Evaluating Labor Market Biases Reflected in German Word Embeddings

Leander Rankwiler and **Mascha Kurpicz-Briki**
Applied Machine Intelligence, Bern University of Applied Sciences
Biel/Bienne, Switzerland
leander.rankwiler@students.bfh.ch

Abstract

An unsolved issue in the domain of Natural Language Processing (NLP) is the perpetuation of stereotypical biases inherent in the training data. This has led to increased attention in the research community, but the focus has predominantly been on English models, often neglecting models for other languages. This work aims to counter this trend by investigating bias in German word representations. This analysis includes representations that focus on the word itself, known as static word embeddings, and extends to contextualized embeddings that take into account the context provided by surrounding words. The German datasets for this research are partly derived from a workshop with experts from different fields, including human resources and machine learning in Switzerland. The workshop aimed to identify language-specific biases relevant to the labor market. Our analysis shows that both static and contextualized German embeddings exhibit significant biases along several dimensions.

1 Introduction

Natural Language Processing (NLP) is widely applied in various domains, with its most recent and prominent influence being in language generation. Word embeddings are key components of NLP applications. These vector representations capture semantic meaning in a numerical representation. Studies have demonstrated biases in these embeddings related to gender, race, ethnicity, and other dimensions (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019). For example, science-related words were found to be more correlated with male-related words such as *him, brother*, and *man* compared to female-related words. These stereotypes perpetuate existing social and racial hierarchies (Gao et al., 2020; Bender et al., 2021; Lauscher et al., 2022), leading to unfair treatment and discrimination of certain groups (Köchling and Wehner, 2020). To measure and mitigate these unwanted stereotypes, there has been a surge in NLP bias research. However, as Ramesh et al. (2023) point out, languages other than English do not get the attention they deserve. They should be studied separately (Kurpicz-Briki and Leoni, 2021), especially German with its rich morphology and gender marking (Bartl et al., 2020a). In addition, Zhao et al. (2020) note that since biases in multilingual models are transferred to other languages, it is crucial to understand relevant stereotypes in the respective languages themselves. This paper aims to fill this research gap by providing an analysis of biases in established, pre-trained German word embeddings, considering both static and contextualized variants. While static word embeddings focus on the word in question itself, contextualized word embeddings also take into account the context in which the word is used. Bias is quantified with a metric, that uses topic-specific (e.g., male/female and productivity) wordlists. In the presented research we focus on real-world biases from the labor market. We rely on two sources for this analysis. Firstly, we utilize data from the outcomes of a dedicated co-creation workshop conducted with German speakers from different domains including human resources, machine learning, non-governmental organizations, and the legal field. Secondly, we refer to existing datasets for bias detection in word embeddings (Caliskan et al., 2017; Kurpicz-Briki, 2020). Our research is guided by the following research questions:

(RQ1) Are the identified societal biases at the co-creation workshop reflected in German static word embeddings?

(RQ2) Is there evidence of bias in the German contextualized embeddings:

(a) for the same wordlists used for RQ1?

(b) in the German translations of the wordlists from Caliskan et al. (2017)?

(c) in the German wordlists from Kurpicz-Briki (2020)?

2 Related Work

Common tests used for bias detection in English word embeddings cannot always be reproduced for other languages. Lauscher and Glavaš (2019) examined gender bias in German static embeddings with regard to math/art and gender bias with regard to science/art. They found no significant bias in these dimensions; Kurpicz-Briki (2020) confirms this finding. However, Kurpicz-Briki (2020) presents two German wordlists, both of which show significant biases in static embeddings. GER1 examines gender bias, comparing female versus male study choices. GER2 examines historical gender role perceptions, focusing on stereotypes of rationality versus emotionality. Two other studies of German word embeddings highlight the inherent challenges of detecting biases. Bartl et al. (2020b) created a dataset specifically designed to uncover gender bias in the context of the labor market. However, they encountered limitations with the gender-specific postfix forms of occupations in German (postfix: '-in' for female), which inadvertently distorted the associations. Kraft et al. (2022) developed a German language regard classifier that showed a bias towards positive classifications for female subjects. This finding was initially attributed to positive stereotyping, but on closer inspection the authors found the cause to be sexist stereotyping.

