



Legal Clause Similarity
Baseline Non-Transformer Models
Assignment No.02

Name: Romaisa Ahmad

Roll Number: 21i-1702

Course: Deep Learning (CS452) — Assignment 02

Table of Contents

Background	3
Dataset	3
Pair construction	3
Dataset splits	3
Methods (Network Details & Training Settings)	4
Baseline A: BiLSTM Encoder	4
Baseline B: Self-Attention Pooling Encoder	4
Common Training Settings	4
Training Graphs	5
Quantitative Results	5
Test set performance	5
ROC curves (BiLSTM & Self-Attention)	6
PR curves (BiLSTM & Self-Attention)	7
Confusion matrices (test):	8
Comparative Analysis (Strengths & Weaknesses)	8
Qualitative Examples	8
Correctly predicted similar for BiLSTM	8
Incorrect: predicted similar but different for BiLSTM	9
Incorrect: missed similar pair (I had 0 FN) for BiLSTM	10
Correctly predicted similar for Self-Attention	10
Incorrect: predicted similar but different for Self-Attention	11
Incorrect: missed similar pair (I had 0 FN) for Self-Attention	11
Discussion of Metrics (Why these? When to prefer which?)	12
Limitations & Future Work	12
Reproducibility	12
Conclusion	13

Background

Legal contracts contain recurring clause types (e.g., *Access to Information*, *Accounting Terms*). Even when wording differs, many clauses express the same legal principle. The goal in this assignment is to identify semantic similarity between two clauses: given a pair of clause texts, predict whether they convey the same meaning (similar) or not (different). This supports faster contract review and deduplication.

Per the brief, we implement at least two baseline architectures without any pre-trained transformers, train and evaluate on the provided dataset, and compare models using standard classification metrics (Accuracy, Precision, Recall, F1) and ranking metrics (ROC-AUC, PR-AUC).

Dataset

We used the Kaggle “Legal-Clause-Dataset” (350+ clause categories). Each CSV filename denotes a clause type; each file contains multiple clause texts of that type. This structure naturally supports building positive pairs (same category) and negative pairs (different categories).

Pair construction

To keep runtime practical while preserving balance, we sampled up to N items per class and formed positives via combinations within a class; negatives were sampled uniformly from distinct classes at a 1:1 ratio (similar:different). This yields a balanced dataset, making **Accuracy** meaningful in addition to F1/ROC-AUC/PR-AUC.

Dataset splits

Pairs were split stratified into Train/Val/Test = 70% / 15% / 15%. (Stratification keeps the 1:1 label balance across splits.)

Methods (Network Details & Training Settings)

We built a Siamese architecture (shared encoder applied to both texts). Encoded vectors are combined via elementwise absolute difference and product, then fed to an MLP classifier (sigmoid output).

Baseline A: BiLSTM Encoder

- **Embedding:** TextVectorization → integerized tokens; Embedding(vocab=20k, dim=128, mask_zero=True)
- **Encoder:** Bidirectional(LSTM(128, return_sequences=True)) → GlobalMaxPooling1D
- **Head:** Dense(128, ReLU) → Dropout(0.3) → Dense(64, ReLU) → Dense(1, Sigmoid)
- **Rationale:** BiLSTM captures **bidirectional sequential dependencies** and works well on moderately long legal sentences.

