README *for* Topic Classification for Political Texts with Pretrained Language Models

Yu Wang
University of Rochester
w.y@alum.urmc.rochester.edu

This README file contains five 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 and the estimated running time for each script. The fourth section, Miscellaneous, is optional and provides some details about data processing and model loading. The last section, Package Versions, reports the versions of the packages used in the paper. For any question regarding the scripts, please reach out to Yu Wang.

1 General Description

The replication package contains the following files:

- README.pdf
- Table_1.ipynb
- Table_2.ipynb
- Figure_1.ipynb
- Figure_1_Visualization.ipynb
- Appendix_Figure_1.ipynb
- Appendix_Figure_1_Visualization.ipynb
- Appendix_Table_1.ipynb
- Appendix_Sequence_Length_Statistics.ipynb

For the main paper, Table_1.ipynb generates the results for Table 1. Table_2.ipynb generates the results for Table 2. Figure_1.ipynb generates the raw results for Figure 1; Figure_1_Visualization.ipynb plots the raw results into Figure_1.pdf.

For the online appendix, Appendix_Figure_1.ipynb generates the raw results, which are passed into Appendix_Figure_1_Visualization.ipynb for plotting. Appendix_Table_1.ipynb generates the results for Table 1. Appendix_Sequence_Length_Statistics.ipynb generates summary statistics about sequence lengths reported on Page 2 of the appendix.

2 Hardware Requirements

The replication package has been tested for running on Tesla T4, Tesla P100, and A100 GPUs, all on Google Colab with Python 3.8.15.[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 are replicable on the same type of GPUs. The results reported in the main paper and in the online appendix are based on A100.

3 Instructions

All the raw results can be generated by running each of the following self-contained scripts:

- Table_1.ipynb (262 minutes)
- Table_2.ipynb (52 minutes)
- Figure_1.ipynb (245 minutes)
- Appendix_Figure_1.ipynb (472 minutes)
- Appendix_Table_1.ipynb (52 minutes)

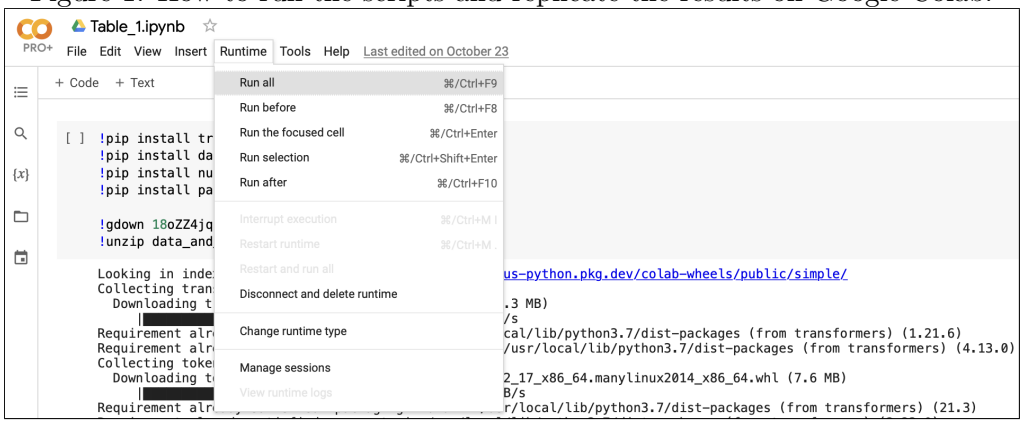
Each of the scripts above can be run independently. In Google Colab, we only need to hit the **Run all** button under **Runtime** (Figure 1).[2] Each script will print out the results as well as the time spent.

[1] https://colab.research.google.com.

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

The cumulative running time is around 1,083 minutes or 18 hours, with Figure_1_Appendix.ipynb taking the longest with 472 minutes. Currently, the scripts use five seeds and 20 epochs. If the goal is faster experimentation rather than 100% replication, readers could consider reducing the number of seeds and the number of epochs.

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



4 Miscellaneous: Further Information on Data and Models

(a) The data and cross-domain classifiers are loaded when we run the scripts. Specifically, this line below downloads the zipped file, data_and_models.zip, which contains the raw data and the cross-domain classifiers from Osnabrügge et al. (2021).[3]

```
!gdown 18oZZ4jqRK-uF-Nz6ftRdgNjKix88hrnO
```

When unzipped, the folder contains the following files:

logistic_model_8.pkl	Oct 18, 2022 at 9:14 PM	1.3 MB	Document
logistic_model_44.pkl	Oct 18, 2022 at 9:14 PM	6.9 MB	Document
target_corpus.csv	Oct 18, 2022 at 9:15 PM	4.7 MB	CSV Document
tfidf_8.pkl	Oct 18, 2022 at 9:15 PM	31.2 MB	Document
tfidf_44.pkl	Oct 18, 2022 at 9:15 PM	31.2 MB	Document

[3] An alternative is to upload the unzipped files to the reader's Google Drive. The zipped file can be downloaded directly from our shared file on Google Drive: https://drive.google.com/file/d/18oZZ4jqRK-uF-Nz6ftRdgNjKix88hrnO/view?usp=sharing. Readers can also access the same data and models from https://doi.org/10.7910/DVN/CHTWUB. Specifically, it is the data folder within capsule-9e05a29c-4b3d-457e-a53a-90f4601cda2f.zip in Osnabrügge et al. (2021).

(b) The pre-trained language model is downloaded when we call the following line:

```
RobertaForSequenceClassification.from_pretrained("roberta-base",
    num_labels=tasks[task]["number_of_labels"])
```

5 Package Versions

```
datasets==2.7.1
gdown==4.5.4
matplotlib==3.2.2
numpy==1.21.6
pandas==1.3.5
scikit-learn==1.0.2
torch==1.12.1+cu113
torchvision==0.13.1+cu113
torchaudio==0.12.1+cu113
transformers==4.25.1
```

References

Osnabrügge, M., Ash, E., & Morelli, M. (2021). Cross-Domain Topic Classification for Political Texts. *Political Analysis*.