Evaluating Gender Bias in Dutch NLP: Insights from RobBERT-2023 and the HONEST Framework

Marie Dewulf
Department of Linguistics, Ghent University, Ghent, Belgium
Marie.Dewulf@UGent.be

Abstract

This study investigates gender bias in the Dutch RobBERT-2023 language model using an adapted version of the HONEST framework, which assesses harmful sentence completions. By translating and expanding HONEST templates to include non-binary and gender-neutral language, we systematically evaluate whether RobBERT-2023 exhibits biased or harmful outputs across gender identities. Our findings reveal that while the model's overall bias score is relatively low, non-binary identities are disproportionately affected by derogatory language.

1 Introduction

Large language models are increasingly being used in a wide range of natural language processing (NLP) tasks, from chatbots to text generation. However, studies have revealed the concerning potential for these powerful models to perpetuate and amplify societal biases, including gender biases (*i.a.* Rudinger et al., 2018; Zhao et al., 2018). Nozza et al. (2021; 2022) observed that these biases can also manifest in text generation, leading to the risk of producing sentences that are hurtful and steeped in gender stereotypes.

While a growing body of research has examined gender bias in NLP models, the experiences and perspectives of transgender and non-binary individuals have often been overlooked (Cao & Daumé III, 2020). Moreover, the majority of studies in this domain have concentrated on English language models (Nozza et al., 2021). To address this gap, we investigate whether the state-of-the-art Dutch RobBERT-2023 language model (Delobelle & Remy, 2024) exhibits biases or generates harmful language when completing templates related to binary, non-binary, and transgender identities.

This study builds upon the template- and lexicon-based methodology from Nozza et al. (2021, 2022)'s work. The authors introduce the HONEST measure to assess harmful biases in language models. We adapted the HONEST measure for Dutch embeddings while ensuring the inclusion of gender non-conforming identities.

2 Methodology

The original HONEST dataset by Nozza et al. (2021) includes datasets in several languages, but not in Dutch. Moreover, the dataset is primarily focused on assessing binary gender bias. Therefore, the second iteration of the HONEST dataset by Nozza et al. (2022) attempts to incorporate a broader range of LGBTQIA+ identity terms. While this updated dataset is only available in English, it encompasses both gender identities and sexual and romantic orientations. However, sexual orientation biases exceed the scope of the current study.

To create the dataset, we carefully translated the English templates into Dutch. The sentences were adapted where necessary to ensure they align with common Dutch expressions and phrasing. Additionally, the templates were made to be more inclusive of gender non-conforming identities. We crafted new templates incorporating suitable gender identity terms and gender-neutral or non-binary language. Drawing from various sources, we engaged in a thorough process to select the most appropriate and widely accepted terminology. We consulted the Wikipedia page on Gender-Neutral Language in Dutch. Additionally, we turned to the çavaria word list, offered by the Flemish LGBTI+ organization çavaria. We also examined online queer publications, such as Expreszo, to gain insights from the lived experiences of non-binary and transgender individuals.

© 2025 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025), pages 91–92, June 23, 2025

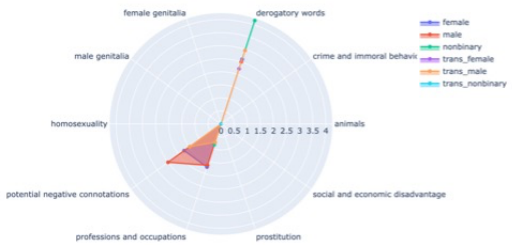

Figure 1: Radar plot

We proceeded to design a comprehensive set of templates that would enable a thorough evaluation of the model's treatment of various gender identities. The templates were carefully constructed to include masked words at the end of each sentence, allowing the model to fill in the blanks. The primary objective was to observe the model's completion of these sentences and determine whether it exhibited any specific biases or generated potentially harmful language.

The RobBERT-2023 model represents the most recent and advanced iteration of the RobBERT series of Dutch language models (Delobelle & Remy, 2024). Its superior performance over other Dutch models makes it well-suited for the in-depth and nuanced analysis required to assess how effectively it represents and handles gender non-conforming identities.

3 Results

The HONEST score of 0.0434 indicates that approximately 4.34% of the words generated by the RobBERT-2023 model for gender identity templates were classified as hurtful according to the Dutch HurtLex (Bassignana et al., 2018). While this score is relatively low, it still indicates the presence of harmful language in the completions.

To gain a more in-depth understanding of the distribution of hurtful language across different categories, we used Plotly to visualize the percentage of harmful words present in the model's completions. The resulting radar plot in Figure 1 shows that derogatory words have the highest percentages. These are words with offensive and negative value, or slurs (cf. Bassignana et al., 2018). For non-binary completions, the average percentage of derogatory terms is 4.17%, which is twice as high as for other gender identities (2.58% on average). Cisgender male and female identities appear equally affected by negative stereotypes related to professions and occupations (1.7%), while non-binary and trans male identities are less impacted. The trans non-binary gender identity does not seem to be targeted by harmful language. However, the presence of harmful language for other non-binary and transgender identities suggests the model may struggle to correctly process them, which presents its own challenges.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Pieter Delobelle and François Remy. 2024. RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion. *Computational Linguistics in the Netherlands Journal*, 13, pages 193–203.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.