

- This is a summary of a paper about sustainability in NLP.
- Your task is to read the summarised version of the paper and identify which aspects belong to the Triple Bottom Line (TBL): People, Planet and Performance. If you do not identify all aspects in your paper, leave it blank.
- Based on the findings from this paper, provide ideas to how that could be implemented in Machine Translation Quality Assessment.
- And remember: Actively do good! After this class, I suggest you read the full paper and reflect about your practice.

Title: Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning

Reference: Henderson, P., Hu, J., Romoff, J., Straszak, E., Yusuf, H., Bacon, T., Jain, S. and Jaques, N., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, [online] Available at: <https://arxiv.org/abs/2002.05651>

Framework for Reporting Energy and Carbon Footprints in ML

This paper presents a framework that aims to address the environmental impact of machine learning models, focusing specifically on the need to systematically track the energy consumption and carbon emissions involved in machine learning (ML) experiments. The authors underscore that training large ML models, particularly in natural language processing (NLP), is highly resource-intensive and often requires considerable energy, which in turn results in a significant carbon footprint.

The proposed framework is built around the need for transparency in reporting the energy used during model training and deployment, alongside the associated carbon emissions. The authors recommend that AI researchers systematically report the total energy consumed, typically measured in kilowatt-hours, while also specifying whether the energy comes from renewable or non-renewable sources. In



Module: Machine Translation Quality Assessment
Lecturer: João Lucas Cavalheiro Camargo
Instructions - Lab 4 - Ethics in Evaluation

In addition to energy tracking, the framework stresses the importance of calculating the carbon emissions that result from ML processes. This involves estimating the emissions based on the amount of energy consumed and the type of electricity used by the data centres responsible for hosting the computations.

An essential component of the framework is the development of standardised green metrics to allow for fair comparisons of the environmental costs between different models and experiments. These metrics should evaluate not just the performance of the models, such as their accuracy, but also their energy efficiency and overall environmental impact. The paper further proposes the introduction of leaderboards to rank machine learning models publicly, based not only on their performance metrics like accuracy or speed but also on their energy consumption and carbon footprint. The goal of these leaderboards is to encourage researchers to optimise their models in a way that balances both high performance and low environmental cost, promoting the development of models that deliver good results while reducing their environmental impact.

Lastly, the framework also addresses the issue of accessibility in AI research, noting that the significant resource requirements for training large models currently limit participation in the field to well-funded institutions. By advocating for energy-efficient models, the framework aims to democratise AI research, making it more accessible to a broader group of researchers, including those with fewer resources. This emphasis on energy-efficient model design not only benefits the environment but also fosters a more inclusive AI research community.



MTQA
Training
with
Teacher João Lucas
Cavalheiro Camargo

Module: Machine Translation Quality Assessment

Lecturer: João Lucas Cavalheiro Camargo

Instructions - Lab 4 - Ethics in Evaluation

From this paper, what could fall under the category “People”?

From this paper, what could fall under the category “Planet”?

From this paper, what could fall under the category “Performance”?

How could we adapt this to Machine Translation Quality Assessment?