Exercise: Fine-Tuning a Pre-trained BERT Model for Sentiment Classification

1. **Fine-tuning a Pre-trained BERT Model**
   Repeat Exercise 2 of Part 5 (sentiment classifier), by fine-tuning a pre-trained BERT model.
   - Tune the hyper-parameters (e.g., sizes of any task-specific layers on top of BERT, number of BERT encoder blocks to keep frozen) on the development subset of your dataset.
   - Monitor the performance of your models on the development subset during training to decide how many epochs to use.
   - If the texts of your experiments exceed BERT's maximum length limit, you may want to truncate them at the maximum allowed length of BERT or use a BERT-like model that can handle longer texts (e.g., Longformer).
2. **Experimental Results**
   - Include experimental results of a baseline majority classifier, as well as experimental results of your best classifiers from Exercise 15 of Part 2, Exercise 9 of Part 3, Exercise 1 of Part 4, Exercise 2 of Part 5, now treated as additional baselines.
   - Otherwise, the contents of your report should be as in Exercise 2 of Part 5, but now with information and results for the experiments of this exercise.
3. **Optional Bonus: Test Set Results**
   - You may optionally include (for extra bonus) indicative experimental results on a small subset of the test set (e.g., 10 test examples) obtained by prompting an LLM (e.g., Chat-GPT), using appropriate instructions and possibly including few-shot examples (demonstrators).

Assert whether `PyTorch` can use an available GPU card

Creating a Dataset

We will use the `Dataset` class from `PyTorch` to handle the text data. We will pad the text sequences with 0 to a pre-defined length (the average number of tokens in the training split).

```
Using device: cuda
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

```
True
```

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	1
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day	News	December 23, 2017	1

	title	text	subject	date	label
		a...			
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	1
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	1
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	1

	text	label
0	donald trump just couldn t wish all americans ...	1
1	house intelligence committee chairman devin nu...	1
2	on friday it was revealed that former milwauke...	1
3	on christmas day donald trump announced that h...	1
4	pope francis used his annual christmas day mes...	1

```
DatasetDict({
    train: Dataset({
        features: ['text', 'label', '__index_level_0__'],
        num_rows: 35918
    })
    val: Dataset({
        features: ['text', 'label', '__index_level_0__'],
        num_rows: 4490
    })
    test: Dataset({
        features: ['text', 'label', '__index_level_0__'],
        num_rows: 4490
    })
})
```

Define the model

We will create a model class and parameterize our neural network with several choices

```
Loading model: distilroberta-base
```

```
Some weights of RobertaForSequenceClassification were not initialized
from the model checkpoint at distilroberta-base and are newly
initialized: ['classifier.dense.bias', 'classifier.dense.weight',
'classifier.out_proj.bias', 'classifier.out_proj.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
{"model_id":"a274b3e7a9d24f128fc2d455c89b50aa","version_major":2,"version_minor":0}
```

```
{"model_id":"ccee207c0818427abd349234f3436e8f","version_major":2,"version_minor":0}
```

```
{"model_id":"168f215822074393b301cdafca0a0dcf","version_major":2,"version_minor":0}
```

[4490/4490 13:01, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Re
1	0.002400	0.002488	0.999777	0.999999	{'0': {'precision': 0.999533364442 'recall': 1.0, 'f1-sco 0.99976662777712 'support': 2142.0} {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy': 0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.999787052810 'f1-score': 0.9997768176130

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Re
					'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}
2	0.000000	0.002163	0.999777	1.000000	{'0': {'precision': 0.9995333644423 'recall': 1.0, 'f1-sco 0.9997666277712 'support': 2142.0} {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy': 0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.9997870528109 'f1-score': 0.9997768176130 'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}

```
Final Classification Report for distilroberta-base:

Training Classification Report:
{'0': {'precision': 1.0, 'recall': 0.9999417588817705, 'f1-score':
0.9999708785928536, 'support': 17170.0}, '1': {'precision':
0.9999466638220705, 'recall': 1.0, 'f1-score': 0.9999733311998293,
'support': 18748.0}, 'accuracy': 0.9999721588061696, 'macro avg':
{'precision': 0.9999733319110353, 'recall': 0.9999708794408853, 'f1-
score': 0.9999721048963415, 'support': 35918.0}, 'weighted avg':
{'precision': 0.9999721602911124, 'recall': 0.9999721588061696, 'f1-
score': 0.9999721587720278, 'support': 35918.0}}
Validation Classification Report:
{'0': {'precision': 0.9995333644423705, 'recall': 1.0, 'f1-score':
0.9997666277712952, 'support': 2142.0}, '1': {'precision': 1.0,
'recall': 0.9995741056218058, 'f1-score': 0.999787007454739,
'support': 2348.0}, 'accuracy': 0.9997772828507795, 'macro avg':
{'precision': 0.9997666822211853, 'recall': 0.999787052810903, 'f1-
score': 0.9997768176130171, 'support': 4490.0}, 'weighted avg':
{'precision': 0.9997773867785207, 'recall': 0.9997772828507795, 'f1-
score': 0.9997772851202321, 'support': 4490.0}}
Test Classification Report:
{'0': {'precision': 0.9995249406175772, 'recall': 0.9995249406175772,
'f1-score': 0.9995249406175772, 'support': 2105.0}, '1': {'precision':
0.99958071278826, 'recall': 0.99958071278826, 'f1-score':
0.99958071278826, 'support': 2385.0}, 'accuracy': 0.999554565701559,
'macro avg': {'precision': 0.9995528267029186, 'recall':
0.9995528267029186, 'f1-score': 0.9995528267029186, 'support':
4490.0}, 'weighted avg': {'precision': 0.999554565701559, 'recall':
0.999554565701559, 'f1-score': 0.999554565701559, 'support': 4490.0}}
Precision-Recall AUC Scores:
Training PR AUC: 1.0000
Validation PR AUC: 1.0000
Test PR AUC: 1.0000
==================================================
Loading model: bert-base-uncased
```

```
{"model_id":"9487e8c74f154fb8ae4bc4e7cd7710c5","version_major":2,"version_minor":0}
```

```
{"model_id":"1b4ed50528494faeb2916fd5f13c4eb4","version_major":2,"version_minor":0}
```

```
{"model_id":"3d39cef654c34e6b9274d333ebeb3922","version_major":2,"version_minor":0}
```

```
{"model_id":"5caf03823dcc44b5a68654ac9cdb6edb","version_major":2,"version_minor":0}
```

{"model_id":"caa52cef2b964ca3b7907adcc9ba9df8","version_major":2,"version_minor":0}

```
Some weights of BertForSequenceClassification were not initialized
from the model checkpoint at bert-base-uncased and are newly
initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

{"model_id":"24c6ca7f015f43d79c6b961774707fe4","version_major":2,"version_minor":0}

{"model_id":"4d94346ec6054e7aa947b0b7c00a3dbb","version_major":2,"version_minor":0}

{"model_id":"09651f547f20475fbc303c3fadb28024","version_major":2,"version_minor":0}