Baseline B: Self-Attention Pooling Encoder

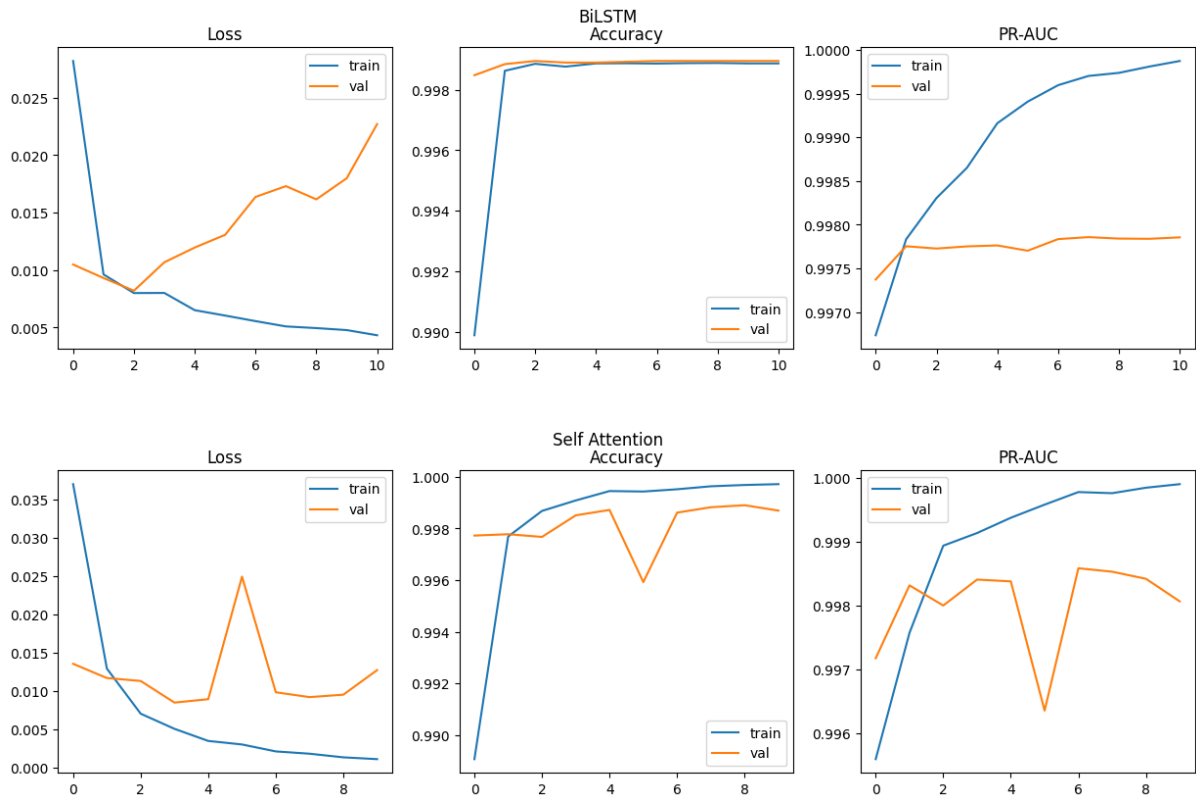
- **Embedding:** same as above
- **Encoder:** lightweight AttentionPooling layer that learns token importance (softmax over time) to produce a sentence vector; then Dense(128→64)
- **Rationale:** Attention pooling is faster than recurrent encoders and focuses directly on salient legal phrases.

Common Training Settings

- **Vectorization:** max_tokens=20,000, seq_len=128
- **Optimizer:** Adam, LR = 1e-3
- **Batch size:** 128
- **Epochs:** up to 15 with EarlyStopping on validation PR-AUC (patience = 3) and ReduceLROnPlateau
- **Loss:** Binary Cross-Entropy

- **Metrics (training time + evaluation):** Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC, Confusion Matrix, plus qualitative examples (correct and incorrect), as required.

