Reproducibility Appendix
Project Report for NLP Course, Winter 2023/4

Dawid Płudowski, Anotni Zajko, Mikołaj Roguski, Piotr Robak

Warsaw University of Technology

{dawid.pludowski, antoni.zajko, mikolaj.roguski, piotr.robak}.stud@pw.edu.pl

supervisor: Anna Wróblewska

Warsaw University of Technology

anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION – We used RoBerTa and ERNIE transformers during our experiments. The versions from HuggingFace are as follows: `roberta-base` and `nghuyong/ernie-2.0-base-en`. Further details are covered in the Methodology section.

- LINK TO CODE – https://github.com/azoz01/nlp-2024-fake/tree/master/notebooks.

- INFRASTRUCTURE – We used a personal computer setup with GPU (RTX 3080 or similar but code is suitable for CPU only), and up to 32GB of RAM.

- RUNTIME PARAMETERS – The fine-tuning of a single model takes up to 2 hours. In total, running all scripts necessary to reproduce our results should take over 24h on a decent PC.

- PARAMETERS – We used RoBerTa with 125M parameters and ERNIE with 110M parameters.

- VALIDATION PERFORMANCE – As a validation we use accuracy metric. Detailed information about the validation performance is provided in the article.

- METRICS – Metrics of evaluation are provided in the Methodology section. The code responsible for computing them is in the notebook: https://github.com/azoz01/nlp-2024-fake/blob/master/notebooks/results.ipynb.

Multiple Experiments:

- NO TRAINING EVAL RUNS – Each model was trained 3 times for each configurations, in total 24 runs.

- HYPER BOUND – We did not tune any hyperparameters (fine-tuning task).

- HYPER BEST CONFIG – We did not tune any hyperparameters (fine-tuning task).

- HYPER SEARCH – We did not tune any hyperparameters (fine-tuning task).

- HYPER METHOD – We did not tune any hyperparameters (fine-tuning task).

- EXPECTED PERF – The results are provided in the article in Table 3.

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – Provided in the article in Table 2.

- DATA SPLIT – split data is provided in `data` directory on the aforementioned repository.

- DATA PROCESSING – We did not exclude any data from the processing. The only processing we applied was transforming raw text into tokens. Other details are covered in the Methodology section of the article.

- DATA DOWNLOAD – get from our repository on GitHub or see links below

  - https://www.kaggle.com/code/hendrixwilsonj/liar-data-analysis
    https://www.kaggle.com/datasets/csmalarkodi/isot-fake-news-dataset
    https://github.com/cuilimeng/CoAID

- NEW DATA DESCRIPTION – In our research, we did not create new datasets

- DATA LANGUAGES – English