Evaluating Gender Bias in BERTić, a Large Language Model Trained on South Slavic Data

The body of research on large language models (LLMs) is growing continually in the field of computational linguistics. These computational models are typically trained using artificial neural networks and very large corpora. Google's BERT is trained on 3.3 billion tokens, for example. The applications of LLMs like BERT cover natural language processing (NLP) tasks such as translation, question answering, and text prediction which are services employed by millions of users every day. Many researchers focus on the harmful consequences of LLMs, including the preservation of biases present in society (Bender et al., 2021). Because of the immense quantity of text available online, the internet is the source of training data for most LLMs which require large amounts of data for optimized performance. Platforms featuring user-generated content like Wikipedia and Reddit are often then part of training data. As a consequence, biases held by users are encoded in the models and perpetuated by them during their output. Recent papers on artificial intelligence seek to characterize and quantify social bias involving gender, class, and sexual orientation in LLMs and the impact of bias on their users (Liang et al., 2021). Current work on the topic has mainly examined bias in English; however, not much has been published regarding languages with smaller speech communities.

This paper will attempt to evaluate gender bias in Croatian for BERTić, a BERT-like LLM trained on 8 billion tokens of Bosnian, Croatian, Serbian, and Montenegrin text. Compared to English, South Slavic languages are morphologically complex and low-resource in terms of training data which poses complex challenges in language modeling (Ljubešić and Lauc, 2021). The template-based method for quantifying bias in BERT-like models is widely accepted (Kurita et al., 2021). In this approach, a template sentence is formulated featuring a gendered target word for bias and a career-related attribute word for which bias will be measured (1).

(1) *She is a programmer,* where *she* is the target and *programmer* is the attribute.

In a highly inflected language like Croatian, *programmer* is marked for gender ("programerica"), nullifying any attempt at measuring gender bias. The template sentences would need to be reworked (2).

(2) *Ona se bavi programiranjem.*

she refl. deal-PRES programming-INS

"She works in programming."

The target is masked and the probability assigned to the sentence (P_{tgt}) is calculated. Next, the attribute is masked and the probability assigned to a sentence containing the same attribute but with the target *he* is calculated. The likelihood of the original sentence is reweighted using the prior bias of the model toward predicting *he* (P_{prior}). Then the difference between the normalized predictions for the target words is used to quantify bias for that attribute, calculated as $\log P_{tgt}/P_{prior}$. With this approach, gender bias in the model may be measured. This paper will address how to reformulate input used to evaluate gender bias in a model trained on South Slavic languages, as well as discuss the means and value of mitigating bias in this and other LLMs.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *Association for Computational Linguistics, Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. https://doi.org/10.18653/v1/w19-3823

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards Understanding and Mitigating Social Biases in Language Models. *International Conference on Machine Learning*, 6565–6576. https://doi.org/10.48550/arXiv.2106.13219

Ljubešić, N., & Lauc, D. (2021). BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *Association for Computational Linguistics, Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 37–42.