Evaluating and Mitigating Gender Bias in Pre-trained Embeddings for ML-based Recruitment

Farnaz Faramarzi Lighvan[1,*], Lynn Houthuys[1]

[1]AI Lab, Vrije Universiteit Brussel, Belgium

Abstract

AI-based recruitment systems that rely on machine learning models trained on historical CV data, risk perpetuating and amplifying social biases. A key challenge arises in unstructured CV text, where pre-trained language model embeddings may infer sensitive attributes such as gender even after explicit indicators are removed. In this paper, we evaluate nine pre-trained embedding models on the synthetic FairCVdb dataset, analyzing the informativeness of their embeddings for applicant scoring and their susceptibility to gender leakage, on both original and gender-scrubbed biographies. We further use a multi-task adversarial learning framework with gradient reversal to predict applicant suitability while suppressing gender information from learned representations. Finally, we use a multi-objective Pareto-front-based model selection to balance predictive utility and fairness. Our experimental results show that explicit gender scrubbing substantially reduces but does not eliminate gender leakage, while adversarial learning improves fairness mainly on original biographies and acts as a complementary strategy rather than a substitute for text-level debiasing.

Keywords

automatic recruitment, gender bias, fairness, text embedding models, adversarial debiasing

1. Introduction

Artificial intelligence is increasingly being adopted in recruitment systems to support candidate screening and hiring decisions, and reducing manual effort when evaluating large numbers of applicants [1]. In many such systems, applicant CVs are used to predict suitability scores or generate hiring recommendations. However, these systems raise serious concerns about fairness and discrimination toward demographic groups such as defined by gender and ethnicity [2], further reflected in regulation such as the EU AI Act [3], classifying recruitment and candidate evaluation as high-risk and subjecting them to bias-mitigation, transparency, and human-oversight obligations.

AI is commonly used in recruitment in two ways. The first relies on large language models (LLMs) to match job descriptions with candidate CVs. Although efficient, LLM-based approaches have shown biased output generation [4] and can be less transparent and harder to control in high-stakes settings. Prior work has examined LLM-based resume matching with respect to gender, race, maternity and pregnancy status, political affiliation, and education [5].

The second approach trains machine learning (ML) models on historical recruitment data [6]. These models can be tailored to specific job contexts, but their targets may reflect existing social and organisational biases [7]. In particular, suitability scores assigned by human decision-makers may encode bias, which ML models can reproduce or amplify [8]. A common mitigation strategy is to remove sensitive attributes from the input. While this is relatively straightforward for structured CV features, it is more difficult for unstructured text. For gender bias, names, pronouns, and titles can be removed or replaced with gender-neutral alternatives [9, 10], but remaining proxy words may still reveal gender [9, 11]. Amazon's discontinued recruitment model illustrates this risk, as it penalised resumes containing terms associated with female applicants after being trained on historically male-dominated hiring data [12].

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.

*Corresponding author.

Farnaz.faramarzi.lighvan@vub.be (F. F. Lighvan); Lynn.houthuys@vub.be (L. Houthuys)

0009-0009-7505-7128 (F. F. Lighvan); 0000-0003-0867-4168 (L. Houthuys)

Therefore, textual CV inputs often require careful scrubbing of explicit gendered terms, such as replacing "mother" and "fireman" with "parent" and "firefighter" [13].

For unstructured text, natural language processing methods encode CV text into vector representations. Pre-trained embedding models can capture rich semantic information for tasks such as applicant ranking or score prediction, but they may also infer gender from implicit cues such as word choice, writing style, and linguistic patterns. Petkova et al. [13] proposed round-trip translation to reduce such implicit gender cues, while Pena et al. [14] proposed explainability and adversarial learning as training-time alternatives for mitigating gender bias in LLM embeddings.

The main objective of this paper is to determine whether multi-task adversarial training can substitute for, or complement, preprocessing-stage gender scrubbing when pre-trained embeddings are used for recruitment scoring. To study this, we train multilayer perceptron (MLP) models on the synthetic FairCVdb dataset [15], which provides applicant CVs together with gender-biased and unbiased suitability scores. We deliberately train on gender-biased scores, since unbiased labels are rarely available in real-world recruitment; the unbiased scores are used only as a reference for ranking-based fairness evaluation. Using nine pre-trained embedding models on both original and gender-scrubbed biographies, we first train a model to predict the suitability score for a candidate, as well as a separate gender classification model to analyze gender leakage. We then use a multi-task adversarial framework with a gradient reversal layer [16] to jointly predict suitability and gender while suppressing gender-related information in the shared representation. The main contributions are:

