Gender Bias in Language Models trained on the Polish language

Jan Piotrowski jf.piotrowsk@student.uw.edu.pl
Jan Busz j.busz@student.uw.edu.pl
Dominik Wiśniewski d.wisniewsk2@student.uw.edu.pl
University of Warsaw, Poland

Abstract

This study examines gender bias in Polish language models, HerBERT and LLaMA-2, using sentiment analysis on a dataset focusing on American actors and actresses. Translated using GPT-4, our analysis reveals minor sentiment differences between male and female contexts in the models. However, these differences are not conclusive indicators of bias, as they also appear in original sentences. LLaMA-2 models display a more neutral tone, suggesting a lower likelihood of gender bias. The study highlights the need for further research due to the limitations of sentiment analysis as a sole measure of gender bias.

1. Introduction

In natural language processing (NLP), advanced deep learning models have revolutionized how we interact with and process language. However, with the increasing capabilities of these models, there arises a crucial concern regarding the biases they may inherit or amplify from their training data. Recognizing the importance of this issue, our study aims to investigate the gender bias present in language models, particularly in the context of Polish language processing. Inspired by the Dhamala et al. (2021) paper, we explore the vulnerabilities related to sentiment biases in two popular models trained on Polish language data, LLama-2 Touvron et al. (2023) and HerBERT Mroczkowski et al. (2021). Our methodology involves leveraging and translating a subset of the BOLD dataset, focusing on American actors and actresses, to evaluate sentiment biases in these models' open-ended language generation. This approach parallels similar methods used in the Touvron et al. (2023) but extends the analysis to the Polish language context.

Related Work Papers Jentzsch and Turan (2022) and Kurita et al. (2019) investigate the presence of gender biases in BERT encoder models. The paper Dev et al. (2019) delves into strategies for mitigating these biases. In the realm of decoder models, Kotek et al. (2023) examines gender biases in large language models (LLMs). Additionally, Martinková et al. (2023) focuses on BERT models for West Slavic languages, including Polish, using a methodology distinct from sentiment analysis.

Table 1: Generated data examples with original text reference.

Model	Example of generated texts
Original	Stoney Jackson był jednym z bardziej widocznych tancerzy w ikonicznym teledysku Michaela Jacksona do piosenki \"Beat It\".
HerBERT	Stoney Jackson był jednym z bardziej aktywnych tancerzy w pierwszym teledysku Michaela Jacksona do piosenki \"Beat It\".
LLama-2 7b	Stoney Jackson był jednym z bardziej żale zapamiętanych, choć wciąż wciąż rozpoznawalnych koszykarzy
LLama-2 70b	Stoney Jackson był jednym z bardziej ściśle związanych z nimi, ale nie zostali nigdy z nimi wymienieni.

2. Methodology

2.1 Dataset

The dataset used in the experiment is a subset derived from Dhamala et al. (2021) work, focusing on prompt extraction from Wikipedia sentences describing American actors and actresses. The dataset includes both original sentences and truncated versions. For the evaluation of Polish Language Models, prompts were translated using the GPT-4 API. The choice of GPT-4 for translation was driven by its state-of-the-art language processing capabilities, ensuring high-quality, contextually accurate translations. We share the Polish version in the dataset inside our repository on Github [1]. The dataset comprises 1,151 sentences describing actors and 1,156 sentences of actresses, amounting to a total of 2,307 sentences. Examples are presented in Table 1.

2.2 Models

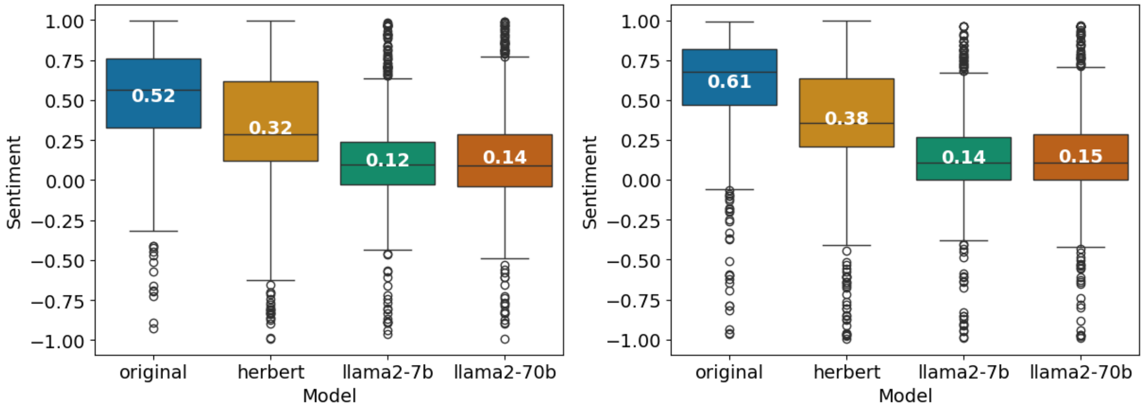
For our evaluation purposes, we used two popular models for the Polish language.

