Gender and Ethnicity Bias Mitigation in Automatic Recruitment

Angela Petkova, advisor Farnaz Faramarzi Lighvan, and promotor Lynn Houthuys

Vrije Universiteit Brussel, Brussels, Belgium
angela.s.petkova@gmail.com, lynn.houthuys, farnaz.faramarzi.lighvan@vub.be

Keywords: Machine Learning · Fairness in AI · Automatic Recruitment

Due to many advancements in machine learning research in recent years, decision-making algorithms are increasingly used in recruitment processes to assist hiring managers in finding successful candidates. An example is Amazon's Recruitment tool, an AI-driven hiring system that was trained on resumes submitted to the company over 10 years (Dastin Jeffrey, 2018). There was no explicit gender variable available on the resumes; it learned to associate certain words with female candidates, such as "women's chess club captain" and women's colleges. Because the majority of the training data came from men who previously applied for jobs at Amazon, the algorithm learned to prefer male candidates, leading to discriminatory hiring decisions (Cpollnow & Sornik, 2023). This shows that machine learning models can detect and exploit sensitive attributes, such as gender or ethnicity, by identifying patterns in unstructured resume data, even when those attributes are not explicitly provided. This bachelor thesis focuses on creating fairer hiring outcomes by debiasing the unstructured textual component of a resume: the biography. This component has not been previously addressed in isolation and represents a complementary modality to the visual debiasing approach introduced by Peña et al. (2020).

The biography was debiased for the sensitive gender attribute. The approach combined two techniques: scrubbing and round-trip translation. Scrubbing removes or replaces explicit gender- or ethnicity-related indicators. Explicit gender indicators included first names, gendered titles, -pronouns, -occupational terms, and familial terms. Gendered occupational terms, such as "fireman," and familial terms, such as "mother," were replaced with gender-neutral alternatives like "firefighter" and "parent," rather than being removed, as they carry relational or professional information. Round-trip translation was applied to remove implicit gender indicators by translating the text into the pivot language, French and then back to the original using the Google Translate API, thereby stripping away stylistic features while preserving the original sentence's meaning (Rabinovich et al., 2017).

To evaluate this debiasing approach, we used Peña et al. (2020)'s framework on the FairCVtest dataset, which contains 24,000 resumes with 12 structured

features and a biography section. Each candidate has a Target Score from 0 to 1, indicating predicted job suitability. Two types of scores are provided: Unbiased and Gender-biased. The biased scores simulate real-world scenarios where certain demographic groups are disadvantaged. We create a biased regression model, using original biographies, and an agnostic regression model, using debiased biographies (scrubbing + round-trip translation). Both models will be trained on biased scores, along with the structured features [1]. In all cases, the model is used to rank candidates by predicting a score for each candidate. The top 100 are selected based on the predicted scores, and the fairness of this ranking is analysed. The goal is not to accurately identify the top 100 most qualified candidates, because the target scores the model is trained on are intentionally biased. The first version, using raw biography embeddings, evaluates whether the model can uncover sensitive attributes purely from unstructured data, without access to explicit gender or ethnicity attributes. In contrast, the agnostic version is specifically made to assess whether applying the bias mitigation techniques can prevent the model from reproducing biased outcomes, even when trained on biased scores. All models were implemented using an early fusion strategy. The MLP regression models were feedforward neural networks, and the original and debiased biographies were encoded using Bert.

The removal of sensitive information using the debiasing techniques improved model fairness. The agnostic model achieved a perfectly balanced 50–50 male-to-female ratio in the top-10 selection, better representation in the top-100 selection, and a top-10 ranking more closely aligned with the true Unbiased top-10 (higher NDCG score). MAE calculated against the Unbiased scores and KL-divergence also improved (see Table 1). Based on these findings, it can be concluded that scrubbing and using round-trip translation effectively helps correct the gender bias in the FairCVtest dataset.

Table 1: Performance and fairness results between the biased and agnostic model

Metric	**Biased model**	**Agnostic model**
MAE (Gender-biased)	0.018	0.043
MAE (Unbiased)	0.050	0.049
KL (female score distribution to male)	0.237	0.015
% Female in top 100	23%	44%
% Female in top 10	10%	50%
NDCG (against Unbiased top 10)	0.977	0.995

Finally, the same debiasing pipeline was also applied with ethnicity as the sensitive attribute. However, since the biographies contain limited ethnicity-related information, scrubbing and back-translation did not significantly reduce ethnic bias in the FairCVtest dataset.

[1] Excluding the sensitive feature of gender

References

1. Cpollnow, K., Sornik, K.: Hiring bias gone wrong: Amazon recruiting case study. Cangrade (2023). https://www.cangrade.com/blog/hr-strategy/hiring-bias-gone-wrong-amazon-recruiting-case-study/
2. Dastin, J.: Insight: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018). https://www.reuters.com/article/business/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1VB1FQ/
3. Peña, A., Serna, I., Morales, A., Fierrez, J.: Bias in multimodal AI: testbed for fair automatic recruitment. arXiv preprint arXiv:2004.07173 (2020)
4. Rabinovich, E., Mirkin, S., Nath Patel, R., Specia, L., & Wintner, S. (2017). Personalized Machine Translation: Preserving original author traits. In Proceedings Of The 15th Conference Of The European Chapter Of The Association For Computational Linguistics: Volume 1, Long Papers (pp. 1074–1084).