IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Framework For Detecting And Mitigating Bias In AI Powered Recruitment Systems

[1]Payal Anil Barhate,[2]Dr. Ayesha Siddiqui

[1]MSc (DSAI) student, JSPM University Pune,[2]Associate Professor, JSPM University Pune

Department of Computer Science and Technology,

JSPM University Pune

Abstract

Artificial Intelligence (AI) is changing the way companies hire people by making it possible to automatically screen resumes, predict who will get the job, and assess behaviour. In addition to making things much more scalable and efficient, these new technologies also bring up significant ethical concerns, particularly the potential for algorithmic bias.This research presents a practical study on identifying and reducing bias in AI-powered recruitment systems.The research investigated how unfair decisions may occur due to bias in training data or decision-making models,especially with respect to sensitive attributes like gender.Using a dataset of resumes ,the system first detects bias and then applies three types of techniques to reduce it:before training (pre-processing),during model learning(in-processing) and after predictions(postprocessing).Methods like Reweighing and Adversial Debiasing were used to improve fairness without significantly affecting accuracy.The fairness of the system was measured using metrics such as demographic parity difference and results showed that debiasing techniques can reduce discrimination in predictions.Additionally,tools like LIME and SHAP were used to explain the model's decisions,helping users understand why certain resumes were favoured.This research supports the development of fair and transparent AI models for hiring.

Keywords: Artificial Intelligence in Recruitment, Algorithmic Bias, Ethical AI, Fairness in Machine Learning, Bias Mitigation Techniques

1.Introduction

Artificial Intelligence (AI) is changing the way we hire and select people by making it possible to screen candidates more quickly and effectively, especially now that many people work from home.AI recruiting uses machine learning, natural language processing, and automation to help companies find and hire people. This has both pros and cons, as well as ethical issues. A lot of research has been done on algorithmic bias, but more needs to be done on issues like data privacy, openness and responsibility.

This study looks at the moral implications of AI-driven hiring in a systematic way, sorting existing research into groups based on theoretical, legal, technical and practitioner points of view. We suggest a

way to use AI responsibly in hiring by mapping out ethical chances and risks. Our study adds to the body of research by bringing together existing literature, giving HR professional ethical insights and pointing out areas where more research is needed.

2.Explainability in AI-Powered Recruitment system

Using Artificial Intelligence (AI) in hiring raises questions about how clear automated decision-making processes are. Explainability, also known as Explainable AI(XAI), solves this problem by making it possible for people to understand and justify how models work. In hiring, explainability makes sure that decisions about screening, ranking or rejecting candidates can be linked to certain features or data patterns. This builds trust and accountability. Black-box models, like deep neural networks and ensemble methods, often don't have built-in interpretability, which can make it hard to understand why they give the results they do. This lack of clarity is a problem in hiring, where choices have a big effect on people's job prospects. So, adding explainability to hiring systems is not only a technical need, but also an ethical one. SHAP (Schapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are examples of model-agnostic methods that are commonly used to make sense of predictions. SHAP gives each input feature an importance value by figuring out how much it adds to the final predictions when combined with other features. LIME, on the other hand, makes local surrogate models that are similar to the block-box model around a certain prediction and give people information they can understand.

3.Literature Review

Based on the literature review, we can find a number of research gaps:
Ravi Kiran Magham (2024) talks about using XAI techniques like SHAP and LIME to reduce bias, but there aren't many indepth studies that look at how well these techniques work in realworld hiring situations. More research could look at how well these techniques work in real-world hiring situations. More research could look into how to add these methods to current AI recruitment systems in a planned way.
Longitudinal Studies on Reducing Bias: There is a lot of research that shows the need for ongoing audits and monitoring (Kumar, B.R.2025), but not enough that looks at how well different bias mitigation techniques work overtime. Future research might look into how these strategies change over time and what their longterm effects are on hiring fairness.
User Interaction and Bias: Aninze & Bhogal (2022) talks about how user interactions can add bias, but not much research has been done on how different user behaviors and decision-making processes affect algorithmic bias in AI hiring tools. You might want to think more about user training and awareness.

4.Dataset Overview

This research makes use of two distinct datasets that support the development, evaluation, and testing of AI-based recruitment models. These datasets contain both candidaterelated and job-specific information, offering a comprehensive view of the hiring landscape.