[4490/4490 24:55, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Re
1	0.006200	0.000051	1.000000	1.000000	{'0': {'precision': 1 'recall': 1.0, 'f1-sco 1.0, 'support': 214 '1': {'precision': 1.0 'recall': 1.0, 'f1-sco 1.0, 'support': 2348.0}, 'accuracy 1.0, 'macro avg': {'precision': 1.0, 'recall': 1.0, 'f1-sco 1.0, 'support': 4490.0}, 'weighte avg': {'precision': 'recall': 1.0, 'f1-sco 1.0, 'support': 4490.0}}
2	0.000000	0.001179	0.999777	1.000000	{'0': {'precision': 0.999533364442 'recall': 1.0, 'f1-sco 0.999766627771 'support': 2142.0} {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy': 0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.999787052810 'f1-score': 0.9997768176130 'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}

```
Final Classification Report for bert-base-uncased:

Training Classification Report:
{'0': {'precision': 1.0, 'recall': 0.999883517763541, 'f1-score':
0.9999417554895451, 'support': 17170.0}, '1': {'precision':
0.9998933333333333, 'recall': 1.0, 'f1-score': 0.9999466638220705,
'support': 18748.0}, 'accuracy': 0.9999443176123393, 'macro avg':
{'precision': 0.9999466666666667, 'recall': 0.9999417588817705, 'f1-
score': 0.9999442096558078, 'support': 35918.0}, 'weighted avg':
{'precision': 0.9999443235517939, 'recall': 0.9999443176123393, 'f1-
score': 0.9999443174756854, 'support': 35918.0}}
Validation Classification Report:
{'0': {'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0, 'support':
2142.0}, '1': {'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0,
'support': 2348.0}, 'accuracy': 1.0, 'macro avg': {'precision': 1.0,
'recall': 1.0, 'f1-score': 1.0, 'support': 4490.0}, 'weighted avg':
{'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0, 'support': 4490.0}}
Test Classification Report:
{'0': {'precision': 0.9995249406175772, 'recall': 0.9995249406175772,
'f1-score': 0.9995249406175772, 'support': 2105.0}, '1': {'precision':
0.99958071278826, 'recall': 0.99958071278826, 'f1-score':
```

```
0.99958071278826, 'support': 2385.0}, 'accuracy': 0.999554565701559,
'macro avg': {'precision': 0.9995528267029186, 'recall':
0.9995528267029186, 'f1-score': 0.9995528267029186, 'support':
4490.0}, 'weighted avg': {'precision': 0.999554565701559, 'recall':
0.999554565701559, 'f1-score': 0.999554565701559, 'support': 4490.0}}
Precision-Recall AUC Scores:
Training PR AUC: 1.0000
Validation PR AUC: 1.0000
Test PR AUC: 1.0000
==================================================
Loading model: FacebookAI/xlm-roberta-base
```

```
{"model_id":"901453777c374d6ba40c9ed5427409f8","version_major":2,"version_minor":0}

{"model_id":"30e8cf0e933f42b9baff3bf0cb4792f7","version_major":2,"version_minor":0}

{"model_id":"5f14182455624801b48845c4f66b1e9e","version_major":2,"version_minor":0}

{"model_id":"a0acd66954e64e5cb8c57d255002c29e","version_major":2,"version_minor":0}

{"model_id":"4ae36abe26ac4b10a83a871b578af4a5","version_major":2,"version_minor":0}
```

```
Some weights of XLMRobertaForSequenceClassification were not
initialized from the model checkpoint at FacebookAI/xlm-roberta-base
and are newly initialized: ['classifier.dense.bias',
'classifier.dense.weight', 'classifier.out_proj.bias',
'classifier.out_proj.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
{"model_id":"c04f9d6c07bf40a1925c1e9d3942279b","version_major":2,"version_minor":0}

{"model_id":"50b309f8cc014b9aac334684cdb6a355","version_major":2,"version_minor":0}

{"model_id":"d4622c2cadc4468c9d94b9ef674ec389","version_major":2,"version_minor":0}
```

[4490/4490 25:57, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Re
1	0.003000	0.002192	0.999777	0.999997	{'0': {'precision': 0.9995333644423 'recall': 1.0, 'f1-sco 0.9997666277712 'support': 2142.0}, {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy': 0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.9997870528109 'f1-score': 0.9997768176130 'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}
2	0.002000	0.001886	0.999777	0.999999	{'0': {'precision': 0.9995333644423 'recall': 1.0, 'f1-sco 0.9997666277712 'support': 2142.0}, {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy': 0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.9997870528109

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Rep
					'f1-score': 0.99977681761301 'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}

```
Final Classification Report for FacebookAI/xlm-roberta-base:

Training Classification Report:
{'0': {'precision': 0.9996506346803308, 'recall': 0.999883517763541,
'f1-score': 0.9997670626601444, 'support': 17170.0}, '1':
{'precision': 0.9998932991890739, 'recall': 0.9996799658630254, 'f1-
score': 0.9997866211458445, 'support': 18748.0}, 'accuracy':
0.9997772704493568, 'macro avg': {'precision': 0.9997719669347023,
'recall': 0.9997817418132833, 'f1-score': 0.9997768419029944,
'support': 35918.0}, 'weighted avg': {'precision': 0.9997772974736354,
'recall': 0.9997772704493568, 'f1-score': 0.9997772715384201,
'support': 35918.0}}
Validation Classification Report:
{'0': {'precision': 0.9995333644423705, 'recall': 1.0, 'f1-score':
0.9997666277712952, 'support': 2142.0}, '1': {'precision': 1.0,
'recall': 0.9995741056218058, 'f1-score': 0.999787007454739,
'support': 2348.0}, 'accuracy': 0.9997772828507795, 'macro avg':
{'precision': 0.9997666822211853, 'recall': 0.999787052810903, 'f1-
score': 0.9997768176130171, 'support': 4490.0}, 'weighted avg':
{'precision': 0.9997773867785207, 'recall': 0.9997772828507795, 'f1-
score': 0.9997772851202321, 'support': 4490.0}}
Test Classification Report:
{'0': {'precision': 0.9995251661918328, 'recall': 1.0, 'f1-score':
0.9997625267157445, 'support': 2105.0}, '1': {'precision': 1.0,
'recall': 0.99958071278826, 'f1-score': 0.9997903124344726, 'support':
2385.0}, 'accuracy': 0.9997772828507795, 'macro avg': {'precision':
0.9997625830959165, 'recall': 0.9997903563941299, 'f1-score':
0.9997764195751085, 'support': 4490.0}, 'weighted avg': {'precision':
0.9997773886044118, 'recall': 0.9997772828507795, 'f1-score':
0.9997772859449574, 'support': 4490.0}}
Precision-Recall AUC Scores:
Training PR AUC: 1.0000
Validation PR AUC: 1.0000
Test PR AUC: 1.0000
==================================================
Loading model: jy46604790/Fake-News-Bert-Detect
```

```
{"model_id":"4b9415ca48074d6e85d70ec6d559baac","version_major":2,"version_minor":0}

{"model_id":"66dc5c37cd4447599b9b8fb8136a7036","version_major":2,"version_minor":0}

{"model_id":"363fe3d3deb342fb9708702dc977c106","version_major":2,"version_minor":0}

{"model_id":"6358ee35a5d64fafb8b6a2f0cf7fbde6","version_major":2,"version_minor":0}

{"model_id":"47294f3815874e3f80d352b7f28b9e12","version_major":2,"version_minor":0}

{"model_id":"49d0482ea86f4f47b2d0b55217dfdfb2","version_major":2,"version_minor":0}

{"model_id":"5b000a51cbc940deab2514c2789a384e","version_major":2,"version_minor":0}

{"model_id":"30924069036145e3be4dfbf789a7184d","version_major":2,"version_minor":0}

{"model_id":"31c6f97461464c2f93e04654f96873b9","version_major":2,"version_minor":0}

{"model_id":"a692c00df5ef4597b891a745df7e231c","version_major":2,"version_minor":0}
```

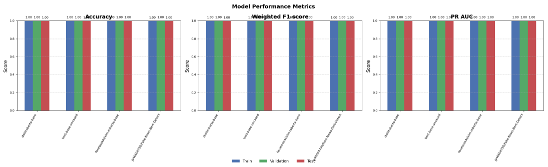
[4490/4490 24:50, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Rep
1	0.005400	0.002002	0.999777	1.000000	{'0': {'precision': 0.9995333644423 'recall': 1.0, 'f1-sco 0.9997666277712 'support': 2142.0}, {'precision': 1.0, 'recall': 0.9995741056218 'f1-score': 0.9997870074547 'support': 2348.0} 'accuracy':

