Task2 Report

This is the report for task2 of the assignment2 in NLP-DL course.

Jiahao Li

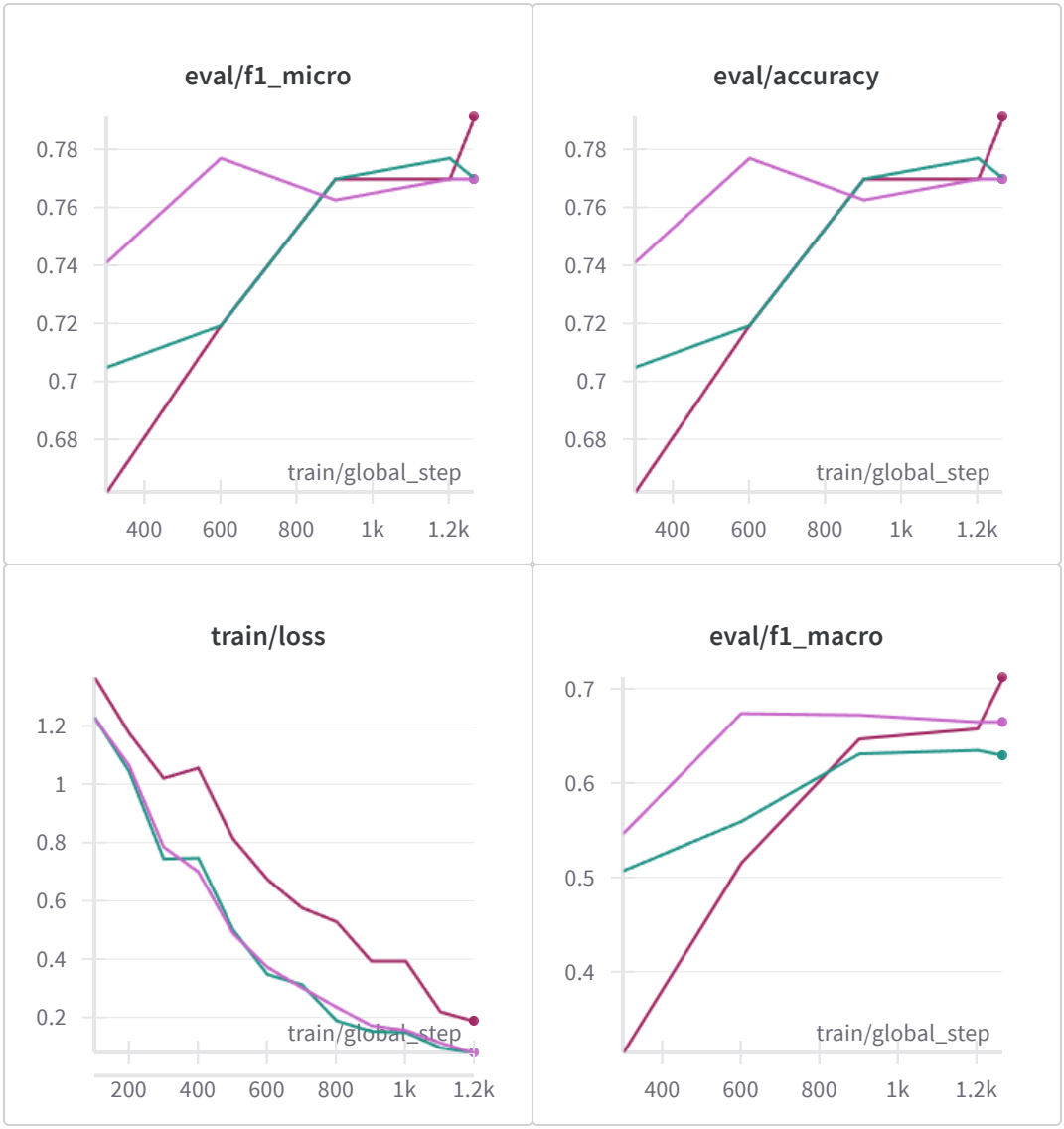
Created on October 22 | Last edited on October 26

I have run several experiments on the three assigned datasets with three assigned models, and the sections below are divided according to the types of datasets. In every section, I chose one of my runs to demostrate the results. Most of the configurations are identical except the number of labels and epochs.

