Gender Bias Detection in Pretraining Data and Language Models

- **Supervisor**: Abdullatif Köksal
- **Examiner**: Prof. Schütze
- **BSc, MSc, Open**: BSc/MSc
- **Title**: Gender Bias Detection in Pretrained Language Models
- **Summary**: In this project, we aim to investigate gender bias in pretrained language models and their pretraining data across **multiple languages**. Given that different languages have varying gendered features, from gender-neutral languages like Turkish to languages with grammatical gender and gender agreement like French, we will propose a set of language-agnostic tests to evaluate both unannotated data and pretrained (monolingual) language models. By doing so, we will be able to compare the gender bias in PLMs according to language features, model size, and the bias present in the pretraining data.
- **Prerequisites**: Experience/interest in PLMs and gendered languages.

Recommended Readings:

1. Gender Bias in Coreference Resolution, https://aclanthology.org/N18-2002/
2. Investigating gender bias in language models using causal mediation analysis, https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html
3. Fewer errors, but more stereotypes? the effect of model size on gender bias, https://aclanthology.org/2022.gebnlp-1.13/

If you are interested in this topic, I can also share our paper, currently under review, on the detection of nationality bias in pretraining data and pretrained language models for a wide range of languages, which can serve as a good starting point.