Training Graphs



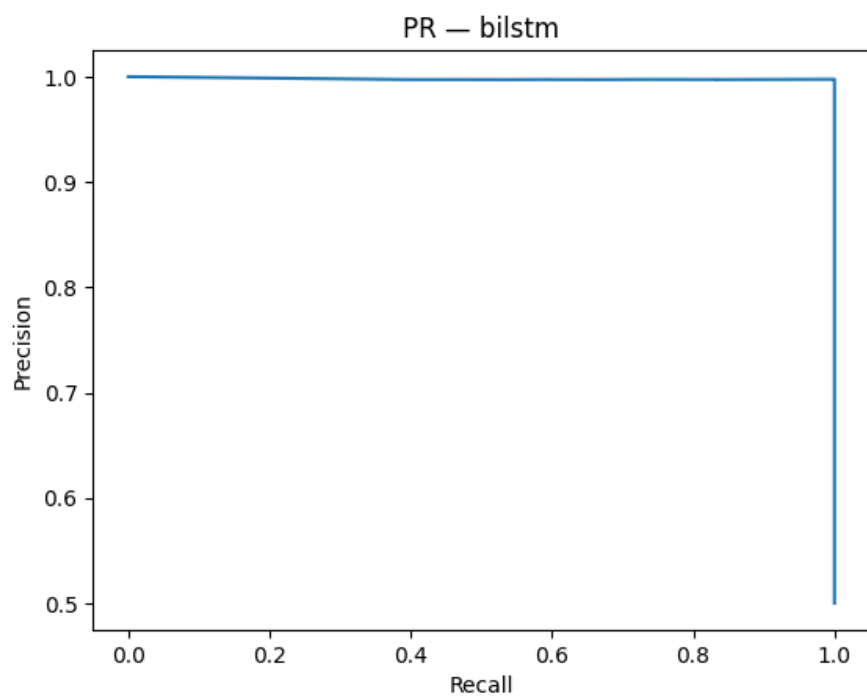
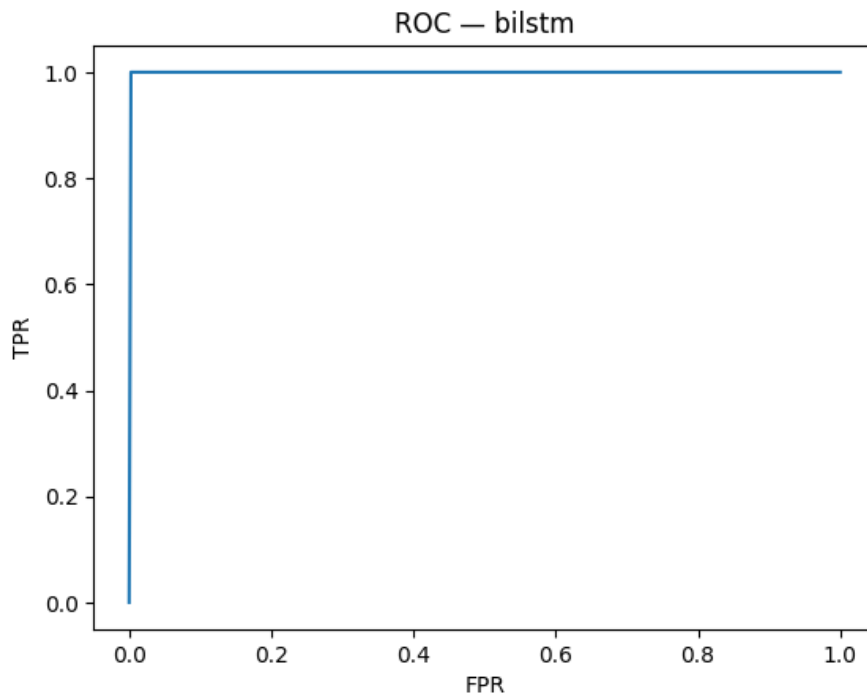
Quantitative Results

