Since encoder-based transformer models are best suited to classification task such as token classification, I chose to work with BERT model for this assignment. An initial analysis of the English part of the dataset showed that the dataset has a skewed class/entity distribution, with entities (not counting O, overrepresented) belonging to PER, LOC and ORG being by far the most common.

NER is a token classification task. For evaluation, I report overall accuracy, but also precision, recall and F1 (harmonic mean of the former two) since the distribution is not balanced. System B (see README) achieved a slight increase in overall accuracy compared to system A on the test set, but a greater increase in F1 (as well as precision and recall). This is likely due to there being fewer classes/kinds of entities and fewer classes/entities with comparatively little data that the model must predict. Entity level scores for entities considered in both A and B are similar, however. The O tag is overrepresented in both systems.

Limitations: I did not experiment much with hyperparameters. If one wants a model that performs more similar over classes/entities, resampling techniques to balance the distribution could be investigated.