3 Methods

3.1 Static Embeddings

Embedding: Fasttext We use Fasttext (Bojanowski et al., 2017) as our pre-trained static word embedding because it is available for multiple languages, thereby allowing us to test our German wordlists. As it uses sub-words (parts of words, or characters) it is well suited to morphologically rich languages like German (Bojanowski et al., 2017). The model is trained on Common Crawl and Wikipedia datasets (Grave et al., 2018).

Metric: WEAT To assess bias in word embeddings, researchers have developed a range of metrics specifically designed to assess bias in word embeddings. A well-known example is the Word Embedding Association Test (WEAT), developed by Caliskan et al. (2017), which we use in our analysis of static embeddings. Their method is based on the Implicit Association Test (IAT), a well-established psychological method for measuring implicit biases (Greenwald et al., 1998). Its widespread use in research e.g., (Chaloner and Maldonado, 2019; May et al., 2019; Chávez Mulsa and Spanakis, 2020), and its adaptability to languages beyond English are additional reasons to use it in our tests. Caliskan et al. (2017) test this method with ten wordlists derived from the underlying psychological literature (Greenwald et al., 1998), they are referred to as WEAT1-WEAT10. The metric WEAT quantifies bias by comparing the vector representations of the assumed bias topics, which are captured in the wordlists. For a detailed explanation, refer to Appendix C.

3.2 Contextualized Embeddings

Embedding: BERT Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is an open-source architecture for contextualized embeddings, which is available in several languages, and widely used in research. Because of the above, we perform our analysis on BERT. We use the version “bert-base-german-cased”, updated in 2020, which is trained on data from Wikipedia, German court citations, and news articles[1].

Metric: SEAT For contextualized embeddings we use the Sentence Embedding Association Test (SEAT) proposed by May et al. (2019). The underlying method is based on the methodology of WEAT, with the difference that SEAT is able to measure bias in sequences of words, i.e. sentences. We chose SEAT because the underlying sentence templates can be built from WEAT-like wordlists in languages other than English. In addition, SEAT is the most widely used metric for contextualized word embeddings (He et al., 2022), thus allowing comparison with other research. The concept of SEAT is to insert the words of interest into unspecific contexts, which the authors call 'semantically bleached sentences' - sentences that are deliberately empty of much meaning to allow the inserted word to stand out. Examples are:

- This is <word>.
- <word> is here.

To generate a single representation of the sentence they use the <CLS> token of the embedding (in the case of BERT). [CLS] stands for classification and represents a vector containing the semantic meaning of the whole sentence (McCormick, 2020).

[1] https://huggingface.co/google-bert/bert-base-german-cased

4 Data

To quantify the biases, the metrics WEAT and SEAT are dependent on wordlists, which for our purposes originate from three different sources. The wordlists CW1-CW4 are derived from a **c**o-creation **w**orkshop with German speaking domain experts from human resources, machine learning, non-governmental organizations, and the legal field[2]. The co-creation workshop was organized as an activity in the EU research project BIAS to discuss the topic of bias and discrimination with interdisciplinary stakeholders. During the workshop, participants engaged in structured discussions and exercises to identify language-specific biases in the labor market. The resulting data is the foundation for the creation of the wordlists CW1-CW4. They cover biases related to gender, hobbies, family status, immigration, and productivity. The following are some words from each of the four categories of CW2, which captures the bias that productivity is related to age:

CW2: Productivity / Age:
Productive: *Effizienz, Leistungswille, Ambition, . . .*
Unproductive: *Ablenkung, Ineffizienz, . . .*
Old: *Älterer, Ältere, Lebensmitte, Erwachsene, . . .*
Young: *Jugendlicher, Jugendliche, Jugend, . . .*

Inspired by the bias results of CW1-CW4, we additionally suggest the wordlist CW5, which is a combination of CW1 and CW4 and suggests a gender bias related to productivity. CW1-CW5 are listed in the Appendix E. They are evaluated for static as well as for contextualized embeddings. The wordlists GER1 and GER2 and the German translations of WEAT7 and WEAT8 have been created and investigated by Kurpicz-Briki (2020) for static embeddings. We extend this study by investigating these wordlists on contextualized embeddings. The sentences for the contextualized analysis are created by integrating the words into semantically neutral sentences. For instance, instead of the standalone word *Frau* (woman), sentences such as *Dies ist eine Frau* (This is a woman) are formulated. This is done with the help of GPT-4 (Achiam et al., 2023), the prompt for the creation can be found in the Appendix B, and the sentences are publicly available[3].