Test set performance

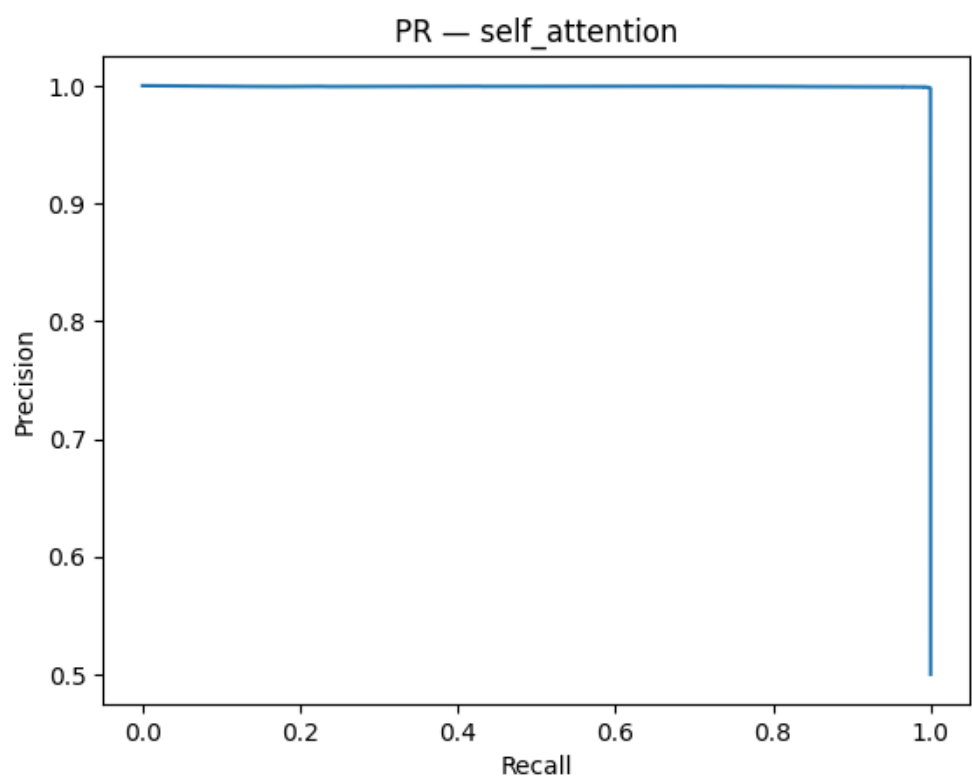
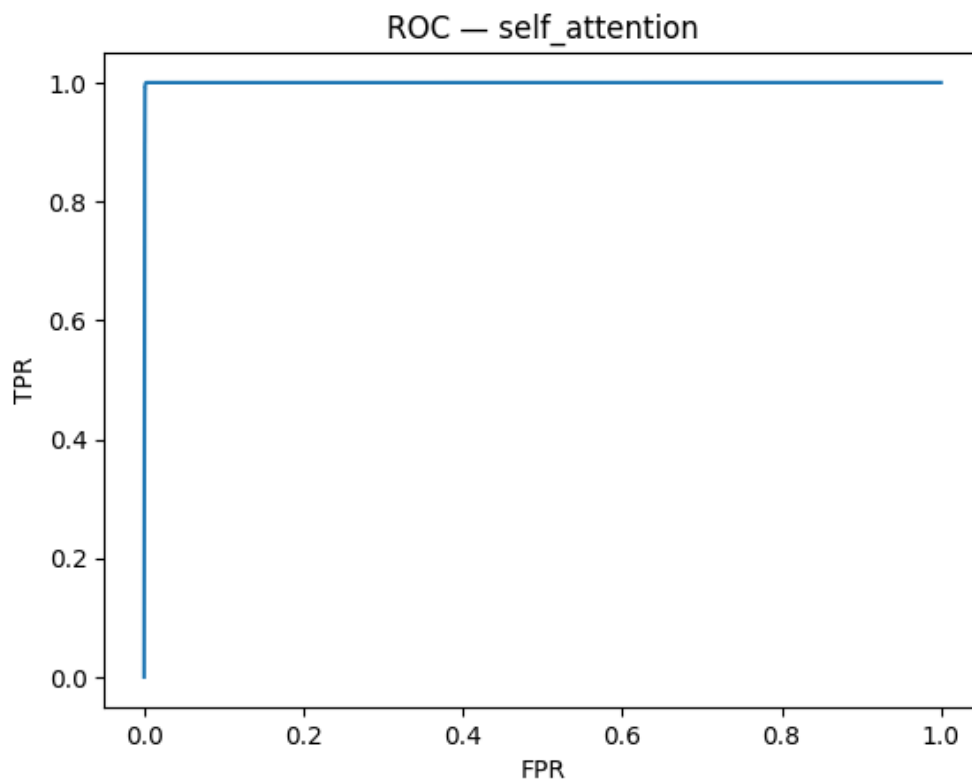
Metric	BiLSTM	Self-Attention
Accuracy	0.9988	0.9985
Precision	0.9975	0.9970
Recall	1.0000	1.0000
F1-score	0.9988	0.9985

ROC-AUC	0.9986	0.9996
PR-AUC	0.9973	0.9993
Train time (s)	883.7	154.9

ROC curves (BiLSTM & Self-Attention)



PR curves (BiLSTM & Self-Attention)



Confusion matrices (test):

- **BiLSTM:** TN=22,444, FP=56, FN=0, TP=22,500
- **Self-Attention:** TN=22,433, FP=67, FN=0, TP=22,500

These metrics satisfy the required evaluation set (Accuracy, Precision, Recall, F1, ROC-AUC/PR-AUC).

