**Training with bert-base-uncased:**

Explicit dev set used, train sequences were shuffled, train and dev weren’t.

Epochs: **3**

Duration: **13:34 min**

Sequence length: **64 (words), 128 (tokens)**

Batch size: **16**

Train samples: **19285 (shuffled)**

Test samples: **2411 (not shuffled)**

Dev samples: **2411 (not shuffled)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | support |
| **BOS** | **0.78** | **0.83** | **0.80** | **6870** |
| **O** | **0.99** | **0.99** | **0.99** | **147434** |
|  |  |  |  |  |
| **accuracy** |  |  | **0.98** | **154304** |
| **Macro avg** | **0.89** | **0.91** | **0.90** | **154304** |
| **Weighted avg** | **0.98** | **0.98** | **0.98** | **154304** |

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | BOS | **5673** | **1197** |
| True | O | **1610** | **145824** |
|  |  | BOS | O |
|  |  | Predicted |  |

Sentences in test data: **6871**

Number of 1:1 correctly predicted sentences: **4570**

Accuracy: **66.51%**