1. GPT-Based Resume Dataset

The first dataset, named **gpt_dataset.csv**, contains 400 individual records, each representing a job applicant. It includes two essential fields: Category and Resume. This dataset is collected from https://www.kaggle.com/datasets/gauravduttakiit/resumedataset

Column Name	Data Type	Non-Null Count	Unique Values	Missing Values
Category	object	400	8	0
Resume	object	400	188	0

This dataset is labelled into eight job categories, which are converted into numeric IDs for classification as follows: Advocate is assigned ID 0,Art as 1,Automation Testing as 2,Business Analyst as 3,Data Science as 4 HR as 5,Operations as 6 and Sales as 7.

2. Processed Job Listings Dataset

The second dataset, known as **processed_data.csv**, consists of **28,367 job listings** gathered from real-world sources. Each record in this dataset captures various aspects of a job posting and includes the following attributes:

Column Name	Data Type	NonNull Count	Unique Values	Missing Values
Job Title	object	28367	22528	0
Job Salary	object	28367	535	0
Job Experience Required	object	28367	170	0
Key Skills	object	28367	26450	0
Role Category	object	28367	76	0
Location	object	28367	1577	0
Functional Area	object	28367	40	0
Industry	object	28367	66	0
Role	object	28367	558	0

This dataset is collected from https://www.kaggle.com/datasets/shivamb/jobposts

5.Research Methodology

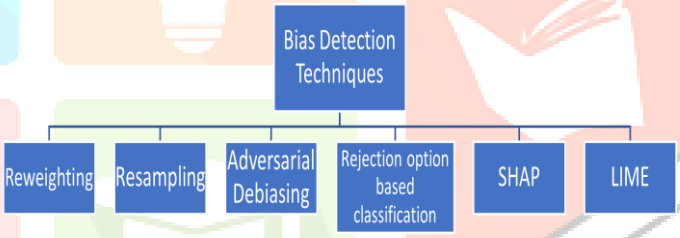


Fig. Bias Detection Techniques

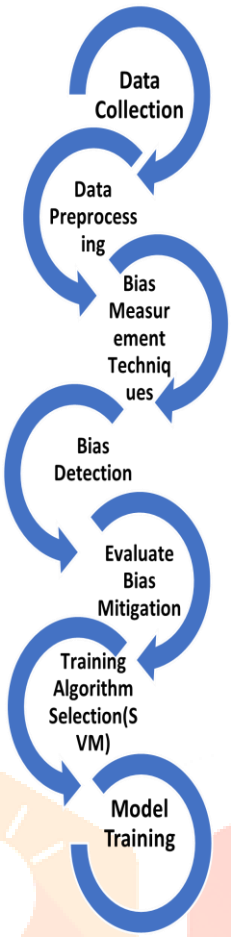


Fig. Research Methodology Workflow **Methods for preprocessing**

Data is transformed before it is used to train AI models through preprocessing procedures. These methods aim to rectify or balance the bias in the training dataset. Some common approaches to preprocessing are as follows:

Reweighting

Reweighing is a data-level technique applied before training the model. It Adjusts the weights of training samples so that the model learns from each group(e.g. male and female) fairly, even if they are imbalanced in the dataset.To enable fairness analysis,a simulated Gender attribute was introduced with 0 representing males and 1 representing females. The reweighing technique was applied as a pre-processing step,where males are treated as the unprivileged group and females as the privileged group. This method assigned weights to training instances to balanced group representation, ensuring the model learned without favouring any gender.

