Language Models from the Sweatshop?

Hendrik Erz[1], Sebastian Gießler[2]

[1]: Institute for Analytical Sociology, Linköping University; hendrik.erz@liu.se

[2]: International Center for Ethics in the Sciences and Humanities (IZEW), University of Tübingen; sebastian.giessler@izew.uni-tuebingen.de

With the advent of large language models (LLMs) such as BERT or ChatGPT, CSS researchers have a novel resource for working with the increasing amounts of text available in the digital era. This has clear benefits, as off-the-shelf models are readily available and easy to deploy. By using such pre-trained models, however, researchers waive control over both methods and data they use. This work explores the implications of using pre-trained language models for CSS on several dimensions.

What are Pre-Trained Language Models?

A pre-trained language model is a neural network that has been trained, e.g., in next-word prediction and next-sentence prediction with a large training dataset. Such a model can either be used as-is, or fine-tuned for a specific downstream task [2; 10].

Instruments based on Pre-Trained Models

- **Active Learning**: BERT-style transformer models for corpus prediction [3]
- **POS-Tagger**: LSTM-style networks, trained on so-called treebanks [4]
- **Word Embeddings**: Shallow SGNS-models, trained on general corpora [5]; but also SVD-based (GloVe [6])
- **OCR software:** e.g., tesseract; includes a pre-trained LSTM-network

Models that can be Pre-Trained

- Word Embeddings (SGNS-models)
- LSTM-networks
- Transformer models

Introduction

- With the increasing popularity & power of LLMs and the increasing prominence of text data, computational social scientists use more and more pre-trained language models [1]
- Usage of pre-trained language models embeds a **secondary data analysis pipeline** into the research
- This can have implications from benign degradations of explanatory power to invalidation of results
- The severity of these implications is not yet fully understood
- This project focuses on three dimensions: **ethics and legal issues**, **theory**, and **methods**.

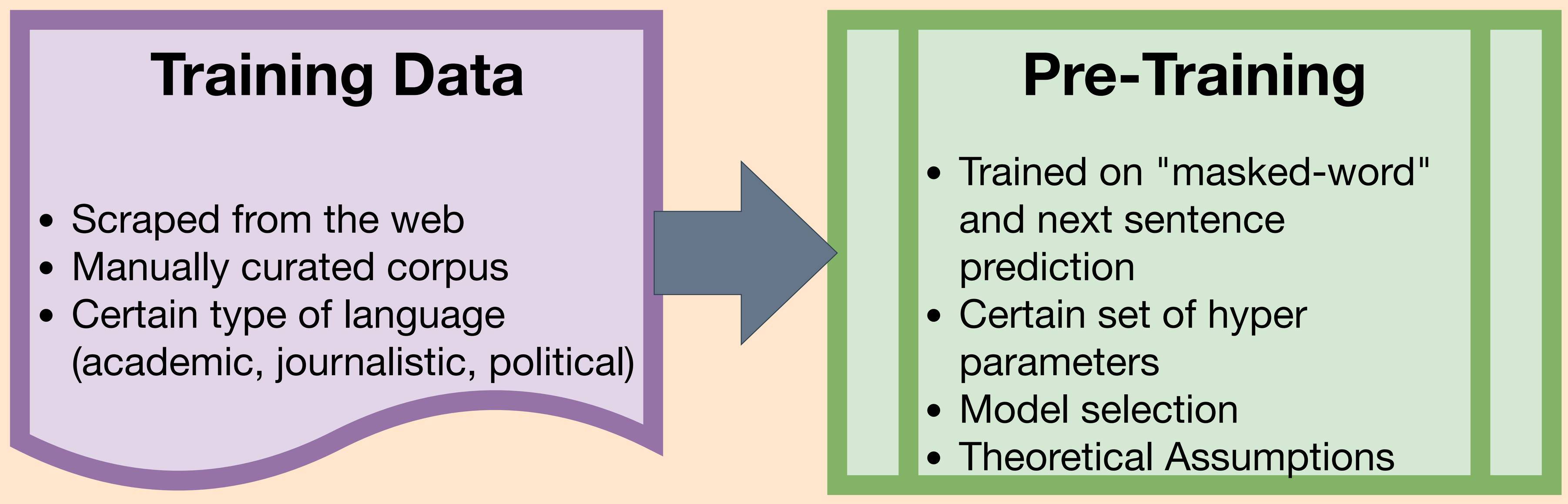

Fig. 1: A Secondary Data Analysis Pipeline As Part of CSS Research

The dataset parrots the condition of its upbringing

Implications I
Ethical & Legal

- Ethical and legal issues arise both at the data collection and model selection stages [7]
- The data is scraped from the web (e.g., CommonCrawl):
  - violates many websites's ToS
  - includes PII (violates GDPR)
  - violates copyright
- The models then need to be fine-tuned:
  - Especially large generative models may employ "clickworkers" from AMT or developing countries, paying far below a living wage

