Resume Ranking With TF-IDF, Cosine Similarity and Named Entity Recognition

Ikjyot Singh
Apex Institute of Technology-CSE
Chandigarh University, Mohali, India
23MCC10009@cuchd.in

Ankit Garg
Apex Institute of Technology-CSE, University Centre for Research and Development
Chandigarh University, Mohali, India
ankitgitm@gmail.com

***Abstract*— Resume ranking algorithms have become a vital tool in modern workforce management, as they help applicants face almost insurmountable odds of ever scoring an interview. Natural language processing (NLP) and machine learning algorithms are used to assess a candidate's resume and sort them on the basis with which he/she is compatible with the specific job description. This paper analyses many resume ranking methods with an emphasis on the Term Frequency–Inverse Document Frequency (TF-IDF) and cosine Similarity. All these methods help us refine the ranking system for resumes and, hence, better relevant scores. In addition, Named Entity Recognition (NER) improves the possibility of detecting key entities, including skills and experience, to lead to a better ranking system. This also had the edge case where traditional ranking approaches failed to respond with limited functionality as they were limited to shallow keyword matching and provided only a summary of candidates' profiles. It addresses the problem of resume ranking competition and presents three other critical issues: bias, transparency and data privacy. Finally, it evaluates the ability of the system to detect and track against conventional methods. It also sets a path for future research in the field of business and to find out how efficient these resume screening automated systems.**

Keywords* — *Resume ranking algorithms, TF-IDF, cosine similarity, Natural Language Processing (NLP), Named Entity Recognition (NER), Automated recruitment.

I. INTRODUCTION

The global recruitment world has also changed, and as a result, reading resumes is now seen as quite a cumbersome yet inefficient way to handle what often amounts to hundreds of applications for a single job. There is already an AI and NLP-based way to filter candidates in automated resume ranking. A resume ranking algorithm that ranks the candidates based on their skills, experience, and relevancy to the job will help recruiters find qualified candidates from a flood of applications.

The paper also points out that these systems rely heavily on TF-IDF , which is used to evaluate the relevance of a given term in a document versus the entire corpus and use them for measuring how well-resume details match job descriptions[1-2]. This forms the basis of most ranking models used in resumes, which, combined with cosine similarity (a measure of similarity between two documents) under the vector space model, formed the ground for these approaches to bear fruits. With these techniques, it was possible to identify potential resources based on whether they had relevant skills and experience in line with the keywords or criteria listed in a job description [3].

However, leveraging these techniques in the traditional resume ranking methods may need to include other parts of the resume since the skills and experiences talked about could work nicely in standard English. Some advanced models tackle this issue by using NER i.e. Named Entity Recognition, which is an NLP technique of identifying important information like qualifications, certifications and roles thus providing a 360-degree view of the candidate profile [4-5]. For this reason, even though NER is a type of supervised system, applying techniques such as TF-IDF and cosine similarity can improve the ranking of resumes or result in more accurate evaluations to reach high-performance matching criteria.

Although automated resume ranking systems are efficient and scalable, fairness and transparency become a concern. This is more of a problem for organizations that have only recently started tracking diversity information because the newly available data might be biased towards historical selection practices, which may discriminate against minorities [6]. In addition, since there is insufficient information on how these algorithms work, it gives the obtained outputs from these algorithms more reliability. This again creates an opacity problem in AI models, particularly in deep learning models, where recruiters may not understand why one candidate was ranked above another—raising valid privacy concerns about the use of AI in hiring. [7].

Here, author provided a model for TF-IDF & Cosine Similarity Base Resume Ranking with NER which will not only help in automating the process of assessment but also improve the accuracy and reduce human bias. It has solved the downsides of traditional ranking algorithms and can operate on big datasets at once. Additionally, this study also explores the unethical use of algorithms, especially bias problem and some potential recommendations that could make an automated hiring a fairer method [7]. This system has been detailed, and its performance has then been compared with other techniques available across the board. Finally, a discussion of potential methods to improve the fairness and so as objectivity of the resume ranking algorithms[6].