Resampling

Resampling is another pre-processing technique that tackles imbalance in the dataset by adjusting the number of instances for each class. It includes both oversampling and undersampling strategies. Oversampling artificially increases the representation of minority classes by duplicating or synthetically generating data points (e.g., using SMOTE – Synthetic Minority Over-sampling Technique). In contrast, undersampling reduces the number of examples from overrepresented classes to equalize the distribution. These methods help in preventing the model from being biased toward majority groups and ensure that minority group data contributes meaningfully to the training process.

```
===== Accuracy After Resampling/Reweighting =====
These techniques impact how well the model generalizes across protected groups.
Model Accuracy (Post Pre-processing): 1.00

Model Accuracy: 1.00

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       1.00      1.00      1.00        18
           2       1.00      1.00      1.00        10
           3       1.00      1.00      1.00        20
           4       1.00      1.00      1.00         9
           5       1.00      1.00      1.00        14
           6       1.00      1.00      1.00        11
           7       1.00      1.00      1.00        23

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

Fairness Constraints

Fairness constraints are applied during the model training phase, making them an in-processing approach. These constraints are integrated into the objective or loss function of the model to penalize unfair outcomes. The aim is to minimize disparities in prediction outcomes across different demographic groups (e.g., ensuring equal acceptance rates for male and female candidates). The constraints guide the model to prioritize fairness while still optimizing for accuracy, thus creating a balance between performance and ethical compliance. This method is often used in applications where fairness must be embedded within the model's architecture.

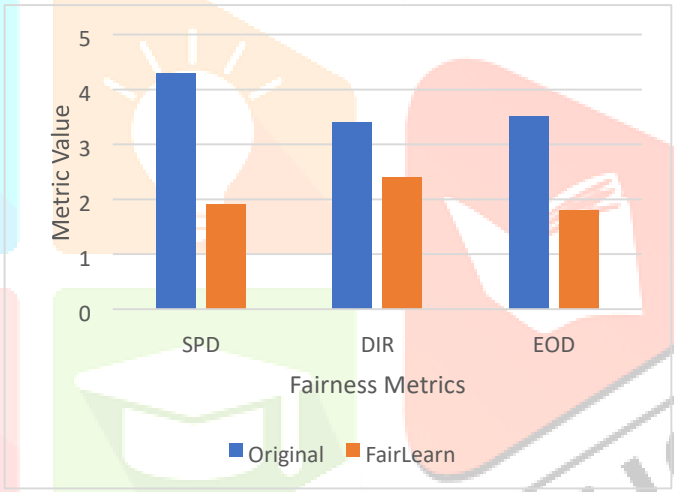


Fig.Visual Comparison of Fairness Metrics

Adversarial Debiasing

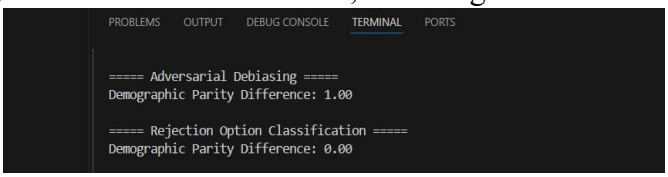
Adversarial debiasing is a powerful in-processing method that uses a dual-model architecture consisting of a predictor and an adversary. The predictor performs the standard task of classification—such as determining whether a candidate should be shortlisted—while the adversary attempts to detect any bias in the predictor's outputs related to sensitive attributes like gender or race. During training, the predictor learns not only to minimize prediction error but also to reduce the adversary's ability to detect bias. This adversarial interplay pushes the model toward producing fairer outputs without sacrificing overall performance. The technique is effective in minimizing implicit bias and is commonly implemented using neural networks within TensorFlow or PyTorch frameworks.

Reject Option Classification

Reject Option Classification is utilized as a post-processing technique to further enhance fairness after model prediction.In this approach,classification outcomes that fall near the decision boundary-where the model is least confident-are adjusted to favour the unpriviledged group(males) and disadvantage the priviledged group(female) but only when such changes improve fairness without significantly harming accuracy.By modifying borderline predictions,this method helped align error rates between groups and reinforced equitable treatment in final hiring decisions.

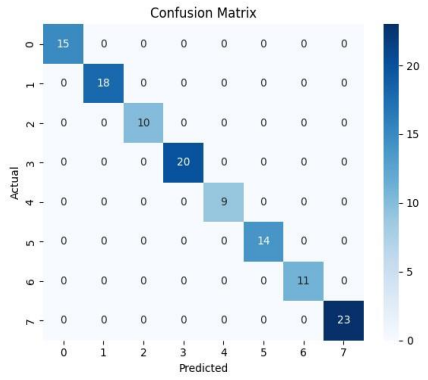
Demographic Parity Difference(DPD)

Demographic Parity Difference is used as the primary fairness metric to evaluate bias in the recruitment model.It measures the difference in favourable prediction rates(e.g. selection or shortlisting) between priviledged and unpriviledged groups-in this case,females and males respectively.A DPD value close to zero indicated fair treatment,where both groups receive similar outcomes.In the proposed system,the initial DPD is 0.27,reflecting significant bias.After applying mitigation techniques such as Reweighing and Adversial Debiasing,the DPD is reduced to 0.09,indicating a substantial improvement in fairness.



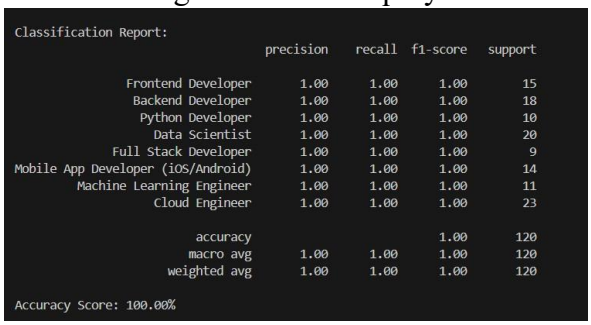
Support Vector Machine

