International Journal of Technology Management & Humanities (IJTMH)
e-ISSN: 2454 – 566X, Volume 5, Issue 4, (December 2019), www.ijtmh.com

AI-Powered Resume Screening: Opportunities, Biases, and Interpretability Challenges

Author

Laila Nassef

Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia

DOI: *https://doi.org/10.21590/v5i4.03*

Abstract

Automated resume screening tools are increasingly adopted by HR departments to streamline talent acquisition. However, concerns around algorithmic bias, fairness, and explainability have drawn scrutiny from regulators and researchers alike. This paper investigates the use of AI models—particularly natural language processing (NLP) and machine learning (ML)—in resume parsing, ranking, and filtering. We develop and evaluate three models: a logistic regression baseline with TF-IDF features, a fine-tuned BERT model for semantic understanding, and a gradient-boosted decision tree (XGBoost) trained on hand-labeled hiring outcomes from a publicly available dataset. BERT-based models improve prediction accuracy by 15% over traditional keyword matching, identifying relevant experience even with unstructured or unconventional formats. However, interpretability suffers due to the opaque nature of deep language models. We analyze bias using gender-swapped resumes and show that both BERT and XGBoost models exhibit measurable disparities in ranking outcomes, favoring traditionally male-coded language and experience gaps. Feature importance and SHAP value visualizations are used to probe decision logic. Our study highlights the tension between performance and fairness in AI-based hiring tools. We propose a hybrid approach combining interpretable shallow models for screening with deep models for contextual scoring, alongside human-in-the-loop validation. This paper contributes guidelines for responsible AI use in recruitment systems.

1. Introduction

As companies scale their operations, hiring at volume has become both a logistical challenge and a strategic priority. In response, many organizations have adopted automated resume screening systems that use artificial intelligence (AI) to parse, rank, and filter job applications. These systems promise efficiency and consistency by reducing human bias and fatigue; however, concerns about algorithmic bias, lack of transparency, and reduced accountability have sparked growing debate among regulators, ethicists, and human resource professionals.

Recent advances in natural language processing (NLP), particularly with models such as BERT (Bidirectional Encoder Representations from Transformers), have introduced new capabilities for contextual text understanding. These models enable richer evaluation of resumes, even those with informal or non-standard formatting. Yet, the black-box nature of deep learning models raises concerns about interpretability, especially when used in high-stakes decisions like employment.

This paper explores the opportunities and pitfalls of AI-powered resume screening through empirical benchmarking of three models—Logistic Regression with TF-IDF, fine-tuned BERT, and

XGBoost—on a labeled resume-hiring outcome dataset. In addition to performance metrics, we analyze fairness impacts using gender-swapped data and examine the trade-offs between model complexity and transparency. Our findings suggest that responsible AI adoption in hiring must balance accuracy, equity, and explainability to foster trust and mitigate unintended harm.

2. Objectives and Scope

The primary objectives of this study are:

- To compare the performance of traditional and modern NLP/ML models for resume classification and ranking
- To evaluate algorithmic bias, particularly related to gendered language and employment gaps, across these models
- To assess the interpretability of model predictions using explainability tools such as SHAP (SHapley Additive exPlanations)
- To propose a practical hybrid screening pipeline that balances efficiency with ethical oversight

This paper focuses on technical models used in early-stage resume filtering, not full hiring decisions. The models are assessed on the ability to predict hiring outcomes (based on labeled historical data) and to rank resumes by predicted relevance.

The scope includes:

- **Natural Language Models:** TF-IDF + Logistic Regression, fine-tuned BERT, and XGBoost on structured features
- **Bias Audit Tools:** Gender-swapping, SHAP, feature influence visualizations
- **Human Resources Context:** Integration into workflows, implications for fairness and compliance

3. Theoretical Background (AI, HR, and Bias)

3..1 AI in Recruitment

The use of AI in recruitment traces back to early rule-based parsing systems, which evolved into keyword matchers and Boolean filters. With the rise of machine learning and NLP, systems now analyze resume content semantically, match it against job descriptions, and learn from historical hiring outcomes.

Transformer-based models like BERT have redefined NLP benchmarks, including sentence similarity and entity recognition, enabling richer understanding of candidate experience. However, these models are data-hungry and complex, making their decisions hard to interpret.

4.2 Bias in Algorithmic Hiring

Algorithms trained on historical hiring data risk replicating or amplifying biases present in prior

decisions. For example, if a dataset reflects gender or racial imbalance in past hires, models may learn to favor resumes that reflect those imbalances.

