ACL Tutorial Proposal: Towards Reproducible Machine Learning Research in Natural Language Processing

Ana Lucic[1], Maurits Bleeker[1], Samarth Bhargav[1], Jessica Zosa Forde[2],
Koustuv Sinha[3], Jesse Dodge[4], Sasha Luccioni[5] and Robert Stojnic[6]

[1]University of Amsterdam
[2]Brown University
[3]McGill University
[4]Allen Institute for AI
[5]HuggingFace
[6]Meta AI

1 Motivation & Objectives

While recent progress in the field of ML has been significant, the reproducibility of these cutting-edge results is often lacking, with many submissions lacking the necessary information in order to ensure subsequent reproducibility (Hutson, 2018). Despite proposals such as the Reproducibility Checklist (Pineau et al., 2020) and reproducibility criteria at several major conferences (NAACL, 2021; Dodge, 2020a; Beygelzimer et al., 2021), the reflex for carrying out research with reproducibility in mind is lacking in the broader ML community. We propose this tutorial as a gentle introduction to ensuring reproducible research in ML, with a specific emphasis on computational linguistics and NLP.

2 Target Audience and Prerequisites

This tutorial targets senior researchers in academic institutions who want to include reproducibility initiatives in their coursework, and well as junior researchers who are interested in participating in reproducibility initiatives. The only prerequisite for this tutorial is a basic understanding of the scientific method.

3 Outline of Tutorial Content

The tutorial will cover four parts over the course of three hours:

1. Introduction to reproducibility (45 mins)
2. Reproducibility in NLP (45 mins)
3. Mechanisms for Reproducibility (45 mins)
4. Reproducibility as a Teaching Tool (45 mins)

3.1 Introduction to reproducibility (45 mins)

We will start the tutorial by motivating the overall problem: what does reproducibility mean and why is it important? What does it mean for research results to (not) be reproducible? What are some examples of important results that were (not) reproducible? Why is there a reproducibility crisis in ML (Hutson, 2018)? What would it look like if we, as a community, prioritized reproducibility?

We will explain how reproducibility works in fields outside of computer science, such as medicine or psychology, explain the mechanisms they use, and the criteria for achieving reproducible results. Next, we will discusses successes and failures of reproducibility in these fields, the reasons why the research was (not) reproducible, and the resulting consequences. We will follow with a similar discussion of fields within computer science, specifically in ML, before diving into reproducibility in NLP.

3.2 Reproducibility in NLP (45 mins)

In this part of the tutorial, we will focus on reproducibility in NLP, including examples of results that were reproducible and those that were not reproducible. For the latter, we will categorize reproducibility failures in NLP. We will also discuss the specific challenges with reproducibility in NLP and how they differ from the challenges in ML, and in science more broadly.

3.3 Mechanisms for Reproducibility (45 mins)

After explaining what reproducibility is and what the challenges are, we will examine existing mechanisms for reproducibility in ML and NLP, such as reproducibility checklists (Pineau et al., 2020; NAACL, 2021; Dodge, 2020a; Beygelzimer et al., 2021), ACM's badging system (ACM, 2019), and reproducibility tracks at conferences (ECIR, 2021). We will follow with an in-depth discussion on the ML Reproducibility Challenge[1], where the objective is to investigate the results of papers at top ML conferences by reproducing the experiments. Finally, we will discuss in length on useful tips, methodologies and tools researchers and practitioners in NLP can use to enforce and encourage reproducibility in their own work.

3.4 Reproducibility as a Teaching Tool (45 mins)

To improve the scientific process, scientific discourse, and science in general, it is imperative that we teach the next generation of academics and researchers about conducting reproducible research. In the final part of the tutorial, we will provide recommendations for using reproducibility as a teaching tool based on our experiences with incorporating a reproducibility project into a graduate-level course (Lucic et al., 2022; Lucic, 2021; Dodge, 2020b). We will share our experiences and reflect on the lessons learned, with the goal of providing instructors with a playbook for implementing a reproducibility project in a computer science course. Next to that, we will also give an overview of how reproducibility has been used as a tool in other academic courses.

4 Breadth of the tutorial

In the tutorial, we introduce and contrast reproducibility (Drummond, 2009), discuss papers reflecting on the reproducibility crisis in ML and NLP (Pedersen, 2008; Mieskes et al., 2019; Belz et al., 2021a,b), including possible reasons for this crisis (Hutson, 2018). This includes barriers to reproducibility, such as lack of code availability (Pedersen, 2008; Wieling et al., 2018) and the influence of different experimental setups (Fokkens et al., 2013; Bouthillier et al., 2019; Picard, 2021).

