Training with bert-base-uncased usage of "final data":

Train/dev/test datasets, data was shuffled as described in general information
Epochs: **3**
Duration: **24:12 min**
Sequence length: **64 (words), 128 (tokens)**
Batch size: **16**

Train samples: **34156**
Test samples: **4142**
Dev samples: **3848**

	Precision	Recall	F1-score	support
BOS	0.85195	0.85431	0.85313	19789
O	0.98727	0.98703	0.98715	226483
accuracy			0.97636	246272
Macro avg	0.91961	0.92067	0.92014	246272
Weighted avg	0.97639	0.97636	0.97638	246272

Confusion matrix:

True		Predicted	
	BOS	16906	2883
	O	2938	223545
		BOS	O

Accuracy: **85.43%**

Sentences in test data: **19790**
Number of 1:1 correctly predicted sentences: **14912**
Accuracy: **75.35%**