9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Language Models from the Sweatshop? Towards Guidelines For CSS Researchers to Avoid Ethical and Legal Issues With Off-The-Shelf Software.

Keywords: Natural Language Processing (NLP), Ethics of NLP, Ethics of computational research on human behavior, Large Language Models, ChatGPT

Extended Abstract

With the increasing pace of success in the area of Natural Language Processing (NLP), language models make their way into the toolkit of Computational Social Sciences (Grimmer et al., 2022) in various forms and for various purposes (Bonikowski et al., 2022; Knight, 2022).

Almost unnoticeably, this process sees researchers waiving direct control over both methods and data they use. Since models become more and more computationally demanding (Bender et al., 2021), the "pre-training" of these models is commonly performed by a few institutions that possess the necessary compute power (Whittaker, 2021). In this situation, researchers take on the role of consumers of off-the-shelf software. One cannot assume, however, that the training data fulfills ethical standards and that the resulting models do not pose legal risks.

The most recent iteration of large language models (LLMs) by OpenAI – ChatGPT – serves as a case in point. Not only are model architecture and data set inaccessible, the company allegedly used underpaid Kenyan workers to fine-tune their model (Xiang, 2023). This is ethically questionable. Some LLMs also raise legal questions. The software company GitHub, for example, faces a class-action lawsuit over their AI assistant's purported copyright infringements (Vincent, 2022).

LLMs are highly valuable, however (Do et al., 2022), and with the increasing amount of textual data produced everyday by people around the world (Lazer and Radford, 2017), Computational Social Scientists cannot afford to ignore the opportunities opened up by these models (Bonikowski and Nelson, 2022).

But several reasons speak for explicitly vetting pre-trained models for ethical and legal issues. First, there could be biases that negatively affect conclusions drawn from research (Akter et al., 2021). Second, malicious actors can perform supply-chain attacks by either attacking these models directly (Szegedy et al., 2014) or "poisoning" the data used to train them (Carlini et al., 2023). Third, researchers are responsible for real-world impact of their work, and thus vetting models averts harm for both research subjects and the researchers.

This work develops a set of guidelines that researchers can use as a checklist to verify pre-trained language models. These guidelines do not expect researchers to fully understand a model's architecture and do not significantly impede the researchers' ability to utilize LLMs in their analysis. Instead, they ensure that researchers can confidently use LLMs in their work, minimizing the ethical and legal risks for research subjects and themselves that could arise with the use of improperly curated off-the-shelf models.

The guidelines cover ethical and legal issues (accountability, dual-use, data-laundering, adversarial attacks) for methods (algorithmic bias, machine reasoning) and data (acquisition, consent, data bias) at various stages of the analysis (discovery, measurement, inference).

References

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N., 2021. Algorithmic bias in data-driven innovation in the age of AI. International Journal of Information Management 60, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Bender, E.M., Gebru, T., McMillan-Major, A., Mitchell, M., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? , in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, Virtual Event Canada, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bonikowski, B., Luo, Y., Stuhler, O., 2022. Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models. Sociological Methods & Research 51, 1721–1787. https://doi.org/10.1177/00491241221122317

Bonikowski, B., Nelson, L.K., 2022. From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. Sociological Methods & Research 51, 1469–1483. https://doi.org/10.1177/00491241221123088

Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., Tramèr, F., 2023. Poisoning Web-Scale Training Datasets is Practical. https://doi.org/10.48550/arXiv.2302.10149

Do, S., Ollion, É., Shen, R., 2022. The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. Sociological Methods & Research 00491241221134526. https://doi.org/10.1177/00491241221134526

Grimmer, J., Roberts, M.E., Stewart, B.M., 2022. Text as data: a new framework for machine learning and the social sciences. Princeton University Press, Princeton, New Jersey Oxford.

Knight, C., 2022. When Corporations Are People: Agent Talk and the Development of Organizational Actorhood, 1890–1934. Sociological Methods & Research 00491241221122528. https://doi.org/10.1177/00491241221122528

Lazer, D., Radford, J., 2017. Data ex Machina: Introduction to Big Data. Annu. Rev. Sociol. 43, 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks. arXiv:1312.6199 [cs].

Vincent, J., 2022. The lawsuit that could rewrite the rules of AI copyright. The Verge.

Whittaker, M., 2021. The steep cost of capture. interactions 28, 50–55. https://doi.org/10.1145/3488666

Xiang, C., 2023. OpenAI Used Kenyan Workers Making $2 an Hour to Filter Traumatic Content from ChatGPT. Vice. URL https://www.vice.com/en/article/wxn3kw/openai-used-kenyan-workers-making-dollar2-an-hour-to-filter-traumatic-content-from-chatgpt (accessed 2.17.2023).