RNN HW3

Basic :

1. Model Architecture

The implemented Named Entity Recognition (NER) system consists of three main components:

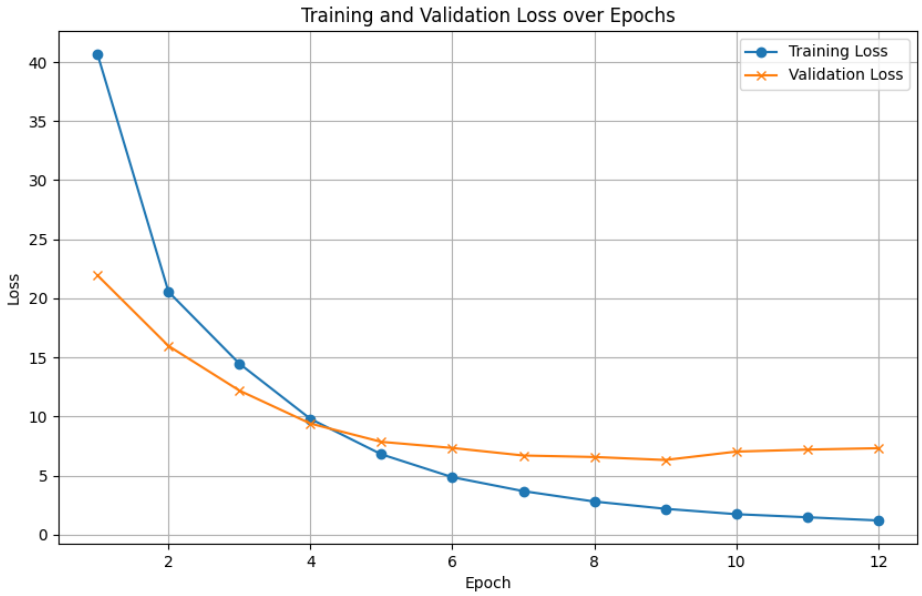
1.1 Model Components

1. **BERT Layer**: Uses bert-base-cased as the foundation to provide contextual word embeddings.
2. **BiLSTM Layer**: Processes BERT embeddings to capture bidirectional sequential dependencies.
    - Input size: 768 (BERT's hidden size)
    - Hidden size: 384 per direction
    - Layers: 2
    - Bidirectional: True
3. **CRF Layer**: Models tag dependencies to ensure valid prediction sequences, particularly important for multi-token technical entities.

1.2 Implementation Details

- Maximum sequence length: 128 tokens
- Batch size: 16
- Learning rate: 2e-5
- Optimizer: AdamW with weight decay (0.01 for regular parameters, 0 for bias/LayerNorm)
- Training epochs: 10
- Special handling for subword tokens to align with word-level entity tags

2. Experimental Results



2.1 Performance Metrics

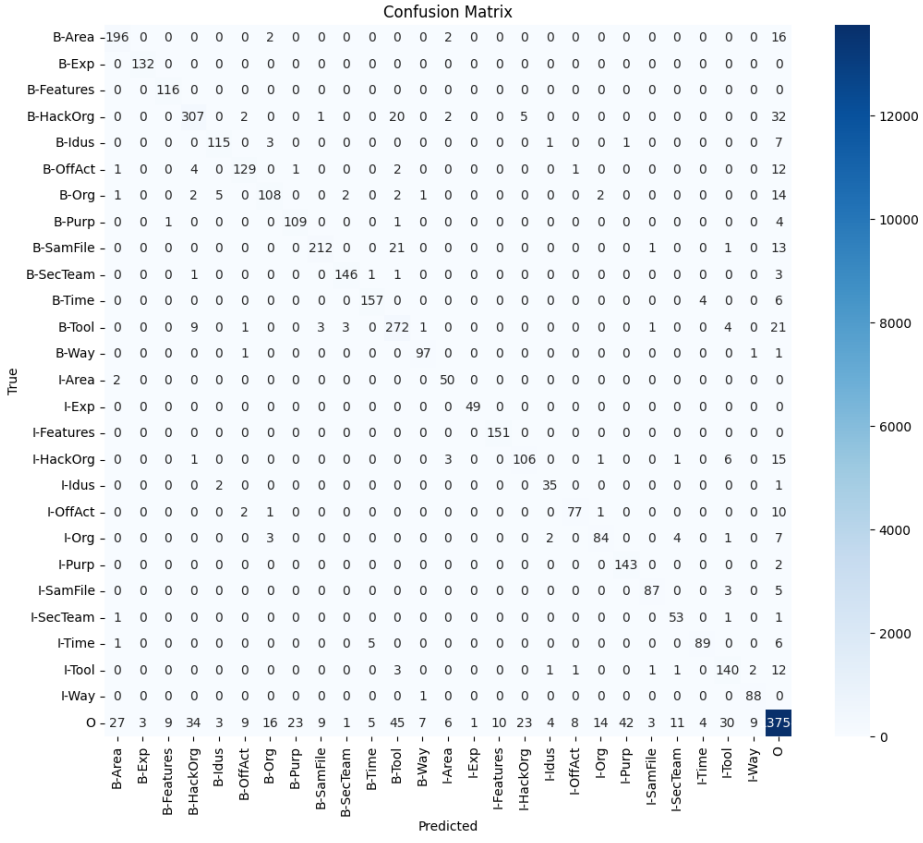
The model achieved strong overall performance on the DNRTI cybersecurity dataset:

	precision	recall	f1-score	support
micro avg	0.9614	0.9614	0.9614	17205
macro avg	0.9593	0.9369	0.9465	17205
weighted avg	0.9617	0.9614	0.9611	17205

2.2 Entity-Level Performance

- Strongest performance (F1 ≥ 0.96):
  - I-Exp, I-Features (F1 = 1.0000)
  - I-Area (F1 = 0.9804)
  - B-Way (F1 = 0.9749)
  - B-SecTeam (F1 = 0.9739)
- Most challenging entities:
  - B-Org (F1 = 0.8000)
  - B-HackOrg (F1 = 0.8650)
  - B-Tool (F1 = 0.8757)

2.3 Error Analysis



Key error patterns:

- Confusion between related technical entities (B-HackOrg and B-Tool)
- B-SamFile tokens misclassified as B-Tool (21 instances)
- Non-entity tokens occasionally misclassified as B-HackOrg (34 instances)

2.4 Training Dynamics

- Rapid decrease in both training and validation loss during the first 4 epochs
- Training loss continues to decline throughout all 12 epochs
- Validation loss stabilizes around epoch 6-7
- Slight overfitting observed after epoch 7, but validation loss remains stable
- Best model performance likely occurs around epochs 6-8

3. Conclusion

The BERT-BiLSTM-CRF architecture demonstrates strong performance for cybersecurity NER, with a weighted F1 score of 0.9611. The model excels at recognizing most cybersecurity-specific entities but faces more challenges with organization entities and technical tools. The architecture effectively combines BERT's contextual understanding with BiLSTM's sequential learning and CRF's structured prediction capabilities, making it well-suited for the technical language patterns found in cybersecurity text.

Bonus :

The architecture is composed of three main components:

- **SecBERT Encoder**: A pretrained transformer model (jackaduma/SecBERT) tailored for cybersecurity text, used for extracting contextual embeddings for each token.
- **Bidirectional LSTM (BiLSTM)**: A two-layer LSTM with hidden dimension 768, which captures forward and backward dependencies in sequences.
- **CRF Layer**: A Conditional Random Field layer placed on top of the BiLSTM output to model the label dependencies and produce the most likely tag sequence.

Hyperparameters:

- Maximum sequence length: 128
- Batch size: 16
- Learning rate: 2e-5
- Epochs: 30
- Dropout: 0.1
- LSTM layers: 2
- Hidden dimension: 768
- Optimizer: AdamW with weight decay
- Learning rate scheduler: ReduceLROnPlateau
- Early stopping: patience = 3, delta = 0.001

Performance :

```
Evaluating on test set...
Evaluating: 100%|██████████| 42/42 [00:04<00:00,  8.44it/s]
Entity-level metrics - F1: 0.8471, Precision: 0.8314, Recall: 0.8633
Entity-Level Test Report:
```