Raff (2019) investigates the reproducibility of ML papers without accessing provided code, relying on only details provided in the paper. (Belz, 2021) attempt to quantify reproducibility in NLP and ML. We also discuss reproducibility checklists from multiple venues (Pineau et al., 2020; NAACL, 2021; Dodge, 2020a; Beygelzimer et al., 2021; ACM, 2019; ECIR, 2021). Finally, we discuss coursework focused on teaching through reproducibility in ML (Yildiz et al., 2021) and FACT-AI (Lucic et al., 2022; Lucic, 2021).

5 Reading List

We briefly describe recommended reading for participants in this section.

5.1 General Background

Heaven (2020) (link) provides an overview of the replicability/reproducibility crisis in AI, noting common barriers, potential solutions and their drawbacks. Interested readers can also refer to (Baker, 2016) for a general discussion of the replicability/reproducibility crises in science.

5.2 NLP

We recommend participants read the following papers about reproducibility in NLP: (Mieskes et al., 2019; Belz et al., 2021a).

5.3 Teaching Reproducibility

Yildiz et al. (2021) introduce a portal[2], focusing on teaching AI/ML through 'low-barrier' reproducibility projects. They show that this can help develop critical thinking skills w.r.t. research, and that participants placed more value on scientific reproductions.

6 Sharing of Tutorial Materials

All of our tutorial materials will be publicly available at `https://acl-reproducibility-tutorial.github.io`.

7 Ethics Statement

Reproducibility and ethics are inherently related, since ensuring that research is reproducible by members of the community that are not its original authors contributes to making the field more inclusive (e.g. providing the code and hyperparameters needed to replicated a state-of-the-art ML model can help researchers build and expand upon it). Furthermore, being transparent about the costs of the model, both in terms of the computational power need to train it as well as the data involved, helps members of the community be more equitable in evaluating it: for instance, if two models achieved similar accuracy on the same dataset, with one requiring 10x more computation than the other,

[1] `https://paperswithcode.com/rc2021`

[2] `https://reproducedpapers.org/`

that could help researchers choose which one to use given their constraints. Finally, progress in the field of computational linguistics specifically is being led by large organizations that are the ones training and deploying equally large language models that are difficult to replicate without having access to the same resources that they do; being more transparent and ensuring that even large language models are replicable is important for making the field more democratic as a whole.

8 Pedagogical Material

As mentioned in Section 3.4, we want instructors to be able to use content from our tutorial in order to design reproducibility projects for graduate-level coursework. The content will largely be based on the following components: (i) a blog post on how to use the ML Reproducibility Challenge as an educational tool (Dodge, 2020b), (ii) blog post on one university's experience in using the ML Reproducibility Challenge as an educational tool (Lucic, 2021), and (iii) the corresponding paper (Lucic et al., 2022). We hope this can function as a starter pack for any instructor who is interesting in incorporating reproducibility projects in their coursework.

9 Presenter Information

Ana Lucic is a PhD Candidate at the University of Amsterdam. Her work primarily focuses on developing and evaluating methods for explainable machine learning (ML). She co-developed a graduate-level course called *Fairness, Accountability, Confidentiality and Transparency in Artificial Intelligence (FACT-AI)* that is centered around reproducing existing FACT-AI algorithms. Her email is a.lucic@uva.nl.

Maurits Bleeker is PhD Candidate at the University of Amsterdam who co-developed the FACT-AI course. His main interest lies in the development of new optimization functions for image-text matching, by taking task- and data-specific inductive priors into account. This with the goal to improve the computational efficiency of multi-modal optimization. He also co-developed and coordinated two iterations of the FACT-AI course at the University of Amsterdam. His email is m.j.r.bleeker@uva.nl.

Samarth Bhargav is a PhD Candidate at the University of Amsterdam. Samarth's research focuses on representation learning for information retrieval, with a goal of making IR systems (e.g recommenders) more amenable to user control, for example, through conversational interfaces. His secondary interests include recommendation in a cross-market or cross-domain setting, known-item retrieval, FACT in IR and teaching IR. He has co-developed and taught multiple iterations of graduate IR courses at the University of Amsterdam. His email is s.bhargav@uva.nl.

Jessica Zosa Forde is a PhD Candidate at Brown University. Jessica's research focuses on the empirical study of deep learning models, to improve their reliability in high stakes domains such as healthcare. She has also studied the inductive bias of overparameterized models, and model pruning. She believes that the open science movement is important for improving transparency and accountability in ML. She is also am a co-organizer of the ML Reproducibility Challenge (MLRC) and the ML Retrospectives workshop. Her email is jessica_forde@brown.edu.

