Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, Anne-Laure Ligozat

presented at SustaiNLP – November 2021

firstname.lastname@lisn.upsaclay.fr

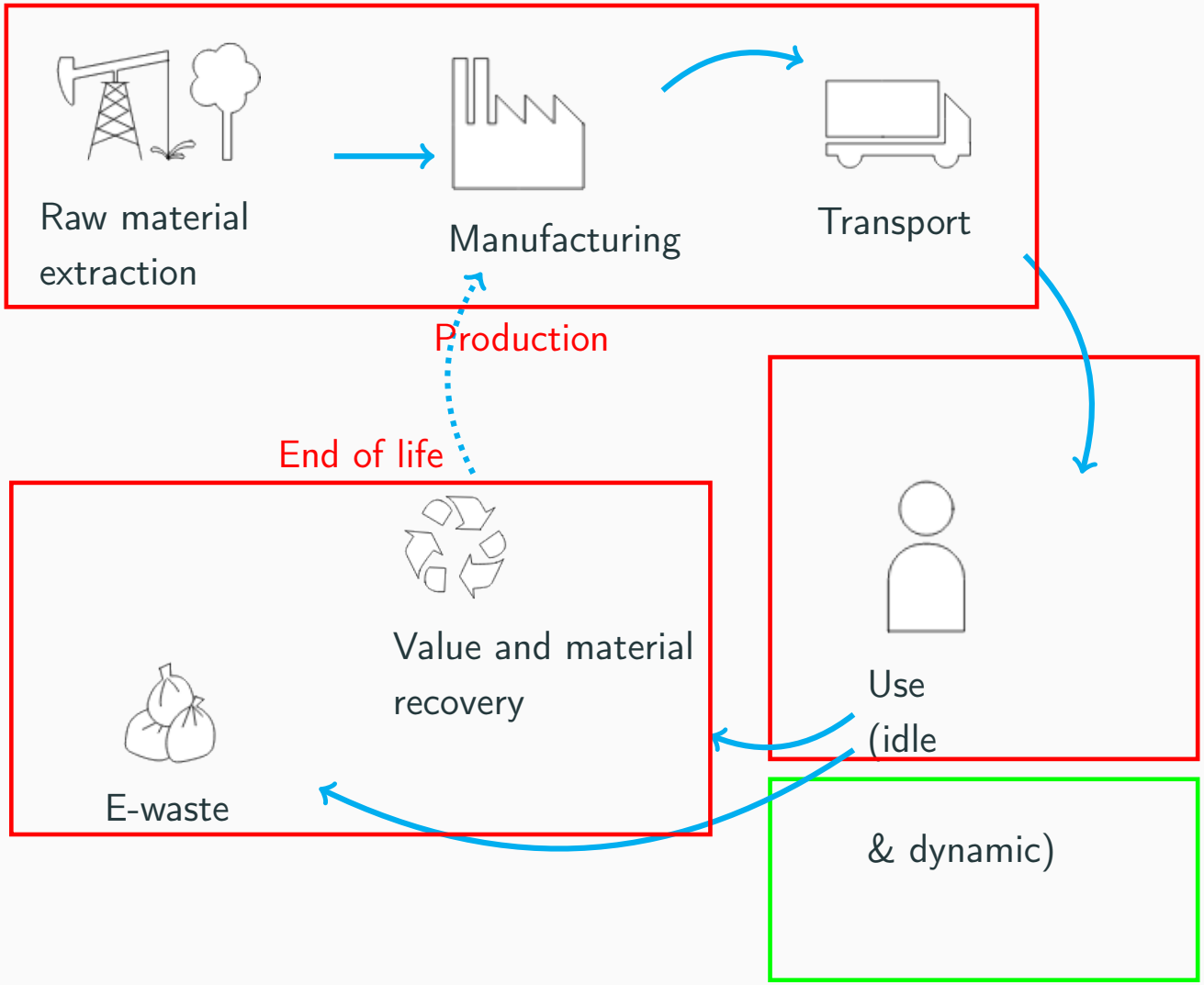
Why measure the impact of NLP experiments?

