Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Adnan Al Ali
Thesis title	Gender Stereotypes in Neural Sentence Representations
Year submitted	2023
Study program	Computer Science
Study branch	Computer Science with the Artificial Intelligence Specialization
Review author	Mgr. Ondřej Dušek, Ph.D. Advisor
Department	Ústav formální a aplikované lingvistiky

Overall

	better	OK	worse	insufficient
Assignment difficulty	X			
Assignment fulfilled	X			
Total size *… text and code, overall workload*	X			

The thesis of Adnan Al Ali focuses on a very timely topic – biases in pretrained language models (PLMs). The author specifically examines how the PLMs deal with gender stereotypes, and whether any bias with respect to gender exists due to biases in the underlying training data. In addition to generic biases, they also aim at gender differences in terms of political values. As a welcome move, the thesis focuses on PLMs available for Czech, which is both underexplored in this area and a highly gendered language.

The thesis includes two experimental parts, aimed at discovering biases in general and exploring the political values. The first part features an interactive web-based tool to assist users/researchers in finding gender-stereotype sentences – example sentences are easily semi-automatically converted from a male to female variant and subsequently checked for a probability score from a PLM. The biased sentences can be stored for further use. The second part analyzes multiple PLMs by using constructed sentences ("he/she agreed/disagreed that...") based on a political values questionnaire and checking how their probability scores differ between genders. As PLMs tend to be biased towards positive sentences, a special correction is applied, based on probability values on (supposedly unbiased) GPT3-generated calibration data.

The evaluation shows that while multiple annotators were able to find gender-stereotypical sentences using the tool, finding them was challenging due to Czech generic masculine forms, which make masculine forms more probable overall. The results from the second experiment suggest that PLMs do not have any significant connection between gender and political values and generally tend to regress to the mean it terms of ratings.

Overall, the thesis presents a very timely topic, evaluated in a novel way with a reasonable methodology. Even though the results are mixed, I believe their scientific value exceeds the general requirements for a bachelor thesis. I have no reservations to the experimental contents of the thesis in general, my only reservations are to the writing, and they are rather minor (see below). I believe the thesis deserves to be defended with the top grade.

Thesis Text

	better	OK	worse	insufficient
Form *… language, typography, references*	X	X		
Structure *… context, goals, analysis, design, evaluation, level of detail*	X			
Problem analysis	X			
Developer documentation		X		
User documentation		X		

The thesis fulfills the usual basic formal requirements. It is written in very good academic English and contains all the necessary sections, showing not only the author's experiments but also theoretical background and related works. I have only two minor reservations: (1) The main thesis text itself is a little too terse – I believe that part of what is now included in the appendices (figures, raw evaluation of the first experiment) should be part of the main text. However, the reader can find all the necessary information. (2) The theoretical chapters includes a few statements where I would not necessarily agree, not all of the introduced concepts are necessarily used in the experiments (CNNs and RNNs are not as far as I can see), so it could use a little pruning. On the other hand, embeddings and subword tokens could have been explained better. However, these are all minor details and they do not hamper the understanding of the experiments for a reader familiar with common NLP concepts.

The included documentation is perfectly sufficient for making the author's code run and re-running the experiments.

Thesis Code

	better	OK	worse	insufficient
Design *… architecture, algorithms, data structures, used technologies*	X			
Implementation *… naming conventions, formatting, comments, testing*	X			
Stability	X			

In terms of the experiments, I am happy with almost all the steps and choices the author made. I only have two comments: Regarding the probability computation from masked PLMs in the first experiment, I think this is essentially equivalent to the approach of Salazar et al. (`https://aclanthology.org/2020.acl-main.240`), but could have involved a little more discussion (also note that there is a recent paper by Kauf & Ivanova proposing improvements: `https://aclanthology.org/2023.acl-short.80`). The GPT-3 generated "unbiased" sentences used for calibration in the second experiment are not really evaluated in the thesis, so I have to take the author's word that there are no apparent biases in them.

As to the results of the second experiment, I agree with the conclusions, but I wonder if the models are simply too crude to take a subtle cue such as a masculine/feminine grammatical form, which is the only distinction between the template sentences input into the PLMs. I also wonder if decoder-type and/or larger PLMs would be more sensitive, but this is purely hypothetical as these are generally not available for Czech, and multilingual models do not seem to perform particularly well.

Otherwise, I am very happy with the technical implementation quality, everything works straight out of the box, with no apparent issues.

Overall grade Excellent

Award level thesis No

Date: 28 August 2023 Signature