- A comparative analysis of gender-leakage potential across nine embedding models on original and gender-scrubbed biographies in a recruitment scoring setting.
- A mid-fusion neural network architecture that processes biography embeddings before combining them with structured CV features for increased flexibility for information fusion of both modalities.
- A multi-task adversarial learning framework with gradient reversal for suppressing gender information from shared representations while preserving scoring utility.
- A Pareto-front-based multi-objective optimization strategy for principled model selection under an explicit accuracy–fairness trade-off.
- An empirical investigation into whether adversarial training can substitute for, or complement preprocessing-stage gender scrubbing.

The most closely related work is the recent study by Pena et al. [14], which uses a similar multi-task framework for bias mitigation in ML-based recruitment. While this shows the potential of the approach, this study effectively extends on this in several key areas. More specifically: (1) by considering both original and scrubbed biographies and a more diverse set of embeddings, allowing for a more thorough analysis on the importance of scrubbing and selection of embedding model.; (2) by using a Pareto-front-based model selection instead of using a fixed fairness penalty weight, allowing for an explicit and principled accuracy–fairness trade-off.

2. Methodology

This section describes the experimental pipeline used to study bias-aware score prediction on FairCVdb. The task is formulated as a regression problem in which each candidate is represented by structured CV attributes and a textual biography, and the model predicts a hiring suitability score in the interval $[0, 1]$. Models are trained on gender-biased targets, reflecting realistic conditions where unbiased labels are unavailable; however, the goal is to achieve accurate score prediction while reducing unfair differences between male and female applicants.

We conduct three sets of experiments using the same mid-fusion strategy to combine the input modalities of the dataset. First, we train a single-task gender classifier on both original and gender-scrubbed biographies to measure the gender-leakage potential of each embedding model. Second, we train a single-task regressor for score prediction using both biography variants, evaluating the effect of

text-level debiasing and the utility of each embedding model. Third, we use a multi-task adversarial neural network with shared fusion layers, task-specific output heads, and a gradient reversal layer to learn representations that are informative for score prediction while reducing gender information. For model selection, we use a Pareto-front-based strategy to balance predictive accuracy and fairness instead of fixing a fairness penalty weight manually.

2.1. FairCVdb dataset

We conduct all experiments on FairCVdb [17], a synthetic dataset of 24,000 resume profiles designed to study fairness in automated recruitment. Each profile includes structured candidate information, demographic attributes, occupation information, a name, and a short biography. The structured features contain 12 competency attributes derived from education, availability, previous experience, recommendation letter status, and language proficiency, together with a suitability attribute measuring the affinity between the candidate's occupation sector and the target job. The dataset is split into 19,200 training and 4,800 test profiles, balanced across gender, ethnicity, and occupational sector.

FairCVdb provides unbiased scores T^U, gender-biased scores T^G, and ethnicity-biased scores. In this work, we focus on gender bias and train all models using T^G, since real-world recruitment data would typically contain biased historical labels rather than unbiased ground truth. The unbiased scores T^U are not used for training or model selection; they are reserved only for computing ranking-based fairness metrics during evaluation, as described in Section 2.6.

The dataset provides two biography versions: original biographies, which include explicit gender indicators such as names, pronouns, and titles, and gender-scrubbed biographies, where these explicit indicators are removed. We use only the structured competency-related features and biography text, excluding face image embeddings.

2.2. Embedding models

We evaluate nine pre-trained text representation models spanning four model groups to encode the biography sections of the CVs.

Static word embeddings. Static word embeddings assign one fixed vector to each word, independent of context. We include fastText [18](`crawl-300d-2M`) and GloVe [19] (`glove.6B.300d`), both using 300-dimensional pre-trained word vectors. For each biography, token-level vectors are averaged to produce a fixed-size biography representation.

Contextual transformer models. Contextual transformer models produce token representations that depend on the surrounding text. We include `bert-base-uncased` model of BERT [20], `FacebookAI/roberta-base` model of RoBERTa [21], and `microsoft/mpnet-base` of MPNet [22]. Biography embeddings are obtained by mean-pooling the last hidden states over non-padding tokens, using a maximum sequence length of 256 tokens.

