TU Dortmund University
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2024/25
01/21/2025

Exercise sheet 12

Natural Language Processing

Hand-in (voluntarily): 01/27/2025 until 10:00 a.m. via Moodle

---

Task 1

In Moodle, you can find the files `train.csv` and `test.csv`. Load the data set into your console. The data sets contain questions and their associated coarse- and fine-grained labels from the TREC data set. The dataset consists of 6 coarse-grained categories and 50 fine-grained categories. In the coming tasks, we will compare different the classification rates of different models for this multi-label classification task. Preprocess the data so that it is suitable for the different pipelines in the upcoming tasks.

Task 2

Train a Doc2Vec model on the data set and train a classifier based on the document embeddings of the train data set and the coarse labels.

Task 3

Train a lora adapter on top of BERT base uncased as a supervised model using the coarse labels of the training data set.

Task 4

Fine tune a BERT base uncased model as a supervised model using the coarse labels of the training data set.

Task 5

Evaluate your three models on the test data set using a macro f1 score and compare their performances as well as the time it took to train them. Which model would you chose in which situation?

Task 6

Repeat tasks 2-5 with the fine grained labels instead. Has anything changed?

Recommended packages & functions

Python: `adapters`, `sklearn.metrics.f1_score`