On Bias and Influence

by Elaiza Bolislis, Dianne Del Pilar, and Neptune Sy

Introduction

Language models (LMs) are machine learning models designed to understand, process, and generate language. These models utilize statistical techniques and are trained on massive datasets in order to determine patterns and relationships in language. In other words, language models are optimized to mirror language systems.

While these language models allow machines to help humans perform and automate linguistic tasks–such as topic modeling, text classification, text translation, text summarization and more–it can also have its disadvantages. As language models' goal is to mimic language systems, these models inevitably imitate, and possibly perpetuate, biases and stereotypes present in languages as well. Specifically, biases and stereotypes can manifest in language models through the training data, model design and architecture, and data labeling and annotation (Hovy & Prabhumoye, 2021; Søby Borgqvist, 2023).

Biases and Stereotypes from Training Data

Biases in training data present itself in two main ways: either the data collected reflects the existing prejudices present in languages, or it is unrepresentative of reality because of selection bias (Hao, 2020).

As mentioned earlier, language models are trained on vast amounts of data largely from the internet, including websites, books, articles, and more. As a result, these models learn the patterns and relationships that appear in the training data. Hence, in the first case, if the data

itself contains biased languages–regardless of the type of bias it possess–there is a high possibility that the model adopts those prejudices.

Meanwhile, in the second case, the training data can either introduce or amplify biases in the model if the dataset has an unbalanced distribution of observations, which may be in terms of domain and genre, time of creation, demographics, or language and cultures. This happens because of data selection bias, where the data collection includes more information about a specific class of data than the others. Because of this, the model may perform better (and can be biased) towards a certain type of data compared to others (Hovy & Prabhumoye, 2021; Navigli et al., 2023).

Biases and Stereotypes from the Models

As language models recognize patterns and relationships between words and grammar, these models also get an idea of the general structure and meaning of language. This generalization capability of language models allows them to understand texts they did not encounter before and use the existing knowledge they have from training in a new context. While this serves as a great feature of language models, it also poses a problem. This ability enables models to always make a prediction even though they might not have sufficient information when given an input they are unfamiliar with, which, consequently, may lead to risks of stereotyping and making biased inferences (Hovy & Prabhumoye, 2021; Søby Borgqvist, 2023).

Biases and Stereotypes from Data Labeling and Annotations

In types of machine learning methods where labels and annotations are required in order to train the model–such as supervised, semi-supervised and reinforcement learning–the labels

and annotations of data itself can introduce bias as well. This bias arises either when human annotators are distracted, uninterested or lazy about their annotation task, causing them to choose wrong labels, or when the annotators reinforce their own biases into data consciously or not (Hovy & Prabhumoye, 2021; Søby Borgqvist, 2023).

Influence of Biases and Stereotypes to Users of LMs

The influence of biases and stereotypes of language models extends far beyond behavioral aspects and is now affecting opinions among its users. A study about co-writing with opinionated language models that affect users' views introduces a paradigm called *latent influence* that happens when language models persuade users with their generated views more commonly than others. Findings from the study show that interactions with such large language models, like GPT-3, can affect what users' write and think, even if unintended, and it raises concerns about new targeted opinion influence methods (Jakesch et al., 2023).

In another sense, large-scale language models can pick up a lot of language that causes serious offenses, or encourage social norms that discriminate against an individual's perspectives, even personalities, and raises stereotypes that continues the unfair treatment of marginalized groups. This is particularly true when the models come from historical instances of systemic injustice or environments where inequality is the norm (Weidinger et al., 2022). As a result of these biases, members from these marginalized groups may have fewer opportunities and representation, which may influence societal perceptions and impede the development of more inclusive and equitable ways of living.

LM Gopher, a massive language model developed by DeepMind, showed through a counterfactual evaluation that such NLP links negative sentiment to various social groups and

exhibits gender stereotypes with regards to occupation (Huang et al., 2020). For instance, nuances such as referring to *women doctors* as though being a doctor implies not being a woman are examples of how exclusionary norms in language models can produce (Weidinger et al., 2022) or the word *nurse* which might be defined as "a woman who is compassionate and caring" as if there are no existent male nurses and that caregiving roles are primarily for women. As a consequence, language models may use language that silences, rejects, or excludes identities that do not fit into these categories and those individuals who are affected by a language model like this may also suffer from representational or allocational harm, placing unreasonable burdens, mental and emotional well-being or "psychological tax" (Weidinger et al., 2022) to those who face discrimination due to noncompliance with societal norms or when they actively seek to change those norms.

