Unveiling Gender Bias: Transformer Models and Explainability Techniques in Dutch Job Ad Analysis

Megi Dragoti[1], Eliza Hobo[2][0009−0003−7344−4112], and Maaike de Boer[2][0000−0002−2775−8351]

[1] Tilburg University, Tilburg, The Netherlands
[2] TNO, The Hague, The Netherlands
megidragoti95@gmail.com, eliza.hobo@tno.nl, maaike.deboer@tno.nl

Abstract. Explicit bias in job advertisements perpetuates systemic discrimination, challenging fairness and equity in labor markets. The contributions of this paper are: 1) the evaluation of the effectiveness of transformer-based Natural Language Processing (NLP) in detecting explicit gender bias in Dutch job postings; 2) the interpretability of these models to uncover discriminatory terms and contextual biases. Our experiments compare monolingual models (BERTje, RobBERT) and a multilingual model (XLM-RoBERTa), as well as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). Results show that RobBERT leverages Dutch linguistic nuances effectively, achieving balanced performance, while XLM-RoBERTa demonstrates superior precision-recall balance and achieves the highest AP score. Learning curve analysis highlights the importance of larger datasets for improved model generalization, and SHAP and LIME analyses reveal critical linguistic features driving predictions, emphasizing transparency in bias detection systems. By addressing gaps in explicit bias detection and advancing transparency, this research not only shows novel insights into transformer-based models and scalabilty, but also supports societal efforts to foster equitable hiring practices.

Keywords: Gender bias · Transformer models · SHAP · LIME · Fairness · Dutch NLP

1 Introduction

Hiring discrimination is a global challenge that perpetuates systemic inequalities and reduces workforce diversity. According to the United Nations Development Programme (UNDP), nearly 90% of people harbor at least one bias against women, affecting recruitment outcomes across sectors [23]. In the Netherlands, despite progressive legislation like the Dutch Equal Treatment Act, job advertisements continue to feature explicit gendered language that subtly or overtly discourages certain groups from applying. Terms like “enthousiaste jongeman” (“enthusiastic young man”) remain commonplace, signaling persistent cultural

and structural bias [22]. In fact, discrimination in Dutch vacancy texts has been empirically demonstrated across multiple dimensions, including age, as shown by Fokkens et al. [13].

Job advertisements play a critical gatekeeping role in the hiring process. Research shows that gendered wording in these texts influences who applies, thereby reinforcing occupational segregation and limiting access to opportunity [3]. As women hold only around 25% of engineering roles in the Netherlands, compared to over 90% of caregiving roles, the gendered formulation of job advertisements may be a significant factor in this imbalance [24].

While traditional methods for detecting bias, such as keyword matching, lack nuance and scalability, transformer-based language models offer a more powerful alternative. Yet questions remain about their performance on Dutch-language data, their ability to generalize across data sizes, and the transparency of their decisions.

This study investigates how monolingual (BERTje, RobBERT) and multilingual (XLM-RoBERTa) transformer models perform in detecting explicit gender bias in Dutch job advertisements. It also explores how explainability techniques such as SHAP and LIME can uncover the linguistic features driving the predictions of these models. Learning curve analysis is used to assess how data availability affects model performance, with implications for scalable, low-resource deployment.

By integrating model performance, interpretability, and scalability, this work contributes to a comprehensive framework for bias detection in Dutch recruitment texts, supporting both academic inquiry and real-world fairness in hiring practices.

2 Related Work

Early approaches to detecting bias in text relied on rule-based systems and keyword dictionaries [3], which were effective for overt cases but lacked scalability and failed to capture contextual nuance. Vectorization methods such as Bag-of-Words (BoW) and TF-IDF improved representation but remained limited by sparsity and the inability to model semantic meaning.

The introduction of static word embeddings like Word2Vec [1] and FastText [6] enabled models to capture semantic relationships between words. However, these embeddings are context-independent and often perpetuate biases present in training corpora [15]. Contextualized embeddings with transformer architectures [2] overcame this limitation and have since become the state of the art in NLP.