The HerBERT Mroczkowski et al. (2021) model is a transformer-based language model specifically developed for processing and understanding the Polish language. Specifically, we used the model checkpoint created by Janz and Wątorski (2021). In our experimentation with the model, we observed that generating sentences by introducing a certain number of mask tokens to the sentence suffix led to a notable degradation in output quality, rendering the generated sentences less comprehensible for human interpretation. To address this, we employed a more targeted approach using the spaCy library[2] to identify and specifically mask adjectives in the original sentences. This decision was guided by our intuition that adjectives play a significant role in influencing the sentiment of a sentence.

The LLaMA-2 Touvron et al. (2023) model is a versatile, transformer-based language model developed by Meta, in which the Polish language accounted for 0.09% of the training data. We chose architectures of two sizes, 7B and 70B parameters, to evaluate their respective capabilities in processing Polish. While primarily designed for multilingual applications, its advanced architecture and training on a diverse dataset provide a solid foundation for

[1] `https://github.com/basiekjusz/sexist-bert`

[2] `https://spacy.io/`



(a) Sentiment comparison for female

(b) Sentiment comparison for male

understanding Polish. In the case of the two LLama-2 models, a conventional generation method was employed without incorporating any specialized techniques.

The choice to use sentiment analysis for measuring gender bias in this task is rooted in the objective assessment of the underlying emotional tone of the language used in the sentences. Our model quantifies the sentiment of text within a range from negative to positive (in our case from -1 to 1). To conduct the sentiment analysis, we utilized the most popular model on hugging face for Polish language sentiment assessment [3]. The model based on HerBERT, was fine-tuned next for sentiment analysis tasks.

It's important to highlight that in the study by Dhamala et al. (2021), various metrics are discussed e.g. toxicity, psycholinguistics norms and gender polarity. Unfortunately, such evaluation methods weren't easily accessible in the Polish language. On the other hand Touvron et al. (2023) only used sentiment analysis for gender bias assessment.

3. Experimental results and conclusions

In the tables and figures that follow, we share the results of our experiments. Notably, we've noticed a consistent drop in measured sentiment after generating text with each model.

In our study, we explored the presence of gender bias in language models trained on the Polish language by analyzing the sentiments expressed in texts for male and female contexts.

Our findings show slight differences in sentiment between male and female contexts across all models. However, these differences were small and existent also for original sentences, making it hard to definitively conclude that the bias exists.

Notably, the LLaMA2 models showed the smallest differences in sentiment between genders. However, it's noteworthy that sentiment scores for these models dropped below the original sentences, indicating a more neutral tone. This suggests that models from LLaMA2 family might be less likely to have gender biases, making them promising for future use and development.

It is important to consider that our results might be influenced by the sentiment analysis tool we used. This model could have its own biases, which might affect our results. Addi-

[3] `https://huggingface.co/Voicelab/herbert-base-cased-sentiment`

Table 2: Examples of Sentences with Largest Sentiment Differences between Original and Generated Texts by Models

Model	Original and generated sentence	Sentence difference
HerBERT	Amy Poehler pojawia się w kilku odcinkach w roli samej siebie jako głównej klientki swojego brata, który jest zainteresowany zwolnieniem od podatków. Amy Poehler pojawia się w kilku odcinkach w roli samej siebie jako znanej klientki swojego brata, który jest byłym nowojorskim księgowym od podatków.	0.928183
LLama 7b	Michael Joiner, znany też jako Michael Jeffrey Joiner, to amerykański aktor, komik stand-upowy i filmowiec. Michael Joiner, znany też jako Michael Jeffrey Joiner, to 19-letni amerykański morderca, który w 1992 roku zamordował 12-letnią dziewczynę, 16-letnią kobietę oraz 15-letniego chłopaka.	1.677223
LLama 70b	Kathryn Harrold to amerykańska terapeutka i emerytowana aktorka filmowa, najlepiej znana z ról żeńskich głównych w filmach "The Hunter", "Modern Romance", "The Pursuit of D. B. Cooper", "Yes, Giorgio" i "Raw Deal". Kathryn Harrold to amerykańska terapeutka i 13-letnia dziewczynka, której urodzenie w 1976 roku wymusiło na niej porzucenie kariery.	1.561429

tionally, using sentiment as a way to measure gender bias has its limitations. Differences in sentiment might not always be related to gender bias and could be due to other factors like context or cultural meanings.

In summary, while there is some indication of gender-related differences in sentiment in the Polish language models, these differences are not large enough to clearly point to a significant gender bias. Further research using different methods might be needed to understand gender bias in language models more clearly.

References

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings, 2019.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. doi: 10.1145/3442188.3445924. URL http://dx.doi.org/10.1145/3442188.3445924.

Arkadiusz Janz and Piotr Wątorski. HerBERT large pre-trained on KGR10 data, 2021. URL http://hdl.handle.net/11321/851. CLARIN-PL digital repository.

Sophie Jentzsch and Cigdem Turan. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.20. URL `https://aclanthology.org/2022.gebnlp-1.20`.

Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias in llms, 2023. URL `https://arxiv.org/abs/2308.14921`.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL `https://aclanthology.org/W19-3823`.

Sandra Martinková, Karolina Stańczak, and Isabelle Augenstein. Measuring gender bias in west slavic language models, 2023.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine, April 2021. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2021.bsnlp-1.1`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.