README *for* A Tutorial on the Pretrain-Finetune Paradigm for Natural Language Processing

This README file contains three sections. The first section, *General Description*, describes the contents in this replication folder. The second section, *Hardware Requirements*, describes what hardware is used for running the scripts. The third section, *Instructions*, reports on how to run the scripts.

1 General Description

The replication package contains the following files:

- README.pdf
- Table_1_finetune.ipynb: this file generates last row of Table 1 in the main manuscript (81 minutes).
- Table_1_zero_shot.ipynb: this file generates second to last row of Table 1 in the main manuscript (104 minutes).
- Table_2_bert.ipynb: this file generates the BERT results in Table 2 in the main manuscript.
- Table_2_roberta.ipynb: this file generates the RoBERTa results in Table 2 in the main manuscript.
- Study-1.ipynb: this file generates Table 1 in the online appendix (74 minutes).
- Study-2a.ipynb: this file generates Column 3 of Table 2 in the online appendix (49 minutes).
- Study-2b.ipynb: this file generates Column 4 of Table 2 in the online appendix (97 minutes).

– Table_1_few_shot.ipynb: this files generates the few-shot results in Table 1.

– distilRoBERTa_comparison_python.ipynb: results from distilRoBERTa using Python.

– distilRoBERTa_comparison_R.R: results from distilRoBERTa using R for comparison.

2 Hardware Requirements

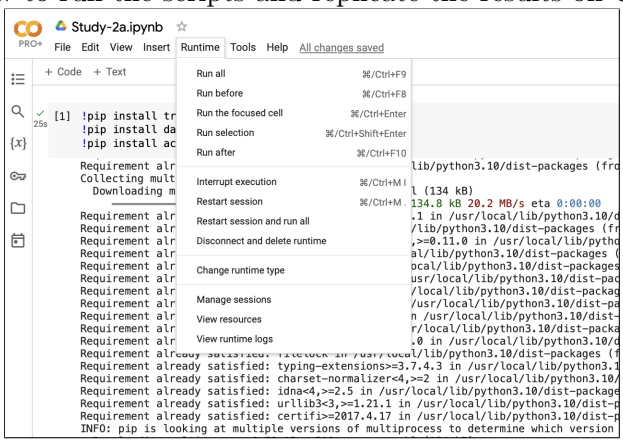
The replication package has been tested for running on A100 GPUs on Google Colab.[1] Exact performance metrics would vary to some extent depending on the type of GPUs used, but they would not affect the key results. Exact performance metrics should be replicable on the same type of GPUs.

3 Instructions

Execution

Each of the scripts above can be run independently. In Google Colab, we only need to hit the **Run all** button under **Runtime** (Figure 1).[2] All the key results will be printed out during execution.

Figure 1: How to run the scripts and replicate the results on Google Colab.



[1] Note that usage of A100 GPUs on Google Colab is not free. Usage of some other types GPUs (e.g. T4) is free.

[2] We can also run the scripts interactively if readers so prefer.