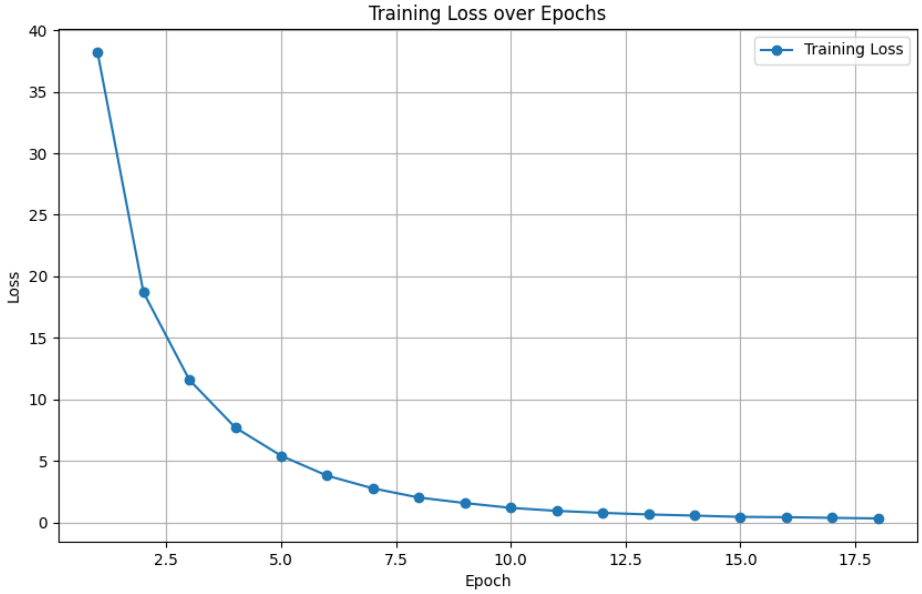
	precision	recall	f1-score	support
Area	0.6973	0.8426	0.7631	216
Exp	0.9706	1.0000	0.9851	132
Features	0.9508	1.0000	0.9748	116
HackOrg	0.7538	0.8130	0.7823	369
Idus	0.8905	0.9457	0.9173	129
OffAct	0.8478	0.7800	0.8125	150
Org	0.7021	0.7226	0.7122	137
Purp	0.8394	1.0000	0.9127	115
SamFile	0.9364	0.8911	0.9132	248
SecTeam	0.8973	0.8618	0.8792	152
Time	0.8864	0.9231	0.9043	169
Tool	0.7926	0.7524	0.7720	315
Way	0.8919	0.9900	0.9384	100
micro avg	0.8314	0.8633	0.8471	2348
macro avg	0.8505	0.8863	0.8667	2348
weighted avg	0.8344	0.8633	0.8472	2348

Token-Level Validation Report:

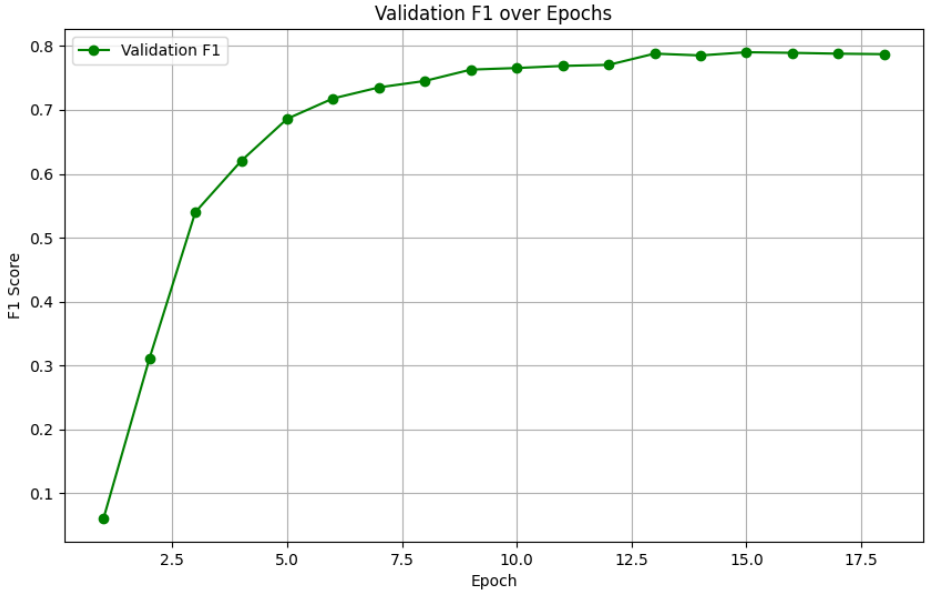
	precision	recall	f1-score	support
B-Area	0.6056	0.8958	0.7227	144
B-Exp	0.9850	1.0000	0.9924	131
B-Features	0.8435	1.0000	0.9151	97
B-HackOrg	0.8172	0.8498	0.8331	426
B-Idus	0.9412	0.9320	0.9366	103
B-OffAct	0.7059	0.7200	0.7129	100
B-Org	0.7526	0.6577	0.7019	111
B-Purp	0.8351	0.9878	0.9050	82
B-SamFile	0.9048	0.8953	0.9000	191
B-SecTeam	0.9461	0.8876	0.9159	178
B-Time	0.9080	0.8087	0.8555	183
B-Tool	0.7794	0.6909	0.7324	317
B-Way	0.8455	0.8532	0.8493	109
I-Area	0.7705	0.9792	0.8624	48
I-Exp	0.9762	1.0000	0.9880	41
I-Features	0.8614	1.0000	0.9255	87
I-HackOrg	0.6517	0.8227	0.7273	141
I-Idus	0.9000	0.9730	0.9351	37
I-OffAct	0.7925	0.6462	0.7119	65
I-Org	0.6667	0.7536	0.7075	69
I-Purp	0.8667	1.0000	0.9286	130
I-SamFile	0.9195	0.8989	0.9091	89
I-SecTeam	0.8704	0.7231	0.7899	65
I-Time	0.7216	0.8140	0.7650	86
I-Tool	0.6852	0.7115	0.6981	156
I-Way	0.8087	0.8692	0.8378	107
O	0.9776	0.9682	0.9729	14309
accuracy			0.9451	17602
macro avg	0.8273	0.8644	0.8419	17602
weighted avg	0.9477	0.9451	0.9457	17602