In the domain of bias detection, Van Zandweghe et al. [**?**] showed that BERT-based models outperform dictionary-based tools such as the Gender Decoder for English STEM job postings, highlighting the advantages of contextual modeling. For Dutch, Vethman et al. [4] critically evaluated AI for discrimination detection in job ads. They created a 5,947-sentence dataset with expert annotations for Highly Suspected Discrimination (HSD), showing that BERTje achieved an

Average Precision of 83.3%, outperforming BoW and Word2Vec baselines. Importantly, BERT generalized well to unseen discriminatory terms, suggesting that transformers can detect new patterns beyond pre-defined keyword lists.

At the same time, several limitations have been identified. The annotation task revealed subjectivity (moderate inter-annotator agreement), and experts emphasized that efficiency gains must be weighed against issues of accountability, transparency, and proportionality. These concerns echo broader critiques of language models. Bender et al. [12] caution that larger models do not automatically lead to societal benefit, while Blodgett et al. [17] highlight the limitations of narrow bias definitions in NLP. Similarly, Delobelle et al. [19] argue for actionable fairness metrics that move beyond abstract benchmarking. Industry has also begun experimenting with inclusive writing tools such as Textio [20] and Develop Diverse [21], showing that bias-aware NLP has both practical demand and societal visibility.

While these studies confirm the promise of transformers for bias detection, two gaps remain. First, the comparative performance of monolingual versus multilingual transformer models on Dutch bias detection tasks has been underexplored. Second, explainability methods such as SHAP [8] and LIME [9], though widely applied in NLP, have not been systematically used to interpret Dutch models in this context. Addressing these gaps, our study contributes a comparative evaluation of BERTje, RobBERT, and XLM-RoBERTa, combined with interpretability techniques, to assess performance, scalability, and transparency in detecting explicit gender bias in Dutch job advertisements.

3 Methodology

3.1 Dataset

The dataset used in this study originates from Vethman et al. [22], who investigated explicit discrimination in Dutch job advertisements. It was derived from a large-scale scrape of 2.4 million vacancies posted online in 2018. Domain experts from the Netherlands Labour Authority (NLA), the Dutch Employment Insurance Agency (UWV), and the Netherlands Institute for Human Rights (NIHR) provided guidance in curating and annotating the data.

Sentences were first flagged using gender-related search terms aligned with the Dutch Equal Treatment Act, then annotated independently by five experts. Each sentence was labeled as either *Highly Suspected of Discrimination (HSD)* or *Non-discriminatory*, with ambiguous cases removed to ensure binary classification. Inter-annotator agreement on a shared subset reached a Fleiss' κ of 0.557[3], indicating moderate reliability and substantial agreement for positive (HSD) cases[18].

[3] Fleiss' κ is a statistical measure for assessing the reliability of agreement between multiple raters when assigning categorical ratings to a fixed number of items. Values between 0.41 and 0.60 typically indicate moderate agreement.

After preprocessing and filtering, the final dataset contains 5,947 sentences, of which 28.8% are labeled as HSD and 71.2% as non-discriminatory. This class imbalance not only reflects the relative rarity of explicit bias in real-world postings, but also underscores its persistence: despite being legally prohibited, a substantial proportion of vacancies still contain discriminatory wording. At the same time, the dataset provides a representative basis for training and evaluating models.

3.2 Models and Setup

We compare three transformer models: BERTje [22], RobBERT [10], and XLM-RoBERTa [11]. BERTje and RobBERT are Dutch-specific models trained on large national corpora, while XLM-RoBERTa is a multilingual model pretrained on text from over 100 languages. Together, these models represent strong monolingual baselines alongside a widely adopted multilingual architecture, enabling a direct comparison between language-specific and cross-lingual approaches.