5 Results

As Schröder et al. (2024) suggest, we report effect size (ES), as well as p-values (p) of the WEAT and SEAT metrics, shown in Table 1.

	FastText		GermanBERT	
	ES	p	ES	p
CW1	1.26	**0.003**	1.04	<**0.001**
CW2	0.91	**0.023**	0.99	<**0.001**
CW3	0.67	0.141	-0.14	0.744
CW4	1.46	**0.003**	1.11	<**0.001**
CW5	1.10	**0.002**	0.55	<**0.001**
GER1	***1.74***	*<**0.001***	0.58	**0.005**
GER2	***1.43***	***0.002***	0.98	<**0.001**
WEAT7	*0.23*	*0.65*	-0.04	0.594
WEAT8	*0.11*	*0.83*	-0.36	0.98

Table 1: ES = effect size; p = p-value. **Bold** results are significant at the < 0.05 level. *Italic* results are from Kurpicz-Briki (2020)

.

6 Discussion

6.1 CW1 - CW5

We demonstrated that, of the four dimensions of bias identified during the co-creation workshop (CW1-CW4), three exhibit significant bias in both static and contextual analyses, thereby affirmatively addressing RQ1 (CW: bias in static embeddings?) and RQ2a (CW: bias in contextualized embeddings?). This supports existing research that argues the efficiency of language-specific bias identification. CW1 and CW2 highlight the stereotype that productivity conflicts with family and old age, respectively. This finding is echoed outside the domain of NLP by researchers that investigate real-world bias in the labor market: Pärli (2018) found that older people are disadvantaged in the Swiss professional environment, and Kleinert (2006) found that women with children are disadvantaged in obtaining managerial positions. The CW3 wordlist does not yield significant results, possibly because 'traditionalists' are not the direct counterparts of 'communicators', which could reduce the effect size of the wordlist. Hobbies like *Backen, Kunst, Ballett* are related to female terms, as shown by CW4. When combined with the findings of CW5, these results could perpetuate problematic stereotypes. CW5 finds a link between productivity and gender. Taken together, these findings could suggest a correlation whereby typical male hobbies are associated with productivity and typical female hobbies with unproductivity. However, a direct experiment did not confirm this speculation.

[2] https://www.biasproject.eu/
[3] https://github.com/BFH-AMI/BIAS

	Dutch	English	German (ours)
Static (WEAT)	**7, 8: FastText**	**7, 8: Glove, word2vec** **7, 8: FastText**	7, 8: FastText
Contextualized (SEAT)	**7: BERTje, RobBERT** 8: BERTje, RobBERT	**7: BERT** 8: BERT	7, 8: German BERT

Table 2: Comparison of presence of bias in different languages and embeddings. **Bold** = significant at < 0.05. 7,8: number of WEAT wordlist. Dutch results are equal for BERTJe (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020). GloVe: static word embedding from Pennington et al. (2014), Word2Vec: static word embedding from Mikolov et al. (2013). Dutch results by Chávez Mulsa and Spanakis (2020). English static results for Glove and Word2Vec by Caliskan et al. (2017) and for FastText by Lauscher and Glavaš (2019); Kurpicz-Briki (2020). English contextualized results by May et al. (2019). The BERT models are available online with the identifier (bert-base-dutch-cased, robbert-2023-dutch-large, bert-base-cased, bert-base-german-cased) on https://huggingface.co/.

As the results of Kraft et al. (2022) demonstrate, we should be cautious to draw conclusions too quickly in the domain of bias detection.

