Exercise 05

# 1 Named Entity Recognition using BERT

## Warning Message with BertForTokenClassification and BERT-base:

## The warning message encountered during the initialization of the BertForTokenClassification class using BERT-base concerns the addition of an untrained linear layer on top of the BERT model. BERT itself is originally trained for tasks like Masked Language Modeling and does not use specific labels for token classification. Therefore, when adapting BERT for a token classification task (such as Named Entity Recognition), an additional linear classifier is appended to the model. This classifier does not have pre-trained weights and must be trained from scratch for the specific task at hand. This is the primary reason behind the warning, as the base BERT model remains unchanged in terms of label output; the adjustment is in the added, untrained classification layer.

## Best-Performing Model on Evaluation Set

The model trained with 3,000 examples performed the best on the evaluation set. It achieved an F1-micro score of 91.05% and an F1-macro score of 78.81%, which are higher than the scores of the model trained with 1,000 examples (F1-micro: 89.26%, F1-macro: 73.99%) and the model with frozen embeddings (F1-micro: 71.93%, F1-macro: 42.68%).

## Differences Between F1-micro and F1-macro Score

The F1-micro score calculates the metric globally by considering the total true positives, false negatives, and false positives. It's more reflective of the model's performance on frequent labels. In contrast, the F1-macro score calculates the metric for each label independently and then takes the average. It treats all labels equally, making it more sensitive to the model's performance on rare labels. The difference in these scores can indicate how well the model performs across different label frequencies.

## Performance Gap Between 1,000 and 3,000 Sentences for Fine-Tuning

Comparing the F1 scores of models trained with 1,000 and 3,000 sentences, we observe a performance improvement. For the 3,000 sentence model, the F1-micro score increased from 89.26% to 91.05%, and the F1-macro score increased from 73.99% to 78.81%. This indicates a significant improvement in performance, highlighting the benefits of using more training data for fine-tuning.

## Freezing vs. Not Freezing Embeddings

Based on the F1 scores, not freezing the embeddings (allowing the pre-trained layers of BERT to update during training) results in better model performance. The model trained with 3,000 examples and unfrozen embeddings outperformed the one with frozen embeddings by a considerable margin (F1-micro: 91.05% vs. 71.93%, F1-macro: 78.81% vs. 42.68%).

# 2 Emotion Regression

We chose the bert-base-uncased model as it was the most suitable model to fine-tune for our sentiment regression task. It provides contextual understanding of relationships between words enabling the understanding of sentiment nuances. As it is already pre-trained on general language patterns, the fine-tuning enables transfer learning to make use of this pre-existing, general understanding. The uncased variant of BERT ignores the case of letters, making it robust to different capitalization styles evident in Tweets. Overall BERT-based models have demonstrated state-of-the-art performance on various NLP tasks, i.e. applying fine-tuned layers to make use of transfer learning as as this task prescribed.

We experimented with some of the hyperparameters and found more epochs as well as larger batch-sizes a valuable adjustment to our architecture. We decided against preprocessing the texts, as the Tweets are actually quite unstructured naturally and also contain many special characters or grammatical errors by nature. Because of this and the good performance on the default and unprocessed setup (96.8% on train-set), we decided not to further preprocess. In this, it could also be the case that normally pruned errors such as grammatical errors or special characters might actually indicate 'anger'. Therefore the results supported our hypothesis. Furthermore, our hyperparameter tuning resulted in a almost maximal train-set Pearson Correlation of 98.92%.

The Pearson Correlation is a very precise form of model evaluation. It tells us how closely our model's predictions follow the true values, addressing the linear relationship between them. In this it is a useful instrument that measures the alignment of our predictions with reality i.e. the actual observations. In this, the values range from -1 to 1. 0 means no correlation, -1 means perfect negative correlation and 1 means perfect positive correlation. Therefore the aim is to maximize the Pearson Correlation on the datasets with 1 or 100% being the perfect score. For our dataset, our best performing model with 98.92% Pearson correlation on the train-set translated to 73.61% Pearson correlation on the test-set (which was not regarded during experimentation/optimization. Therefore we can state in other words, that our fine-tuned BERT model achieved a 73% alignment of its predictions with the actually observed datapoints from the test-set.