Four ethical principles [8, 9]

- **Respect for persons**
  - Basis for the principle of "informed consent"; e.g., in GDPR § 14(5b) and § 9(2e)
- **Beneficience**
  - Maximize benefits and minimize risks for research subjects
  - For LLMs, the exclusion of vulnerable or marginalized populations makes the models misclassify or outright miss any factors that are specific to these groups
- **Justice**
  - The curation of datasets is frequently made in an ad-hoc fashion without reflection of societal biases that the researchers who collect the data have internalized
- **Respect for Law and Public Interest**
  - First, datasets should not infringe upon laws such as GDPR-rights
  - Second, no copyrighted material or otherwise legally protected information ends up in the training data without contractual consent
  - Third, it demands transparency and accountability for both training data and model

Implications II
Theory

- The training of each pre-trained model is based on a certain model of the world
- This means that no model is bias- or assumption-free
- If the theoretical backing of a pre-trained model is incongruent with the theory behind a research task utilizing this model, this can make results hard to defend
- **Example 1: Dependency parsers**
  - Computational Linguists frequently train similar languages together, which assumes that the differences between various languages are not meaningful [11]
  - This is problematic for work that wants to find language differences in social groups, even if they speak different languages
- **Example 2: Training Data**
  - The larger the models, the more training data they need. This data, however, only comes from the internet; digitized resources are scarce.
  - This means that LLMs often only encode a part of culture, rendering them theoretically inadequate for working with any text older than, say, mid 20th century [12]
- **Example 3: Time-Series Data**
  - Pre-trained LLMs encode the data without time-awareness.
  - In order to reflect time-based invariance; several models need to be trained on slices of the data [13]
  - However: how does one select adequate ranges?

Implications III
Methods

- Using Pre-Trained Language Models means that researchers implement a secondary data analysis pipeline within the model selection step of the original research (Fig. 1).
- Ergo, any assumptions of the pre-trained model become assumptions of the research.
- Model selection involves metrics such as BLEU, GLUE, F1, and others. These metrics may not capture what the research task at hand requires.
- **Example 1: Word embeddings**
  - Depending on the window size, word embeddings either encode synonyms or topically similar words [14]; which radically changes the interpretation of any downstream instrument that is being calculated based off it
- **Example 2: Active Learning/BERT**
  - BERT-models are pre-trained on masked word prediction and next-sentence prediction [2]. Hence, they have a specific assumption on language generation that might not hold for some classification tasks; for example scientific paper abstracts
- **Example 3: Prediction vs. Causation**
  - Even if models correctly classify text, one should not rely on the model output, as its black-box nature means that it could have arrived at the right conclusions based on a wrong model of the process [15]

Glossary

- **BERT**: Bidirectional Encoder Representations from Transformers
- **Fine-Tuning**: Taking a pre-trained language model and continuing training for a specific task
- **GDPR**: General Data Protection Regulation
- **GPT**: Generative Pre-Trained Transformer
- **LLM**: Large Language Model
- **LSTM**: Long Short-Term Memory
- **PII**: Personally Identifying Information
- **ToS**: Terms of Service
- **Transfer Learning**: The process of taking a pre-trained LLM and fine-tuning it on a downstream task

Sebastian Gießler

Hendrik Erz

Contact us: Scan the QR codes for our VCards

References

[1]: Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, New Jersey Oxford: Princeton University Press, 2022.
[2]: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*, May 24, 2019. http://arxiv.org/abs/1810.04805.
[3]: Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., & Slonim, N. (2020). Active Learning for BERT: An Empirical Study. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7949–7962. https://doi.org/10.18653/v1/2020.emnlp-main.638
[4]: Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *ArXiv:2003.07082 [Cs]*. http://arxiv.org/abs/2003.07082
[5]: Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.
[6]: Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
[7]: Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
[8]: Ryan, K. J. (1978). *The Belmont Report. Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* (No. 78–0012; U.S. Department of Health, Education, and Welfare (DHEW) Publications). The National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research. https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf
[9]: Kenneally, E., & Dittrich, D. (2012). The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2445102
[10]: Whittaker, M. (2021). The steep cost of capture. *Interactions*, *28*(6), 50–55. https://doi.org/10.1145/3488666
[11]: Li, Y. (n.d.). *Selecting Languages for Cross-Lingual Dependency Parsing*.
[12]: Martin, J. L. (2010). Life's a beach but you're an ant, and other unwelcome news for the sociology of culture. *Poetics*, *38*(2), 229–244. https://doi.org/10.1016/j.poetic.2009.11.004
[13]: Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135
[14]: Stoltz, D. S., & Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics*, 101567. https://doi.org/10.1016/j.poetic.2021.101567
[15]: Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726