Towards Reproducible Machine Learning Research in Information Retrieval

Ana Lucic
University of Amsterdam
a.lucic@uva.nl

Maurits Bleeker
University of Amsterdam
m.j.r.bleeker@uva.nl

Maarten de Rijke
University of Amsterdam
m.derijke@uva.nl

Koustuv Sinha
McGill University
koustuv.sinha@mail.mcgill.ca

Sami Jullien
University of Amsterdam
s.jullien@uva.nl

Robert Stojnic
Facebook AI Research
rstojnic@fb.com

ABSTRACT

While recent progress in the field of machine learning (ML) and information retrieval (IR) has been significant, the reproducibility of these cutting-edge results is often lacking, with many submissions failing to provide the necessary information in order to ensure subsequent reproducibility [20, 21, 32]. Despite the introduction of self-check mechanisms before submission (such as the Reproducibility Checklist [31]), criteria for evaluating reproducibility during reviewing at several major conferences [4, 11, 28], artifact review and badging framework [18], and dedicated reproducibility tracks and challenges at major IR conferences [8, 14–17], the motivation for executing reproducible research is lacking in the broader information community. We propose this tutorial as a gentle introduction to help ensure reproducible research in IR, with a specific emphasis on ML aspects of IR research.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**.

KEYWORDS

Information retrieval, Reproducibility

ACM Reference Format:
Ana Lucic, Maurits Bleeker, Maarten de Rijke, Koustuv Sinha, Sami Jullien, and Robert Stojnic. 2022. Towards Reproducible Machine Learning Research in Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3477495.3532686

1 MOTIVATION

Reproducibility of scientific results is a crucial component of scientific progress. It underpins trust in science. Reproducibility has been a primary concern in IR for many decades [22]. As a discipline that is strongly rooted in experimentation, it has long since stressed the importance of repeatability of experiments, for instance, by relying on the common Cranfield paradigm [9], developing shared experimental collections, and running broadly supported collaborative benchmarking activities [17]. With today's increased role for machine learning-based approaches to information retrieval (IR), significant new challenges have emerged for reproducibility in IR, as the experimental conditions in which modern IR experimentation have become far more complex, in terms of data, libraries, dependencies, baselines, and the sheer volume of publications from very diverse technological (sub)communities in our discipline.

Two key dimensions emerge in the IR community's thinking about reproducibility. One has to do with the *mechanics* (i.e., practices and resources), the other has to do to with the notion of *generalizability* (how can we make our scientific findings generalizable, and in which dimensions?). Through a large number of bottom-up initiatives over the past decade — badging, special conference tracks, workshops, etc. — good practices and principles for reproducibility in IR are being discovered and shared, and good (as well as not so good) examples of reproducible research are being generated, both concerning the mechanics dimension and concerning the generalizability dimension. The time is right to bring these many advances together in the form of a tutorial — to help the IR community learn about these advances *and* to help the IR community advance the reproducibility of its own science.

Our focus will be on machine learning aspects of IR, as that is where we believe a large number of lessons and best practices have been learned and can be shared.

As reproducibility is not a widely taught aspect in most curricula in which IR is being taught, in addition to sharing principles and practices about IR reproducibility, we also offer ways of using reproducibility as a teaching tool as part of this tutorial.

2 OBJECTIVES

The objective of this tutorial is to first impart the basic tenets of reproducibility, using which the audience can improve their own research in IR. After attending our tutorial, we expect the audience to be familiar with the processes required to conduct reproducible research, and be aware of the broader efforts in the community to improve the state of reproducible research.

Another objective of this tutorial is to showcase the use of reproducibility as a teaching tool, in order to equip the audience to further impart the knowledge and best practices of reproducible research in their own setting, through course offerings or educational programs at their home institutions.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3532686

Our final objective is to facilitate a discussion between our tutorial participants and several members of the SIGIR community about reproducibility in IR through a panel discussion.

2.1 Introduction to Reproducibility

The objective of the introduction is to 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. For example, 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 IR and in ML [7, 21]? What would it look like if we, as a community, prioritized reproducibility?

After the introduction, the audience will be able to provide examples of successes and failures of reproducibility in non-CS 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 IR.