Visualizations :

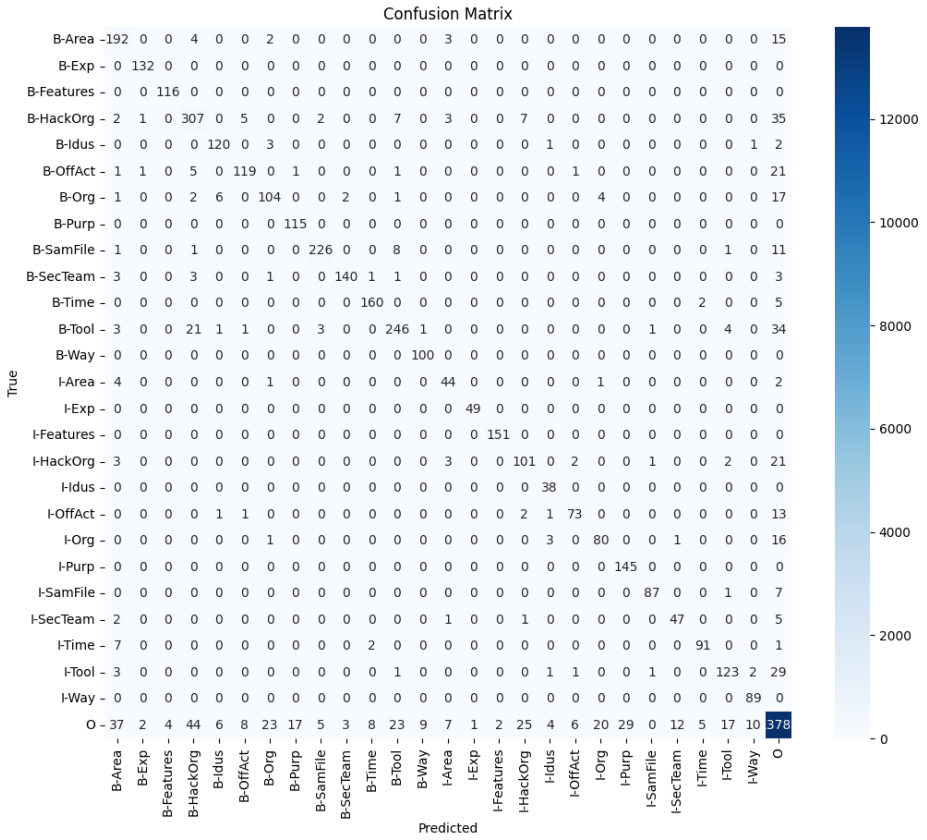
Training loss :



Validation fl score :



confusion matrix :



Conclusion

The SecBERT + BiLSTM + CRF model significantly improves NER performance in the cybersecurity domain by combining:

- domain-specific contextual representations (SecBERT),
- sequential modeling (BiLSTM), and
- structured output prediction (CRF).

This approach demonstrated high accuracy and robust generalization through the use of advanced training strategies including early stopping, learning rate scheduling, and CRF-based decoding.