Sentence transformer models. Sentence transformer models are optimized to produce semantically meaningful sentence-level embeddings. We include `all-mpnet-base-v2`, a fine-tuned version of MPNet [22], and `all-MiniLM-L6-v2`, a fine-tuned version of MiniLM [23], both adapted within the SentenceTransformers framework [24] on large sentence-pair datasets using a contrastive objective, and optimised specifically for producing semantically meaningful fixed-size sentence representations.

OpenAI API embeddings. These are large-scale proprietary embedding models trained on broad web-scale corpora, being among the current state-of-the-art in dense text representation. We include `text-embedding-3-large` and `text-embedding-3-small`, accessed via the OpenAI API [25].

2.3. Single-Task Gender Classification

We train an MLP for binary gender classification using BCE loss. The input is the concatenation of standardized structured CV features and the biography embedding. Gender classification accuracy is used as a proxy for gender leakage: higher accuracy indicates that the representation retains more

gender-revealing information. We run this experiment on both original and gender-scrubbed biographies to assess how much gender information remains after scrubbing.

2.4. Single-Task Score Prediction

We train an MLP regression model to predict the gender-biased hiring score T^G using MAE loss. The input is the concatenation of standardized structured features and the biography embedding. Fairness is considered during model selection by tuning hyperparameters using the Pareto-front-based procedure in Section 2.7. Experiments are run on both original and gender-scrubbed biographies to evaluate the effect of text debiasing on accuracy and fairness.

2.5. Multi-Task Adversarial Learning with Gradient Reversal

To reduce gender information in the learned representation, we employ a multi-task adversarial learning architecture with a Gradient Reversal Layer (GRL) [16]. The model consists of a shared encoder, a score prediction head, and an auxiliary gender prediction head.

Given the input representation $\mathbf{x}_i$, the shared encoder computes a latent representation:

$$\mathbf{z}_i = f_\theta(\mathbf{x}_i). \tag{1}$$

The score prediction head h_ϕ predicts the hiring score, while the adversarial gender head a_ψ predicts the gender attribute from the reversed representation:

$$\hat{y}_i = h_\phi(\mathbf{z}_i), \qquad \hat{g}_i = a_\psi(\mathrm{GRL}(\mathbf{z}_i)). \tag{2}$$

The GRL acts as an identity function in the forward pass, but reverses the gradient during backpropagation:

$$\frac{\partial \mathrm{GRL}(\mathbf{z})}{\partial \mathbf{z}} = -\mathbf{I}. \tag{3}$$

where $\mathbf{I}$ is identity matrix. The joint training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \lambda\, \mathcal{L}_{\text{gender}}. \tag{4}$$

Because of the gradient reversal, minimizing this objective updates the gender head to improve gender prediction, while simultaneously updating the shared encoder in the opposite direction, encouraging it to discard gender-predictive information from $\mathbf{z}$. The score prediction head is trained with MAE loss and the gender head with Binary Cross-Entropy (BCE) loss. The adversarial weight λ is treated as a tunable hyperparameter. As with the single-task setting, both prediction accuracy and fairness are considered during hyperparameter tuning using the Pareto front-based procedure described in Section 2.7, and experiments are conducted on both original and gender-scrubbed biographies across all nine embedding models. Figure 1 shows the single-task and multi-task architectures used in our experiments.

2.6. Performance and Fairness Metrics

For score prediction, we report the following performance and fairness metrics.

MAE measures the average absolute deviation between predicted scores and gender-biased ground-truth scores T^G, reflecting predictive utility.

KL divergence. We use Kullback–Leibler (KL) divergence [26] to measure the distributional disparity between the predicted score distributions of male and female candidates. Lower values indicate that the male and female predicted-score distributions are more similar, suggesting greater distributional parity.

P-percent rule (P%) measures demographic parity as the ratio of female-to-male (or male-to-female, whichever is smaller) representation in the top-100 predicted scores. A value of 1.0 indicates perfect parity; lower values indicate greater disparity.