2.2 Reproducibility in Information Retrieval

The objective of this part is to focus on reproducibility in IR specifically and understand the challenges that the IR community is facing, and how these differ from the challenges in ML, and in science more broadly.

We will discuss examples of results that were reproducible and those that were not reproducible. For the latter, we will categorize reproducibility failures in IR, such as the work by Dacrema et al. [10] as well as work that has been published in reproducibility tracks at IR conferences [14, 15, 33].

2.3 Mechanisms for Reproducibility

The purpose of this part of the tutorial is to understand the various existing initiatives to tackle the reproducibility problem in ML, NLP, and IR, such as reproducibility checklists [4, 11, 28, 31], and ACM's badging system [1, 18].

Another objective is to introduce the ML Reproducibility Challenge,[1] where researchers investigate the results of papers at top ML conferences by reproducing the experiments and writing a report about their experiences. There are several university-level courses which have incorporated a reproducibility project via the ML Reproducibility Challenge, which is the subject of the following part of the tutorial.

2.4 Reproducibility as a Teaching Tool

Our objective in this section is to discuss how reproducibility can be used as a tool in education to improve the scientific process, scientific discourse, and science in general. It is imperative that we teach the next generation about conducting reproducible research.

After this part of the tutorial, attendees will have the tools to be able to set up a reproducibility project in a university-level computer science course. We will provide recommendations for using reproducibility as a teaching tool based on our experiences [12, 23, 24], and reflect on the lessons learned.

2.5 Panel Discussion

We will conclude our tutorial with a panel discussion about reproducibility in IR with one moderator from our teaching team and three invited panelists from the SIGIR community with diverse backgrounds in reproducibility. The moderator and invited panelists will be on-site at the SIGIR conference.

3 RELEVANCE TO THE INFORMATION RETRIEVAL COMMUNITY

In the tutorial, we introduce and contrast reproducibility [13], discuss papers reflecting on the reproducibility crisis in ML and IR [2, 3, 6, 7, 26, 29], including possible reasons for this crisis [21]. This includes barriers to reproducibility, such as lack of code availability [29, 34] and the influence of different experimental setups [5, 19, 30].

Our focus is on the reproducibility of ML-based research in IR, as we believe that that is the area where the community has made most progress, with the ECIR reproducibility track as one of the primary outlets for this type of research, since 2015. Unfortunately, other areas of importance to IR have (so far) witnessed less work devoted to reproducibility, such as work on understanding users, either in the small through users studies [35] or at scale through online surveys [25].

We note that parts of this tutorial will be part of a half-day tutorial at ACL 2022. The focus of that tutorial will be on ML research reproducibility in NLP.[2]

4 FORMAT AND DETAILED SCHEDULE

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

I: Introduction to Reproducibility (35 mins)
- 1.1 Definitions and challenges
- 1.2 Reproducibility crisis in ML
- 1.3 Reproducibility in fields outside of computer science
- 1.4 Best practices for conducting reproducible research

II: Reproducibility in IR (35 mins)
- 2.1 Reproducibility challenges in the IR community
- 2.2 Reproducibility failures in IR
- 2.3 Reproducibility tracks at SIGIR,ECIR [8, 14, 15, 33]

III: Mechanisms for Reproducibility (35 mins)
- 3.1 Reproducibility checklists [4, 28]
- 3.2 ACM badging system [1]
- 3.3 ML Reproducibility Challenge [27]

IV: Reproducibility as a Teaching Tool (35 mins)
- 4.1 How to incorporate a reproducibility project in a university-level course [12, 23]
- 4.2 Courses that have used reproducibility as a teaching tool [24, 36]

V: Panel Discussion (40 mins)
- 5.1 Discussion
- 5.2 Closing

5 TYPE OF SUPPORT MATERIALS TO BE SUPPLIED TO ATTENDEES

We will share the following materials with participants:

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

[2] https://acl-reproducibility-tutorial.github.io

(1) **Slides:** All slides will be made publicly available.
(2) **Annotated bibliography:** An annotated compilation of references will list all works discussed in the tutorial and should provide a good basis for further study.