Epoch	Training Loss	Validation Loss	Accuracy	Pr Auc	Classifica Re
					0.9997772828507 'macro avg': {'precision': 0.9997666822211 'recall': 0.9997870528109 'f1-score': 0.9997768176130 'support': 4490.0} 'weighted avg': {'precision': 0.9997773867785 'recall': 0.9997772828507 'f1-score': 0.9997772851202 'support': 4490.0}
2	0.001900	0.002233	0.999555	0.999999	{'0': {'precision': 0.9995331465919 'recall': 0.9995331465919 'f1-score': 0.9995331465919 'support': 2142.0}, {'precision': 0.999574105621 'recall': 0.999574105621 'f1-score': 0.999574105621 'support': 2348.0} 'accuracy': 0.999554565701 'macro avg': {'precision': 0.9995536261068 'recall': 0.9995536261068 'f1-score': 0.9995536261068 'support': 4490.0} 'weighted avg': {'precision': 0.999554565701 'recall': 0.999554565701 'f1-score': 0.999554565701 'support': 4490.0}

```
Final Classification Report for jy46604790/Fake-News-Bert-Detect:

Training Classification Report:
{'0': {'precision': 0.9995339081799114, 'recall': 0.9991846243447874,
'f1-score': 0.9993592357429952, 'support': 17170.0}, '1':
{'precision': 0.9992534925882478, 'recall': 0.9995732878173672, 'f1-
score': 0.9994133646205535, 'support': 18748.0}, 'accuracy':
0.9993874937357314, 'macro avg': {'precision': 0.9993937003840796,
'recall': 0.9993789560810773, 'f1-score': 0.9993863001817744,
'support': 35918.0}, 'weighted avg': {'precision': 0.9993875405783603,
'recall': 0.9993874937357314, 'f1-score': 0.9993874892146934,
'support': 35918.0}}
Validation Classification Report:
{'0': {'precision': 0.9995333644423705, 'recall': 1.0, 'f1-score':
0.9997666277712952, 'support': 2142.0}, '1': {'precision': 1.0,
'recall': 0.9995741056218058, 'f1-score': 0.999787007454739,
'support': 2348.0}, 'accuracy': 0.9997772828507795, 'macro avg':
{'precision': 0.9997666822211853, 'recall': 0.999787052810903, 'f1-
score': 0.9997768176130171, 'support': 4490.0}, 'weighted avg':
{'precision': 0.9997773867785207, 'recall': 0.9997772828507795, 'f1-
score': 0.9997772851202321, 'support': 4490.0}}
Test Classification Report:
{'0': {'precision': 0.9990498812351544, 'recall': 0.9990498812351544,
'f1-score': 0.9990498812351544, 'support': 2105.0}, '1': {'precision':
0.9991614255765199, 'recall': 0.9991614255765199, 'f1-score':
0.9991614255765199, 'support': 2385.0}, 'accuracy':
0.9991091314031181, 'macro avg': {'precision': 0.9991056534058371,
'recall': 0.9991056534058371, 'f1-score': 0.9991056534058371,
'support': 4490.0}, 'weighted avg': {'precision': 0.9991091314031181,
'recall': 0.9991091314031181, 'f1-score': 0.9991091314031181,
'support': 4490.0}}
```

```
Precision-Recall AUC Scores:
Training PR AUC: 1.0000
Validation PR AUC: 1.0000
Test PR AUC: 1.0000
==================================================
distilroberta-base:
Final Training Accuracy: 0.9998
Final Validation Accuracy: 0.9998
Final Precision-Recall AUC (Test): 1.0000
Classification Report (Test):
{'0': {'precision': 0.9995249406175772, 'recall': 0.9995249406175772,
'f1-score': 0.9995249406175772, 'support': 2105.0}, '1': {'precision':
0.99958071278826, 'recall': 0.99958071278826, 'f1-score':
0.99958071278826, 'support': 2385.0}, 'accuracy': 0.999554565701559,
'macro avg': {'precision': 0.9995528267029186, 'recall':
0.9995528267029186, 'f1-score': 0.9995528267029186, 'support':
4490.0}, 'weighted avg': {'precision': 0.999554565701559, 'recall':
0.999554565701559, 'f1-score': 0.999554565701559, 'support': 4490.0}}
==================================================
bert-base-uncased:
Final Training Accuracy: 1.0000
Final Validation Accuracy: 1.0000
Final Precision-Recall AUC (Test): 1.0000
Classification Report (Test):
{'0': {'precision': 0.9995249406175772, 'recall': 0.9995249406175772,
'f1-score': 0.9995249406175772, 'support': 2105.0}, '1': {'precision':
0.99958071278826, 'recall': 0.99958071278826, 'f1-score':
0.99958071278826, 'support': 2385.0}, 'accuracy': 0.999554565701559,
'macro avg': {'precision': 0.9995528267029186, 'recall':
0.9995528267029186, 'f1-score': 0.9995528267029186, 'support':
4490.0}, 'weighted avg': {'precision': 0.999554565701559, 'recall':
0.999554565701559, 'f1-score': 0.999554565701559, 'support': 4490.0}}
==================================================
FacebookAI/xlm-roberta-base:
Final Training Accuracy: 0.9998
Final Validation Accuracy: 0.9998
Final Precision-Recall AUC (Test): 1.0000
Classification Report (Test):
{'0': {'precision': 0.9995251661918328, 'recall': 1.0, 'f1-score':
0.9997625267157445, 'support': 2105.0}, '1': {'precision': 1.0,
'recall': 0.99958071278826, 'f1-score': 0.9997903124344726, 'support':
2385.0}, 'accuracy': 0.9997772828507795, 'macro avg': {'precision':
0.9997625830959165, 'recall': 0.9997903563941299, 'f1-score':
0.9997764195751085, 'support': 4490.0}, 'weighted avg': {'precision':
0.9997773886044118, 'recall': 0.9997772828507795, 'f1-score':
0.9997772859449574, 'support': 4490.0}}
==================================================
jy46604790/Fake-News-Bert-Detect:
Final Training Accuracy: 0.9998
Final Validation Accuracy: 0.9998
Final Precision-Recall AUC (Test): 1.0000
Classification Report (Test):
{'0': {'precision': 0.9990498812351544, 'recall': 0.9990498812351544,
'f1-score': 0.9990498812351544, 'support': 2105.0}, '1': {'precision':
0.9991614255765199, 'recall': 0.9991614255765199, 'f1-score':
0.9991614255765199, 'support': 2385.0}, 'accuracy':
0.9991091314031181, 'macro avg': {'precision': 0.9991056534058371,
'recall': 0.9991056534058371, 'f1-score': 0.9991056534058371,
'support': 4490.0}, 'weighted avg': {'precision': 0.9991091314031181,
'recall': 0.9991091314031181, 'f1-score': 0.9991091314031181,
'support': 4490.0}}
==================================================

                              Model  Accuracy_train  Accuracy_val  \
0                distilroberta-base        0.999972      0.999777
1                 bert-base-uncased        0.999944      1.000000
2       FacebookAI/xlm-roberta-base        0.999777      0.999777
3  jy46604790/Fake-News-Bert-Detect        0.999387      0.999777

   Accuracy_test  Weighted_F1_train  Weighted_F1_val  Weighted_F1_test
\
0       0.999555           0.999972         0.999777          0.999555
1       0.999555           0.999944         1.000000          0.999555
2       0.999777           0.999777         0.999777          0.999777
3       0.999109           0.999387         0.999777          0.999109

   PR_AUC_train  PR_AUC_val  PR_AUC_test
0      1.000000    1.000000     1.000000
1      1.000000    1.000000     1.000000
2      0.999999    0.999999     0.999999
3      1.000000    1.000000     1.000000
```



Model Performance Results

The following table presents the performance metrics for different models evaluated on training, validation, and test datasets. The metrics include Accuracy, Weighted F1-Score, and Precision-Recall AUC (PR AUC).

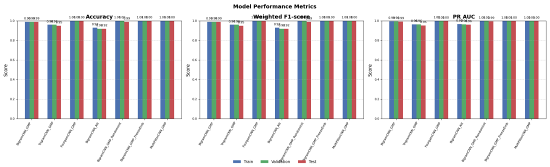
Model	Accuracy (Train)	Accuracy (Val)	Accuracy (Test)	Weighted F1 (Train)	Weighted F1 (Val)
distilroberta-base	0.999972	0.999777	0.999555	0.999972	0.999777
bert-base-uncased	0.999944	1.000000	0.999555	0.999944	1.000000
FacebookAI/xlm-roberta-base	0.999777	0.999777	0.999777	0.999777	0.999777
jy46604790/Fake-News-Bert-Detect	0.999387	0.999777	0.999109	0.999387	0.999777

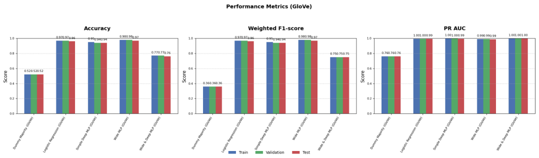
Observations:

- **Accuracy**: All models perform excellently across the datasets, with accuracy values consistently near 1. The `bert-base-uncased` model achieves perfect accuracy on the validation set.
- **Weighted F1-Score**: The F1-scores are very high for all models, showing strong balance between precision and recall. The `bert-base-uncased` model stands out with a perfect F1-score on the validation set.
- **PR AUC**: All models have perfect PR AUC values for training, validation, and test sets, indicating exceptional model performance in distinguishing between classes, with no false positives or false negatives across all datasets.

Conclusion:

These models, particularly `bert-base-uncased` and `distilroberta-base`, demonstrate outstanding performance across all evaluated metrics. The results indicate excellent generalization to unseen data, especially on the validation and test sets.





Exercise 2

- Repeat Exercise 3 of Part 5 (POS tagger), by fine-tuning a pre-trained BERT model.

- Tune the hyper-parameters on the development subset of your dataset.

- Monitor the performance of your models on the development subset during training to decide how many epochs to use.

- If the sentences of your experiments exceed BERT's maximum length limit, you may want to truncate them at the maximum allowed length of BERT or use a BERT-like model that can handle longer texts (e.g., Longformer).

- Include experimental results of a baseline that tags each word with the most frequent tag it had in the training data; for words that were not encountered in the training data, the baseline should return the most frequent tag (over all words) of the training data.

- Also include experimental results of your best method from exercise 10 of Part 3, exercise 2 of Part 4, exercise 3 of Part 5, now treated as additional baselines.
- Otherwise, the contents of your report should be as in exercise 3 of Part 5, but now with information and results for the experiments of this exercise.
- You may optionally include (for extra bonus) indicative experimental results on a small subset of the test set (e.g., 10 test examples) obtained by prompting an LLM (e.g., Chat-GPT), using appropriate instructions and possibly including few-shot examples (demonstrators).

