Generative AI Technical and Social: Homework 1

September 7, 2025

In this exercise, we will study how to do text classification with LLMs. Please read the code from Colab and solve the following problems. We also leave the Kaggle competition as a Bonus.

Problem 1

Using the simple model provided in the Colab notebook, which is the same model discussed in class, modify the code by adding a hidden layer between the averaged embedding and the final logits. Use a ReLU activation function for this hidden layer, and set its dimension to 128.

1. [2 pts] Calculate the number of trainable parameters in your modified model.
2. [1 pt] Run the classification pipeline with your modified model and report the **accuracy** and **loss** after 10 epochs.
3. [2 pts] Increase the batch size from 128 to 1024 and run the classification pipeline again (ensure to reinitialize your model). What differences do you observe in the results, and how would you explain these changes?

Problem 2

Modify the code for fine-tuning using BERT.

1. [2 pts] Freeze the base model (making its parameters non-trainable) and add a hidden layer (with a dimension size of 128) with a ReLU activation function, similar to what was done in Problem 1. Report the accuracy and loss after 2 epochs.
2. [1 pt] Provide an explanation for any differences observed in Problem 2 compared to Problem 1.

Problem 3

Modify the code for Zero-shot LLM text classification. You may check the code in the notebook from Lecture 2 as well.

1. [2 pts] Optimize the prompt as much as possible (e.g., by applying the chain of thought method discussed in the lecture). Run your optimized prompt with 1000 test samples. Report both the optimized prompt and the resulting accuracy.

Problem 4 [Kaggle Competition, Bonus]

1. **[2 pts]** Copy and rerun the following Kaggle notebook: GATS LLM Finetuning Notebook, and submit your results to the competition: LLM Classification Finetuning. Record your scores in the shared spreadsheet: Google Sheet.

2. **[3 pts]** Attempt one approach to improve your score (it is acceptable to report unsuccessful attempts). Describe what you tried and record your results in the same spreadsheet: Google Sheet.