6.2 WEAT7, WEAT8

The German WEAT7 and WEAT8 results from Kurpicz-Briki (2020) are consistent with our SEAT results from BERT, thus not confirming RQ2b (WEAT7, WEAT8: bias in contextualized embeddings?). Comparing these results with WEAT results in other languages, we find no obvious correlation across languages, refer to Table 2. This could be an indication that language specificity is relevant.

6.3 GER1, GER2

The results for the static embeddings from Kurpicz-Briki (2020) are confirmed in our contextualized setting with BERT, thus answering RQ2c (GER1, GER2: bias in contextualized embeddings?) positively.

6.4 Correlation Static to Contextualized

The correlation between static and contextualized results (from Table 1) is high, i.e. wordlists with low p-values in FastText analyses tend to show low p-values in BERT tests (the same counts for high p-values). To quantify this correlation, we perform a meta-analysis of our results, refer to Appendix D for the numerical results. This finding supports the hypothesis that WEAT wordlists can be effectively transferred to SEAT through contextualization with semantically bleached sentences. The two tested models are partly trained on the same data, which could explain the high correlation. These results support the validity of our approach and suggest that the two models have at least some common bias directions.

6.5 Static WEAT Results

The numerical results of Lauscher and Glavaš (2019) and Kurpicz-Briki (2020) for FastText of the German translations of WEAT7 and WEAT8 differ slightly, but the conclusions are the same. Their difference might be due to different translation approaches. For example, the term *dance* was translated as *tanzen* by Lauscher and Glavaš (2019) and as *Tanz* by Kurpicz-Briki (2020). We use the results of Kurpicz-Briki (2020) for comparison with ours, as the p-values are reported. See Appendix A for both numerical results.

6.6 Future Work

Further exploration of additional datasets may prove fruitful. For example, dividing CW3 into two distinct wordlists (e.g., comparing immigration status to traditionalists vs. progressives). In addition, to further explore the importance of language specificity in bias assessment, the CW1-CW4 wordlists could be translated into other languages and tested with corresponding models. The significance of these results would be even greater if more languages were considered to provide a quantitative analysis of language specificity.

7 Conclusion

Our investigation of bias within German word embeddings, covering both static and contextualized models, reveals bias along several dimensions. Gender, age and family status biases were particularly prevalent, reflecting societal stereotypes found in the real world. We also found a strong correlation between bias results in static and contextualized embeddings. Furthermore, our results suggest that language specificity is important for identifying and understanding bias.

Limitations

Language Specificity Our research only suggests that language specificity is important, it does not prove it. To make a stronger statement, more languages need to be considered. This could be done by translating the datasets into different languages and comparing the biases in the respective models.

Missing Extrinsic Metrics Intrinsic metrics - as the ones used in this study - measure the bias in the pre-trained representations of the model. Extrinsic metrics quantify the bias that appears in the results of the model's downstream task. Recent studies, including those by Cao et al. (2022); Kaneko et al. (2022); Orgad et al. (2022), have shown that the correlation between intrinsic and extrinsic metrics is very limited. As highlighted by Orgad and Belinkov (2022), the inclusion of extrinsic metrics is critical for several reasons, including the greater relevance of these metrics to bias mitigation efforts. While we fully acknowledge these recommendations, we only use intrinsic metrics because the availability of extrinsic datasets for non-English languages is very limited (Ramesh et al., 2023). This finding is echoed by Wambsganss et al. (2022), who analyze the bias in German embeddings at different stages along the NLP pipeline. They find that when a pre-trained model that shows no bias on a particular metric is fine-tuned with unbiased data (on the same metric), it can produce biased output (measured again on the same metric). This underlines that the intrinsic evaluation done with WEAT and SEAT can at best be a signal of bias, a sentiment reflected by Goldfarb-Tarrant et al. (2021).

Missing Replicability of Sentences The use of GPT-4's chat interface to generate sentences for the SEAT metric introduces a replicability limitation, as it is not possible to consistently generate exactly the same model output. To enable replicability, future research could use GPT-4's API to generate sentences, setting the temperature parameter to zero to ensure deterministic output.

