CS475 Fall 2022 Homework 4

Team 0: Juhee Son, Rifki Afina Putri, and Alice Oh

October 24, 2022

Language models like BERT [2] has become a standard architecture for NLP research ever since it was published. However, they still exhibit some social bias, such as gender, racial, or ethnic bias [1].

In this homework, your team will **review a paper** about ethnic bias in BERT models and **conduct an exploration** on bias in language models using `fill-mask` pipeline of `huggingface` library.

1 Tasks

The tasks in this homework are as follows:

1. Read "Mitigating Language-Dependent Ethnic Bias in BERT" paper [1] and put your review in the report. The review must contain several points:
   (a) What is this paper about and what contributions does it make?
   (b) What do you think are the main strengths and weaknesses of this paper?

   Please put your own words to make sure that you properly understand the contents of the paper. Your score will automatically be zero if we found that some plagiarism is involved. If you need to use any expression in the paper, please make sure that you use proper quotation for such expression.

2. Try `fill-mask` task on BERT or any other language models provided in `huggingface` library, and analyze the result. We provide a starter code (`fill_mask.py`) for your exploration. Specifically, you have to:
   (a) Try different `targets`, `attribute`, or `template` other than presented in the starter code and the main contents of the paper. You can conduct an exploration of any languages, such as Korean, Arabic, Chinese, Indonesian, etc. We encourage you to use the readily available language models. Some models for languages other than English like Korean BERT or Indonesian BERT are available at `https://huggingface.co/models`.
   (b) Explain the modification that you use in your `fill-mask` exploration, show the model's output in the report[1], and analyze: is the model present some bias in that particular setting and language? If yes, what is the reason? If not, what do you think makes the model do well on that setting?
   (c) Paper [1] provides some bias mitigation methods. Do you think the methods will work well on the languages in your exploration? Provide your reason and analysis. Also, what other improvements can potentially be applied to minimize the language-dependent ethnic bias of the model, specifically on your chosen languages?

2 Submission

The files you should submit are

1. Your team's report `report_{team_no}.pdf` (e.g., `report_0.pdf`). **Use this LaTeX file as a template**, and **do not change style attributes** in this file. The report must consists of **2-4 pages**. References are not included in the page-limit. You will get a **penalty (-10 points)** if you fail to follow this rule.

[1] Please do not directly copy-and-paste the raw output. Present the result properly using some tables or graphs.

2. Your team's `fill_mask_{team_no}.py` (e.g, `fill_mask_0.py`). Note that the code submission is not for checking the functionality of the code, but it is for checking whether we can reproduce the exploration results presented in your report. If you use other external libraries, please also provide your `requirements.txt` file.

3 Grading

Comprehensive evaluation based on clarity, validity, and interestingness of your report. You will get zero points if you violate academic integrity (e.g., plagiarism and data manipulation). Late submission will be penalized (-5 points/day).

References

[1] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.