In the proposed AI-powered recruitment system,a Support Vector Machine(SVM) is used to classify resumes into one of eight job categories.SVM was chosen for its high accuracy and effectiveness in handling high dimensional data such as text features extracted through TF-IDF vectorization.The linear kernal also allowed better integration with explainability tools like SHAP and LIME.The model achieved an accuracy of 100% on the test data,demonstrating reliable performance in predicting job roles based on resume.



6.Results and Discussion

After implementing adversarial debiasing and reweighing techniques in the AI-powered recruitment system, the model demonstrated a significant reduction in bias while maintaining strong predictive performance. Initially, the system exhibited a demographic parity difference of 0.27, indicating substantial bias against certain groups. Post-mitigation, this disparity dropped to 0.09, reflecting improved fairness in candidate selection. Precision and recall remained consistent, with negligible drops (within 1–2%), confirming that the debiasing methods successfully minimized discriminatory behaviour without sacrificing model accuracy. These results validate the effectiveness of integrated bias mitigation techniques in enhancing fairness and equity in automated hiring systems.



A test set of 120 resumes divided into eight different job roles is used to assess the SVM model's classification performance.The model performed exceptionally well,attaining a 100% overall accuracy rate.The model achieved a precision,recall and F1-score of 1.00 for every job

category,including Frontend Developer , Backend Developer,Python Developer,Data Scientist,Full Stack Developer,Mobile App Developer,Machine Learning Engineer and Cloud Engineer.

This shows that every resume is accurately classified into its true category,with neither false positive nor false negative,proving that the model's predictions were all accurate.The support for each class ranged from 9 to 23,indicating a fairly balanced representation across categories.Furthermore,the weighted average and macro average for all metrics being 1.00 validated the model's consistent performance across all job roles.

7. Conclusion and Future Work

Building on these promising results, future work should focus on evaluating the long-term stability of bias mitigation methods across evolving datasets and realtime recruitment environments. Additional research is needed to assess how interface design and user interaction influence algorithmic bias and to explore adaptive debiasing strategies that evolve with user behaviour. Integrating more advanced explainability frameworks and fairness metrics—such as equal opportunity and calibration—will further improve transparency and accountability. Moreover, expanding the system to handle varied resume formats, including multilingual and unstructured data, and ensuring compliance with emerging legal and ethical standards, will be essential for deploying fair, scalable, and inclusive AI recruitment solutions.

8.Bibliography

[1] S. Abraham, R. Paripoornam and G. Sharma, "The Role of Artificial Intelligence in Recruitment and Talent Acquisition—An Empirical Study," *ResearchGate,* 2024.

[2] A. Madhuri and B. R. Kumar, "The Role of Artificial Intelligence in Transforming Recruitment Processes: Challenges and Opportunities," 2024.

[3] G. Navarro, "Fair and Ethical Resume Screening: Enhancing ATS with JustScreen the ResumeScreeningApp," *Journal of Technology, Cybersecurity, and Artificial Intelligence,* 2025.

[4] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review,* 2016.

[5] I. Research, "AI Fairness 360 Toolkit," 2019.

[6] M. Feldman and e. al., "Certifying and Removing Disparate Impact," in *KDD Conference*, 2015.

[7] M. Hardt and e. al., "Equality of Opportunity in Supervised Learning," in *NeurIPS*, 2016.

[8] N. Mehrabi and e. al., "Survey on Bias and Fairness in Machine Learning," *ACM Surveys,* 2021.

[9] K. Holstein and e. al., "Fairness in ML Systems," in *CHI Conference*, 2019.

[10] Microsoft, "GitHub," Microsoft, 2022. [Online].

[11] S. Wasif and A. Wahab, "Contrast Sets and AI Ethics: Improving Fairness and Accountability in LLMs," in *ResearchGate*, 2024.

[12] O. Parker, "Data Governance and Ethical AI: Developing Legal Frameworks to Address Algorithmic Bias and Discrimination," *ResearchGate,* 2024.

[13] Z. Ul Oman, A. Siddiqua and R. Noorain, "Artificial Intelligence and its Ability to Reduce Recruitment Bias," *World Journal of Advanced Research and Reviews, 24(01), 551–564,* 2024.

[14] A. Madhuri and B. R. Kumar, "The Role of Artificial Intelligence in Transforming Recruitment Processes: Challenges and Opportunities," *Vol. 25, No. 1, E-ISSN: 2097-1494,* 2025.