Each model has been validated in prior research. BERTje has proven effective in detecting explicit discrimination in Dutch job advertisements [22], establishing its relevance for bias detection. RobBERT, introduced by Delobelle et al. [10], achieved state-of-the-art performance across Dutch NLP benchmarks such as sentiment analysis and part-of-speech tagging, demonstrating its robustness across tasks. XLM-RoBERTa, developed by Conneau et al. [11], set new standards on cross-lingual understanding benchmarks, showing its ability to generalize effectively across diverse languages.

By comparing these three models, we aim to evaluate whether Dutch-specific pretraining offers advantages in capturing linguistic nuance, or whether a large multilingual model provides superior generalization. This rationale directly supports our research objective of analyzing the trade-offs between monolingual and multilingual architectures for detecting explicit gender bias in Dutch recruitment texts.

The overall experimental workflow, including preprocessing, model fine-tuning, hyperparameter optimization, and evaluation, is illustrated in Figure **??**. All models were fine-tuned for binary classification with a dense classification head on top of the transformer encoder. To maintain label balance, we applied stratified sampling to split the dataset into 70% training (4,163 samples) and 30% test data (1,784 samples). Hyperparameters were optimized using Optuna's Tree-structured Parzen Estimator (TPE) across a search space including learning rate, batch size, weight decay, and number of epochs, with optimization conducted through 3-fold cross-validation on the training data to ensure robustness.

Model performance was evaluated using metrics suited to imbalanced data: Average Precision (AP), which emphasizes the precision–recall trade-off, and ROC AUC, which measures class separability across thresholds. These metrics provide a fair assessment of each model's ability to detect minority-class (biased) cases.

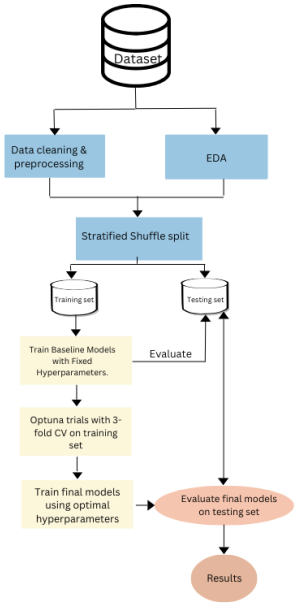

Fig. 1. Overview of the experimental pipeline: from dataset to interpretability.

3.3 Explainability

To address the "black-box" nature of transformer models, we employ two complementary post-hoc explainability methods. Shapley Additive Explanations (SHAP) [8] provide global interpretability by quantifying the overall contribution of individual tokens to model predictions across the dataset. Local Interpretable Model-agnostic Explanations (LIME) [9], in contrast, generate local explanations by perturbing inputs and approximating the model decision boundary around specific instances.

Using SHAP, we identify the most influential gendered terms (e.g., *vrouwelijk*, *jongeman*) driving classifications, while LIME highlights token-level contributions for individual job advertisements. Together, these methods ensure both systemic and instance-level transparency, offering insights into how biased language shapes predictions and supporting trust in automated bias detection.

4 Results

4.1 Model Performance

Table 1 summarizes the performance of the three transformer models. XLM-RoBERTa achieves the highest Average Precision (AP) at 86.5%, indicating superior performance in detecting minority-class instances. RobBERT shows the best balance, combining high AP (84.3%) with stable ROC AUC (93.1%).

BERTje lags slightly behind, but remains competitive with an AP of 83.6% and the highest ROC AUC at 93.3%.

Table 1. Performance of transformer models on Dutch job ad bias detection.

Model	AP (%)	ROC AUC (%)
BERTje	83.6	93.3
RobBERT	84.3	93.1
XLM-RoBERTa	**86.5**	93.2

While quantitative metrics highlight overall model performance, they do not reveal where and why the models succeed or fail. To gain deeper insight, we examined individual job advertisement sentences using SHAP and LIME explanations. This qualitative error analysis shows how models handle explicit gendered terms and how token importance differs across architectures. Table 2 presents illustrative examples, including cases where models disagreed or misclassified, and highlights how reliance on specific tokens (e.g., *mannen*, *vrouwelijke*) influenced predictions. These examples complement the aggregate results by demonstrating the strengths and limitations of each model in practice.

