Multilingual Gender Bias

Keywords: Gender Bias, Multilingual, Data Collection, Social NLP
Skills: Python

It is a well-known phenomenon that language models learn and replicate social biases that are present in the training data. This has sparked a new wave of research in NLP, focusing on measuring and debiasing pretrained language models. However, the vast majority of these proposed methods were developed specifically for English, thus much work still remains to be done in creating techniques that can be applied to other languages.

In this project we will transfer a popular technique[1] for measuring gender bias from English, to a target language. The target language depends on the preferences of the student. The project will involve building sentence templates in the target language and examining occupational gender stereotypes. Comparisons in stereotypes will be drawn between English and the target language. Additionally, the bias scores will be compared to employment data in a country where the target language is widely spoken.

The first goal of the project would be to translate and modify sentence structures such that they are suitable to examine gender bias in the target language.
Subsequently, the second goal would be to measure the gender bias encapsulated in language models.
Finally, for the data analysis, employment data would need to be scraped from the internet. The correlations between the bias scores and the scraped employment data would also be investigated.

[1] Marion Bartl, Malvina Nissim, and Albert Gatt, 'Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias', *ArXiv:2010.14534 [Cs]*, 27 October 2020, http://arxiv.org/abs/2010.14534.