Ethical considerations

We only consider binary gender bias, and therefore do not consider non-binary gender identities. This does not reflect what is found in the real world (Devinney et al., 2022). The BERT model has been shown to fail to represent non-binary gender in a meaningful way (Dev et al., 2021), which further complicates matters.

The data from the co-creation workshop was collected anonymously and with the individual's consent following the approved data protection research protocol of the project.

Acknowledgements

This work is part of the Europe Horizon project BIAS funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020a. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020b. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. ArXiv:1607.04606 [cs].

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. Publisher: American Association for the Advancement of Science.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating Bias In Dutch Word Embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. ArXiv:1912.09582 [cs].

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of “Gender” in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. ArXiv:2101.00027 [cs].

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74:1464–1480. Place: US Publisher: American Psychological Association.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. MABEL: Attenuating Gender Bias using Textual Entailment Data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing Isn’t Enough! – on the Effectiveness of Debiasing MLMs and Their Social Biases in Downstream Tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Corinna Kleinert. 2006. Frauen in Führungspositionen: Karriere mit Hindernissen. Research Report 9/2006, IAB-Kurzbericht.

Angelie Kraft, Hans-Peter Zorn, Pascal Fecht, Judith Simon, Chris Biemann, and Ricardo Usbeck. 2022. *Measuring Gender Bias in German Language Generation.*

Mascha Kurpicz-Briki. 2020. *Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings.*

Mascha Kurpicz-Briki and Tomaso Leoni. 2021. A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. *Frontiers in Big Data*, 4.

Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3):795–848.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. ArXiv:2202.11923 [cs].

Anne Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris McCormick. 2020. *The Inner Workings of BERT*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How Gender Debiasing Affects Internal Model Representations, and Why It Matters. arXiv. Version Number: 2.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Kurt Pärli. 2018. Altersdiskriminierung - von der Anstellung bis zur Kündigung. *Schulthess Juristische Medien 2018*, pages 1 – 10.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Sarah Schröder, Alexander Schulz, Fabian Hinder, and Barbara Hammer. 2024. Semantic Properties of cosine based bias scores for word embeddings. ArXiv:2401.15499 [cs].

Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, and Tanja Käser. 2022. Bias at a Second Glance: A Deep Dive into Bias for German Educational Peer-Review Data Modeling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1344–1356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

A WEAT7, WEAT8 German Results

The results of Lauscher and Glavaš (2019) and Kurpicz-Briki (2020) point to the same conclusion with slightly different values. Table 3 shows a comparison of the reported WEAT scores for the FastText embedding for the German WEAT lists. The difference can be attributed to differences in the translations of the WEAT wordlists.

	Lauscher 2019	Kurpicz-Briki 2020
WEAT 7	0.46 (p>0.05)	0.23 (p=0.65)
WEAT 8	0.05 (p>0.05)	0.11 (p=0.83)

Table 3: Comparison of WEAT scores.

B GPT-4 Prompt

The following prompt is used to create the sample sentences for SEAT. This prompt contains plural sample sentences. Target words sometimes do not make sense in the plural form, so we used a shorter version of the prompt, only containing singular sample sentences (the first seven sentences). Using GPT-4 instead of manually coding sentences eliminates the need to define various linguistic elements. These include determining the gender of words (neuter, masculine, or feminine), distinguishing between things and people, knowing the plural forms of words, and deciding whether to use articles in specific sentences (nos. 1-4, 12, 13).

```
Dies ist (der/die/das ) XX.
Das ist (der/die/das ) XX.
Dort ist (der/die/das ) XX.
```

```
Hier ist (der/die/das ) XX.
(Der/Die/Das) XX ist hier.
(Der/Die/Das) XX ist dort.
(Der/Die/Das) XX ist (ein/eine)
(Sache/Mensch).
Es ist (der/die/das) XX.
Dies sind XX.
Das sind XX.
Sie sind XX.
(Die/ ) XX sind hier.
(Die/ ) XX sind dort.
XX sind (Sachen/Menschen).

Ersetze in der oben genannten Vorlage die
Sätze "XX" mit den untenstehenden
Wortlisten. Dies ergibt 14 x 5 Sätze.
Passe die Sätze an, damit sie
grammatikalisch korrekt sind. Wenn nötig,
ändere das Wort ins Plural, damit es zur
Vorlage passt. Verwende beim Satz Nr. 7
und 14 "Sache(n)", ausser "Mensch(en)"
passt offensichtlich besser ("Sache" wird
manchmal nicht passen, verwende es
trotzdem). Schreibe keinen code um dies
zu erreichen. Gib dies im CSV-Format
zurück, jeder Satz auf einer neuen Zeile:

"Dies ist das XX.",
"Das ist das XX.",
"Dort ist das XX.",
usw.

Wortliste:
Mann,Junge,Bruder,Sohn,Vater
```