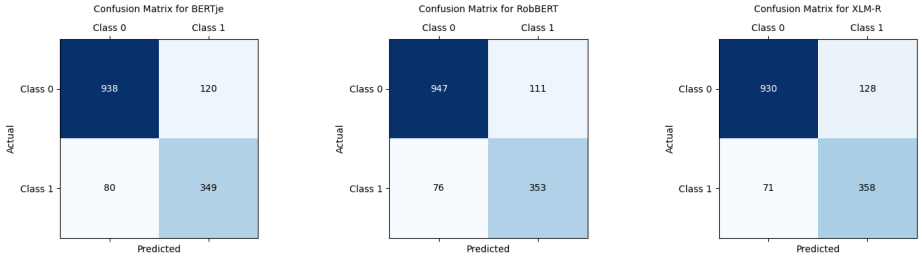

Fig. 2. Confusion matrices for BERTje, RobBERT, and XLM-RoBERTa on the held-out test set. The matrices illustrate true positives (bottom right), true negatives (top left), false positives (top right), and false negatives (bottom left) for the binary classification task (Class 0 = Non-discriminatory, Class 1 = HSD).

4.2 Learning Curve

Figure 3 illustrates how model performance scales with increasing amounts of training data. Both monolingual models, BERTje and RobBERT, show steady improvements and reach stable performance with relatively modest amounts of data. In contrast, XLM-RoBERTa exhibits greater variability at smaller training sizes but continues to improve with more data, reflecting its reliance on larger

Table 2. Illustrative examples of model predictions with token-level importance differences.

Sentence (Dutch → English)	**True Label**	**Model Predictions**
"Beschrijving vacature ... op zoek naar 2 sterke **mannen** ... glaszetter." (Job description ... looking for 2 strong **men** as glaziers.)	HSD	BERTje: Non-discriminatory RobBERT: HSD XLM-R: HSD (strong weight on *mannen*)
"Vacature omschrijving **Vrouwelijke assistente/secretaresse** gezocht ..." (Job description: **Female assistant/secretary** wanted ...)	HSD	BERTje: HSD RobBERT: HSD (high weight on *vrouwelijke*) XLM-R: HSD (distributed weights)
"... op zoek naar een representatieve **dame** als Administratief Medewerkster ..." (... looking for a representative **lady** as an administrative assistant ...)	HSD	BERTje: Non-discriminatory / borderline RobBERT: HSD XLM-R: HSD
"... op zoek naar renovatie**timmermannen** ..." (... looking for renovation **carpenters** ...)	HSD	BERTje: Non-discriminatory / borderline RobBERT: HSD XLM-R: HSD (strong weight on *timmermannen*)
"Onze organisatie zoekt medewerkers die goed kunnen samenwerken met **mannen en vrouwen**." (Our organization is looking for employees who can work well with **men and women**.)	Non-HSD	BERTje: Non-discriminatory RobBERT: HSD XLM-R: HSD (emphasis on *mannen* / *vrouwen*)

corpora for stability. Across all models, performance gains plateau around 80% of the training data, suggesting diminishing returns beyond this threshold.

4.3 Explainability Findings

To better understand why the models classify a sentence as biased or unbiased, we applied two complementary interpretability methods. SHAP provides *global interpretability*, identifying which tokens most strongly contribute to predictions across the dataset, while LIME offers *local interpretability* by highlighting the specific tokens driving individual classifications.

Using SHAP, we found that explicitly gendered terms—particularly *vrouwelijk* (female)—consistently received strong positive attribution across all models. In contrast, masculine-coded tokens such as *man* or *jongeman* showed more varied, and occasionally even negative, SHAP contributions. This suggests that models internalized a stronger association between feminine markers and discriminatory language.