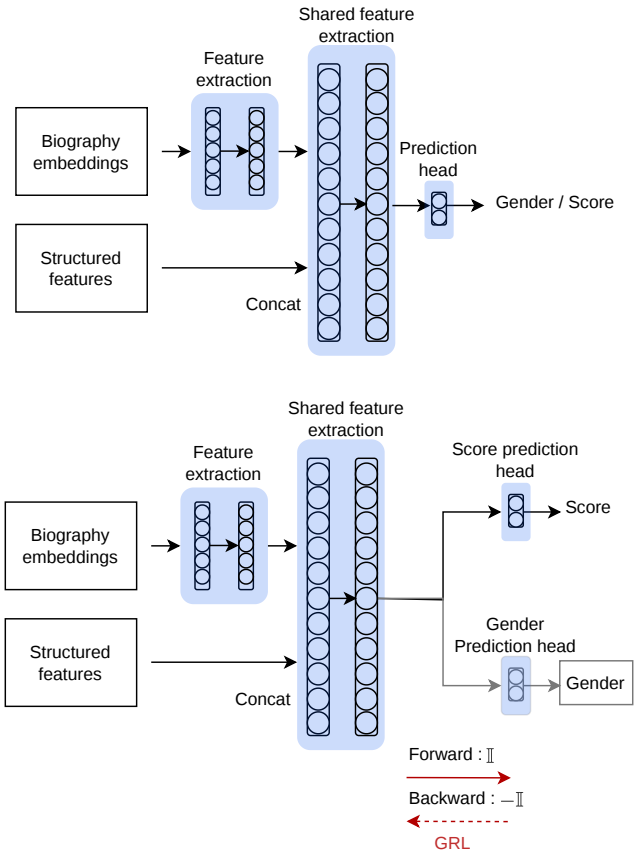

Figure 1: Architectures used for single-task and multi-task experiments. Both use the same mid-fusion strategy: biography embeddings are processed by a feature extractor, concatenated with structured features, and passed through a shared feature extractor. (Top) Single-task architecture for gender classification and score prediction with a single prediction head. (Bottom) Multi-task adversarial architecture with a score prediction head and a gender prediction head connected through a gradient reversal layer (GRL), which acts as the identity in the forward pass and reverses the gender-head gradient during backpropagation.

Recall@100 by gender. Using top-100 candidates from unbiased ground-truth scores T^U, we measure the fraction of these candidates that are recovered in the model's predicted top-100 list. At top 100 candidates, we report Male Recall, Female Recall, Average Recall, and Recall Gap, where Recall Gap is the absolute difference between male and female recall. Although T^U would generally not be available in a real-world deployment, it is available in the synthetic FairCVdb dataset. This allows us to evaluate how well each model recovers truly qualified candidates across gender groups. A high Average Recall together with a small Recall Gap indicates a ranking model that is both accurate and more balanced across gender groups.

2.7. Fairness-Aware Multi-Objective Optimization

For both the single-task and multi-task settings, we use a fairness-aware validation objective during hyperparameter tuning:

$$\mathcal{J}_\alpha = \mathrm{MAE}_{\mathrm{val}} + \alpha D_{\mathrm{KL,val}}, \quad \alpha \in \mathcal{A}, \tag{5}$$

where $\mathrm{MAE}_{\mathrm{val}}$ is the validation prediction error, $D_{\mathrm{KL,val}}$ is the KL divergence between male and female predicted score distributions on the validation set, and α controls the relative importance of distributional fairness during model selection. Rather than fixing a single α manually, we evaluate a grid of candidate values $\mathcal{A}$. For each α, model hyperparameters are tuned independently via Optuna

[27], producing one best validation configuration associated with a pair $(\mathrm{MAE}_{\mathrm{val}}^{(\alpha)}, D_{\mathrm{KL,val}}^{(\alpha)})$.

After all α values have been evaluated, we identify the Pareto front in the objective space of these pairs. A configuration is Pareto-efficient if no other configuration achieves both lower validation MAE and lower validation KL simultaneously, with at least one strictly improved. Because the MAE is computed against the gender-biased targets, it measures fidelity to biased labels rather than correctness with respect to any unbiased notion of suitability. The accuracy axis of the Pareto front should therefore not be read as predictive quality in an absolute sense: a configuration with a slightly higher MAE can be preferable, since some deviation from the biased labels is expected when gender-related information is suppressed. To ensure that fairness gains are not achieved at an unreasonable deviation from the biased labels, we apply an MAE tolerance constraint. Let $\mathrm{MAE}_{\mathrm{val}}^{(0)}$ denote the validation MAE of the selected configuration when $\alpha = 0$. We retain only Pareto-efficient configurations satisfying:

$$\mathrm{MAE}_{\mathrm{val}}^{(\alpha)} \leq (1+\delta)\,\mathrm{MAE}_{\mathrm{val}}^{(0)}. \tag{6}$$

where δ is the acceptable MAE deviation threshold. Among the remaining candidates, we select the configuration with the lowest validation KL:

$$\alpha^* = \arg\min_{\alpha \in \mathcal{P}_\delta} D_{\mathrm{KL,val}}^{(\alpha)}, \tag{7}$$

where $\mathcal{P}_\delta$ denotes the set of Pareto-efficient configurations satisfying the MAE tolerance constraint. The final model is then retrained on the full training set using the hyperparameters corresponding to α^* and evaluated once on the held-out test set, ensuring that model selection reflects an explicit and principled accuracy–fairness trade-off rather than an arbitrary choice of fairness weight.