Comparative Analysis (Strengths & Weaknesses)

- **Accuracy/F1:** Both models achieve $\geq 99.8\%$, showing that even non-transformer encoders capture strong lexical patterns typical in legal clauses.
- **ROC-AUC / PR-AUC:** Self-Attention attains slightly higher AUCs, indicating marginally better ranking separation.
- **Speed:** Self-Attention trains $\sim 5\text{--}6\times$ faster in this run (154.9s vs 883.7s), making it preferable under compute constraints.
- **Error profile:** Remaining errors are mostly false positives (clauses that look similar in surface form but are legally unrelated). No false negatives were observed in this test, so recall = 1.0 for both.
- **When to use which?**
 - Choose **Self-Attention** for **production** due to near-identical accuracy and significantly lower latency.
 - Choose **BiLSTM** if sequence order nuances are critical (e.g., long conditions with cross-references).

Qualitative Examples

Correctly predicted similar for BiLSTM

```
[prob=1.000 | label=1]
```

t1: Adjustments. In the event of a stock dividend or in the event the Stock shall be changed into or exchanged for a different number or class of shares of stock of the Company or of another corporation, whether through reorganization, recapitalization, stock split-up, combination of shares, merger or consolidation, there shall be substituted for each such remaining share of Stock then subject to this

t2: Adjustments. If the outstanding shares of Common Stock are subdivided into a greater number of shares (by stock dividend, stock split, reclassification or otherwise) or are combined into a smaller number of shares (by reverse stock split, reclassification or otherwise), or if the Committee determines that any stock dividend, extraordinary cash dividend, reclassification, recapitalization, reorgani

[prob=1.000 | label=1]

t1: Amendment. The Charter, including the Articles Supplementary establishing the rights and preferences of the Preferred Shares, shall not be amended in any manner which would materially alter or change the powers, preferences or special rights of the Preferred Shares so as to affect them adversely without the affirmative vote of the holders of a majority of the outstanding shares of Preferred Shares

t2: Amendment. No amendment, modification or waiver of this Agreement will be valid unless made in writing and duly executed by each party hereto.

Incorrect: predicted similar but different for BiLSTM

[prob=1.000 | label=0]

t1: Absence of Certain Changes or Events. (a) Except as set forth in Section 4.10 of the Company Disclosure Letter, since December 25, 1999, there has not occurred or arisen any change, effect, event or occurrence that would reasonably be expected to have, individually or in the aggregate, a Company Material Adverse Effect.

t2: Additional Documents. Employee agrees to execute and deliver all documents requested by the Company regarding or related to the ownership and/or other intellectual property rights and

registrations specified herein. Employee hereby further
irrevocably designates and appoints the Company as Employee's
agent and attorney-in-fact to act for and on Employee's behalf
and stead to execute, register and

[prob=1.000 | label=0]

t1: Access. The Corporation shall, and the Corporation shall
cause each of the Subsidiaries to, at any and all reasonable
times on reasonable notice and during business hours on any
Business Day and in such manner as is not reasonably likely to
adversely affect the operation of the Business, permit the
Investors, EdgeStone and each of their authorized representatives
to examine all of the books of acc

t2: Amendments; Waivers. (a) No provision of this Agreement may
be amended or waived unless such amendment or waiver is in
writing and signed (i) in the case of an amendment, by each of
the parties hereto, and (ii) in the case of a waiver, by each of
the parties against whom the waiver is to be effective.