C WEAT Method

The metric is based on the Implicit Association Test (IAT), where subjects are presented with two concepts, for example school subjects (e.g., *Science*, *Arts*), and gender (*Male*, *Female*). Short reaction times to classify e.g., *Science* and *Male* in a given class indicate cognitive proximity of *Male* and *Science*. In the context of static word embeddings, WEAT uses cosine similarity as a proxy for reaction time in the IAT. Cosine similarity measures the cosine of the angle between two vectors, serving as an indicator of their semantic proximity in vector space. In the following example, the association between school subjects (target words) and gender (attribute words) is compared. The attribute and target words are also referred to as stimuli. In the example, a smaller angle between *Science* and *Male* (represented by a blue dotted line in Figure 1) indicates that these two concepts are closely related. The angle from *Science* to *Female* (represented by a blue dashed line) is then subtracted from the angle *Science* to *Male* (represented by a blue dotted line). This results in an angle that quantifies the degree of relationship between the concept *Science* and the gender attributes *Male* and *Female*. This calculation is then performed for another target word (here: *Arts*) and its relation to gender (marked in green in Figure 1. The output of the calculation of the word *Arts* is compared to its counterpart for *Science*. In a perfectly unbiased embedding, these two angles should be identical. In the provided example this would clearly not be the case, as the two results of the dotted minus the dashed angles are not equal. The described procedure is done for a set of target words (e.g., *programmer, engineer, scientist,* ... and *nurse, teacher, librarian,* ...) and a set of attribute words (e.g., *man, male, he,* ... and *woman, female, her*, ...). The mean of the angles is used to aggregate the sets. The null hypothesis is that the relative similarity of the two sets of target words to the two sets of attribute words is identical. For the formulas used to compute the effect size and the p-value, we refer the reader to the original paper by Caliskan et al. (2017).

D Correlation Static to Contextualized

The Table 4 shows p-values for different dataset combinations. The consistently low values across combinations of CW1-5, GER1-2, and WEAT7-8 datasets reinforces the observed correlation between static and contextualized word embeddings.

Configuration	p-value
CW1-CW5	0.016
CW1-CW5 + GER1,2	<0.001
CW1-CW5 + GER1,2 + WEAT7,8	0.001

Table 4: Correlation between static and contextualized word embeddings

E CW1-CW5 Wordlists

The Tables 5,6,7,8 and 9 list the wordlists CW1 - CW5 used for the WEAT metric. These lists are also used for the creation of the sentence templates for the SEAT metric via the GPT-4 prompt.