Imports and Pip Installs

```
Mon Mar 10 18:29:38 2025
+-----------------------------------------------------------------------
+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15
CUDA Version: 12.4     |
|-----------------------------------------+------------------------
+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A |
Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage |
GPU-Util  Compute M. |
|                                         |                        |
MIG M. |
|
========================================+========================+===
|   0  Tesla T4                       Off |   00000000:00:04.0 Off |
0 |
| N/A   54C    P8             10W /   70W |       0MiB /  15360MiB |
0%      Default |
|                                         |                        |
N/A |
+-----------------------------------------+------------------------
+----------------------+

+-----------------------------------------------------------------------
+
| Processes:
|
|  GPU   GI   CI        PID   Type   Process name
GPU Memory |
|        ID   ID
Usage      |
|
=======================================================================
|  No running processes found
|
+-----------------------------------------------------------------------
+
```

Data Download

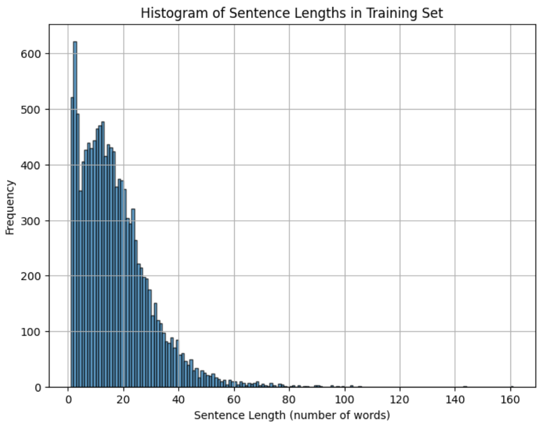
```
Downloading en_ewt-ud-train.conllu...
Downloaded en_ewt-ud-train.conllu
Downloading en_ewt-ud-dev.conllu...
Downloaded en_ewt-ud-dev.conllu
Downloading en_ewt-ud-test.conllu...
Downloaded en_ewt-ud-test.conllu
```

Data Parsing

```
[('This', 'PRON'),
 ('is', 'AUX'),
 ('not', 'PART'),
 ('a', 'DET'),
 ('post', 'NOUN'),
 ('about', 'ADP'),
 ('fault', 'NOUN'),
 ('-', 'PUNCT'),
 ('finding', 'NOUN'),
 ('or', 'CCONJ'),
 ('assigning', 'VERB'),
 ('blame', 'NOUN'),
 ('.', 'PUNCT')]
```

Exploratory Data Analysis

```
Mean train sentence length:  16.520248724489797
```



Downloading the equivalent dataset from Hugging Face

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/
_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your
settings tab (https://huggingface.co/settings/tokens), set it as
secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to
access public models or datasets.
  warnings.warn(
```

```
{"model_id":"a3766d3b4ea845f995d7b5da76879730","version_major":2,"version_minor":0}

{"model_id":"85b7e70b148349c5b05c1ad27680ddeb","version_major":2,"version_minor":0}

{"model_id":"d7720e13e8c947ca81cadfb6e0605d8c","version_major":2,"version_minor":0}

{"model_id":"5f6a133a16ad4c4584208d1a9c8629ec","version_major":2,"version_minor":0}

{"model_id":"6c1e10703813400899e097b3123de217","version_major":2,"version_minor":0}

{"model_id":"c736340ca601432b91475f09ce4162a5","version_major":2,"version_minor":0}

{"model_id":"9ee303b4c1dc4fb9a8bf45309a7cc60c","version_major":2,"version_minor":0}

{"model_id":"422d119a313d4d94b32d12b464112f46","version_major":2,"version_minor":0}
```

```
DatasetDict({
    train: Dataset({
        features: ['idx', 'text', 'tokens', 'lemmas', 'upos', 'xpos',
'feats', 'head', 'deprel', 'deps', 'misc'],
        num_rows: 12543
    })
    validation: Dataset({
        features: ['idx', 'text', 'tokens', 'lemmas', 'upos', 'xpos',
'feats', 'head', 'deprel', 'deps', 'misc'],
        num_rows: 2002
    })
    test: Dataset({
        features: ['idx', 'text', 'tokens', 'lemmas', 'upos', 'xpos',
'feats', 'head', 'deprel', 'deps', 'misc'],
        num_rows: 2077
    })
})
```

```
[('The', 'DET'),
 ('third', 'ADJ'),
 ('was', 'AUX'),
 ('being', 'AUX'),
 ('run', 'VERB'),
 ('by', 'ADP'),
 ('the', 'DET'),
 ('head', 'NOUN'),
 ('of', 'ADP'),
 ('an', 'DET'),
 ('investment', 'NOUN'),
 ('firm', 'NOUN'),
 ('.', 'PUNCT')]
```

```
{'idx': 'weblog-
juancole.com_juancole_20051126063000_ENG_20051126_063000-0006',
 'text': 'The third was being run by the head of an investment firm.',
 'tokens': ['The',
  'third',
  'was',
  'being',
  'run',
  'by',
  'the',
  'head',
  'of',
  'an',
  'investment',
  'firm',
  '.'],
 'lemmas': ['the',
  'third',
  'be',
  'be',
  'run',
  'by',
  'the',
  'head',
  'of',
  'a',
  'investment',
  'firm',
  '.'],
 'upos': [8, 6, 17, 17, 16, 2, 8, 0, 2, 8, 0, 0, 1],
 'xpos': ['DT',
  'JJ',
  'VBD',
  'VBG',
  'VBN',
  'IN',
  'DT',
  'NN',
  'IN',
  'DT',
  'NN',
  'NN',
  '.'],
 'feats': ["{'Definite': 'Def', 'PronType': 'Art'}",
  "{'Degree': 'Pos', 'NumType': 'Ord'}",
  "{'Mood': 'Ind', 'Number': 'Sing', 'Person': '3', 'Tense': 'Past',
'VerbForm': 'Fin'}",
  "{'VerbForm': 'Ger'}",
  "{'Tense': 'Past', 'VerbForm': 'Part', 'Voice': 'Pass'}",
  'None',
  "{'Definite': 'Def', 'PronType': 'Art'}",
  "{'Number': 'Sing'}",
  'None',
  "{'Definite': 'Ind', 'PronType': 'Art'}",
  "{'Number': 'Sing'}",
  "{'Number': 'Sing'}",
  'None'],
 'head': ['2', '5', '5', '5', '0', '8', '8', '5', '12', '12', '12',
'8', '5'],
 'deprel': ['det',
  'nsubj:pass',
  'aux',
  'aux:pass',
  'root',
  'case',
  'det',
  'obl',
  'case',
  'det',
  'compound',
  'nmod',
  'punct'],
 'deps': ["[('det', 2)]",
  "[('nsubj:pass', 5)]",
  "[('aux', 5)]",
  "[('aux:pass', 5)]",
  "[('root', 0)]",
  "[('case', 8)]",
  "[('det', 8)]",
  "[('obl:by', 5)]",
  "[('case', 12)]",
  "[('det', 12)]",
  "[('compound', 12)]",
  "[('nmod:of', 8)]",
  "[('punct', 5)]"],
 'misc': ['None',
  'None',
  'None',
  'None',
  'None',
  'None',
  'None',
  'None',
  'None',
  'None',
  'None',
  "{'SpaceAfter': 'No'}",
  'None']}
```

```
[('The', 'DT', 8),
 ('third', 'JJ', 6),
 ('was', 'VBD', 17),
 ('being', 'VBG', 17),
 ('run', 'VBN', 16),
 ('by', 'IN', 2),
 ('the', 'DT', 8),
 ('head', 'NN', 0),
 ('of', 'IN', 2),
 ('an', 'DT', 8),
 ('investment', 'NN', 0),
 ('firm', 'NN', 0),
 ('.', '.', 1)]
```

