A Reproducibility Analysis of Portuguese Computational Processing Conferences: A Case Study

Daniel A. Leal and **Anthony Irlan M. Luz** and **Rafael T. Anchiêta**
Artificial Intelligence, Robotics and Automation Laboratory (LIARA)
Federal Institute of Piauí (IFPI), Picos, PI, Brazil
danielaraujoleal985@gmail.com, marquesanthony62@gmail.com,
rta@ifpi.edu.br

Abstract

The Association for Computing Machinery (ACM) considers an experiment reproducible when a different and independent group obtains the same result using the artifacts from the author's investigation. Reproducibility is an increasing concern in the scientific community. Several attempts have been made to mitigate the reproducibility crisis, such as calls, chairs' blogs, special themes, and shared tasks. In this paper, we present a reproducibility analysis in the Portuguese computation processing conferences. We analyzed sixty-five papers from the STIL and PROPOR conferences and found that only eight were reproducible. The non-reproducible papers were due to the lack of complete documentation, broken links, and the available source code not working. To improve the reproducibility at these conferences, we suggest a reproducibility review process and an award category for the best reproducible papers.

1 Introduction

In an era marked by an unprecedented surge in data generation and computational capabilities, Artificial Intelligence (AI) has emerged as a transformative force, reshaping industries, economies, and the very fabric of society itself. The pervasive influence of AI technologies is evident in diverse domains, ranging from healthcare (Huang et al., 2020) and finance (Cohen, 2022) to autonomous vehicles (Gandhi et al., 2019) and personalized digital assistants (Campagna et al., 2019). In the scientific community, machine learning, a subset of AI, has proven invaluable for analyzing complex data, making predictions, and extracting meaningful insights. As researchers harness the potential of machine learning to unravel intricate problems, the demand for reproducible and transparent research practices becomes increasingly pronounced[1].

This paper delves into a critical facet of contemporary scientific inquiry, which is the reproducibility of research. A brief questionnaire on reproducibility with 1,576 researchers administered by Nature's Survey revealed that 70% of researchers have tried and failed to reproduce another scientist's experiments. Also, more than half have been unable to reproduce their own experiments. This problem has been called the "reproducibility crisis" (Baker, 2016).

There are some definitions of reproducibility (Belz, 2022). For example, the Association for Computing Machinery (ACM)[2] considers an experiment reproducible when a different and independent group obtains the same result using the artifacts from the author's investigation. The International Vocabulary of Metrology (VIM)[3] defines reproducibility as a measurement precision under reproducibility conditions of measurement. These conditions must be known and recorded and include but are not limited to the source code, hyperparameters, dependencies, and runtime environment. However, this is complicated by the field's recent reliance on deep learning models that are challenging to interpret, have billions of hyperparameters, and are highly sensitive to small changes in architecture and environment. These distinctive characteristics hinder reproducibility, as do the substantial computing resources often required for replication (Hutson, 2018; Abaho et al., 2021).

Faced with this challenge, there were workshops and checklist initiatives, tutorials (Lucic et al., 2022), conferences promoting reproducibility via calls, chairs' blogs, and special themes, and the first shared tasks, including REPROLANG'20 (Branco et al., 2020) and ReproGen'22 (Belz et al., 2022). These initiatives emerged as a need to improve

[1] https://crfm.stanford.edu/fmti/

[2] https://www.acm.org/publications/policies/artifact-review-and-badging-current

[3] https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf

reproducibility in machine learning and computational linguistics studies. Moreover, Wieling et al. (2018) have shown that the median citation count for studies with working links to the source code is higher.

In this context, this paper aims to investigate reproducibility in research at conferences with a focus on Portuguese Computational Processing. To achieve this objective, we analyzed and tested a series of works published at two major events in the Portuguese language area: the Symposium on Information Technology and Human Language (STIL) and the International Conference on Computational Processing of the Portuguese Language (PROPOR). The first is the main event supported and organized by the Special Committee on Natural Language Processing of the Brazilian Computing Society (SBC)[4]. PROPOR is the main conference in the area of Computational Processing of Portuguese. More specifically, PROPOR is held every two years, alternating between Portugal and Brazil[5,6].

We analyzed sixty-five papers, and only eight (12%) were reproducible. This result indicates the need to define strategies to improve the reproducibility of these conferences. Most of the non-reproducible papers were due to the lack of clear documentation indicating the steps to be followed and broken links, i.e., links no longer available, demonstrating a lack of maintenance of the artifacts produced in the scientific research.

The rest of this paper is organized as follows. Section 2 briefly presents related works. In Section 3, we outline the approach to attempt to reproduce scientific papers. Section 4 details our analysis and results, highlighting the main findings. Finally, in Section 5, we conclude the paper and propose future work.