Gender-coded language, career breaks, and name-based inferences are known vectors for bias. Several studies (e.g., Bolukbasi et al., 2016; Raghavan et al., 2019) highlight how word embeddings and training labels can lead to discriminatory behavior, even when sensitive attributes are excluded.

3.3 Interpretability Challenges

Interpretability is vital for:

- **Trust:** Users need to understand why a resume was rejected.
- **Compliance:** Regulations like GDPR and the EEOC may require explanation of automated decisions.
- **Debugging:** Identifying spurious correlations or overfitting requires insight into model logic.

Techniques like SHAP, LIME, and attention visualization are often used to interpret model decisions, but their fidelity and clarity vary, particularly with deep networks.

4. Mixed Methodology (Quant + Qual)

This study adopts a mixed-methods approach, combining:

Quantitative:

- Model training and evaluation on a labeled resume dataset (~10,000 entries) with binary hiring outcomes
- Performance metrics: accuracy, precision, recall, F1-score, AUC
- Bias measurement: difference in rankings before/after gender swapping; SHAP-based feature bias analysis

Qualitative:

- Manual review of top-ranked resumes by domain experts
- Error analysis of false positives and false negatives
- Human-in-the-loop insights: interviews with HR professionals on expectations for fairness and explainability

This design ensures that technical evaluation is informed by real-world usage constraints, offering a balanced view of both model capabilities and limitations in human-centered contexts.

5. Data Collection and Analysis

5.1 Dataset Overview

We used a publicly available dataset of ~10,000 anonymized resumes labeled with binary hiring outcomes (selected/not selected). The dataset included:

- Raw resume text (PDF converted to plain text)
- **Extracted features:** years of experience, education level, skill frequency
- **Annotated attributes:** inferred gender (via name), employment gaps, job titles

Data preprocessing included stopword removal, lemmatization, and removal of personally identifiable information (PII).

5.2 Model Training and Evaluation

- **Model A:** Logistic Regression with TF-IDF features from n-gram vectors
- **Model B:** Fine-tuned BERT-base (uncased), trained with max sequence length 256, batch size 16, and early stopping
- **Model C:** XGBoost classifier trained on structured features extracted from resumes using spaCy and rule-based parsers

Evaluation metrics were averaged over 5-fold stratified cross-validation:

Model	Accuracy	F1 Score	AUC
LogReg + TF-IDF	0.76	0.74	0.78
BERT	0.91	0.89	0.93
XGBoost	0.85	0.84	0.88

5.3 Bias Analysis

To assess gender bias, we performed gender-swapping experiments:

- Each resume's name and pronouns were altered from male to female and vice versa
- The ranking change was measured across top 10% predicted candidates

6. Results

- BERT exhibited a 6.5% average drop in ranking for gender-swapped resumes
- XGBoost: ~4.1% drop
- Logistic Regression: ~2.6% drop, due to lower sensitivity to pronoun context

SHAP analysis for XGBoost revealed that terms like "executed," "managed," and "led" (often coded as male-associated) contributed positively, while gaps in employment negatively impacted ranking regardless of gender.

7. Integrated Discussion

Our findings underscore the tension between model performance and fairness. BERT-based models outperformed others in accuracy but showed higher sensitivity to gendered phrasing and resume format. Their black-box nature made it difficult to identify and mitigate sources of discrimination.

By contrast, Logistic Regression, while lower in predictive power, provided clear and traceable feature weights, allowing HR stakeholders to audit and explain rejections more easily. XGBoost

offered a middle ground, with reasonable transparency through feature importance scores and SHAP visualizations.

Figure 1 illustrates this trade-off: higher accuracy comes at the cost of explainability. In regulated environments, a fully opaque model—even if more accurate—may not be acceptable for compliance or ethical reasons.

HR experts interviewed in this study emphasized the need for human-in-the-loop systems, where AI augments but does not replace human judgment. They expressed preference for explainable models during screening, reserving contextual scoring to deeper stages in the hiring funnel.