3. Experimental Setup

All experiments are conducted on the official FairCVdb train–test split. All models are trained on multi-modal inputs consisting of standardised structured features and biography embeddings (original or scrubbed). Biography embeddings are pre-computed and cached before training.

For all experiments, we use a mid-fusion neural network architecture to combine structured CV features and biography embeddings. The biography embedding is first processed through two dense layers with ReLU activations. The resulting biography representation is concatenated with the standardized structured features and passed through two dense shared layers to obtain a fused representation. For single-task gender classification, this representation is connected to a binary classification output and trained with a BCE loss. For single-task score prediction, it is connected to a sigmoid regression output and trained to predict the gender-biased suitability score T^G using MAE loss. For multi-task adversarial learning, the fused representation is shared by two task-specific heads: a regression head for predicting T^G and a gender-classification head connected through a gradient reversal layer. The multi-task model is trained with the weighted combination of MAE and BCE losses in Eq. 4.

Hyperparameters are optimized using Optuna [27] with the TPE sampler. For score-prediction experiments, model selection follows the Pareto-front-based procedure described in Section 2.7. For each value of α, hyperparameters are optimized independently over 50 trials. We tune learning rate, dropout, batch size, number of epochs, and hidden-layer sizes; for the multi-task model, we additionally tune the adversarial strength λ. We search over a grid of fairness weights $\alpha \in \mathcal{A}$, with values ranging from 0 to 5 across all experiments. The acceptable MAE degradation threshold is set to $\delta = 0.3$ for original biographies and $\delta = 0.1$ for scrubbed biographies. A larger tolerance is used for original biographies because they contain stronger gender cues and therefore require a wider accuracy–fairness trade-off region, while scrubbed biographies are already partially debiased and use a stricter tolerance.

4. Results and Discussion

Table 1 shows that gender is highly predictable from original biographies across all embedding models, as expected. After gender scrubbing, accuracy drops substantially to 0.745–0.802. However, per-

Table 1
Gender prediction results across embedding models using original and scrubbed biographies.

	Original biographies			Scrubbed biographies		
Embedding model	**Accuracy**	**M Recall**	**F Recall**	**Accuracy**	**M Recall**	**F Recall**
fastText	0.985	0.986	0.983	0.745	0.799	0.689
GloVe	0.976	0.974	0.979	0.757	0.727	0.788
bert-base-uncased	0.990	0.995	0.983	0.774	0.778	0.769
FacebookAI/roberta-base	0.969	0.976	0.961	0.768	0.771	0.765
microsoft/mpnet-base	0.993	0.992	0.993	0.770	0.810	0.729
all-mpnet-base-v2	0.977	0.977	0.977	0.754	0.738	0.770
all-MiniLM-L6-v2	0.984	0.986	0.981	0.749	0.718	0.781
openai-text-embedding-3-large	0.995	0.996	0.994	0.802	0.830	0.774
openai-text-embedding-3-small	0.992	0.995	0.988	0.766	0.766	0.766

formance remains clearly above random guessing, indicating that scrubbed biographies still contain implicit gender cues that can be captured by the embeddings. Among the scrubbed results, `openai-text-embedding-3-large` gives the highest gender prediction accuracy, possibly due to its rich, high-dimensional semantic embeddings. In addition, male and female recall values are not perfectly balanced across embedding models. Some models recover male examples better, while others recover female examples better. This suggests that the remaining gender cues are not captured uniformly across embedding models. Since the imbalance is not consistently biased toward one gender group, it may reflect that the embedding models themselves encode gender-related associations differently, potentially producing biased representations after explicit gender indicators are removed. This is consistent with prior work showing that static word embeddings can exhibit different levels of social and gender bias depending on the model, training setting, bias dimension, and evaluation metric [28, 29].

