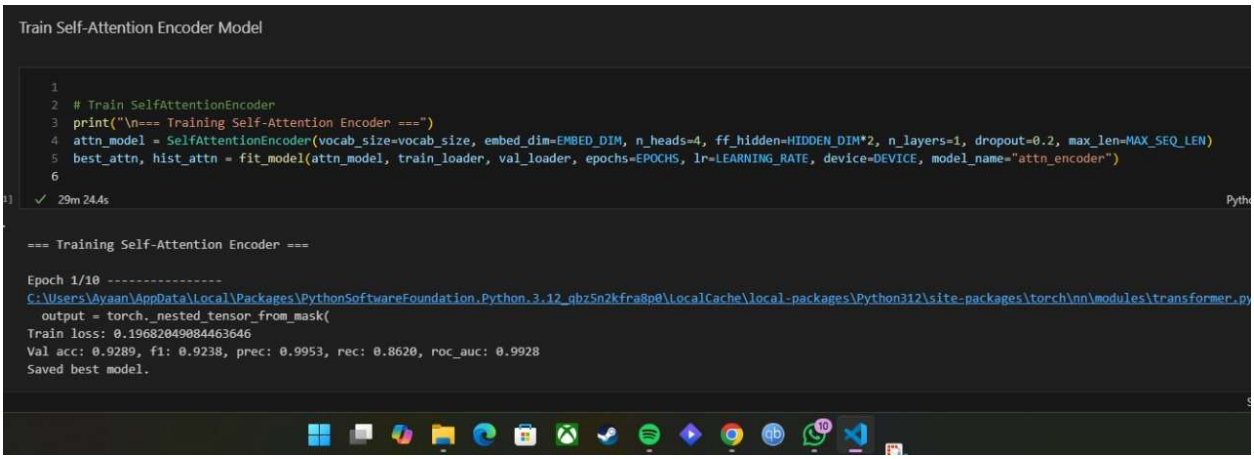
Model	Best Val Accuracy	Val F1	Val Precision	Val Recall	ROC AUC
Siamese BiLSTM	0.9997	0.9997	0.9994	1.0000	1.0000
Attention Encoder	0.8413	0.8122	0.9945	0.6864	0.9864

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





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.

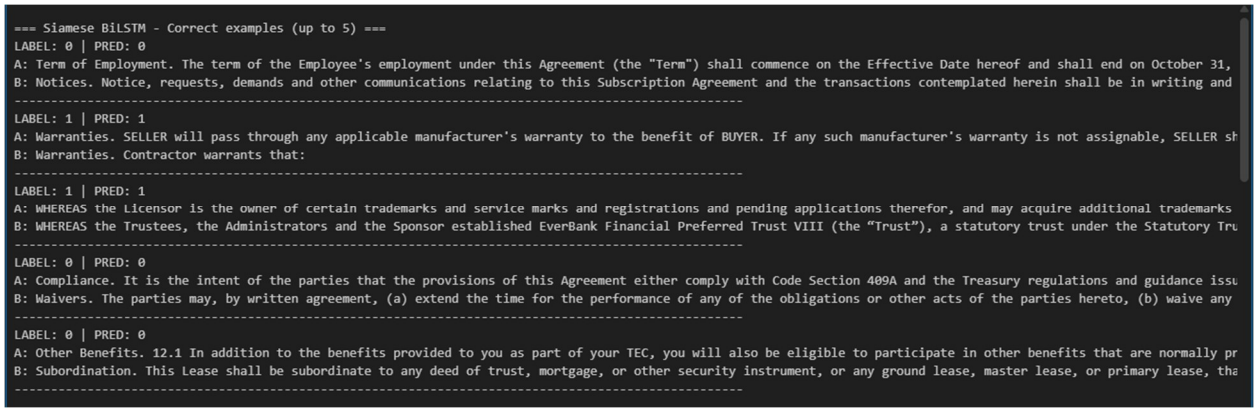
4.3 Real-World Suitability (“In the Wild”)

A real-world clause matching system should:

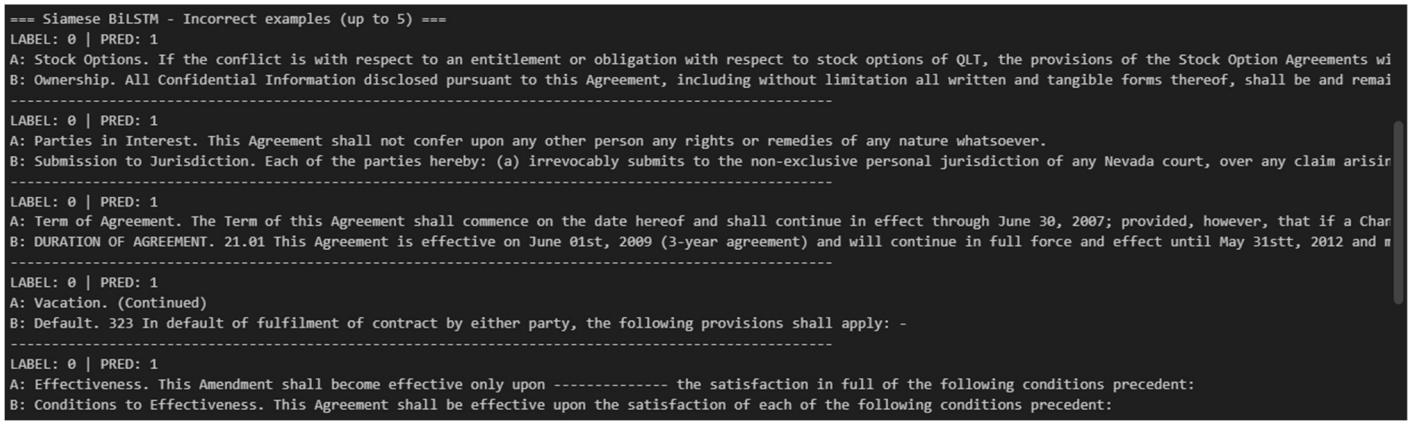
- Prioritize **high recall** (catch all true matches) to avoid missing legally critical similarities.
- Maintain **high precision** to reduce false alarms.
  Thus, **F1 and ROC-AUC** are the most robust evaluation measures for deployment scenarios.

5. Qualitative Results (4 pts)

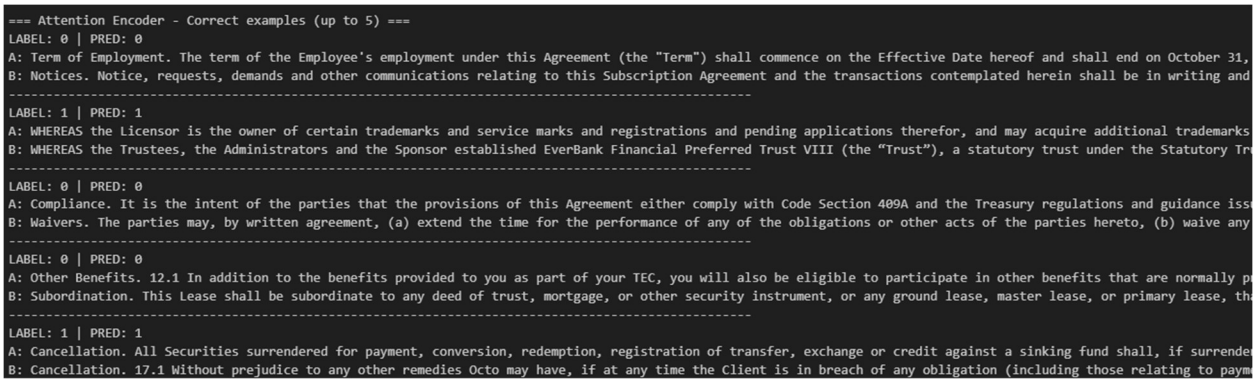
Siamese BiLSTM Correct:



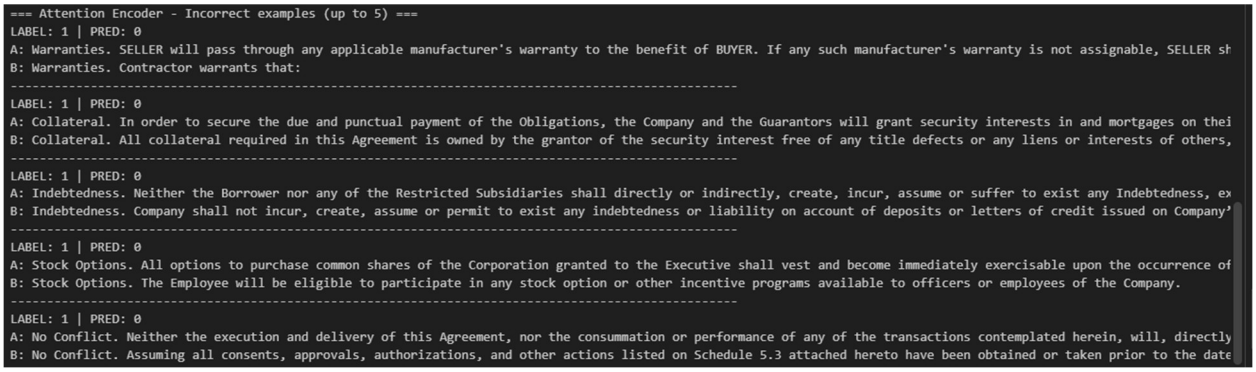
Siamese BiLSTM Incorrect:



Attention Encoder Correct:



Attention Encoder Incorrect:



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