Koustuv Sinha is a PhD Candidate at McGill University/Mila. He is the lead organizer of the annual ML Reproducibility Challenge (MLRC), which has had five iterations since 2018 (at ICLR 2018, ICLR 2019, NeurIPS 2019, MLRC 2020, MLRC 2021). He also serves as an associate editor of ReScience, a journal promoting reproducibility reports in various fields of science. Koustuv's research focuses on investigating systematicity in natural language understanding (NLU) models, especially the state-of-the-art large language models. His research goal is to develop methods to analyze the failure cases in robustness and systematicity of these NLU models, and develop methods to alleviate them in production. His email is koustuv.sinha@mail.mcgill.ca.

Jesse Dodge is a research Scientist at AllenNLP, Allen Institute for AI. Jesse created the NLP Reproducibility Checklist, has been an organizer of the ML Reproducibility Challenge (MLRC) 2020 and 2021, will be a Reproducibility Chair at NAACL 2022, and has published numerous papers in top NLP conferences on reproducibility. Jesse's research focuses on efficient and reproducible NLP and ML. He also has experience building large-scale NLP datasets. His email is jessed@allenai.org.

Sasha Luccioni is a Research Scientist at HuggingFace. She has been an organizer of the ML Reproducibility Challenge (MLRC) since 2021 and is an Area Chair for the Ethics in NLP track at EMNLP 2021. Sasha's research aims to contribute towards understanding the data and techniques used for developing Machine Learning approaches. She is particularly interested in developing tools for analyzing and filtering the data used for training large language models, as well as quantifying their carbon footprint. She has lectured several classes in ML and NLP, and is the main instructor for the forthcoming Deeplearning AI "AI for Social Good" course. Her email is sasha.luccioni@huggingface.co.

Robert Stojnic an Engineering Manager at Meta AI (formerly Facebook AI Research). He is the co-creator of Papers with Code, which has the biggest collection of papers, code, datasets and associated results, and co-organizes the ML Reproducibility Challenge (MLRC). He created the ML Code Completeness Checklist (Stojnic, 2020), which is part of the ML Reproducibility Checklist used by multiple conferences, including NeurIPS. He is a co-organizator for ML Reproducibility Challenge. His email is rstojnic@fb.com.

References

ACM. 2019. Artifact review and badging. https://www.acm.org/publications/policies/artifact-review-badging.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Anya Belz. 2021. Quantifying reproducibility in nlp and ml. *arXiv preprint arXiv:2109.01211*.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021a. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman-Vaughan. 2021. Introducing the neurips 2021 paper checklist. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist.

Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR.

Jesse Dodge. 2020a. Guest post: Reproducibility at emnlp 2020. https://2020.emnlp.org/blog/2020-05-20-reproducibility.

Jesse Dodge. 2020b. The reproducibility challenge as an educational tool. Medium, https://medium.com/paperswithcode/the-reproducibility-challenge-as-an-educational-tool-cd1596e3716c.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science.

ECIR. 2021. Ecir: Call for reproducibility track papers. https://www.ecir2021.eu/call-for-reproducibility-track.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

Will Douglas Heaven. 2020. Ai is wrestling with a replication crisis.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science*, 359:725–726.

Ana Lucic. 2021. Case study: How your course can incorporate the reproducibility challenge. Medium, https://medium.com/paperswithcode/case-study-how-your-course-can-incorporate-the-reproducibility-challenge-76e260a2b59.

Ana Lucic, Maurits Bleeker, Sami Jullien, Samarth Bhargav, and Maarten de Rijke. 2022. Teaching fairness, accountability, confidentiality, and transparency in artificial intelligence through the lens of reproducibility. *AAAI Symposium on Educational Advances in AI (AAAI-EAAI)*.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 768–775, Varna, Bulgaria. INCOMA Ltd.

MLRC. Machine learning reproducibility challenge 2021. https://paperswithcode.com/rc2021.

NAACL. 2021. Naacl 2021 reproducibility checklist. https://2021.naacl.org/calls/reproducibility-checklist/.

Ted Pedersen. 2008. Last words: Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

David Picard. 2021. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, 32:5485–5495.

Robert Stojnic. 2020. Ml code completeness checklist. Medium, https://medium.com/paperswithcode/ml-code-completeness-checklist-e9127b168501.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Burak Yildiz, Hayley Hung, Jesse H Krijthe, Cynthia CS Liem, Marco Loog, Gosia Migut, Frans A Oliehoek, Annibale Panichella, Przemysław Pawełczak, Stjepan Picek, et al. 2021. Reproducedpapers. org: Openly teaching and structuring machine learning reproducibility. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 3–11. Springer.