Experiment Design for Data Science – WS 2023/24
Exercise 2: Reproduce experimental results from a paper

Deadline for submission: Sun, January 28th, 2024, 23:59

For this assignment, you will work in **groups of** (up to) **four** and **reproduce** the experimental setup, experiments, and results as explained in a scientific paper. The task of interest is – based on careful experimental design – to confirm the numbers and findings reported in the paper or, alternatively, uncover inconsistencies and therefore challenge the conclusions drawn.

Specifically, by reproducing the experiments, it can be checked whether

- the information given in the paper is sufficient to reproduce the results reported
- statistically significant differences can be made out between the different settings, in particular when this is not reported in the paper (no significance tests, confidence intervals, p values, etc. or not even variance)
- the values reported in the paper could stem from the distribution sampled by the reproduced experiments

Your efforts need to be documented in a report that should be accompanied by the code and workflow you used to reproduce the experiments. To do so, first, identify the experimental setup and strategy taken in the paper and outline the steps necessary for you to reproduce results. Try to stay to the described steps as close as possible, i.e., use the same implementations and settings wherever possible. If the exact implementations are unknown or unavailable, find a reasonable substitute. You might have to consult additional sources to get the full picture of the used data or methods.

In each step, document which information is given in the paper and which measures you took to implement this step, i.e., which implementation and which parameters you used. Report the numbers obtained in each intermediate step. Identify deviations from the numbers reported in the original paper, their origin, and estimate whether they will have a significant impact on subsequent steps.

For the results, perform adequate tests to test for statistical significance. Justify your choice. Can you confirm the findings of the paper? Did you identify a flaw in their setup? Could you correct the flaw? How did it impact the results and findings of the experiments? Typical problems that could occur are listed at the end of this document (Appendix). Include a table into your report that indicates for the different types whether you have encountered them or not. Also indicate other problems not matching the identified issues.

You need to **reproduce the results of one scientific paper out of ~30 candidates** from recent years of the SIGIR, CIKM, ECIR, and RecSys conferences, to be found in the *Exercise 2 Topics (Papers)* section on TUWEL.

Reports need to be prepared using the ACM Conference Proceedings single-column format using Overleaf,[1] the offline Latex Template,[2] or the Word Template.[3] **Report page limit: Maximum 6 pages!** That is, focus on the key aspects!

Upload your report to the TU Wien ExDDS WS 2023/24 community on Zenodo (https://zenodo.org/deposit/new?c=tuw-exdds-ws23, choose Report as submission type). Please, make it open or embargoed (embargo end = deadline for this exercise). As a result, you and the other students will have a chance to compare and discuss your findings. When writing your report, keep in mind that it will be publicly visible and open to the research community. For submission in TUWEL, apart from your presentation slides (see below), you will only provide the DOI pointing to your report and experimentation scripts/code/workflows.

In the classes on Jan 11/18/25, 2024, each group will have to give a **very(!) short presentation on the status of the work** (strict time limit: 240 seconds!). To this end, prepare a deck of 4-5 slides (including first slide containing group number, members, and chosen option), presenting your strategies, encountered difficulties and key findings, as well as your conclusion about the experimental design, intermediate results, conclusions so far and remaining work. Your group will be assigned a presentation time slot by Jan 4, 2024. **Upload your slides in PDF format 24h before your presentation the latest.** PDFs will be consolidated into one presentation to avoid switching overheads. Rehearse the presentation beforehand to ensure meeting the time limit.

The final report should detail the same aspects: strategies, difficulties, key findings, as conclusion about the experimental design, results, and conclusions.

Further information:

- If the chosen paper is very easily reproducible, i.e. all code is available and runs out of the box, producing exactly the results as described in the paper, focus even more on performing a detailed examination of the experimental setup and put the results under scrutiny: were experiments run multiple times, e.g. using different seeds for data splits; was variance reported; were significance tests performed (applicable ones); do conclusions hold if you rerun experiments using different seeds?

- Existing attempts to reproducing your paper should not influence or prime your investigations. The goal of the assignment is to investigate the reproducibility of published scientific papers, and not necessarily to smoothly confirm the numbers presented.

- If you are lacking resources for running the experiments, we recommend to make use of Google Colaboratory (https://colab.research.google.com) as a free service.

[1] https://authors.acm.org/proceedings/production-information/overleaf
[2] https://authors.acm.org/proceedings/production-information/preparing-your-article-with-latex
[3] https://authors.acm.org/proceedings/production-information/preparing-your-article-with-microsoft-word

Grading scheme:

- Reproduction of results (code and report): max. 40%
- Description, interpretation, statistical testing: max. 25%
- Quality/clarity of report and presentation: max. 30%
- Formal and complete submission to Zenodo: max. 5%

Appendix

13 categories of possible problems structured into 3 families:

A. Code/Setup Difficulties

1. Code not provided
2. Code partially not provided
3. Faulty code provided
4. Information about software versions missing
5. Specific hardware and software setup necessary

B. Data

1. Data not provided
2. Data partially not provided
3. Metadata not provided

C. Process

1. Comparison metrics or baselines not discussed
2. Experimental setup not described
3. Experiment workflow incomplete
4. Inconsistencies in paper, code and data
5. Results differ significantly