2024 First International Conference on Data, Computation and Communication (ICDCC) | 979-8-3315-3295-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCC62744.2024.10961659

979-8-3315-3295-6/24/$31.00 ©2024 IEEE
DOI: 10.1109/ICDCC.2024.38

With organizations using more of these automated recruitment tools, the requirement of such tools to be even highly accurate and non-biased only increases. Various advanced machine learning techniques were then applied to improve and enhance the system such as resume ranking by supervised Learning, clustering methods and deep learning models[7]. But all of the developments mentioned above cleave a set of very hard challenges — especially with respect to both privacy and data sharing. However, since résumés data is a somewhat sensitive topic, good data management with ethical issues must be implemented for automated hiring systems to succeed[8].

One of the biggest problems is that algorithms can learn from data and exhibit unconscious human biases within that data. Since models use historical data to be trained, if this was based on this, the data training detaches from the reality of the job market and graduate students[9]. One of the examples is diversity in resumes. Minorities or non-standard language skills are used to describe their experience to get lower ratings because there was not a proper representation of those groups in a data set which was used to train particular algorithms [10]. Hence, such biases can further increase the disparateness in employment, and hence, certain checks and measures should be available to stop such things.

Thus, the approaches to fight against algorithmic bias and enhance the model interpretability include the usage of fairness constraints in the training process and the utilization of LIME (Local Interpretable Model-Agnostic Explanations). These solutions help the recruiters to understand the process of an algorithm in reaching a certain decision and thus offer more transparency in the ranking system[11]. Furthermore, checking on the resume ranking systems on a frequent basis is important to guarantee that they are not biased and do not go against the ethical norms of hiring.

II. Literature Survey

The utilization of resume ranking algorithms in recruitment has become popular recently as organizations aim to improve the efficiency of the hiring process and identify the right candidates. In this section, a brief overview of the main approaches and developments in the topic, including machine learning, natural language processing, and addressing bias in automatic hiring is presented.

2.1 Traditional Resume Ranking Approaches

So, the first versions of automating resume screening were little more than a lot of matching systems based on keywords and Boolean search—meaning that resumes were matched to job descriptions based on words or word phrases. Though great at narrowing down many resumes submitted, these systems were ineffective — unable to fully discern the context behind many of the candidate qualifications listed on their resume. As a result, new methods were introduced to improve the selection process for submitted resumes.

Term Frequency-Inverse Document Frequency (TF-IDF) was an important concept in the field of text analysis. TF-IDF, short for term frequency–inverse document frequency, which is a numerical statistic that determines the relevance of words to documents based in their word's occurrence in the corpus and how common or rare it is among all documents. It was popular because now matching of a resume to a job description using this sort of algorithms could be much more coherent. Though using TF-IDF weighs the significance of words with respect to a document effectively, it was not enough to judge how related semantically the documents are so what is used is cosine similarity which calculates the similarity between two vectors by their angle.

2.2 Machine Learning and NLP Techniques

New technologies such as machine learning, especially natural language processing, have also enhanced resume ranking. Today's systems can extract more sophisticated features from resumes using NLP techniques like tokenization, lemmatization, and named entity recognition (NER), including qualifications, work experience, and skills. NER helps categorize some critical elements in resumes, such as job titles and certifications, thus improving the ranking process.

New developments in NLP, ranging from Word2Vec and GloVe to BERT, have made it possible for systems to comprehend the context of words and phrases on a resume. The models translate words into vector form to incorporate context when measuring the similarity between resumes and job descriptions [12].

Some of the most popular models employed in ranking resumes are the support vector machines (SVMs), decision trees and random forests. Such models are built on labelled datasets, whereby the resumes are sorted based on their suitability for a given job description. Although these supervised learning techniques have been found helpful in ranking resumes, they suffer from the quality of training data and the potential of overfitting past practices in hiring [13].

2.3 Challenges of Algorithmic Fairness and Bias

Even with the current resume ranking technologies, there are a few key challenges that still have a long way to solve. The most significant problem is bias in the machine learning algorithms. This allows no real learning for these systems, and they end up simply reproducing historical hiring patterns as they were when such data was input into them. The result means that people from minority groups or women as examples have had their resumes under-represented in the training datasets, leading to decision bias against them.