Comparing the results of single-task score prediction with multi-task adversarial learning on original biographies in Table 2 and Table 3, MAE decreases for most embedding models in the multi-task setting; however, since MAE is computed against the gender-biased scores T^G, a lower MAE does not necessarily indicate a higher unbiased prediction performance. We observe a consistent improvement in fairness-related metrics. For all embedding models, KL divergence decreases and p-percent increases, indicating that the predicted score distributions become more balanced across gender groups. The ranking-based metrics also improve: Average Recall@100 increases for all embeddings, and Female Recall@100 improves consistently. Although Male Recall@100 decreases slightly for some models, it remains relatively high across all embeddings. Most importantly, Recall Gap@100 as another fairness metric, decreases for every embedding model, showing that adversarial learning reduces the disparity between male and female recovery rates among the truly qualified candidates. Among the embedding models, `FacebookAI/roberta-base` provides the strongest overall trade-off in the multi-task setting with original biographies. It is among the top-performing models for four main metrics: KL divergence, p-percent, Average Recall@100, and Recall Gap@100 offering a particularly effective balance between predictive utility and fairness in this setting.

For scrubbed biographies, the results in Table 4 and Table 5 show that the effect of multi-task adversarial learning is less uniform than in the original-biography setting. In terms of prediction performance, MAE decreases and Average Recall@100 increases for almost all embedding models. However, the fairness metrics do not improve consistently: for some embeddings the single-task model achieves lower KL divergence or Recall Gap@100, while for others the multi-task model performs better. Therefore, we cannot conclude that adversarial learning always improves fairness when the input already contains only implicit gender cues. Instead, its effect appears to depend strongly on the embedding model, which is consistent with the single-task gender classification results in Table 1 showing that different embedding models retain and encode residual gender-related information differently after scrubbing, often dissimilarly biased across gender groups. Among scrubbed biographies, `microsoft/mpnet-base` in the multi-task adversarial setting provides the strongest overall balance between predictive utility and fairness, achieving competitive MAE alongside top-three performance

Table 2
Single-task score prediction results on the test set using original biographies. **Bold** values indicate the top-three results per main metric.

Embedding model	α	MAE	KL	P%	Recall@100			
					Male	Female	Avg	Gap
fastText	0.20	0.064	0.249	0.266	0.980	0.412	**0.696**	0.568
GloVe	0.10	**0.029**	0.248	0.235	0.980	0.353	0.667	0.627
bert-base-uncased	0.20	**0.039**	**0.217**	0.266	0.918	0.392	0.655	**0.526**
FacebookAI/roberta-base	0.30	0.058	**0.178**	**0.299**	0.898	0.353	0.625	0.545
microsoft/mpnet-base	0.20	**0.044**	0.273	0.250	0.918	0.373	0.646	0.545
all-mpnet-base-v2	0.50	0.077	0.235	**0.299**	0.898	0.412	0.655	**0.486**
all-MiniLM-L6-v2	0.20	0.064	**0.163**	**0.316**	0.918	0.392	0.655	**0.526**
openai-text-embedding-3-large	0.30	0.074	0.221	**0.333**	0.959	0.470	**0.714**	**0.488**
openai-text-embedding-3-small	0.50	0.062	0.261	0.282	0.939	0.412	**0.675**	0.527

Table 3
Multi-task adversarial learning results on the test set with Pareto-front selection using original biographies. **Bold** values indicate the top-three results per main metric.

Embedding model	α	MAE	KL	P%	Recall@100			
					Male	Female	Avg	Gap
fastText	0.50	0.047	**0.042**	0.562	0.939	0.608	0.773	0.331
GloVe	0.30	0.042	0.082	**0.695**	0.857	0.667	0.762	**0.190**
bert-base-uncased	0.10	**0.034**	0.131	0.351	0.939	0.490	0.714	0.486
FacebookAI/roberta-base	0.20	0.047	**0.034**	**0.667**	0.898	0.686	**0.792**	**0.212**
microsoft/mpnet-base	0.10	0.036	0.100	0.587	0.898	0.647	0.773	**0.251**
all-mpnet-base-v2	0.10	0.038	0.084	0.562	0.918	0.647	**0.783**	0.271
all-MiniLM-L6-v2	0.20	0.044	**0.055**	**0.613**	0.857	0.549	0.703	0.308
openai-text-embedding-3-large	0.10	**0.031**	0.137	0.449	0.959	0.569	0.764	0.391
openai-text-embedding-3-small	0.10	**0.030**	0.160	0.493	0.959	0.608	**0.784**	0.351

on KL divergence, P-percent, Average Recall@100, and Recall Gap@100.