```
Xpos:  NNP  has the following upos:  {0, 10, 12, 6}
Xpos:  HYPH  has the following upos:  {1}
Xpos:  :  has the following upos:  {1}
Xpos:  JJ  has the following upos:  {0, 6, 10, 14, 15}
Xpos:  NNS  has the following upos:  {0, 16}
Xpos:  VBD  has the following upos:  {16, 17}
Xpos:  ,  has the following upos:  {1, 4}
Xpos:  DT  has the following upos:  {8, 11, 6}
Xpos:  NN  has the following upos:  {0, 3, 4, 6, 10, 11, 12, 14, 15}
Xpos:  IN  has the following upos:  {2, 4, 5, 14}
Xpos:  .  has the following upos:  {1}
Xpos:  -LRB-  has the following upos:  {1}
Xpos:  MD  has the following upos:  {17}
Xpos:  VB  has the following upos:  {16, 17}
Xpos:  VBG  has the following upos:  {16, 17, 0, 6}
Xpos:  PRP  has the following upos:  {11}
Xpos:  TO  has the following upos:  {2, 7}
Xpos:  -RRB-  has the following upos:  {1}
Xpos:  VBN  has the following upos:  {16, 17}
Xpos:  RP  has the following upos:  {2, 14}
Xpos:  CD  has the following upos:  {3}
Xpos:  VBZ  has the following upos:  {16, 17}
Xpos:  RB  has the following upos:  {2, 6, 14, 7}
Xpos:  NNPS  has the following upos:  {10}
Xpos:  VBP  has the following upos:  {16, 17}
Xpos:  PRP$  has the following upos:  {11}
Xpos:  CC  has the following upos:  {9, 2, 14}
Xpos:  None  has the following upos:  {13}
Xpos:  WP  has the following upos:  {11}
Xpos:  EX  has the following upos:  {11}
Xpos:  WDT  has the following upos:  {8, 11}
Xpos:  RBR  has the following upos:  {14}
Xpos:  PDT  has the following upos:  {8, 1}
Xpos:  JJR  has the following upos:  {6}
Xpos:  WRB  has the following upos:  {5, 14}
Xpos:  JJS  has the following upos:  {6}
Xpos:  ``  has the following upos:  {1}
Xpos:  ''  has the following upos:  {1}
Xpos:  POS  has the following upos:  {7}
Xpos:  RBS  has the following upos:  {14}
Xpos:  WP$  has the following upos:  {11}
Xpos:  ADD  has the following upos:  {12}
Xpos:  FW  has the following upos:  {12}
Xpos:  LS  has the following upos:  {12}
Xpos:  UH  has the following upos:  {0, 4, 15}
Xpos:  AFX  has the following upos:  {12}
Xpos:  $  has the following upos:  {4}
Xpos:  NFP  has the following upos:  {1, 4}
Xpos:  SYM  has the following upos:  {4}
Xpos:  GW  has the following upos:  {0, 12, 5}
Xpos:  XX  has the following upos:  {12}
```

Xpos Appendix

1. **NNP** - Proper noun, singular (e.g., "John")
2. **HYPH** - Hyphen (used in hyphenated words)
3. **:** - Punctuation mark (colon)
4. **JJ** - Adjective (e.g., "beautiful")
5. **NNS** - Noun, plural (e.g., "dogs")
6. **VBD** - Verb, past tense (e.g., "walked")
7. **,** - Comma
8. **DT** - Determiner (e.g., "the", "a")
9. **NN** - Noun, singular (e.g., "dog")

10. **IN** - Preposition or subordinating conjunction (e.g., "in", "on", "because")
11. **.** - Period (full stop)
12. **-LRB-** - Left round bracket (open parenthesis)
13. **MD** - Modal verb (e.g., "can", "will")
14. **VB** - Verb, base form (e.g., "run")
15. **VBG** - Verb, gerund or present participle (e.g., "running")
16. **PRP** - Personal pronoun (e.g., "I", "he", "she")
17. **TO** - To (used for the infinitive form of a verb, e.g., "to run")
18. **-RRB-** - Right round bracket (close parenthesis)
19. **VBN** - Verb, past participle (e.g., "eaten")
20. **RP** - Particle (e.g., "up" in "give up")
21. **CD** - Cardinal number (e.g., "one", "two", "three")
22. **VBZ** - Verb, 3rd person singular present (e.g., "runs")
23. **RB** - Adverb (e.g., "quickly")
24. **NNPS** - Proper noun, plural (e.g., "Americas")
25. **VBP** - Verb, non-3rd person singular present (e.g., "run")
26. **PRP$** - Possessive pronoun (e.g., "my", "his")
27. **CC** - Coordinating conjunction (e.g., "and", "or")
28. **None** - Used for tokens without a specific UPOS tag
29. **WP** - Wh-pronoun (e.g., "who", "what")
30. **EX** - Existential there (e.g., "there is")
31. **WDT** - Wh-determiner (e.g., "which")
32. **RBR** - Adverb, comparative (e.g., "better")
33. **PDT** - Predeterminer (e.g., "all", "both")
34. **JJR** - Adjective, comparative (e.g., "bigger")
35. **WRB** - Wh-adverb (e.g., "where", "when")
36. **JJS** - Adjective, superlative (e.g., "biggest")
37. `` - Opening quotation mark (e.g., "quote)
38. **''** - Closing quotation mark (e.g., "quote")
39. **POS** - Possessive ending (e.g., "John's")
40. **RBS** - Adverb, superlative (e.g., "best")
41. **WP$** - Possessive wh-pronoun (e.g., "whose")
42. **ADD** - Additional word (usually for things like "etc." or "and so on")
43. **FW** - Foreign word (e.g., "pizza")
44. **LS** - List item marker (e.g., "1.", "a)")
45. **UH** - Interjection (e.g., "wow", "ouch")
46. **AFX** - Affix (e.g., "-ing", "-ly")
47. **$** - Dollar sign (e.g., "$10")
48. **NFP** - Non-final punctuation (e.g., semicolons, ellipses)
49. **SYM** - Symbol (e.g., "%", "&")
50. **GW** - Discourse marker or "general word" (used for words like "thing" or "stuff")
51. **XX** - Unknown word or symbol (e.g., non-lexical tokens)

We can see that the Hugging Face Universal Dependencies dataset contains a whole lot more of POS Tags than the one from the Github repository.

Data Preprocessing

```
DatasetDict({
    train: Dataset({
        features: ['sentence', 'pos_tags'],
        num_rows: 12544
    })
    dev: Dataset({
        features: ['sentence', 'pos_tags'],
        num_rows: 2001
    })
    test: Dataset({
        features: ['sentence', 'pos_tags'],
        num_rows: 2077
    })
})
```

```
Unique tags: {'ADP', 'NUM', 'X', 'VERB', 'INTJ', 'PRON', 'PUNCT',
'PROPN', 'NOUN', 'SCONJ', '_', 'CCONJ', 'ADV', 'AUX', 'PART', 'ADJ',
'DET', 'SYM'}
Label to ID mapping: {'ADJ': 0, 'ADP': 1, 'ADV': 2, 'AUX': 3, 'CCONJ':
4, 'DET': 5, 'INTJ': 6, 'NOUN': 7, 'NUM': 8, 'PART': 9, 'PRON': 10,
'PROPN': 11, 'PUNCT': 12, 'SCONJ': 13, 'SYM': 14, 'VERB': 15, 'X': 16,
'_': 17}
ID to Label mapping: {0: 'ADJ', 1: 'ADP', 2: 'ADV', 3: 'AUX', 4:
'CCONJ', 5: 'DET', 6: 'INTJ', 7: 'NOUN', 8: 'NUM', 9: 'PART', 10:
'PRON', 11: 'PROPN', 12: 'PUNCT', 13: 'SCONJ', 14: 'SYM', 15: 'VERB',
16: 'X', 17: '_'}
```

Converting POS Tag to label numbers.

```
{"model_id":"259033f317f94407aa51877de77be4be","version_major":2,"version_minor":0}
```

```
{"model_id":"c3e570fad4d343fb8e3c7e19b0b8e21f","version_major":2,"version_minor":0}
```

```
{"model_id":"9ff505d4d64f47489a7117e7ace0d09b","version_major":2,"version_minor":0}
```

```
{'sentence': ['A',
  'key',
  'question',
  'is',
  'how',
  'they',
  'acquired',
  'the',
  'anthrax',
  'strain',
  'first',
  'isolated',
  'by',
  'the',
  'Texas',
  'Veterinary',
  'Medical',
  'Diagnostic',
  'Lab',
  'in',
  '1980',
  '.'],
 'pos_tags': [5,
  0,
  7,
  3,
  2,
  10,
  15,
  5,
  7,
  7,
  2,
  15,
  1,
  5,
  11,
  0,
  0,
  0,
  11,
  1,
  8,
  12]}
```

Transformers Training and Evaluation

```
# @title
model, tokenizer, trainer = train_pos_tagger("xlm-roberta-base")
```

```
Some weights of XLMRobertaForTokenClassification were not initialized
from the model checkpoint at xlm-roberta-base and are newly
initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
{"model_id":"ae0a4983b3604fe8a2a64ef1ae069a98","version_major":2,"version_minor":0}
```

```
{"model_id":"d9fa0f78aeef4ba9a2422aadc50932c9","version_major":2,"version_minor":0}
```

```
{"model_id":"87943d588cbb4c1890f912c86c37e929","version_major":2,"version_minor":0}
```

[784/784 09:46, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accu
1	0.110100	0.117467	0.968485	0.968485	0.968485	0.968