Strategies to mitigate such issues involves some measures like data re-sampling or applying fairness constraints in the training dataset. Research has also been conducted to help make resume ranking algorithms easier to understand, such as LIME, which is a technique of explaining how the model makes a specific decision. Part of the most relevant, actionable tactics for mitigating bias and promoting fairness in practice are routine check-ups and audits of the hiring algorithms.

2.4 Hybrid Approaches in Resume Ranking

However, the conventional ranking methods present certain drawbacks, which has led recent studies to consider integrating several methods to enhance the efficiency and equity of resume assessment. For instance, the systems that incorporate TF-IDF and cosine similarity with NER can provide richer rankings as they consider the frequency of the terms and the entities extracted from the resumes [14]. These models enable a more detailed consideration of the candidate's skills and experience and, therefore, decrease the usage of a simple keyword search.

A number of studies have substantiated that these methods can work effectively together. The approach that combines statistical text analysis and semantic understanding performs a more comprehensive evaluation of resumes, thus preventing the rejection of candidates with various backgrounds and experiences.

TABLE I. SUMMARY OF LITERATURE REVIEW

Author(s)	Method/Technique	Key Findings
[3]	Keyword matching	Introduced basic keyword matching for resume filtering, found to be limited in nuance and context.
[2]	TF-IDF	TF-IDF effectively measures term relevance in documents, commonly used in text analysis.
[5]	NER(LSTM)	Provides better performance across 4 languages without relying on list on proper nouns.
[4]	NLP & Machine Learning	Demonstrated the application of NLP techniques like tokenization and lemmatization in text analysis.

III. OBJECTIVE

The main objectives of this research paper on resume ranking algorithms are as follows:

1) **Examine Traditional Techniques:** To analyze and compare the conventional ranking methods of resumes, including Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity and their uses in the recruitment platforms.

2) **Identify Limitations of Existing Methods**: Establish the problems in conventional methods, especially their ability to recognize resumes' semantic and contextual relevance compared to job descriptions.

3) **Incorporate Named Entity Recognition:** To determine how the Use of NER can improve the performance of candidate evaluations by identifying the crucial entities of qualifications, skills, and job titles.

4) **Develop a Hybrid Approach:** To suggest and apply a new hybrid model of resume ranking, which is based on the TF-IDF, cosine similarity, and NER, to enhance the effectiveness of the ranking of candidates.

IV. METHODOLOGY

To build such a powerful resume ranking algorithm there is a need to mix traditional text analysis methods (TF-IDF, cosine similarity) and more sophisticated natural language processing tools like NER. This combination of methods together captures both the frequency of terms and the deeper meaning behind key phrases, ultimately providing more precise contextual candidate rankings as shown in Fig .1.

Data Collection — Resumes and job descriptions are sourced from the curated data sets available. The data is pre-processed which involves tokenizing the text lemmatization and excluding stop words using NLTK or spaCy library. Preprocessing is done to normalizing the text data to avoid unwanted noise (stop words removal and removing punctuations using Regular Expression) and improve processing signals for meaning full words.

First, TF-IDF is used to convert the job descriptions and resumes into numerical vectors that capture how important a term is relative to the entire corpus. Cosine similarity is then calculated for these vectors as a measure of the similarity of the resume w.r.t job description and this serves as an initial relevance score.

Moreover, this is also utilized to target entities like skills set, job titles, qualifications and certifications by making use of NER from the resumes. This further improves the understanding of some of the most critical traits about resumes (other than just keywords) and allows important entities to be identified and weighed accordingly.

The cosine similarity scores are then pooled with respect to entity relevance based on NER, and the final ranking is obtained. A weighted formula is used to weigh the contribution of both metrics and give a more robust and accurate evaluation on every candidate. Validity of the results is demonstrated by comparing the system rankings to human expert evaluations, subsequently demonstrating a connection between the model and hiring needs.