Language evolves with time, reflecting shifts in norms and categories. However, when training language models on their data at a specific point in time, it runs the risk of excluding certain groups and producing a "frozen moment" in which temporary norms are embedded in a model that is unable to adapt to social changes, resulting to models that represent language from a particular group and point in time being "locked in" with the norms, values, classifications from that moment on (Weidinger et al., 2022).

Reducing Biases and Stereotypes in LMs

Making interventions in the data in the corpora that the BERT model uses is one of the solutions in mitigating gender biases which was done in a paper, written by Bartl et al. (2020). In the aforementioned paper, it was pointed out that LMs are mostly trained using corpora with the English language which could contain the biases of those who use it. In this case, it is the bias of

the Americans in gender and professions. Bartl et al. measured the biases and showed how it reflects in the real world. They used Lu et al.'s (2018) Counter Data Augmentation method in order to reduce the gender bias of the BERT. How it works is that, in the dataset, for every instance of data that has a gendered word, a copy will be made to replace that word with its opposite gendered counterpart. For example: actor: actress; king: queen. The only exceptions are gendered words that refer to proper nouns like Queen Elizabeth. However, this method in Bartl et al. 's (2020) did not work that well for a German corpus mainly because of how the language has different rules for gendered things. The method of balancing out the gendered words in a corpus should be done with great fluency in the language used by the model in order for it to be effective.

Another way of mitigating biases is making certain pieces of information impossible to correlate as done by Ravfogel et al. (2020). To put it simply, in the number representations of words, they made it so that some attributes are not connected to other attributes, to make them independent of each other. Those attributes are made functionally independent.

Conclusion

Language is a part of the human experience. It carries our sentiments, opinions, and beliefs. It also includes our biases and stereotypes as these are used to express ideas a society has. Hence, as language models' main purpose is to mirror language systems, these models, consequently, imitate these biases and stereotypes. It is impossible to make a language model without any bias because language itself is biased. However, it should not stop us from trying. Different languages have different biases and we should strive to find a global average language

model. By reducing the bias innate to the language that we are using, we get close to that aforementioned goal.

Sources

Bartl, M., Nissim, M., & Gatt, A. (2020, December). Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), Proceedings of the Second Workshop on Gender Bias in Natural Language Processing (pp. 1–8). Retrieved from https://aclanthology.org/2020.gebnlp-1.1

Hao, K. (2020, April 2). This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review*. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/?truid=&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=08-07-2023

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8). https://doi.org/10.1111/lnc3.12432

Huang, Zhang, Jiang, Stanforth, Welbl, Rae, Maini, Yogatama, & Kohli. (2020, October 8). *Reducing Sentiment Bias in Language Models via Counterfactual Evaluation*. arXiv.org. Retrieved December 2, 2023, from https://arxiv.org/abs/1911.03064

Jakesch, Bhat, Buschek, Zalmanson, & Naaman. (2023, February 1). *Co-Writing with Opinionated Language Models Affects Users' Views*. Arxiv. Retrieved December 1, 2023, from https://arxiv.org/abs/2302.00560

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2019). Gender Bias in Neural Natural Language Processing. arXiv [Cs.CL] (pp. 5–6). Retrieved from http://arxiv.org/abs/1807.11714

Navigli, R., Conia, S., & Roß, B. (2023). Biases in large language models: origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. https://doi.org/10.1145/3597307

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (pp. 7239-7240) https://doi.org/10.18653/v1/2020.acl-main.647

Søby Borgqvist, C. (2023). *The Bias and Fairness of Language Models* [Master's Thesis]. Copenhagen Business School. https://research-api.cbs.dk/ws/portalfiles/portal/98733555/1729063_Reviewing_Bias_and_Fairness_in_the_context_of_language_models.pdf

Weidinger, Uesato, Rauh, Griffin, Huang, Mellor, Glaese, Cheng, Balle, Kasirzadeh, Biles, Brown, Kenton, Hawkins, Stepleton, Birhane, Hendricks, Rimell, Isaac, . . . Gabriel. (2022, June). *Taxonomy of Risks posed by Language Models*. DL. Retrieved December 1, 2023, from https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533088