- Need for sustainable research
- Need for a global approach to evaluation, beyond leaderboards

Sources :
Strubell E, Ganesh A and McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Proc Annual Meeting of the Association for Computational Linguistics (ACL):3645-3650 (2019).
Ethayarajh K and Jurafsky D Utility is in the Eye of the User: A Critique of NLP Leaderboards. Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 4846-53. (2020).

How can we measure the impact of NLP experiments?

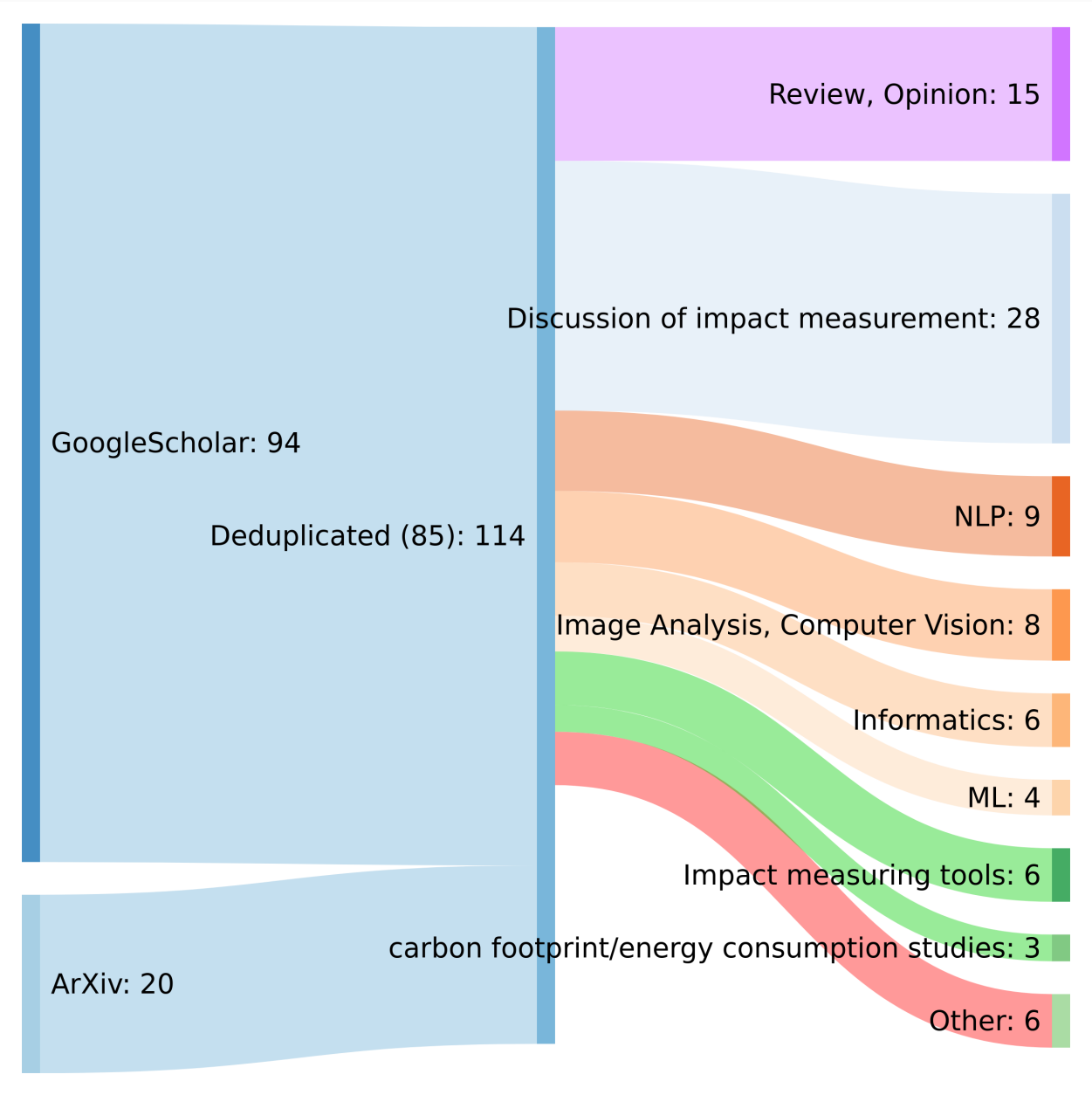
Sources of CO2 emissions include:



Can a tool provide CO2 impact measurement?

- Literature search:
  - Seed tools: Experiment Impact Tracker, Pyjoules, Carbon Tracker
  - Snowballing in Google Scholar + ArXiv "related papers"
- Selection criteria:
  - Freely available
  - usable in linux/mac OS
  - documented in a scientific publication
  - suitable to measure the impact of NLP experiments
  - CO2 equivalent measure

Literature survey



85 publications reviewed lead to identification of 6 tools providing C02 impact measurement for NLP

- Online tools
  1. Green Algorithms
  2. ML CO2 Impact ... newly available as *Code Carbon* toolkit
- Python toolkits
  3. Energy Usage
  4. Experiment Impact Tracker
  5. Carbon Tracker
  6. Cumulator

Criteria for characterizing tools

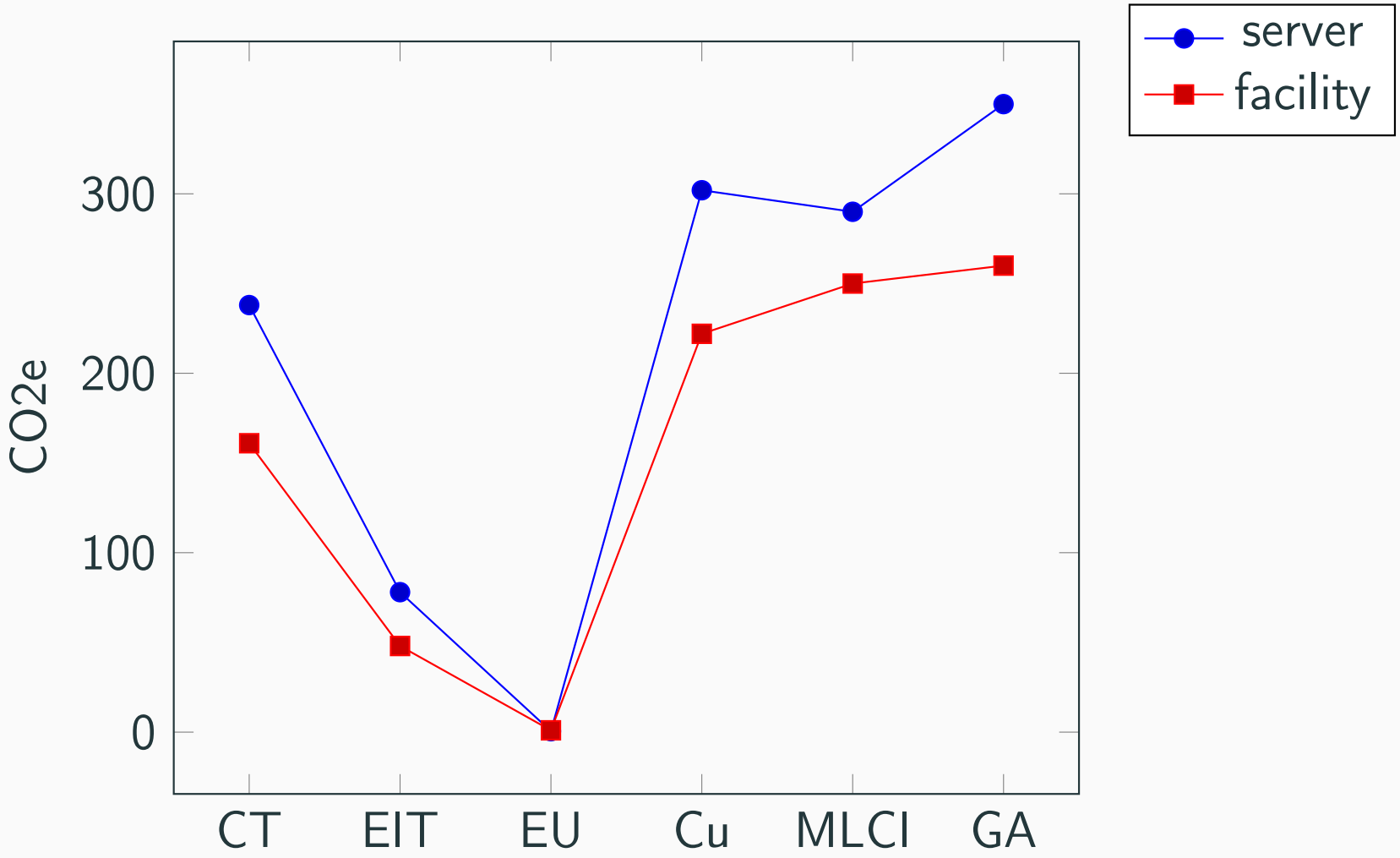
- 3 publication criteria
  1. Publication year
  2. Citations (overall, user studies)
- 7 technical criteria
  1. Availability, ease of installation
  2. Documentation, version
- 5 configuration criteria
  1. Source of carbon intensity and power usage effectiveness values
  2. Equipment covered by the measurements
- 2 functional criteria
  1. Sources of emissions targetted
  2. Type of hardware