```
Classification Report
              precision    recall  f1-score   support

         ADJ       0.94      0.95      0.95      1794
         ADP       0.97      0.99      0.98      2030
         ADV       0.94      0.95      0.94      1183
         AUX       0.99      0.99      0.99      1543
       CCONJ       1.00      0.99      0.99       736
         DET       0.99      0.99      0.99      1896
        INTJ       0.86      0.88      0.87       121
        NOUN       0.96      0.95      0.96      4123
         NUM       0.94      0.99      0.97       542
        PART       1.00      1.00      1.00       649
        PRON       0.99      0.99      0.99      2166
       PROPN       0.92      0.92      0.92      2076
       PUNCT       1.00      1.00      1.00      3096
       SCONJ       0.97      0.94      0.96       384
         SYM       0.84      0.86      0.85       109
        VERB       0.98      0.98      0.98      2606
```

```
          X       1.00      0.00      0.00        42
          _       0.98      0.97      0.98       354

    accuracy                          0.97     25450
   macro avg       0.96      0.91      0.91     25450
weighted avg       0.97      0.97      0.97     25450

AUC scores per tag (class):
ADJ: 0.9977
ADP: 0.9995
ADV: 0.9988
AUX: 0.9997
CCONJ: 0.9997
DET: 0.9996
INTJ: 0.9931
NOUN: 0.9978
NUM: 0.9998
PART: 0.9995
PRON: 0.9998
PROPN: 0.9971
PUNCT: 1.0000
SCONJ: 0.9984
SYM: 0.9991
VERB: 0.9995
X: 0.9292
_: 0.9998
```

```
# @title
model, tokenizer, trainer = train_pos_tagger("bert-base-uncased")
```

{"model_id":"3db8c65a3cb74adb9dc2731929fd5b65","version_major":2,"version_minor":0}

{"model_id":"31ff3996f9c54ad98a44e0775806bff9","version_major":2,"version_minor":0}

{"model_id":"77884c0ca3d0443289492b5aa1d8248a","version_major":2,"version_minor":0}

{"model_id":"15cb8f26fbef402db08ed38d577e80e0","version_major":2,"version_minor":0}

{"model_id":"0be12a88a3d34ed38aa580bc54bf3c55","version_major":2,"version_minor":0}

```
Some weights of BertForTokenClassification were not initialized from
the model checkpoint at bert-base-uncased and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

{"model_id":"858fab6f12e64377bd536075d5783ed7","version_major":2,"version_minor":0}

{"model_id":"a8116d8ed2764370b60c3a3f852ded9d","version_major":2,"version_minor":0}

{"model_id":"429c786e39104de189d28062d9242d82","version_major":2,"version_minor":0}

[784/784 07:05, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accu
1	0.122100	0.135049	0.964526	0.964526	0.964526	0.964

```
Classification Report
              precision    recall  f1-score   support

         ADJ       0.94      0.95      0.95      1794
         ADP       0.98      0.98      0.98      2030
         ADV       0.96      0.94      0.95      1183
         AUX       0.99      0.99      0.99      1543
       CCONJ       1.00      0.99      1.00       736
         DET       0.99      0.99      0.99      1896
        INTJ       0.94      0.84      0.89       121
        NOUN       0.93      0.95      0.94      4123
         NUM       0.94      0.98      0.96       542
        PART       0.99      0.99      0.99       649
        PRON       0.99      1.00      0.99      2166
       PROPN       0.92      0.88      0.90      2076
       PUNCT       0.99      0.99      0.99      3096
       SCONJ       0.95      0.96      0.96       384
         SYM       0.76      0.80      0.78       109
        VERB       0.98      0.99      0.98      2606
           X       1.00      0.00      0.00        42
           _       0.97      0.95      0.96       354

    accuracy                           0.97     25450
   macro avg       0.96      0.90      0.90     25450
weighted avg       0.97      0.97      0.96     25450

AUC scores per tag (class):
ADJ: 0.9976
ADP: 0.9997
ADV: 0.9985
```

```
AUX: 0.9995
CCONJ: 0.9999
DET: 0.9999
INTJ: 0.9948
NOUN: 0.9970
NUM: 0.9997
PART: 0.9999
PRON: 0.9996
PROPN: 0.9949
PUNCT: 0.9999
SCONJ: 0.9984
SYM: 0.9875
VERB: 0.9998
X: 0.8860
_: 0.9985
```

```
# @title
model, tokenizer, trainer = train_pos_tagger("t5-base")
```

```
{"model_id":"bf01290f353842d3a4fb0256a5972fe1","version_major":2,"version_minor":0}

{"model_id":"b1042510b53c4eb09ff93e8eae7d7b9f","version_major":2,"version_minor":0}

{"model_id":"93c4ad72e11c42d380bb2d9e7d73cdc2","version_major":2,"version_minor":0}

{"model_id":"3d0fb4aa825e4bc4b5253f162642fd93","version_major":2,"version_minor":0}

Some weights of T5ForTokenClassification were not initialized from the
model checkpoint at t5-base and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.

{"model_id":"b7602f2875174fd18a16cc4a0cec6b3f","version_major":2,"version_minor":0}

Asking to truncate to max_length but no maximum length is provided and
the model has no predefined maximum length. Default to no truncation.

{"model_id":"7627440794cf481dacd8f290f0b19b09","version_major":2,"version_minor":0}

{"model_id":"ad7947ee17ce4d79b35e41724e8d4738","version_major":2,"version_minor":0}
```

[784/784 08:39, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accur
1	2.528200	2.050722	0.473307	0.473307	0.473307	0.4733

```
There were missing keys in the checkpoint model loaded:
['transformer.encoder.embed_tokens.weight'].

Classification Report
              precision    recall  f1-score   support

         ADJ       0.31      0.15      0.20      1794
         ADP       0.21      0.90      0.34      2030
         ADV       0.13      0.04      0.07      1183
         AUX       0.09      0.01      0.02      1543
       CCONJ       0.32      0.07      0.11       736
         DET       0.64      0.35      0.45      1896
        INTJ       0.00      0.00      0.00       121
        NOUN       0.65      0.61      0.63      4123
         NUM       0.69      0.25      0.37       542
        PART       0.09      0.03      0.04       649
        PRON       0.67      0.61      0.64      2166
       PROPN       0.67      0.55      0.61      2076
       PUNCT       0.87      0.91      0.89      3096
       SCONJ       0.04      0.05      0.05       384
         SYM       0.00      0.00      0.00       109
        VERB       0.49      0.41      0.45      2606
           X       0.00      0.00      0.00        42
           _       0.00      0.00      0.00       354

    accuracy                           0.47     25450
   macro avg       0.33      0.28      0.27     25450
weighted avg       0.50      0.47      0.45     25450

AUC scores per tag (class):
ADJ: 0.6848
ADP: 0.8514
ADV: 0.7361
AUX: 0.7813
CCONJ: 0.8856
DET: 0.8716
INTJ: 0.5090
NOUN: 0.8751
NUM: 0.8401
PART: 0.8467
```

```
PRON: 0.9086
PROPN: 0.8904
PUNCT: 0.9801
SCONJ: 0.8377
SYM: 0.4686
VERB: 0.8477
X: 0.3508
_: 0.5602
```

```
# @title
model, tokenizer, trainer = train_pos_tagger("nlpaueb/bert-base-
        uncased-eurlex")  # pre-trained on EU legislation
```

```
{"model_id":"743130710f674e9d8f0bb50fb62986c3","version_major":2,"version_minor":0}

{"model_id":"527e517ae75f478186fb68bdb92b32dc","version_major":2,"version_minor":0}

{"model_id":"ec424fd32d7a4800bdbf62b8d0236d05","version_major":2,"version_minor":0}

{"model_id":"4739c470a41647c68891d17d2934564e","version_major":2,"version_minor":0}
```

```
Some weights of BertForTokenClassification were not initialized from
the model checkpoint at nlpaueb/bert-base-uncased-eurlex and are newly
initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
{"model_id":"d64748b0aa4a4b1a97076681ef1df835","version_major":2,"version_minor":0}

{"model_id":"3f285bf13b5847eb80e8b2f8f91ab445","version_major":2,"version_minor":0}

{"model_id":"3514e85eb6a14804ba24b0439887cb64","version_major":2,"version_minor":0}

{"model_id":"6f783bc028be4fd3bf09b1cefa56b3bd","version_major":2,"version_minor":0}
```