Finally, comparing multi-task adversarial learning on original biographies in Table 3 with single-task score prediction on scrubbed biographies in Table 4 shows that preprocessing-stage gender scrubbing leads to stronger fairness improvements; while Average Recall@100 of all embedding models remains in a comparable range for both, the scrubbed single-task setting achieves lower KL divergence and higher p-percent than the multi-task adversarial model trained on original biographies. Recall Gap@100 is also lower for almost all embeddings. This suggests that although adversarial learning reduces the effect of explicit gender information to some extent, it does not suppress gender-related cues as effectively as directly removing them from the text.

5. Conclusion

In this paper, we evaluated nine pre-trained text embedding models for gender bias in ML-based recruitment scoring, comparing single-task and multi-task adversarial learning with gradient reversal on both original and gender-scrubbed biographies. Our results show that all embedding models encode substantial gender information, and that gender scrubbing alone does not fully eliminate implicit gender cues. Multi-task adversarial learning consistently improves fairness on original biographies, but its benefit on scrubbed biographies varies across embedding models, reflecting differences in how residual gender information is encoded after explicit indicator removal. The choice of embedding model plays an important role in both predictive utility and fairness, and practitioners should carefully select embedding models for fairness-critical recruitment applications. Overall, preprocessing-stage gender scrubbing leads to stronger and more consistent fairness improvements, suggesting that adversarial

Table 4
Single-task learning results on the test set with Pareto-front selection using scrubbed biographies. **Bold** values indicate the top-three results per main metric.

Embedding model	α	MAE	KL	P%	Recall@100			
					Male	Female	Avg	Gap
fastText	0.10	**0.051**	0.022	0.724	0.837	0.647	**0.742**	0.190
GloVe	2.00	0.053	**0.014**	**0.754**	0.816	0.608	0.712	0.208
bert-base-uncased	0.50	**0.050**	0.023	0.724	0.816	0.667	**0.741**	0.150
FacebookAI/roberta-base	1.00	0.058	0.023	0.724	0.796	0.667	0.731	**0.129**
microsoft/mpnet-base	1.00	0.054	**0.012**	0.695	0.816	0.667	**0.741**	0.150
all-mpnet-base-v2	1.00	0.052	0.028	0.724	0.796	0.667	0.731	**0.129**
all-MiniLM-L6-v2	0.50	**0.051**	**0.021**	0.667	0.837	0.627	0.732	0.209
openai-text-embedding-3-large	1.00	0.087	0.031	**0.852**	0.857	0.706	**0.782**	0.151
openai-text-embedding-3-small	2.00	0.082	0.025	**0.923**	0.776	0.686	0.731	**0.089**

Table 5
Multi-task adversarial learning results on the test set with Pareto-front selection using scrubbed biographies. **Bold** values indicate the top-three results per main metric.

Embedding model	α	MAE	KL	P%	Recall@100			
					Male	Female	Avg	Gap
fastText	1.00	0.051	**0.015**	0.724	0.837	0.667	0.752	0.170
GloVe	1.00	0.052	**0.017**	**0.818**	0.878	0.706	0.792	0.172
bert-base-uncased	1.00	**0.048**	0.021	0.754	0.918	0.686	**0.802**	0.232
FacebookAI/roberta-base	0.50	0.049	0.021	0.639	0.878	0.627	0.753	0.250
microsoft/mpnet-base	0.50	0.050	**0.015**	**0.786**	0.878	0.745	**0.811**	**0.132**
all-mpnet-base-v2	0.50	0.050	0.021	0.754	0.918	0.686	**0.802**	0.232
all-MiniLM-L6-v2	0.50	0.050	0.028	**0.887**	0.837	0.706	0.771	**0.131**
openai-text-embedding-3-large	0.10	**0.043**	0.048	**0.786**	0.837	0.725	0.781	**0.111**
openai-text-embedding-3-small	0.50	**0.047**	0.028	0.667	0.918	0.667	**0.793**	0.252

training is best treated as a complementary strategy rather than a substitute for text-level debiasing.

This study comes with a few limitations that are left for future work. First, our study relies exclusively on FairCVdb, a synthetic dataset. Real CVs are noisier and contain a wider, less predictable, range of gender proxies. While we believe the overall conclusions will hold, validation on real recruitment data remains necessary before operational use. Second, while we focus on binary gender bias, other types of bias, such as related to ethnicity, are also common in ML-based recruitment. The framework extends naturally to other protected attributes where intersectional groups could be addressed either through a multi-class adversary or through several parallel adversarial heads. Intersectional evaluation does, however, reduce per-group sample sizes, making top-100 ranking metrics noisier, and demographic-parity measures would need to be generalized beyond two groups.