Notably, the discriminatory attribution of gendered terms also appeared sensitive to lexical combinations. For instance, phrases like *enthousiaste jongeman* (enthusiastic young man) received higher SHAP scores than the same terms in isolation, underscoring the role of modifier-noun interactions. However, this

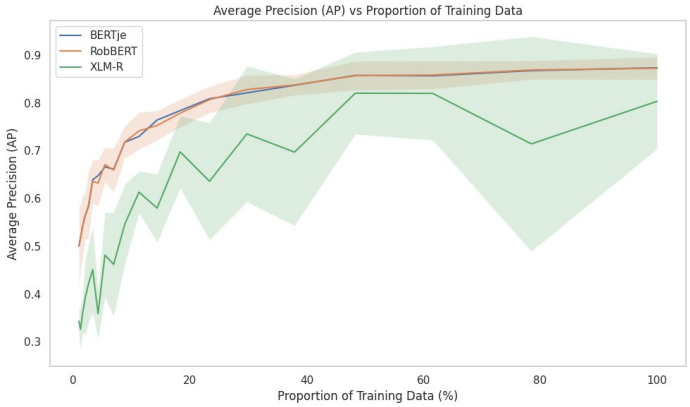

Fig. 3. Learning curves (AP vs. proportion of training data) for BERTje, RobBERT, and XLM-RoBERTa. RobBERT and BERTje scale efficiently, while XLM-RoBERTa requires more data for stable performance.

effect was more consistent for feminine constructions (e.g., *vrouwelijke secretaresse*) than for masculine ones, indicating that feminine tokens may carry a more direct signaling function within these contexts.

LIME analysis reinforced these findings at the sentence level. In one example, a sentence containing the term *mannen* (men) was classified as discriminatory by RobBERT and XLM-RoBERTa—largely due to token attribution—while BERTje down-weighted the same token, leading to a neutral classification. Such inconsistencies highlight that while all models detect overt gender markers, they differ in how strongly they rely on these cues, with feminine-coded terms exerting a more uniform and influential effect.

Overall, these findings suggest that transformer models do not simply memorize biased keywords, but learn context-sensitive patterns of attribution—patterns that are themselves shaped by training data asymmetries. The prominence of feminine terms in model predictions likely reflects both their legal salience and their structural positioning in job ads. Understanding this dynamic is critical if such models are to be used in fairness-sensitive applications.

Together, SHAP and LIME demonstrate that transformer models not only capture explicit gender markers but also contextual interactions. To better understand the model's decisions to classify a sentence as biased or unbiased, we applied these two complementary methods: SHAP for global interpretability, which identifies the tokens that most strongly influence classifications across the entire dataset, and LIME for local interpretability, which highlights the specific tokens that drive the prediction for an individual sentence. By combining global and local perspectives, these methods enhance transparency and provide actionable insights into model behavior, thereby enabling greater trust in automated bias detection.

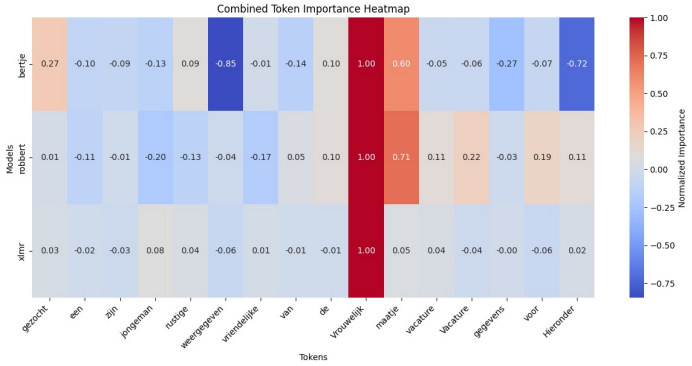

Fig. 4. SHAP summary plot showing influential tokens for bias detection across models. Terms such as *vrouwelijk* receive consistently high importance, while others like *jongeman* show lower or variable contributions.

