

BiasGuard Pro: Auditing and Mitigating Gendered Stereotypes in Career Recommendation Systems

Abstract

Career recommendation systems increasingly shape access to professional opportunities but often reinforce gendered stereotypes embedded in training data. BiasGuard Pro introduces a lightweight, explainable framework for detecting, explaining, and mitigating such bias in career recommendation text. It combines GPT-4-generated synthetic data with real-world corpora (BiasBios, StereoSet) to fine-tune a DistilBERT classifier and integrates SHAP, counterfactuals, and prototype examples for interpretability within a human-centered dashboard. The system achieves an F1-score of 0.828 and 12 ms inference latency, significantly outperforming classical baselines. Statistical tests confirm significance (McNemar $p < 0.001$), with a low disparity gap (0.041) indicating reduced gender asymmetry. BiasGuard Pro exemplifies responsible AI design that embeds fairness as a continuous, participatory process.

Key Results

- * F1-score: 0.828 on BiasBios + Synthetic LinkedIn data
- * Accuracy: 0.923; ROC-AUC: 0.964
- * Latency: 12 ms per prediction (CPU)
- * Disparity Gap: 0.041 (vs. 0.087 for RoBERTa)
- * SHAP + Counterfactual + Prototype explanations integrated in UI
- * Open-source, reproducible pipeline with transparent ethics documentation

Contributions

1. Developed an explainable bias detection model using DistilBERT optimized for fairness and efficiency.
2. Integrated multi-modal interpretability (SHAP, counterfactuals, prototypes) for transparent decision support.
3. Created a hybrid dataset (synthetic + BiasBios) with balanced gender representation.
4. Designed a Gradio dashboard following HCI principles such as progressive disclosure and participatory design.
5. Released a reproducible open-source repository fostering transparency and responsible AI use.

Future Work

- * Extend to multilingual and cross-cultural corpora.
- * Incorporate intersectional fairness (race, age, ethnicity).
- * Integrate fairness-aware loss functions and adaptive learning.
- * Collaborate with HR systems for IRB-approved real-world validation.
- * Evolve into a continuous fairness auditing platform for responsible NLP.

Keywords

algorithmic fairness; gender bias; explainable AI (XAI); human-centered AI (HCAI); NLP; career recommendation systems; SHAP; counterfactual explanations; fairness-aware machine learning; ethical AI