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

Jakub Kozieł
Warsaw University of Technology
jakub.koziel.stud@pw.edu.pl

Jakub Lis
Warsaw University of Technology
jakub.lis2.stud@pw.edu.pl

Bartosz Sawicki
Warsaw University of Technology
01151408@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION – The processing pipeline is described in the report. We used already developed models and described them in the Related work section.
- LINK TO CODE – GitHub repository. External packages are listed in the *requirements.txt* file.
- INFRASTRUCTURE – Experiments were conducted in 3 environments: locally on CPU (Intel i7 7th Gen), locally on GPU (Nvidia GForceGTX 1660ti), and using Google Colab with GPU acceleration.
- RUNTIME PARAMETERS – In our project, we performed different experiments in different environments. Document-based sentiment analysis was calculated on Google Colab and took approx. 2 min per model for the whole dataset (n=1758). Aspect-based sentiment analysis was calculated locally on GPU, and it took approx. three hours for the whole dataset (n=1758). XAI attributions were calculated on CPU with approx. 1 min per instance per explanation method.
- PARAMETERS – Sizes of models used:
  - NER - Babelscape/wikineural-multilingual-ner: 177M
  - Aspect-based sentiment analysis:
    - DeBERTa-ABSA: 435M
    - Flan-T5: 783M
  - Document-based sentiment analysis:
    - SieBERT: 355M
    - FinancialBERT-Sentiment-Analysis: 109M
    - auditor_sentiment_finetuned: 109M
    - twitter-roberta-base-sentiment-latest: 124M
    - distilroberta-finetuned-financial-news-sentiment-analysis: 82M
- VALIDATION PERFORMANCE – We do not have a validation dataset due to manual labeling. We only report metrics for the test set.
- METRICS – Metrics used for evaluation are described in section 4.1 of the report.

Multiple Experiments:

- NO TRAINING EVAL RUNS – We did not train any models. We performed one evaluation run for each model.
- HYPER BOUND, HYPER BEST CONFIG, HYPER SEARCH, HYPER METHOD – We did not tune the hyperparameters.
- EXPECTED PERF – We evaluated the models once, we did not need to summarise the results.

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

- DATA STATS – We analyzed 1,758 English articles from STA API from 2 months (September-October 2023).
- DATA SPLIT – From 1,758 articles, we prepared a test set - 12 articles randomly selected from 6 categories, so in total, the test set contained 72 labeled examples.
- DATA PROCESSING – The processing contains downloading from STA API. The data is converted from JSON to a dataframe. Later, *lede* is concatenated with *text*. Finally, we remove HTML markers from the text.

- DATA DOWNLOAD – We do not provide downloadable data as STA NDA applies.
- NEW DATA DESCRIPTION – The data annotation process is described in details in the report
- DATA LANGUAGES – We prepared the processing pipeline for English news.