2 Related Work

The task of reproducibility often involves attempting to achieve results close enough to the ones reported in the paper with little or no reliance on the released software artifacts, if available.

Raff (2019) attempted to quantify the reproducibility ratio of 255 papers published at NeurIPS from 1984 to 2017. The author selected different thresholds for a minimally acceptable error for algorithmic and empirical claims, ultimately reporting a 63% reproducibility ratio.

Wieling et al. (2018) surveyed 395 papers presented at the Association Computational Linguistics (ACL) 2011 and 2016 conferences and identified whether links to data and code were provided. Then, they attempted to reproduce the results of ten papers using the provided code and data. They ultimately found results close to those reported in six papers.

Arvan et al. (2022) investigated trends in source code availability at computational linguistics conferences, especially those that promote reproducibility. The study analyzed eight papers from the Empirical Methods in Natural Language Processing (EMNLP) 2021 conference. The authors found that source code releases leave much to be desired. They suggest all conferences require self-contained artifacts and provide a venue to evaluate such artifacts at the time of publication, including small-scale experiments and explicit scripts to generate each result to improve the reproducibility of their work.

Storks et al. (2023) conducted a study with 93 students in an introductory Natural Language Processing (NLP) course, where students reproduced the results of recent NLP papers. The authors found that programming skills and comprehension of the students' research papers had a limited impact on their time completing the exercise. The authors also found accessibility efforts by research authors to be the key to success, including complete documentation, better coding practice, and easier access to data files. Finally, the authors recommended that NLP researchers pay close attention to these simple aspects of open-sourcing their work and use insights from beginners' feedback to provide actionable ideas on supporting them better.

Magnusson et al. (2023) provide the first analysis of the Reproducibility Checklist created in 2020 by examining 10,405 anonymous responses. After the Checklist's introduction, the authors found evidence of an increase in the reporting of information on efficiency, validation performance, summary statistics, and hyperparameters. They found that the 44% of submissions that gather new data are 5% less likely to be accepted than those that did not; the average reviewer-rated reproducibility of these submissions is also 2% lower relative to

[4] https://sites.google.com/view/ce-pln/eventos/stil

[5] https://sites.google.com/view/ce-pln/eventos/propor

[6] Coincidently, this year, PROPOR, which will be held in Galicia, will be the first exception.

the rest. Finally, the authors found that only 46% of submissions claim to open-source their code, though submissions that do have an 8% higher reproducibility score relative to those that do not, the most for any item.

Our paper is in the same direction as Arvan et al. (2022). However, we are interested in Portuguese conferences such as STIL and PROPOR to investigate if the scientific papers published at these avenues are reproducible.

In what follows, we present our methodology to investigate reproducibility.

3 Methodology

Aiming to investigate reproducibility in scientific papers from Portuguese conferences, we organized our methodology in three steps.

1. Get the latest published papers at the STIL and PROPOR conferences.

   We chose these conferences because STIL and PROPOR focus on the computational processing of Portuguese, with the former being one of the main events for the Brazilian Portuguese language, while the latter is the main event in the area. Also, we chose the latest published papers to avoid problems with old programming languages and their dependencies and libraries.

2. Extract source code and data from these papers.

3. Attempt reproducing the reported results in the papers from the available source code and data.

   Since deep learning models have thousands of parameters, presenting them in a scientific paper is difficult due to page limitations. Thus, we tried to reproduce only papers that made data and source code available.

We adopt that methodology with the intention of answering the following research question. Are the NLP papers with a focus on the Portuguese language reproducible?

In the following section, we detail our analysis and results.

4 Analysis and Results

Firstly, we gathered some papers from the latest published papers from the STIL and PROPOR conferences. As shown in Table 1, we got 57 papers from STIL and 80 from PROPOR. The published papers in STIL are publicly available at the SOL SBC[7]. PROPOR conference papers are available at Springer[8].

Conference	Period	Number
STIL	2019 - 2021	57
PROPOR	2020 - 2022	80

Table 1: Gathered papers from STIL and PROPOR.

Next, we automatically parsed these papers (and manually checked them) to extract the URLs of the source code and data. As we can see in Table 2, 17 (30%) and 48 (60%) papers from STIL and PROPOR, respectively, have links to code repositories. It is important to say that all of these papers present a strategy for dealing with an NLP task, presenting experiments and results on the developed method. Thus, we believe that it is important for authors to make the data and source code of their strategy available.

Conference	Period	Number
STIL	2019 - 2021	17 (30%)
PROPOR	2020 - 2022	48 (60%)

Table 2: The number of papers with links to code repositories.

After extracting the links from the papers to code repositories, we began reproducibility evaluations by reading the papers. If there were instructions explaining the developed method, we followed them and recorded information about the process and the individual results of each evaluation. We did not allocate limited time and computational resources to each paper. We reported whether we were able to reproduce the experiments of the paper or not. We stopped trying to reproduce the results when some resource was missing or the source code had errors that were too difficult to fix.