Acknowledgments

We would like to thank the Flemish Government under the Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen programme for funding this research.

Declaration on Generative AI

The authors used GPT-5.5 to improve the language and readability of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] W. A. Albassam, The power of artificial intelligence in recruitment: An analytical review of current AI-based recruitment strategies, International Journal of Professional Business Review 8 (2023) 1–4.

[2] D. F. Mujtaba, N. R. Mahapatra, Ethical considerations in AI-based recruitment, in: ISTAS, IEEE, 2019, pp. 1–7. doi:10.1109/ISTAS48451.2019.8937920.

[3] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[4] P. P. Liang, C. Wu, L.-P. Morency, R. Salakhutdinov, Towards understanding and mitigating social biases in language models, in: ICML, PMLR, 2021, pp. 6565–6576.

[5] A. K. Veldanda, F. Grob, S. Thakur, H. Pearce, B. Tan, R. Karri, S. Garg, Investigating hiring bias in large language models, in: R0-foMo, 2023, pp. 1–20.

[6] E. Faliagka, K. Ramantas, A. Tsakalidis, G. Tzimas, Application of machine learning algorithms to an online recruitment system, in: Proc. ICIW, 2012, pp. 215–220.

[7] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke, Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes, in: Proc. of ACM on human-computer interaction, volume 3, 2019, pp. 1–30.

[8] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch, Demographic bias in biometrics: A survey on an emerging challenge, IEEE Transactions on Technology and Society 1 (2020) 89–103.

[9] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, Bias in bios: A case study of semantic representation bias in a high-stakes setting, in: Proc. of ACM FAccT, 2019, pp. 120–128.

[10] H. An, X. Liu, D. Zhang, Learning bias-reduced word embeddings using dictionary definitions, in: Findings of the ACL, 2022, pp. 1139–1152.

[11] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[12] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 296–299.

[13] A. Petkova, F. F. Lighvan, L. Houthuys, Gender and ethnicity bias mitigation in automatic recruitment, in: BNAIC/BeNeLearn, 2025.

[14] A. Peña, J. Fierrez, A. Morales, G. Mancera, M. Lopez-Duran, R. Tolosana, Addressing bias in LLMs: Strategies and application to fair AI-based recruitment, in: Proc. of AAAI/ACM Conference on AI, Ethics, and Society, volume 8, 2025, pp. 1976–1987.

[15] A. Peña, I. Serna, A. Morales, J. Fierrez, A. Ortega, A. Herrarte, M. Alcantara, J. Ortega-Garcia, Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment, SN Computer Science 4 (2023) 434.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, Domain-adversarial training of neural networks, Journal of machine learning research 17 (2016) 1–35.

[17] A. Peña, I. Serna, A. Morales, J. Fierrez, Bias in multimodal AI: Testbed for fair automatic recruitment, in: Proc. of IEEE/CVF on CVPR, 2020, pp. 28–29.

[18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proc. of LREC, 2018.

[19] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proce. of EMNLP, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. of NAACL on human language technologies, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[21] Z. Liu, W. Lin, Y. Shi, J. Zhao, A robustly optimized BERT pre-training approach with post-training, in: Proc. of CCL, Springer-Verlag, 2021, p. 471–484. doi:10.1007/978-3-030-84186-7_31.

[22] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: masked and permuted pre-training for language understanding, in: Proc. of NIPS, 2020, pp. 16857–16867.
[23] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Advances in neural information processing systems 33 (2020) 5776–5788.
[24] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proc. of EMNLP-IJCNLP, 2019, pp. 3982–3992.
[25] OpenAI, OpenAI embeddings documentation, https://developers.openai.com/api/docs/guides/embeddings, 2024. Accessed: 2026-05-15.
[26] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1951) 79–86. doi:10.1214/aoms/1177729694.
[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proc. of ACM SIGKDD, 2019, pp. 2623–2631.
[28] E. Sesari, M. Hort, F. Sarro, An empirical study on the fairness of pre-trained word embeddings, in: Proc. of GeBNLP, 2022, pp. 129–144.
[29] P. Badilla, F. Bravo-Marquez, J. Pérez, WEFE: The word embeddings fairness evaluation framework, IJCAI, 2020.