Application to a named entity recognition task

- 2 NER tools
  - one that addresses flat entity recognition [Ma and Hovy, 2016]
  - one that addresses both flat and nested entity recognition, introduced by [Yu et al., 2020]
- 2 setups
  - GTX 1080 Ti GPUs used on a server
  - Tesla V100 GPUs used on a computing facility
- 2 datasets
  - QUAERO Broadcast News Extended Named Entity dataset [Galibert et al., 2010] (French press)
  - QUAERO French Med dataset [Névéol et al., 2014]
- 2 measures
  - energy consumption
  - carbon footprint

Results

for [Yu et al., 2020] on the French Press corpus



Why are the results so heterogeneous?

- Carbon intensity varies: CT used the average carbon intensity for EU-28 in 2017 (294.21 gCO2eq/kWh), while electricityMap gives around 30 to 40 gCO2eq/kWh
- Hardware options may not be available
- Tools not adapted to a multi-user setting
- Direct measures vs estimations

What did we learn about measuring CO2 impact in NLP?

- It is a recent but global endeavour
- Tools only account for dynamic use of hardware (1 in 4 sources of carbon emission)
- Tools provide different measures for the same experiments
  - direct measure vs. estimation of computation
  - values of Carbon Intensity, Power Usage Effectiveness (PUE)
  - some tools are not sensitive enough to capture small impact
- Server seems more carbon intensive than computing facility

Summary:

- 6 tools to evaluate NLP carbon emissions
- Only account for 1/4 sources of emissions
- Need to better understand measurements

References

Galibert, O., Quintard, L., Rosset, S., Zweigenbaum, P., Nédellec, C., Aubin, S., Gillard, L., Raysz, J.-P., Pois, D., Tannier, X., Deléger, L., and Laurent, D. (2010).
Named and specific entity detection in varied data: The quæro named entity baseline evaluation.
In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Ma, X. and Hovy, E. (2016).
End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014).
The QUAERO French medical corpus: A resource for medical entity recognition and normalization.
In Proc. BioTextM.

Yu, J., Bohnet, B., and Poesio, M. (2020).
Named entity recognition as dependency parsing.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6470–6476. Association for Computational Linguistics.