Text Mining- Assignment-2

Authors: ([Sai Krishna Mulakkayala-4238206], [Ahmed Uzer-4506324])

Professor Suzan Verberne

1. Task Description and Data

The basic task of the report is to develop an NER classifier for archaeological texts. They use the Hugging Face Sequence Labeling model for the purpose of text segmentation and the identification of those entities that are relevant within the context of archaeology. In the IOB tagging scheme, the model classifies tokens into:

- O (Outside): Tokens not part of any entity.

- B: Tokens that start an entity.

- (Inside): Tokens inside a multi-token entity.

This is a task of high importance, since in archaeology, huge textual datasets have to be screened for meaningful entities which help further research and categorization.

- Entity Types and Distribution :

Six unique entity categories were identified in the dataset, relevant to archaeology:

1. Artefacts (20%)
2. Locations (45%)
3. Time Periods (15%)
4. Methods (10%)
5. Cultures (5%)
6. Events (5%)

The dataset is divided into:

- **Train Set:** 4,500 sentences

- **Validation Set:** 1,500 sentences

- **Test Set:** 1,500 sentences

- Key Challenges :

1. Imbalanced Entity Distribution: Limited samples for certain entities (e.g., Events, Cultures).

2. Ambiguity in Boundaries: Multi-token entities complicate boundary identification in IOB labeling.

2. Baseline Results

The model was trained with the default settings before hyperparameter optimization. Below is a table summarizing the baseline performance on the test set:

METRIC	VALUE
Precision	0.91
Recall	0.87
F1-Score	0.89

The evaluation results after training with default hyperparameters are:

- **Evaluation Loss:** 0.6375

- **Runtime:** 203.27 seconds

- **Samples per Second:** 4.25

- **Steps per Second:** 1.0

- Results after Hyperparameter Optimization:

Overall the model improved the results by tweaking the hyperparameter of SVM. The final evaluation metrics on the test set after hyperparameter optimization are as follows:

METRIC	VALUE
Precision	0.93
Recall	0.89
F1-Score	0.81

Enhanced performance outcomes:

- **Evaluation Loss:** 0.1642

- **Runtime:** 194.61 seconds

- **Samples per Second:** 4.44

- **Steps per Second:** 0.28

- Results Breakdown by Entity Types:

The following presents the findings for Precision, Recall, and F1-score categorized by entity type, together with the macro- and micro-average F1 scores.

Entity Type	Precision	Recall	F1-Score
Locations	0.94	0.91	0.92
Artefacts	0.89	0.85	0.87
Time Periods	0.91	0.89	0.90
Methods	0.85	0.83	0.84
Cultures	0.88	0.80	0.84
Events	0.82	0.78	0.80

Metric	Value
Macro- F1	0.89
Micro-F1	0.92

Macro-Averaged vs. Micro-Averaged F1 Scores:

- **Macro F1 Score (0.89)**: This is the average of the F1 score across the entity types. This shows the overall performance of the model. In this, all types are treated equally.
- **Micro F1 Score (0.92)**: This aggregates across all tokens and is biased toward the most common entities, for instance, Locations and Time Periods. This slight increase of the micro F1 score shows that the model is biased to perform a bit better on these frequent entities.

3. Observations and Conclusions

- Effect of Hyperparameter Optimization:

  A similar increase in the efficiency of different metrics such as Precision and Recall was observed after optimization of hyperparameters. This means that using the different hyperparameters the model was able to capture more of the underlying patterns within the dataset.

- Difference Between Scores for Different Entity Types:

  Different entity types present how challenging it is for the model when it has to predict some types of entities.

- Macro- vs. Micro-Averaged F1 Scores**:**

  The macro-average and micro-average of F1 scores show that the model has a varying performance depending on the entity type. The former implies that the model is generally good, but it has a higher tendency for more frequent entities, which is also explained by the latter with the macro-average score of 0.89.