[784/784 06:48, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accur
1	0.133500	0.153926	0.959235	0.959235	0.959235	0.9592

```
Classification Report
              precision    recall  f1-score   support

         ADJ       0.93      0.94      0.94      1794
         ADP       0.97      0.98      0.97      2030
         ADV       0.95      0.93      0.94      1183
         AUX       0.99      0.99      0.99      1543
       CCONJ       1.00      0.99      0.99       736
         DET       0.99      1.00      0.99      1896
        INTJ       0.93      0.83      0.87       121
        NOUN       0.93      0.95      0.94      4123
         NUM       0.92      0.98      0.95       542
        PART       0.99      1.00      0.99       649
        PRON       0.99      0.99      0.99      2166
       PROPN       0.92      0.87      0.89      2076
       PUNCT       0.99      1.00      0.99      3096
       SCONJ       0.96      0.96      0.96       384
         SYM       0.83      0.78      0.81       109
        VERB       0.97      0.99      0.98      2606
           X       1.00      0.00      0.00        42
           _       0.98      0.94      0.96       354

    accuracy                           0.96     25450
   macro avg       0.96      0.89      0.90     25450
weighted avg       0.96      0.96      0.96     25450

AUC scores per tag (class):
ADJ: 0.9971
ADP: 0.9996
ADV: 0.9981
AUX: 0.9995
CCONJ: 0.9998
DET: 0.9999
INTJ: 0.9962
NOUN: 0.9959
NUM: 0.9980
PART: 0.9999
PRON: 0.9994
PROPN: 0.9934
PUNCT: 1.0000
SCONJ: 0.9989
SYM: 0.9991
VERB: 0.9993
X: 0.9075
_: 0.9985
```

```
# @title
```

```
model, tokenizer, trainer = train_pos_tagger("nlpaueb/sec-bert-base")
        # pre-trained on 260,773 10-K filings from 1993-2019
```

```
{"model_id":"9213d1cb8787455a85596d6b6e8e7ff8","version_major":2,"version_minor":0}

{"model_id":"0565fab17f014c93b831b7d23366c5d5","version_major":2,"version_minor":0}

{"model_id":"b0fdcaf41d424156a55e9f46434fd0e8","version_major":2,"version_minor":0}

{"model_id":"a72de74fdaba4df3be7b5a2d5e2b32dc","version_major":2,"version_minor":0}

{"model_id":"2f36639403eb45fc866a468e77374980","version_major":2,"version_minor":0}
```

```
Some weights of BertForTokenClassification were not initialized from
the model checkpoint at nlpaueb/sec-bert-base and are newly
initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
{"model_id":"664d999a367649589a5b493466616d7b","version_major":2,"version_minor":0}

{"model_id":"39f46803d1d4491b99be7195b7a86743","version_major":2,"version_minor":0}

{"model_id":"121604dc658f40e5a333f0562eb83086","version_major":2,"version_minor":0}

{"model_id":"0637ef351947400dbbc0762476eefce7","version_major":2,"version_minor":0}
```

[784/784 08:36, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accu
1	0.148700	0.166297	0.954649	0.954649	0.954649	0.954

```
Classification Report
              precision    recall  f1-score   support

         ADJ       0.91      0.92      0.92      1794
         ADP       0.97      0.97      0.97      2030
         ADV       0.93      0.91      0.92      1183
         AUX       0.99      0.99      0.99      1543
       CCONJ       1.00      0.99      1.00       736
         DET       0.99      0.99      0.99      1896
        INTJ       0.81      0.59      0.68       121
        NOUN       0.93      0.94      0.93      4123
         NUM       0.93      0.98      0.95       542
        PART       0.98      0.99      0.98       649
        PRON       0.99      0.99      0.99      2166
       PROPN       0.88      0.87      0.87      2076
       PUNCT       0.99      0.99      0.99      3096
       SCONJ       0.94      0.94      0.94       384
         SYM       0.80      0.83      0.81       109
        VERB       0.96      0.97      0.97      2606
           X       1.00      0.00      0.00        42
           _       0.97      0.95      0.96       354

    accuracy                           0.95     25450
   macro avg       0.94      0.88      0.88     25450
weighted avg       0.95      0.95      0.95     25450

AUC scores per tag (class):
ADJ: 0.9958
ADP: 0.9993
ADV: 0.9969
AUX: 0.9992
CCONJ: 0.9991
DET: 0.9998
INTJ: 0.9914
NOUN: 0.9952
NUM: 0.9994
PART: 0.9999
PRON: 0.9995
PROPN: 0.9926
PUNCT: 1.0000
SCONJ: 0.9979
SYM: 0.9992
VERB: 0.9991
X: 0.9203
_: 0.9996
```

Comparison Board

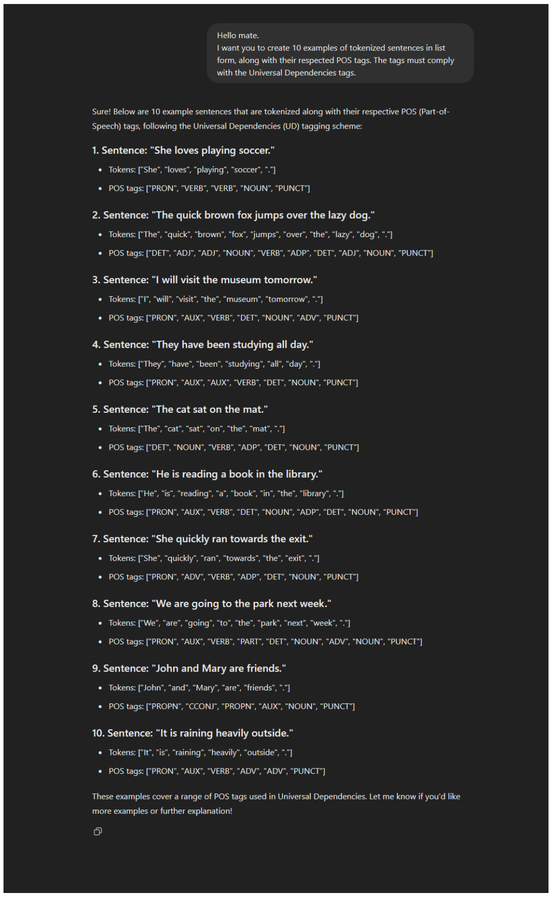
Model	Accuracy	Macro avg F1-score	Weighted avg F1-score	Strengths	V
ShallowPOS_MLP	0.83	0.74	0.83	High performance	S w

Model	Accuracy	Macro avg F1-score	Weighted avg F1-score	Strengths	V
				on **PRON (0.97 precision)**, **AUX (0.96 recall)**, and **VERB (0.91 f1-score)**.	**c C (**
DeepPOS_MLP	0.83	0.74	0.83	Improved recall for **AUX (0.94)** and **PRON (0.97)**, maintains strong performance across categories.	S v c **X**
VeryDeepPOS_MLP	0.82	0.74	0.82	Strong performance on **AUX** and **PRON**, but slightly lower overall performance compared to other MLP models.	S d p e **C (N f**
ShallowPOS_BiGRU	0.84	0.75	0.84	High performance on **PRON (0.97 precision)**, **AUX (0.97 recall)**, and **VERB (0.92 f1-score)**.	S v **c C (**
DeepPOS_BiGRU	0.84	0.75	0.84	High performance on **AUX (0.97)** and **PRON (0.97)**.	S v **C (**
VeryDeepPOS_BiGRU	0.84	0.76	0.84	Strong performance on **AUX (0.97)**, **PRON (0.98)**, and **VERB (0.92)** with a good balance across categories.	S d i p f **(N f** c t n
ShallowPOS_CNN	0.83	0.74	0.83	High performance on **AUX (0.96 recall)** and **PRON (0.97 precision)**, **PUNCT (0.99 recall)**.	S v **c C (**
DeepPOS_CNN	0.84	0.75	0.84	Good performance across most categories, especially **AUX (0.97)** and **VERB (0.92)**.	S v **c C (**

Model	Accuracy	Macro avg F1-score	Weighted avg F1-score	Strengths	V
VeryDeepPOS_CNN	0.82	0.73	0.82	Strong performance on **AUX (0.96)** and **PRON (0.97)**.	S s **C** **(** **N** **f**
Baseline Tagger	0.86	0.80	0.86	High accuracy and strong performance on **CCONJ (0.99 precision)** and **PUNCT (0.99 precision)**.	S v **c** **P** **(**
XLM-RoBERTa Base	0.97	0.91	0.97	Outstanding performance across most categories, especially **AUX (0.99)**, **PUNCT (1.00)**, and **CCONJ (0.99)**. Very high accuracy and AUC scores.	S v **c**
BERT Base Uncased	0.97	0.90	0.96	Strong performance across most categories, particularly **AUX (0.99)**, **CCONJ (1.00)**, and **PUNCT (0.99)**. Excellent AUC scores.	S v **c**
T5 Base	0.47	0.27	0.45	High precision for certain tags like **ADP (0.90 recall)** and **PUNCT (0.91 recall)**.	S s v c i **(** **I** **c** **A** **c**
NLPAueb BERT Base Uncased (EURLEX)	0.96	0.90	0.96	Strong performance across most categories, especially **AUX (0.99)**, **PUNCT (1.00)**, and **CCONJ (0.99)**. High AUC scores.	S v **c**
NLPAueb SEC BERT Base	0.95	0.88	0.95	Strong performance across most categories, particularly **AUX (0.99)**, **PUNCT (0.99)**, and **CCONJ**	S v **c** **I** **f**

Model	Accuracy	Macro avg F1-score	Weighted avg F1-score	Strengths	V
				(1.00). Very good AUC scores.	