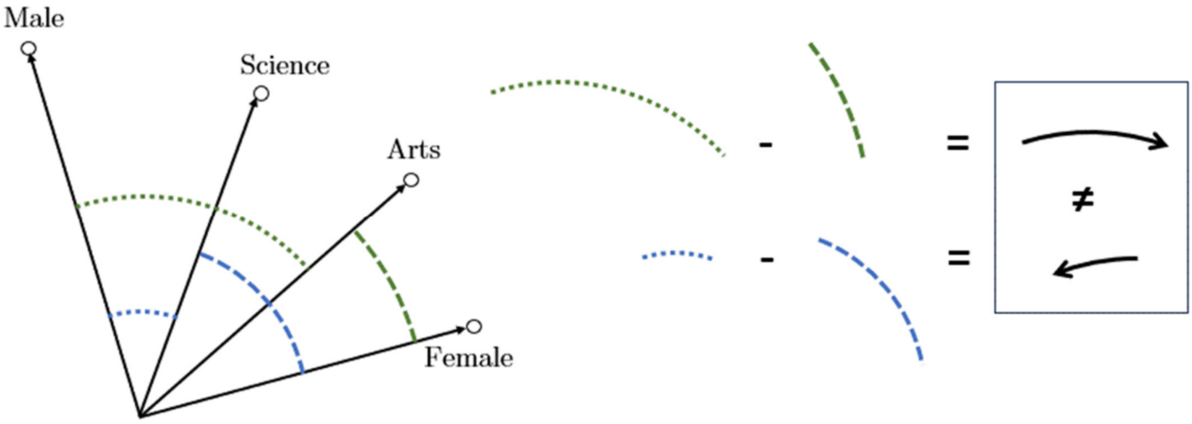

Figure 1: Angles of the attribute words (Male, Female) and target words (Science, Arts), (Own illustration)

Category	**Keywords**
Productive	Effizienz, Leistungswille, Engagement, Ambition, Zielstrebigkeit, Kompetenz, Motivation, Einsatzbereitschaft, Beharrlichkeit, Eifer
Unproductive	Ablenkung, Unaufmerksamkeit, Ineffizienz, Nachlässigkeit, Desorganisation, Überforderung, Zerstreutheit, Desinteresse, Unachtsamkeit, Sorglosigkeit
Business	Geschäftsleitung, Management, Fachkraft, Unternehmen, Gehalt, Büro, Geschäft, Karriere
Children	Elternteil, Kindertagesstätte, Sohn, Tochter, Kinder, Kinderwagen, Schnuller, Windel

Table 5: CW1 wordlists

Category	**Keywords**
Productive	Effizienz, Leistungswille, Engagement, Ambition, Zielstrebigkeit, Kompetenz, Motivation, Einsatzbereitschaft, Beharrlichkeit, Eifer
Unproductive	Ablenkung, Unaufmerksamkeit, Ineffizienz, Nachlässigkeit, Desorganisation, Überforderung, Zerstreutheit, Desinteresse, Unachtsamkeit, Sorglosigkeit
Young	Jugendlicher, Jugendliche, Jugend, Studium, Berufseinstieg, Mittzwanziger, Schülerin, Schüler, Student, Studentin
Old	Älterer, Ältere, Lebensmitte, Erwachsene, Lebenserfahrung, Erwachsener, Rentner, Rentnerin, Senior, Seniorin

Table 6: CW2 wordlists

Category	Keywords
Traditionalists	Bewahrung, Konservatismus, Unflexibilität, Konventionalität, Tradition, Gewohnheit
Communicators	Kommunikation, Vermittlung, Verbindung, Vielfalt, Mehrsprachigkeit, Integration
Native	Einheimischer, Bürger, Staatsangehöriger, Landesangehöriger, Ansässiger, Ortsansässiger
Immigrant	Ausländer, Einwanderer, Immigrant, Migrant, Migrationshintergrund, Gastarbeiter

Table 7: CW3 wordlists

Category	Keywords
Male Hobbies	Gewichtheben, Wrestling, Feuerwehr, Velorennen, Schrauben
Female Hobbies	Backen, Ökologie, Make-Up, Kunst, Ballett
Male	Mann, Junge, Bruder, Sohn, Vater
Female	Frau, Mädchen, Schwester, Tochter, Mutter

Table 8: CW4 wordlists

Category	Keywords
Productive	Effizienz, Leistungswille, Engagement, Ambition, Zielstrebigkeit, Kompetenz, Motivation, Einsatzbereitschaft, Beharrlichkeit, Eifer
Unproductive	Ablenkung, Unaufmerksamkeit, Ineffizienz, Nachlässigkeit, Desorganisation, Überforderung, Zerstreutheit, Desinteresse, Unachtsamkeit, Sorglosigkeit
Male	Mann, Junge, Bruder, Sohn, Vater
Female	Frau, Mädchen, Schwester, Tochter, Mutter

Table 9: CW5 wordlists