Reproducibility Project

CSE 481N: Natural Language Processing Capstone – University of Washington

Spring 2024

This project option is designed to give you an opportunity (1) to engage with the current state of NLP research and (2) to contribute useful knowledge to members of that community. Specifically, your team will choose one paper from the most recent NLP conference (EMNLP 2023) and attempt to reproduce its experiments.[1] In some cases, this may be a straightforward task. In some cases, it may be impossible, for a range of reasons.

Teams: Your team should be composed of three people. Diverse teams are stronger teams; working with people whose perspectives are different from yours (both in that particular dimension and in other dimensions) gives you new opportunities to learn.
Your project will do one of the following:

1. Reproduce the main experiments in the paper. Your report will assess the ease of reproducibility, with respect to the checklist we provide below. In addition, you should attempt at least one additional experiment that isn't in the paper, but that you are able to conduct after having successfully reproduced the main results. For example, you could assess the sensitivity of the model to one or more hyperparameters or to the amount of training data, or measure the variance of the evaluation score due to randomness in initial parameters.

2. Report on your failed attempt to reproduce the paper's main experiments. Not all papers are easily reproducible. Failing to reproduce the exact results of the original paper is not necessarily a bad thing, as long as your experiments are rigorous. Your report should identify all the questions that would need to be answered to reproduce the experiments, or discuss how the findings appear to be in error (if that is what you discover).

Both outcomes (1) and (2) are acceptable and can earn full credit.
Some considerations in choosing a paper to reproduce:

- You should find the problem tackled in the paper interesting.
- You should be able to (immediately) access the data you will need to reproduce the paper's experiments. If doing this requires submitting a special request that will be processed by a human (e.g., it is derived from health records and you must be approved for access), or there is special training you must undergo to gain access to the data, do not choose this paper, because these processes will probably take more time than you have.
- In many cases, the authors may have made code available; this may be a blessing or a curse. You should definitely peruse a paper's codebase before deciding on that paper.
- You should estimate the computational requirements for reproducing the paper and take into account the resources available to you for the project.[2] Some authors will have had access to infrastructure that

[1] The paper you choose must be from either the main EMNLP conference or the *Findings* volume. The ACL Anthology URL for a suitable paper will contain the string "main" or "findings"; if in doubt, check with course staff.

[2] Refer to the course website for details about computational resources.

is way out of your budget; don't choose such a paper.

Some potentially useful links:

- NeurIPS 2019 reproducibility checklist, used as part of the reviewing process for that conference
- ML code completeness checklist
- ML Reproducibility Challenge 2020; accepted papers at NeurIPS 2020 and ICLR 2020
- ML reproducibility tools and best practices

Well before the first deadline, it would be useful for your group to put together a template which has fields for all of the information you want to record from every experiment. (This could be in the form of a spreadsheet, or a rough placeholder draft with information that you fill in as you go, etc.) You should design this so that once it's filled out for all of your proposed experiments, you will meet the overall requirements for the project (described at the end of this document). For example, a group which is going to show training curves as their additional experiment would want a template which has fields for at least:

1. training curves showing dev.-set evaluations every X training steps
2. learning rate, batch size, dropout, size of their model, etc. (they shouldn't have "etc." here; they should be very specific)
3. clear connection showing which hypothesis this experiment will support
4. information on the data
5. computational requirements

(In other words, the list that you construct should consist of the items from the numbered list in the project report description that are specific to your project, with one template filled out for each experiment you run, including baselines.)

Deliverables and Deadlines

Detailed instructions for the report are given in this latex template. It is imperative that you follow the instructions in the template carefully for versions 1 and 2; it is strongly recommended that you familiarize yourself with the template before writing the proposal.
The **deadlines for each deliverable are shown on the course calendar**.

Proposal

Your proposal will be due around week 3. It should be one page and briefly include:

- a (bibtex) citation of the paper whose experiments you plan to reproduce, with a URL
- the hypotheses in the paper you plan to verify by reproducing experiments
- a short description of whether and how you can access the data used in the paper
- whether you will use the existing code (in that case, a link to the code) or implement yourself
- a discussion of the feasibility of the computation you will need to do (essentially, an argument that the project will be feasible)

Estimation does not have to be exact, but you will get more useful feedback if you include specific estimation. There is no specific template for the proposal.

Versions 1 and 2 (Final Report)

Fill out each section in the template by replacing the instructions with the actual content. You must follow this template for both versions 1 and 2 (final report). The final report must not exceed 8 pages, excluding references and the one-page summary that comes first.
For version 1 (due around week 6), you need only complete the following sections, and put a placeholder ("to do") for the other sections: Introduction, scope of reproducibility, methodology, model description, data description, implementation (you only need to say whether you will use existing code or implement yourself), computational requirements (you only need to include an estimate).
For the final report (due at the end of the quarter), all sections in the template must be filled in. The final report must not exceed 8 pages (that limit includes the first page and does not include references).
Please note:

- Grades are shared by your team. Students in this course are expected to work together professionally, overcoming the inevitable challenges that arise in the course of a team project. We recognize that, occasionally, team members behave unreasonably. To help us navigate situations where you feel a shared grade would be unfair, we invite you to submit individual updates on your team's progress at any time during the quarter using this form.
- No late submissions are allowed. Your team will receive zero points for the late submission.
- Instead of submitting code, set it up as a public Github repository and add the link to the project report. If writing your own code, make sure it is documented and easy to use (this project is about reproducibility!). Include a link to a github repository which can be installed and run with a few lines in bash on department machines. Include a description of how difficult the algorithms were to implement. If using public code from the original repository, more of your energy will go into running additional experiments, such as hyperparameter optimization, ablations, or evaluation on new datasets (see below). However, note that it's not always trivial to get a public code release working!
- You may include an appendix in your final report. However, you should include all the important details in the main paper. The appendix is allowed so that your report will be helpful to future researchers; it will not be read by the course staff.
- Submit your reports **both** through Canvas, as a single pdf per submission, submitted once by a single team member, **and** by posting a publicly readable link on the course discussion board in the appropriate thread.

Please see the grading rubric from 2021; it will be similar this year.