# Evaluating gender bias in Dutch embeddings of a multilingual NLP model

Marie Dewulf
*Ghent University, Belgium*
Marie.Dewulf@UGent.be

Previous work on evaluating gender bias in word embeddings and contextualized representations of natural language processing (NLP) models has primarily focused on the English language. Moreover, these analyses have often treated gender as a binary construct, failing to account for the full spectrum of gender identities. Gender is a complex, multifaceted concept that extends beyond the male-female dichotomy, encompassing a diverse range of identities such as transgender, nonbinary, genderfluid, and agender, among others. By overlooking these nuances, the existing literature provides an incomplete understanding of the manifestation and impact of gender biases in NLP models. In this study, we present an evaluation of gender bias in the Dutch embeddings of the multilingual BERT (mBERT) model (Devlin et al., 2019). Moreover, we embrace a more inclusive conceptualization of gender, examining a broader spectrum of gender identities beyond the traditional male-female binary.

The CrowS-Pairs dataset has become a widely utilized benchmark for assessing bias in NLP models. It contains English sentence pairs targeting different bias types, including gender, religion, and ethnicity. Reusens et al. (2023) translated a portion of the CrowS-Pairs dataset to Dutch to evaluate bias in the multilingual BERT model. However, for gender, the dataset is limited to binary gender categories. Our contribution lies in extending this resource to cover a broader spectrum of gender identities, by integrating stereotype-based examples derived from the WinoQueer benchmark (Felkner et al., 2023), originally developed to assess anti-LGBTQ+ biases in large language models. For these additional test sentences, we translated the English sentences into Dutch, while taking into account the Dutch linguistic and cultural context (e.g., gendered forms, names, pronouns).

The resulting dataset combines translated examples from WinoQueer with the Dutch CrowS-Pairs test set. We reused and extended the evaluation pipeline of Reusens et al. (2023), which computes a bias score following Meade et al. (2022). This metric measures the percentage of sentence pairs where the model prefers a more stereotypical over a less stereotypical sentence. An unbiased model would achieve a score of 50% (signifying equal tendencies towards both stereotypical and non-stereotypical statements). Conversely, higher scores indicate stronger bias.

Our experiments reveal that the model has a mean bias score of 64.44 for the CrowS-Pairs benchmark enriched with stereotypes on non-binary and transgender people. This indicates mBERT exhibits high levels of gender bias in its Dutch embeddings. The reported score is an improvement over the 67.66% reported by Reusens et al. (2023). However, this interpretation must be considered in light of the distinctive characteristics of the dataset used. By incorporating representations of non-binary and transgender identities, the dataset introduces additional complexities that may have impacted the model's performance. We further analyze token-level outputs for the additional sentences representing non-binary and transgender identities to offer qualitative insights into the model's capture of the nuances and complexities of gender identity.

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896.