Configurations shared across tasks and among models

- learning_rate: 5e-05
- lr_scheduler: linear scheduler(SchedulerType.LINEAR)
- optimizer: AdamW(OptimizerNames.ADAMW_TORCH)
- adam_beta1: 0.9,
- adam_beta2: 0.999,
- adam_epsilon: 1e-08,
- weight_decay: 0

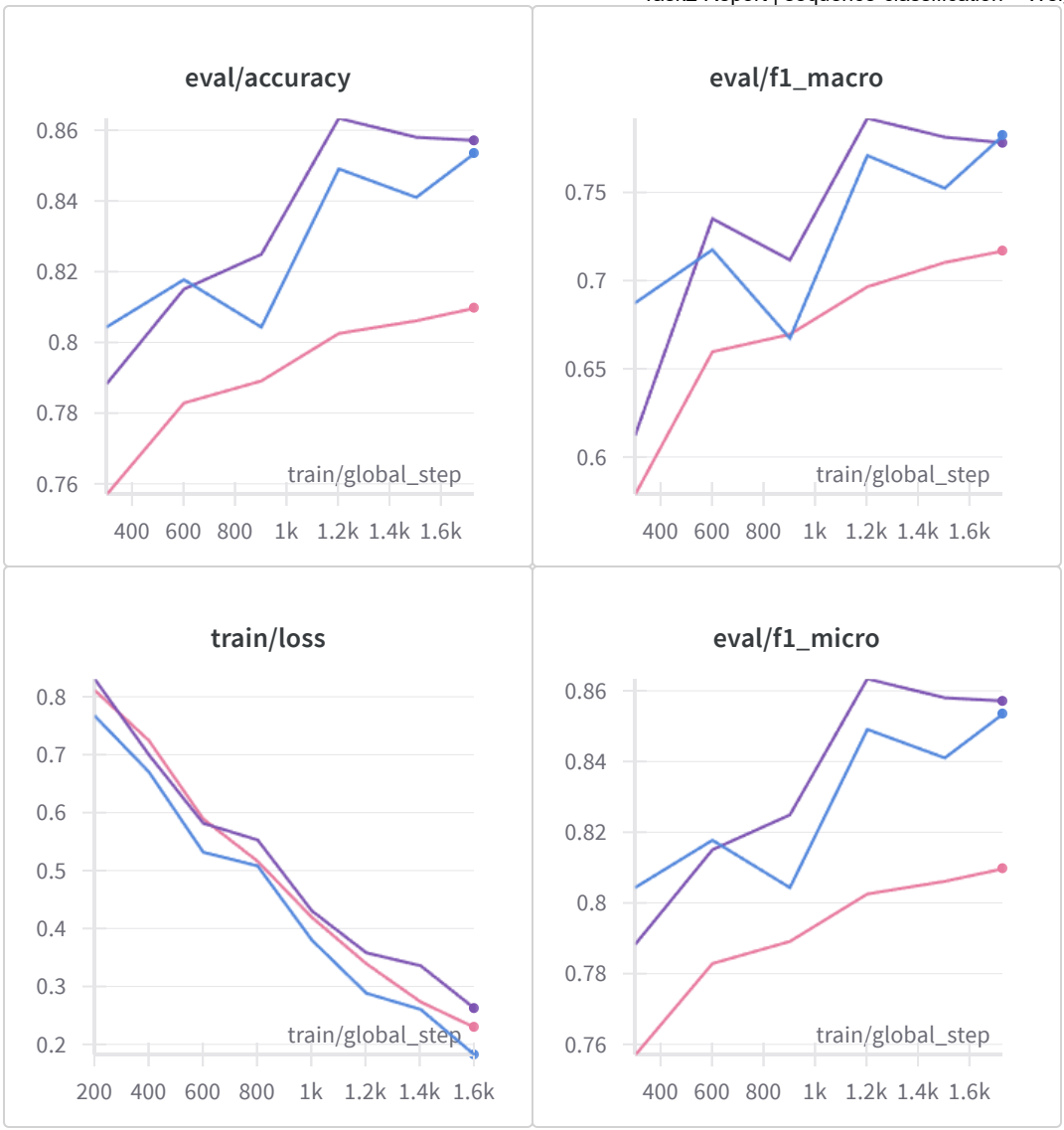
Section 1: ACL



Since the ACL dataset is relatively small, I have set the number of epochs to 6. The training loss curves of the three models are similar, and the accuracy and f1 scores are close, too. Nevertheless, partly because of the larger pre-training data and better tokenizer, Roberta model demonstrates

the best performance. Furthermore, scibert uses a vocabulary especially for scientific texts, so its performance is better than Bert's.

