Master Thesis Topic: Probing Language Models: A Study Case on Biases using the Census Dataset

Supervisor: Ana Ozaki
University of Oslo, Norway

1 Introduction and Motivation

Language models are being widely applied, ranging from applications in virtual personal assistants, web search, various text generation tasks such as text summarization, language translation, production of speeches, and others. The improvement of user interfaces for interacting with language models has contributed to the rapid popularization of language models among the society as a whole.

One of the main issues when applying these models to real world applications is that they are often used as black boxes and trained on datasets that reflect harmful societal biases, which in turn can be present or even amplified in the model, after the training process [5]. Such biases are difficult to detect and measure in a systematic way and can be towards individuals with a certain age, nationality, ethnic origin, gender, sexual orientation, disability, mental health condition, combination of multiple of these factors, among others. Harmful biases in machine learning models have already caused damages in the society [4], when applied for automating the legal system, negatively affecting black people, and when applied for automating the recommendation of job positions, preferring men over women for suggesting jobs with leadership roles.

2 Research questions

Given this scenario where language models are being applied for various tasks while still being trained on data with societal biases, how one can investigate such harmful biases and detect them? There has been recent work on probing language models [3] in a systematic way using Angluin's exact learning framework [1]. It provides the first steps towards a flexible approach for probing language models to detect group biases but it can be enriched in many ways.

First of all, regarding the selection of groups, which should include the possibility of focusing on range of factors that are known to affect certain groups of the population negatively. Considering Census data from countries such as France, Norway, United Kingdom, and USA, where a large range of factors can be analysed in a combined way, which groups are most affected by harmful biases in language models? Secondly,

can we create a platform that simplifies the process of selecting language models, the factors to be analysed, and the extent to which the probing process is applied?

3 Thesis Work

In this master thesis topic, the candidate of a master degree will study an algorithm from the literature for systematically probing language models. This is Angluin's exact learning algorithm for extracting Horn rules [2]. The main tasks of this project are

- to study reference papers from the literature on the topic of the project (suggested by the supervisor and found by the candidate),
- to study how to apply Angluin's algorithm on language models for detecting harmful biases, following the initial steps by Blum et al. [3], and
- to implement a basic library (suggested programming language: python) for probing language models in a systematic way, using an adaptation of Angluin's exact learning algorithm for Horn rules and sentence templates created using various factors that can negatively affect certain groups of the population.

It is also expected that the candidate performs experiments showcasing the applicability of the library for detecting biases against various groups known to be affected by harmful societal biases. In the experiments, it is suggested that the candidate uses Census data and language models such as GPT and BERT-based models.

References

[1] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[2] Dana Angluin, Michael Frazier, and Leonard Pitt. Learning conjunctions of horn clauses. *Machine Learning*, 9:147–164, 1992.

[3] Sophie Blum, Raoul Koudijs, Ana Ozaki, and Samia Touileb. Learning horn envelopes via queries from large language models. *CoRR*, abs/2305.12143, 2023.

[4] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Inc., 2019.

[5] Samia Touileb, Lilja Øvrelid, and Erik Velldal. Measuring normative and descriptive biases in language models using census data. In Andreas Vlachos and Isabelle Augenstein, editors, *EACL*, pages 2234–2240. Association for Computational Linguistics, 2023.