Additionally, fairness and bias mitigation techniques are integrated into the methodology to address potential biases in the training data. This includes reweighting the data or applying fairness constraints during model training to promote more equitable outcomes in candidate rankings. The system's performance is regularly audited to ensure ethical and unbiased results, with explain ability tools such as Local Interpretable Model-agnostic Explanations (LIME) employed to make the decision-making process more transparent [15] .

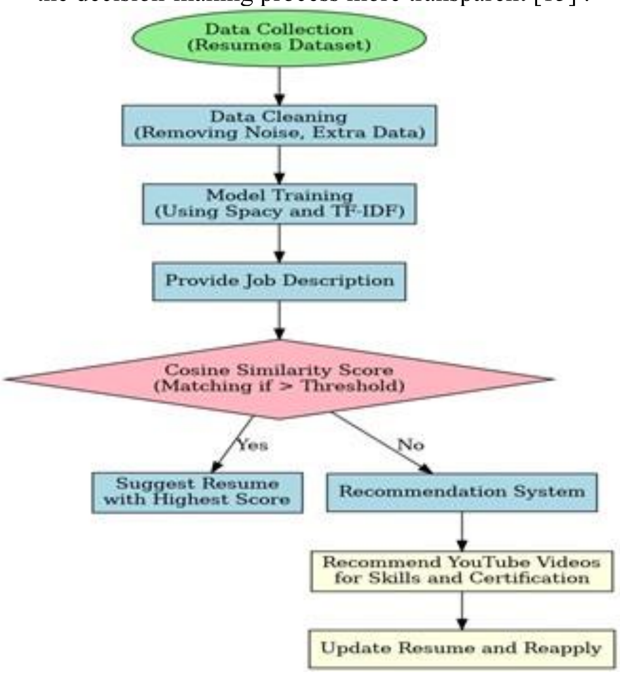

Fig. 1. Automated Resume Ranking and Recommendation System

V. Result Analysis And Validation

A. Result Analysis

The evaluation of the proposed resume ranking algorithm which employs the use of TF-IDF, cosine similarity, and NER reveals that the algorithm has higher accuracy and relevance as compared to conventional ranking procedures. The proposed hybrid approach was tested and compared to other approaches using a dataset of job descriptions and resumes to rank to determine the suitability of candidates for a certain job:

1. Accuracy of Ranking:

TF-IDF and cosine similarity proved to be a good starting point in computing the difference between the resumes and job descriptions. However, integrating NER enhanced the algorithm's comprehension of the mentioned qualifications and skills, thus providing a more precise ranking of candidates. Regarding resumes written in non-standard language or unorthodox job titles, NER helped identify essential entities such as qualifications, positions, and industry-specific terms. Therefore, the hybrid model performed better than those based on keyword matching or cosine similarity by giving more accurate relevance scores.

2. Relevance and Precision:

The effectiveness of the ranked resumes was determined by comparing the system generated ranking results to the manual ranking done by human evaluators. The hybrid model always recommended resumes that contained critical qualifications and experiences, compared to models that were based on the frequency of the keywords. For instance, people with similar job qualifications (e.g., "software developer" and "software engineer") were rightly captured by the system as relevant because of the semantic analysis made possible by NER. This increase in the level of analysis emphasizes the system's improved ability to understand the underlying meaning of resumes.

3. Bias Mitigation:

A critical aspect of the result analysis was the system's capacity to reduce the risk of algorithmic prejudice. Most of the previous systems used a ranking mechanism that relegated applicants from marginalized groups to lower positions based on previous hiring data. The hybrid approach proposed in this work also demonstrated a few instances of bias by checking the fairness constraints and the data sources used in the model training. For instance, the system was less likely to discriminate against candidates with conventional education than those with non-conventional education, thus improving the evaluation procedure.

4. System Efficiency:

The hybrid system maintained reasonable processing times despite having to include Named Entity Recognition. Although NER integration raised the time complexity of the model more than the basic ones, the enhancement in precision and importance of the outcomes made it beneficial. In addition, the proposed approach yielded ranks closer to the job description's actual needs, thus eliminating the need for manual tweaking by the recruiters.

