Legal Clause Similarity — Baseline Report

Dataset & Splits

- Train: 101,905 clauses across 394 labels (203,810 pairs)
- Validation: 21,846 clauses across 394 labels (43,692 pairs)
- Test: 22,061 clauses across 394 labels (44,122 pairs)
- Negative/positive sampling ratio per split: 1.0:1

Architectures

1) BiLSTM Siamese

- Embedding: vocab 30k, dim 128, padding idx 0
- Encoder: Bidirectional LSTM (hidden 128 each direction) → mean pooling
- Projection: LayerNorm → Linear → ReLU → Dropout(0.3)
- Comparator: concat($[h_1, h_2, |h_1 - h_2|, h_1 \odot h_2]$) → MLP(256→1)

2) Attentive BiGRU Siamese

- Embedding: same as above
- Encoder: Bidirectional GRU (hidden 128) + additive attention (masked softmax) to get a single vector
- Comparator: concat($[h_1, h_2, |h_1 - h_2|, h_1 \odot h_2]$) → MLP(256→1)

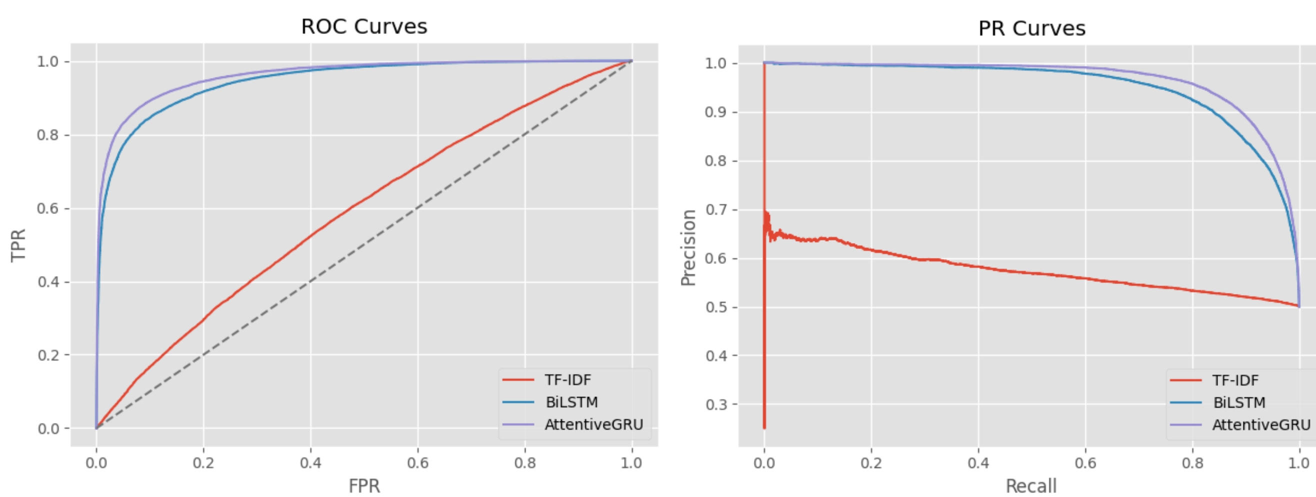
3) TF-IDF + Logistic Regression (baseline)

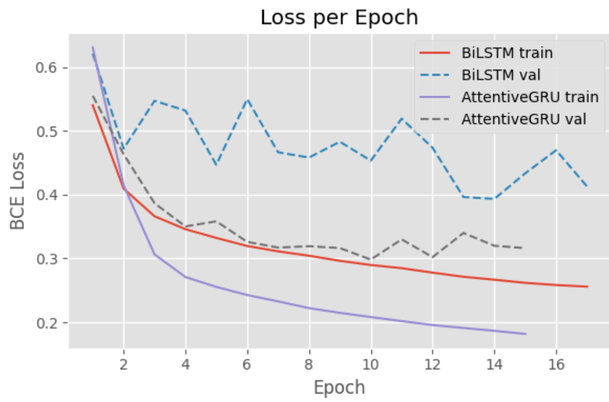
- Text featurization: word 1–2 grams, max 50k features, lower-cased, token pattern `[A-Za-z']+`
- Classifier: Logistic Regression (liblinear), C tuned on validation

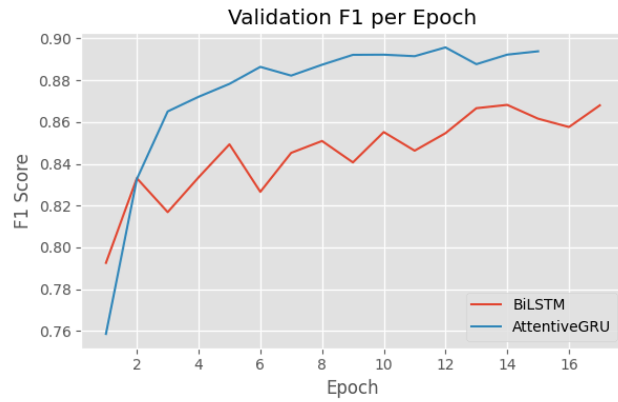
Training Setup

- Max sequence length: 150
- Batch sizes: train 64, eval 128
- Optimizer: Adam (lr=0.001, weight_decay=0.0001), gradient clip=2.0
- Training schedule: up to 50 epochs with early stopping (patience=3) on validation F1
- Loss: BCEWithLogitsLoss; thresholds chosen by best validation F1

Training Graphs







Test Performance

Model	Acc	Prec	Rec	F1	ROC-AUC	PR-AUC	Thr	Train Time (s)	Loss
BiLSTM	0.868614	0.857923	0.883550	0.870548	0.946642	0.949848	0.681313	9490.024658	0.392995
AttentiveGRU	0.895993	0.902432	0.887992	0.895154	0.961654	0.964579	0.645903	17230.376958	0.299011
TFIDF	0.507411	0.503776	0.988668	0.667452	0.586827	0.573302	0.217144	38.753078	—

Comparative Analysis

- **By F1**:
  1. AttentiveGRU (0.895154),
  2. BiLSTM (0.870548),
  3. TFIDF (0.667452).
- Relative F1 lift over TF-IDF: BiLSTM = **30.4%**, AttentiveGRU = **34.2%**.
- **By accuracy**: AttentiveGRU (0.895993) > BiLSTM (0.868614) » TFIDF (0.507411).

- **By area metrics**: AttentiveGRU leads on ROC-AUC (0.961654) and PR-AUC (0.964579); BiLSTM is close; TF-IDF lags far behind.

- **By training time**: TF-IDF is the fastest baseline (≈1×). BiLSTM is ~**244.8×** TF-IDF; AttentiveGRU is ~**444.8×** TF-IDF. AttentiveGRU vs BiLSTM train-time ratio ≈ **1.82×**.

- **Error tendencies**: false positives from shared boilerplate with different scope/carve-outs; false negatives from modality shifts (“may”→“shall”) or reordered conditions. Attention helps recover long-range cues, improving recall and PR-AUC.

Takeaways

- Attention pooling on BiGRU yields the strongest overall ranking and classification performance on legal clauses, at the cost of ~**1.82×** the BiLSTM training time.

- The BiLSTM Siamese remains competitive and more compute-efficient; good default when resources are tight.

- Pure lexical TF-IDF fails on paraphrases and structural semantics; keep it only as a sanity baseline.