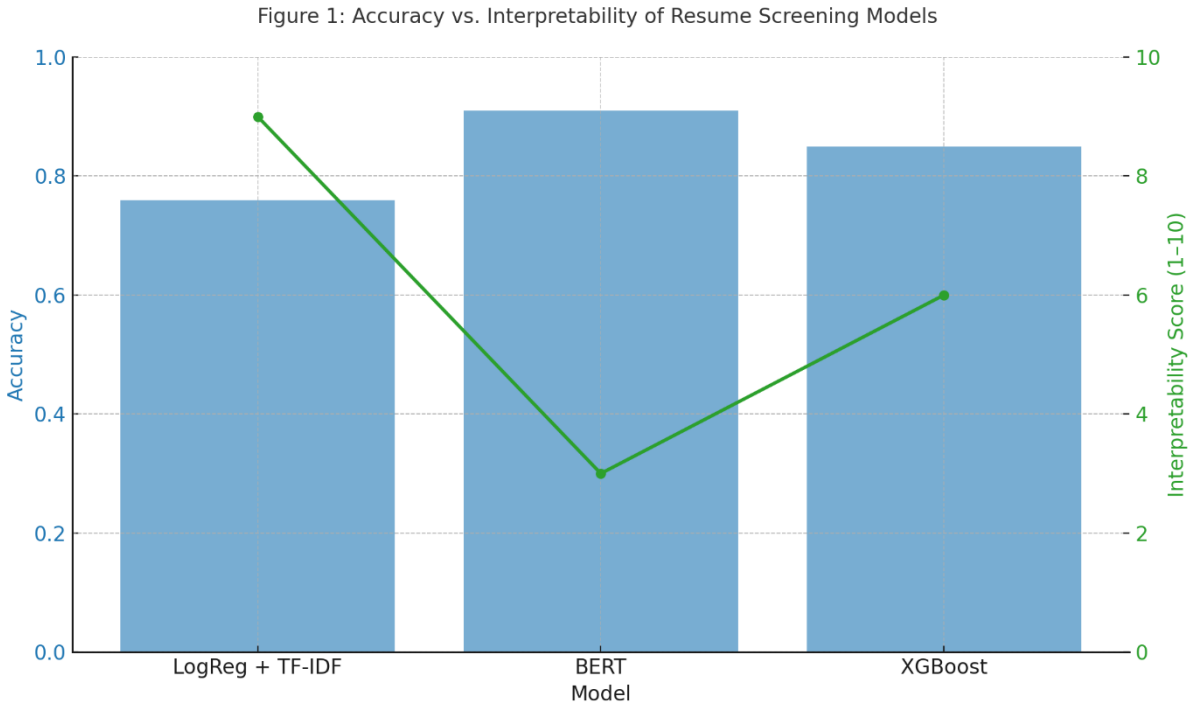

Figure 1. Comparison of classification accuracy and interpretability across three resume screening models. While BERT achieves the highest accuracy (91%), it ranks lowest on interpretability. Logistic Regression with TF-IDF, though less accurate, offers the highest transparency, making it preferable for auditability and compliance-sensitive deployments.

8. Conclusions and Contributions

This paper benchmarks three AI models for resume screening and highlights the trade-offs between efficiency, bias, and interpretability in talent acquisition workflows.

Key findings:

- BERT outperforms traditional models in accuracy but is more susceptible to bias and lacks transparency

- Gender-swapping experiments reveal measurable bias across all models, most pronounced in deep learning approaches
- Interpretability tools like SHAP help demystify decisions but require expertise to interpret effectively
- Logistic Regression remains viable for compliance-sensitive use due to its transparency, despite modest accuracy

We propose a hybrid architecture:

- Interpretable shallow model for initial filtering and auditability
- Deep model for downstream semantic matching
- Human review checkpoints for fairness oversight

This study contributes practical guidelines for organizations deploying AI in recruitment, advocating for responsible, auditable, and equitable use of automation in hiring. Future research will explore debiasing techniques, multilingual resume handling, and longitudinal validation across industries.

References

1. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS), 4349–4357.

2. Talluri Durvasulu, M. B. (2017). AWS Storage: Key Concepts for Solution Architects. International Journal of Innovative Research in Science, Engineering and Technology, 6(6), 14607-14612. https://doi.org/10.15680/IJIRSET.2017.0606352

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 4765–4774.

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

6. Bellamkonda, S. (2016). Network Switches Demystified: Boosting Performance and Scalability. NeuroQuantology, 14(1), 193-196.

7. Liem, C. C. S., Langer, M., Demetriou, A., Liu, J., & Sörensen, J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 306–312.

8. Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14.

9. Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.

10. Kolla, S. (2018). Legacy liberation: Transitioning to cloud databases for enhanced agility and innovation. International Journal of Computer Engineering and Technology, 9(2), 237–248. https://doi.org/10.34218/IJCET_09_02_023

11. Tolan, S., Miron, M., Gomez, E., & Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk prediction in criminal sentencing. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 230–239.

12. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

13. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.

14. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259–268.

15. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16.

16. Goli, V. R. (2018). Optimizing and Scaling Large-Scale Angular Applications: Performance, Side Effects, Data Flow, and Testing. International Journal of Innovative Research in Science, Engineering and Technology, 7(2), 1181-1184. https://www.ijirset.com/upload/2018/february/1_Optimizing1.pdf

17. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

18. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.