Bonus Task



Hello mate.
I want you to create 10 examples of tokenized sentences in list form, along with their respected POS tags. The tags must comply with the Universal Dependencies tags.

Sure! Below are 10 example sentences that are tokenized along with their respective POS (Part-of-Speech) tags, following the Universal Dependencies (UD) tagging scheme:

1. Sentence: "She loves playing soccer."

- Tokens: ["She", "loves", "playing", "soccer", "."]
- POS tags: ["PRON", "VERB", "VERB", "NOUN", "PUNCT"]

2. Sentence: "The quick brown fox jumps over the lazy dog."

- Tokens: ["The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog", "."]
- POS tags: ["DET", "ADJ", "ADJ", "NOUN", "VERB", "ADP", "DET", "ADJ", "NOUN", "PUNCT"]

3. Sentence: "I will visit the museum tomorrow."

- Tokens: ["I", "will", "visit", "the", "museum", "tomorrow", "."]
- POS tags: ["PRON", "AUX", "VERB", "DET", "NOUN", "ADV", "PUNCT"]

4. Sentence: "They have been studying all day."

- Tokens: ["They", "have", "been", "studying", "all", "day", "."]
- POS tags: ["PRON", "AUX", "AUX", "VERB", "DET", "NOUN", "PUNCT"]

5. Sentence: "The cat sat on the mat."

- Tokens: ["The", "cat", "sat", "on", "the", "mat", "."]
- POS tags: ["DET", "NOUN", "VERB", "ADP", "DET", "NOUN", "PUNCT"]

6. Sentence: "He is reading a book in the library."

- Tokens: ["He", "is", "reading", "a", "book", "in", "the", "library", "."]
- POS tags: ["PRON", "AUX", "VERB", "DET", "NOUN", "ADP", "DET", "NOUN", "PUNCT"]

7. Sentence: "She quickly ran towards the exit."

- Tokens: ["She", "quickly", "ran", "towards", "the", "exit", "."]
- POS tags: ["PRON", "ADV", "VERB", "ADP", "DET", "NOUN", "PUNCT"]

8. Sentence: "We are going to the park next week."

- Tokens: ["We", "are", "going", "to", "the", "park", "next", "week", "."]
- POS tags: ["PRON", "AUX", "VERB", "PART", "DET", "NOUN", "ADV", "NOUN", "PUNCT"]

9. Sentence: "John and Mary are friends."

- Tokens: ["John", "and", "Mary", "are", "friends", "."]
- POS tags: ["PROPN", "CCONJ", "PROPN", "AUX", "NOUN", "PUNCT"]

10. Sentence: "It is raining heavily outside."

- Tokens: ["It", "is", "raining", "heavily", "outside", "."]
- POS tags: ["PRON", "AUX", "VERB", "ADV", "ADV", "PUNCT"]

These examples cover a range of POS tags used in Universal Dependencies. Let me know if you'd like more examples or further explanation!

```
Dataset({
    features: ['sentence', 'pos_tags'],
    num_rows: 10
})
```

```
# @title
prompt_dataset = prompt_dataset.map(lambda example: {'pos_tags':
        [label2id[tag] for tag in example['pos_tags']]})
```

{"model_id":"45433c611b394ddb8b1655413c880f37","version_major":2,"version_minor":0}

{"model_id":"74c8e3a3d043441283eae63881bae7b9","version_major":2,"version_minor":0}

[2/2 00:00]

	eval_loss	eval_model_preparation_time	eval_precision	eval_recall	eval_f1	eval_accuracy	eval_runtime	eval_s
0	0.256491	0.0048	0.945205	0.945205	0.945205	0.945205	0.0543	184.159

```
Classification Report
              precision    recall  f1-score   support

           0       0.75      1.00      0.86         3
           1       0.80      1.00      0.89         4
           2       1.00      0.60      0.75         5
           3       1.00      1.00      1.00         7
           4       1.00      1.00      1.00         1
           5       1.00      1.00      1.00        10
           7       0.92      0.92      0.92        13
           9       1.00      0.00      0.00         1
          10       1.00      1.00      1.00         7
          11       0.67      1.00      0.80         2
          12       1.00      1.00      1.00        10
          15       1.00      1.00      1.00        10
```

```
    accuracy                           0.95        73
   macro avg       0.93      0.88      0.85        73
weighted avg       0.96      0.95      0.94        73

AUC scores per tag (class):
ADJ: 0.9952
ADP: 0.9928
ADV: 0.9882
AUX: 1.0000
CCONJ: 1.0000
DET: 1.0000
INTJ: nan
NOUN: 0.9974
NUM: nan
PART: 1.0000
PRON: 1.0000
PROPN: 1.0000
PUNCT: 1.0000
SCONJ: nan
SYM: nan
VERB: 1.0000
X: nan
_: nan
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_ranking.py:379: UndefinedMetricWarning: Only one class is present in
y_true. ROC AUC score is not defined in that case.
  warnings.warn(
```

```
{"model_id":"b43d277780ae493e879aeb533ca8eae9","version_major":2,"version_minor":0}

{"model_id":"cac2dc1c049d453fac88c728a048d87a","version_major":2,"version_minor":0}

{"model_id":"6342ca83aefd40ad8e361b803108d00e","version_major":2,"version_minor":0}

{"model_id":"9e97428e61054cd0b8f3e05e8d556326","version_major":2,"version_minor":0}

{"model_id":"2956fc9cf24a48ab9df2480b8c4c13fc","version_major":2,"version_minor":0}

{"model_id":"f5e6241b010e4fcd9422ed6a71b8e399","version_major":2,"version_minor":0}
```

```
Device set to use cuda:0



Sentence:  She loves playing soccer .
['PRP', 'VBZ', 'VBG', 'NN', 'PUNCT']
['PRON', 'VERB', 'VERB', 'NOUN', 'PUNCT']

Sentence:  The quick brown fox jumps over the lazy dog .
['DT', 'JJ', 'JJ', 'NN', 'VBZ', 'IN', 'DT', 'JJ', 'NN', 'PUNCT']
['DET', 'ADJ', 'ADJ', 'NOUN', 'VERB', 'ADP', 'DET', 'ADJ', 'NOUN',
'PUNCT']

Sentence:  I will visit the museum tomorrow .
['NN', 'MD', 'VB', 'DT', 'NN', 'RB', 'PUNCT']
['PRON', 'AUX', 'VERB', 'DET', 'NOUN', 'ADV', 'PUNCT']

Sentence:  They have been studying all day .
['PRP', 'VBP', 'VBN', 'VBG', 'DT', 'NN', 'PUNCT']
['PRON', 'AUX', 'AUX', 'VERB', 'DET', 'NOUN', 'PUNCT']

Sentence:  The cat sat on the mat .
['DT', 'NN', 'VBD', 'IN', 'DT', 'NN', 'PUNCT']
['DET', 'NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'PUNCT']

Sentence:  He is reading a book in the library .
['PRP', 'VBZ', 'VBG', 'DT', 'NN', 'IN', 'DT', 'NN', 'PUNCT']
```

```
['PRON', 'AUX', 'VERB', 'DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'PUNCT']

Sentence:  She quickly ran towards the exit .
['PRP', 'RB', 'VBD', 'IN', 'DT', 'NN', 'PUNCT']
['PRON', 'ADV', 'VERB', 'ADP', 'DET', 'NOUN', 'PUNCT']

Sentence:  We are going to the park next week .
['PRP', 'VBP', 'VBG', 'TO', 'DT', 'NN', 'JJ', 'NN', 'PUNCT']
['PRON', 'AUX', 'VERB', 'PART', 'DET', 'NOUN', 'ADV', 'NOUN', 'PUNCT']

Sentence:  John and Mary are friends .
['NN', 'NN', 'NN', 'CC', 'NN', 'NN', 'VBP', 'NNS', 'PUNCT']
['PROPN', 'CCONJ', 'PROPN', 'AUX', 'NOUN', 'PUNCT']

Sentence:  It is raining heavily outside .
['PRP', 'VBZ', 'VBG', 'VBG', 'RB', 'RB', 'PUNCT']
['PRON', 'AUX', 'VERB', 'ADV', 'ADV', 'PUNCT']
```