ACKNOWLEDGEMENTS

This research was partially funded by Ahold Delhaize, the Nationale Politie, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, and the Netherlands Organisation for Scientific Research under project nr 652.001.003. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] ACM. 2019. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging.
[2] Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. 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*. Association for Computational Linguistics, Online, 381–393.
[3] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. 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*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 249–258.
[4] 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.
[5] Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *International Conference on Machine Learning*. PMLR, 725–734.
[6] Timo Breuer. 2020. Reproducible Online Search Experiments. In *European Conference on Information Retrieval*. Springer, 597–601.
[7] Timo Breuer, Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, Philipp Schaer, and Ian Soboroff. 2020. How to measure the reproducibility of system-oriented IR experiments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 349–358.
[8] Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu. 2019. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). 1432–1434.
[9] C. W. Cleverdon. 1967. The Cranfield tests on index language devices.. In *Aslib Proceedings*, Vol. 19. 173–192.
[10] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*. 101–109.
[11] Jesse Dodge. 2020. Guest Post: Reproducibility at EMNLP 2020. https://2020.emnlp.org/blog/2020-05-20-reproducibility.
[12] Jesse Dodge. 2020. The Reproducibility Challenge as an Educational Tool. Medium, https://medium.com/paperswithcode/the-reproducibility-challenge-as-an-educational-tool-cd1596e3716c.
[13] Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. (2009).
[14] ECIR. 2021. ECIR: Call for Reproducibility Track papers. https://www.ecir2021.eu/call-for-reproducibility-track.
[15] ECIR. 2022. ECIR: Call for Reproducibility Track papers. https://ecir2022.org/calls/reproducibility/.
[16] Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello. 2016. *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*. Vol. 9626. Springer.
[17] Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on" Reproducibility of Data-Oriented Experiments in e-Science". In *ACM SIGIR Forum*, Vol. 50. ACM New York, NY, USA, 68–82.
[18] Nicola Ferro and Diane Kelly. 2018. SIGIR initiative to implement ACM artifact review and badging. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 4–10.
[19] 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)*. Association for Computational Linguistics, Sofia, Bulgaria, 1691–1701.
[20] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[21] Matthew Hutson. 2018. Artificial Intelligence Faces Reproducibility Crisis. *Science* 359 (2018), 725–726. Issue 6377.
[22] Brian E Lantz. 1981. The relationship between documents read and relevant references retrieved as effectiveness measures for information retrieval systems. *Journal of Documentation* (1981).
[23] 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.
[24] Ana Lucic, Maurits Bleeker, Sami Jullien, Samarth Bhargav, and Maarten de Rijke. 2021. Reproducibility as a Mechanism for Teaching Fairness, Accountability, Confidentiality, and Transparency in Artificial Intelligence. In *Proceedings of the AAAI Symposium on Educational Advances in AI*.
[25] Rishabh Mehrotra, Mounia Lalmas, Doug Kenney, Thomas Lim-Meng, and Golli Hashemian. 2019. Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations. In *The World Wide Web Conference on - WWW '19*. ACM Press.
[26] 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)*. INCOMA Ltd., Varna, Bulgaria, 768–775.
[27] MLRC. 2021. Machine Learning Reproducibility Challenge 2021. https://paperswithcode.com/rc2021.
[28] NAACL. 2021. NAACL 2021 Reproducibility Checklist. https://2021.naacl.org/calls/reproducibility-checklist/.
[29] Ted Pedersen. 2008. Last words: Empiricism is not a matter of faith. *Computational Linguistics* 34, 3 (2008), 465–470.
[30] 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* (2021).
[31] 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* (2020).
[32] Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems* 32 (2019), 5485–5495.
[33] SIGIR. 2022. SIGIR 2022 Call for Reproducibility Track Papers. https://sigir.org/sigir2022/call-for-reproducibility-track-papers/.
[34] Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics* 44, 4 (Dec. 2018), 641–649.
[35] Zijan Xu. 2020. *Generalizability and Reproducibility of Search Engine Online User Studies*. Master's thesis. Virginia Polytechnic Institute and State University.
[36] 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*. Springer, 3–11.