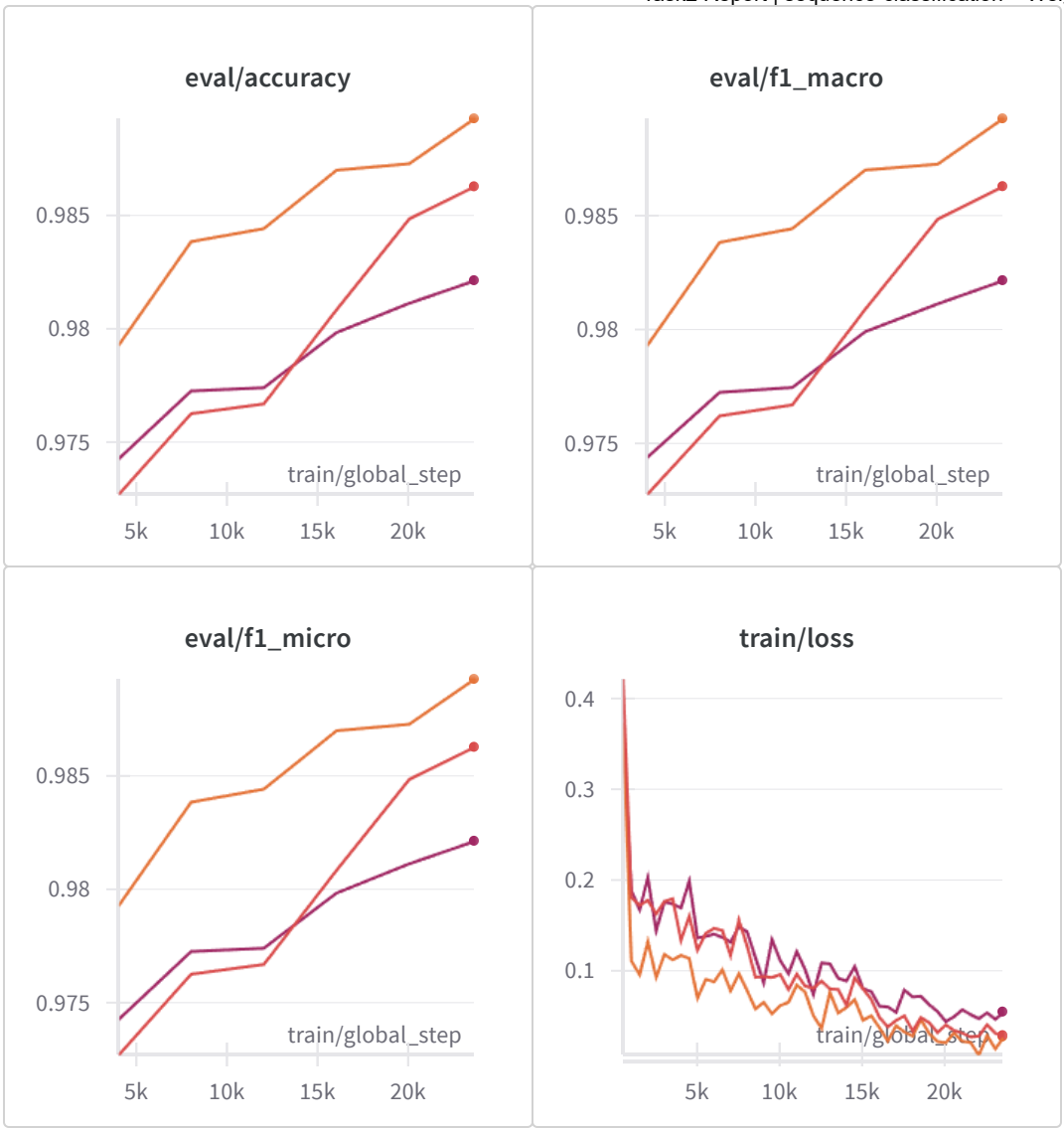
Section 2: restaurant



The task of the restaurant dataset is to classify the sentiments in the comments on restaurants. Unsurprisingly, the performance of Roberta is the best. However, this time, the accuracy and f1 scores of Bert are very close to those of Roberta, but scibert is significantly worse. That is because the vocabulary and pre-training dataset of scibert is not that suitable for restaurant comments

text. Aside from that, the training loss decreases at a steady rate over training steps, which means that the training process is valid.

Section 3: agnews



For the agnews dataset, the number of epochs is 3. Due to limited GPU resources and accidentally identical seed, there is only one valid run for every model. However, we can discover that under our settings, Bert shows the best performance on the agnews dataset, whose task is to classify the domain every news belongs to. However, thanks to the relatively large volume of the training

dataset, all three models achieve an accuracy larger than 98%, which is better than the two datasets above.

Created with 💔 on Weights & Biases.

https://wandb.ai/learnerljh-Peking%20University/sequence-classification/reports/Task2-Report---Vmlldzo5ODIwMzE5