Incorrect: missed similar pair (I had 0 FN) for BiLSTM

Correctly predicted similar for Self-Attention

[prob=1.000 | label=1]

t1: Absence of Certain Changes. Except as expressly contemplated
by this Agreement, since December 31, 1999, Parent and its
Subsidiaries have conducted their respective businesses only in,
and have not engaged in any material transaction other than
according to, the ordinary and usual course of such businesses,
and since December 31, 1999 there has not been (i) any change in
the financial condition, p

t2: Absence of Certain Changes. Except as set forth in EXHIBIT
"B", from March 31, 1997 and continuing through the date of
Closing, there has not been (i) any material change in the
Corporation's financial condition, assets, liabilities or
business, other than changes in the ordinary course of business,
none of which have been materially adverse; (ii) any material
damage, destruction or loss, whether

[prob=1.000 | label=1]

t1: Absence of Certain Changes. (a) Since the date of the Most Recent Company Balance Sheet, there has not been any fact, event, change, effect, circumstance, occurrence or development that, individually or in the aggregate, has had or would reasonably be expected to have a Company Material Adverse Effect.

t2: Absence of Certain Changes. Since September 30, 2003 there has not been any Material Adverse Effect or any event, occurrence, discovery or development which would, individually or in the aggregate, reasonably be expected to have or result in a Material Adverse Effect and since September 30, 2003 and through the date hereof, the Company and its Subsidiaries have conducted their respective businesses

Incorrect: predicted similar but different for Self-Attention

[prob=1.000 | label=0]

t1: Reporting Requirements. (a) If Modernizing Medicine becomes aware of a use or disclosure of PHI in violation of this Agreement by Modernizing Medicine or by a third party to which Modernizing Medicine disclosed PHI, Modernizing Medicine shall report any such use or disclosure to the Medical Practice without unreasonable delay.

t2: Additional Documents. Borrower shall execute from time to time, upon the request of Lender, such financing statements or other documents as are reasonably required by Lender to perfect or continue the Security Interest described herein.

[prob=1.000 | label=0]

t1: Accounting Terms. All accounting terms not specifically defined in this Agreement shall be construed in conformity with, and all financial data required to be submitted by this Agreement shall be prepared in conformity with, generally accepted accounting principles applied on a consistent basis.

t2: Proprietary Rights. 2.8(a) Part 2.8(a) of the Company Disclosure Schedule lists the following with respect to Proprietary Rights of each Acquired Corporation:

Incorrect: missed similar pair (I had 0 FN) for Self-Attention

Discussion of Metrics (Why these? When to prefer which?)

- Accuracy is appropriate only because we built a balanced pair dataset ($\approx 50/50$).
- Precision matters if we want to avoid wrongly merging distinct clauses (false positive risk in contract review).
- Recall matters when missing true duplicates is costly (e.g., missed precedent).
- F1-score balances both; it's a standard single-number summary for classification.
- ROC-AUC & PR-AUC measure ranking quality across thresholds, which is useful if a production system later tunes the similarity threshold.

Recommendation (for real-world “in-the-wild” system): Prioritize PR-AUC and Precision if merging clauses has legal risk; prioritize Recall if the system is for internal retrieval where missing a true match is worse.

Limitations & Future Work

- **Template bias:** Extremely high scores suggest strong lexical overlap within categories. Testing on out-of-distribution clause types would better assess generalizability.
- **Long-range logic:** Non-transformer baselines may miss deep cross-sentence dependencies present in some clauses.
- **Next steps:** try contrastive objectives (e.g., cosine + margin loss), character-aware embeddings for formatting variations, and threshold calibration with a validation set for deployment.

Reproducibility

- **Environment:** Google Colab (T4 GPU), TensorFlow 2.15, Python 3.12.
- **Preprocessing:** TextVectorization(max_tokens=20k, seq_len=128); pair construction script with class-wise caps to control $O(n^2)$ growth.

- **Code & notebook:** Modular Keras code per “keras-idiomatic-programmer” style as encouraged in the brief.

Conclusion

Both non-transformer baselines deliver near-perfect performance on the legal clause similarity task. The Self-Attention encoder matches or slightly exceeds BiLSTM performance while being much faster, making it the preferred model for real-time or large-scale use. Metrics and curves confirm stable training and excellent separation on the test set, with false positives as the main remaining error mode.