After evaluating the reproducibility of the 40 works mentioned above, we found that 57 (88%) could not be reproduced. In only 8 (12%) cases, we can reproduce the results reported by the authors, 6 from PROPOR, and 2 from STIL, as shown in Table 3.

Of the 57 works submitted for analysis, we

[7] https://sol.sbc.org.br/index.php/stil/issue/archive

[8] https://link.springer.com/conference/propor

Number	Reproducible	Non-reproducible
65	8 (12%)	57 (88%)

Table 3: Relationship between reproducible and non-reproducible works.

excluded 22 of them, as they fell into the "non-reproducible nature"(NRN) category. These works, generally related to comparisons, presentation of tools, or construction of corpora, did not fit the research profile that could be easily downloaded and reproduced. After this exclusion, a more in-depth analysis of the remaining 35 works was carried out, aiming to identify trends and obtain insights that could improve their reproducibility.

During the analysis of the remaining 35 works, we identified and labeled them as follows:

- 13 of them were NDOC (No documentation), that is, the documentation provided was insufficient or non-existent, not offering clear guidance for executing the code. Despite several attempts to execute the code, we can not execute them.

- 3 of them were NDEP (No dependencies). The inability to reproduce the results was related to the lack of dependencies or availability of the necessary corpus, which was not made available in the code repository or in another repository. In some cases, it was necessary to request the corpus from the authors, and even then, we could not reproduce the results. It is important to mention that we can not execute the code.

- 12 of them were BLINK (Broken link). The link to the source code repository was broken, i.e., it was not publicly available, making it more difficult to reproduce the results.

- 7 of them were ALLREQ (All requirements). Although the source code has documentation and requirements to execute it, reproducing the results proved unfeasible due to problems in the source code. That is, the available source code was not working.

Number	NRN	NDOC	NDEP	BLINK	ALLREQ
57	22(39%)	13(23%)	3(5%)	12(21%)	7(12%)

Table 4: Results of analysis of non-reproducible papers.

From this analysis, we have learned that only making the source code available does not guarantee that the reported results will be reproducible. It is necessary to have clear documentation showing the steps to be followed to execute the source code. Even with the documentation and requirements, some papers were not reproducible. We believe the authors provided an old source code version for these cases. Thus, testing and maintaining the source code updated is also necessary.

Despite analyzing recently published papers, we had problems with programming language dependencies and broken links. We are aware that experienced authors have little time to test the source code and verify if all the links work. As a suggestion, we also have learned that a solution for these cases is to use containers (e.g., Docker[9]), i.e., a structure that includes all dependencies and libraries necessary to execute the source code, avoiding such problems.

We believe that this result indicates the need to promote the reproducibility of these conferences. Moreover, the small number of reproducible papers may raise a question. Should a non-reproducible paper be rejected? We leave this question for future research.

An alternative for improving reproducibility is to adopt checklists for the submitted papers and request the source code at the time of submission. Therefore, we suggest a review focused on reproducibility, which young researchers could organize.

5 Final Remarks

This paper presented a reproducibility analysis in the Portuguese computational processing conferences. We investigated sixty-five papers from the STIL and PROPOR conferences, which are the main events focused on Portuguese language processing. In our analysis, we found that only eight papers were reproducible. Most non-reproducible papers were due to the lack of complete documentation, broken links, and problems in the source code. This result indicates the need to promote the reproducibility of these conferences and define strategies to improve them.

For future work, we intend to investigate the impact of non-reproducible papers in the scientific community.

[9] https://www.docker.com/

Acknowledge

The authors are grateful to IFPI, CNPq, and Virtex for supporting this work.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022. Reproducibility in computational linguistics: Is source code enough? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2350–2361, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Giovanni Campagna, Silei Xu, Mehrad Moradshahi, Richard Socher, and Monica S Lam. 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 394–410, Phoenix, AZ, USA. Association for Computing Machinery.

Gil Cohen. 2022. Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics*, 10(18):3302.

G Meera Gandhi et al. 2019. Artificial intelligence integrated blockchain for training autonomous cars. In *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 157–161, Chennai, India. IEEE.

Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters*, 471:61–71.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726.

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, and Robert Stojnic. 2022. Towards reproducible machine learning research in natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–11, Dublin, Ireland. Association for Computational Linguistics.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, pages 5486–5496, Vancouver, Canada. Curran Associates, Inc.

Shane Storks, Keunwoo Yu, Ziqiao Ma, and Joyce Chai. 2023. NLP reproducibility for all: Understanding experiences of beginners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10199–10219, Toronto, Canada. Association for Computational Linguistics.

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.