Humboldt-Universität zu Berlin
Institut für Informatik
Prof. Dr. Alan Akbik

Berlin, 14.06.2024
Lehrstuhl Maschinelles Lernen

Introduction to Natural Language Processing

Exercise 8

Submission: Hand your homework until 01.07.2024 (Monday), 23:59 via Github Classroom. Each exercise is worth a total of **10 points** and you need **80%** (of total points over all exercises) to be admitted to the final exam. Sometimes, there are optional tasks that will give you extra points. Exercises in Github Classroom should be completed and submitted individually. However, we encourage you to work on the problems in teams. Please note the information available in Moodle: `https://hu.berlin/nlp24-moodle`.

Github Classroom: `https://classroom.github.com/a/VBjsdHXJ`

This exercise will span two weeks and is worth **10 points + 10 bonus points.**

Task 1 (Train a Classifier) — 7 + 3 + 3 points

The goal of this exercise is to train a classifier which can predict whether a given text is toxic. For this you will need to:

(a) Set up a training script (in `train.py`) using the huggingface ecosystem (using *transformers*, and *datasets*, and *evaluate*). In this script you need to

- Load the toxicity dataset [1]
- Tokenize the dataset
- Initialize a model [2]
- Invoke the `Trainer`
- Save the resulting model

(b) *Bonus Task:* Set up metric tracking using either *tensorboard* or *wandb*. After training, take some screenshots documenting you training and add them to `summary.md`.

(c) Train the classifier and write a few sentences about your process (in `summary.md`): What hyperparameters did you try? How did you select them? What worked for you? etc.

Task 2 (*Bonus Task:* Find Toxic Headlines) — 7 points

Using *fundus* [3] which you may know from Exercise 1, scrape some English newsoutlets and classify the headlines using the classifier you build in the previous task. Find at least two headlines which your classifier tags as toxic and add these to `summary.md`. Add the scripts you used for scraping and classifiying the headlines to your submission.

[1] You can find it here: `https://huggingface.co/datasets/HU-Berlin-ML-Internal/toxicity-dataset`
[2] Use the following pretrained model: `https://huggingface.co/prajjwal1/bert-tiny`
[3] See: `https://github.com/flairNLP/fundus`