5 Discussion

The comparative evaluation highlights important trade-offs between monolingual and multilingual transformer models for Dutch gender bias detection. RobBERT consistently achieved stable and balanced results, demonstrating the benefits of Dutch-specific pretraining for linguistic nuance. BERTje, while slightly weaker, remains a reliable baseline and underscores the value of domain-specific resources. By contrast, XLM-RoBERTa achieved the highest Average Precision, indicating strong capacity to capture discriminatory cues. However, the learning curve analysis (Figure 3) showed that XLM-RoBERTa required substantially more data to reach stability. This can be explained by its larger size and parameterization compared to BERTje and RobBERT: while its capacity allows it to model more complex patterns, it also makes the model more data-hungry and increases the risk of overfitting when trained on smaller datasets. Its multilingual pretraining further contributes to this effect, as representations optimized for over 100 languages may be less efficient when fine-tuned on limited Dutch-specific data.

Explainability analyses provide further insight into model behavior. SHAP revealed that explicit markers such as *vrouwelijk* and *jongeman* were consistently influential across models, while LIME illustrated model-specific differences in token weighting. For instance, RobBERT and XLM-RoBERTa emphasized terms like *mannen* more strongly than BERTje, leading to divergences in classification. These findings show that even high-performing models differ in their reliance on linguistic cues, which has several implications. First, from a fairness perspective, it means that two systems trained on the same data may still disagree on which sentences are discriminatory, raising questions of consistency and accountability in deployment. Second, from a trust perspective, it highlights the need for transparent explanations: recruiters, auditors, and regulators must be able to see why a sentence is flagged, and whether this reasoning aligns with legal and societal

expectations. The combination of global (SHAP) and local (LIME) explanations proved valuable in uncovering both systematic patterns and case-specific decisions, thereby supporting transparency in deployment.

Despite these contributions, several limitations must be acknowledged. First, the dataset is restricted to explicitly gendered terms and a binary label scheme, limiting the ability to detect more subtle or intersectional forms of discrimination. Second, inter-annotator agreement, while moderate, indicates inherent ambiguity in defining discriminatory language. This has two important implications. First, for models, it means the training signal may be noisy—different annotators may interpret the same phrase differently, which can limit achievable performance. Second, for practical deployment, it underscores that even experts may disagree—discrimination in language can be subjective and context-dependent. This echoes recent discussions in NLP disagreement research, which argue that annotator differences are not merely noise but reflect valid, competing perspectives on sensitive language tasks (Fleisig et al. 2023) [16]. Collectively, these findings suggest that automated systems should support rather than replace human judgment in complex evaluative tasks. Third, while transformer models achieved strong results, they remain computationally demanding relative to traditional methods. However, compared to much larger GPT-like language models, BERT-style architectures such as BERTje and RobBERT are already more efficient and thus represent a more realistic option in resource-constrained environments. Future work should further examine efficiency and fairness trade-offs, particularly when deploying models in settings with limited computational or data resources. Finally, this study focused on gender; future research should expand to age, ethnicity, and other legally protected categories to build a more comprehensive framework for bias detection in recruitment.

Overall, the findings demonstrate that transformer-based NLP models can effectively detect explicit gender bias in Dutch job advertisements, but their real-world adoption requires careful consideration of interpretability, annotation quality, and societal context. The discussion underscores the importance of balancing performance with transparency and fairness when deploying automated systems in sensitive domains such as hiring.

6 Conclusion

This study demonstrates the effectiveness of transformer-based models in detecting explicit gender bias in Dutch job advertisements. Among the three models, RobBERT provides stable and balanced performance, while XLM-RoBERTa achieves the highest precision–recall trade-off. BERTje, though slightly behind, remains a competitive Dutch baseline.

Beyond classification performance, the integration of SHAP and LIME strengthens interpretability by revealing the linguistic cues driving model predictions. By making model reasoning visible, interpretability builds trust in automated bias detection, not as a black box, but as a tool that can be inspected, challenged, and improved. This trust is essential for adoption in recruitment, where legal

accountability and fairness are paramount: organizations must be able to justify why a vacancy is flagged as discriminatory, and regulators need transparent evidence to act upon. Interpretable outputs also empower practitioners to identify systematic patterns of bias, refine recruitment texts, and increase confidence that automated systems are supporting, rather than undermining, equitable hiring practices.

Overall, the findings highlight the potential of scalable and explainable NLP pipelines to support fairer hiring practices in the Netherlands and beyond. Future research should expand the scope beyond gender to other forms of bias (e.g., age, ethnicity), evaluate models on larger and more heterogeneous datasets, and explore newer transformer architectures to further advance fairness in recruitment technologies.

Acknowledgments

This research was supported by TNO (Netherlands Organisation for Applied Scientific Research). I thank my supervisors, Dr. Grzegorz Chrupała and Merle Beaujon, for their guidance, as well as my colleagues Eliza Hobo and Maaike de Boer for their invaluable support throughout this process.

References

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the 1st International Conference on Learning Representations (ICLR 2013, Workshop Track) (2013). `https://arxiv.org/abs/1301.3781`
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008 (2017)
3. Gaucher, D. et al.: Evidence that gendered wording in job ads exists and sustains gender inequality. JPSP (2011)
4. Vethman, S., Veenman, C., Adhikari, A., Hobo, E., de Boer, M., van Genabeek, J., de Maaike, M.: Detecting discrimination in job vacancies: A critical reflection on the potential of AI language models. In: Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25), PMLR, pp. 1–20 (2025)
in *Proceedings of [Conference Name]*, June 2025.
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
7. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4349–4357 (2016)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774 (2017)

9. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
10. Delobelle, P., Winters, T., Berendt, B.: RobBERT: a Dutch RoBERTa-based language model. In: Proceedings of the 29th Benelux Conference on Artificial Intelligence, pp. 325–338 (2020)
11. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451 (2020)
12. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: 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), pp. 610–623 (2021)
13. Fokkens, A.S., Beukeboom, C.J., & Maks, E.: Leeftijdsdiscriminatie in vacatureteksten: Een geautomatiseerde inhoudsanalyse naar verboden leeftijd-gerelateerd taalgebruik in vacatureteksten. Rapport in opdracht van het College voor de Rechten van de Mens (2018).
14. Van Zandweghe, N.: Using Bidirectional Transformer Neural Networks for Advancing Gender Bias Recognition in STEM Job Advertisements. National High School Journal of Science (NHSJS), 2024.
15. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4349–4357 (2016)
16. Fleisig, E., Pavlick, E., & Kwiatkowski, T.: Modeling annotator disagreement for subjective tasks. In: Proceedings of EMNLP 2023. (2023)
17. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476 (2020)
18. Fleiss, J.L.: Measuring nominal-scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382 (1971). https://doi.org/10.1037/h0031619
19. Delobelle, P., Attanasio, G., Nozza, D., Blodgett, S.L., & Talat, Z.: Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pp. 21669–21691, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1207
20. Textio: Augmented writing platform. https://textio.com, last accessed 2025/09/10
21. Develop Diverse: Inclusive communication software. https://www.developdiverse.com, last accessed 2025/09/10
22. Vethman, S., Adhikari, A., de Boer, M.H.T., van Genabeek, J.A.G.M., Veenman, C.J.: Context-aware discrimination detection in job vacancies using computational language models. In: ACM Conference Proceedings, 17 pages (2022).
23. United Nations Development Programme (UNDP): Tackling Social Norms: A Game Changer for Gender Inequalities. Human Development Perspectives (2020). https://hdr.undp.org/gender-norms-2020
24. CBS (Centraal Bureau voor de Statistiek): Beroepen naar geslacht, leeftijd en herkomst. (2023). https://opendata.cbs.nl/statline/#/CBS/nl/dataset/82808NED