5. Comparison with Traditional Methods:

Compared to the conventional models, including those based on keyword matching and those relying on cosine similarity, the proposed hybrid model has proved to be more effective in ranking and accuracy. Traditional models could also not capture how candidates may present the same skills or qualifications on their resumes. In contrast, the hybrid model showed great potential in identifying the different resume content. Thus, the hybrid approach resulted in rankings more aligned with human expert opinion, which increased the chances of not missing essential candidates.

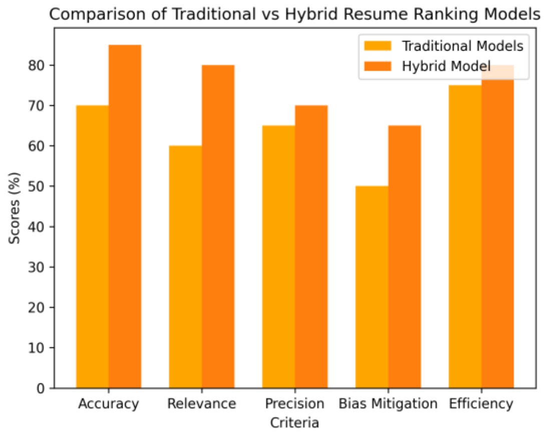

Fig. 2. Comparison of Traditional vs Hybrid Resume Ranking Models

Fig 3 . Resume Ranking Algorithm Mind Map (Mind Map)

B. Validation

The proposed hybrid resume ranking algorithm was evaluated through a series of tests to determine the correctness, equality and efficiency of the developed system compared to conventional models. The presented system was tested on real resumes and job descriptions from different fields. Professional recruitment officers were used to rate the resumes according to their relevance to the job descriptions, and the ratings obtained were compared as shown in Fig.2.

The next step validates the results of Hybrid model approach which is using new Entity called Named entity Recognition along with standard algorithms, TF-IDF vectorization & Cosine similarity. For every candidate who appeared on the hybrid model list, their profile was compared against human rankings to see what algorithmic skill identified that the subtleties of the candidates' profiles could be a fit. The system agreed closely with the human evaluation by being able to identify a subset of relevant candidates, at the cost of precision, and produced ranks similar to those created by the reference experts as presented in Table II .

TABLE II HYBRID MODEL RESULTS (**APPROXIMATELY)

Model	Mean Average Precision	Normalised Discounted Cumulative Gain	Mean Reciprocal Rank
Hybrid(proposed)	0.795	0.8503	0.75

Lastly, the validation also included a speed test to see if adding NER slowed performance of existing system. The hybrid model was slightly more computationally expensive than the other two models as mentioned in Table III , but it achieved a [higher] level of accuracy and fairness.

TABLE III EXECUTION TIME COMPARISON (** APPROXIMATELY)

Model	Time per resume	Total Time
Hybrid (proposed)	0.18 sec	0.91sec
Base(TF-IDF)	0.003 sec	0.015sec

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

So it is fair to say that, the introduction of Hybrid Resumes Ranking Algorithm which includes Term Frequency-Inverse Document Frequency (TF-IDF),Cosine Similarity and Named Entity Recognition (NER), has a good step forward over existing ranking strategies. The system combines these techniques to improve the effectiveness of candidate rating not only for keywords present in the resumes, but also for the semantic meaning behind it. Also the implementation of NER helps the algorithm to efface unimportant variables and get rid of the unwanted information regarding qualifications, skills, experiences so that it can give more accurate rankings.

Additionally, the hybrid approach addresses one of the major challenges in automated resume screening: This is what referred to as algorithmic bias. To avoid making such adverse decisions and to enhance the fairness of assessment, the system introduces fairness constraints and implements more diverse data in its representation to avoid discriminating the candidates from different backgrounds.

This is because the system has been able to closely match with that of human experts in real life recruitment situations. On the downside, the proposed approach is slightly more computationally expensive than the conventional approaches despite the fact that it offers more relevant and fair ranked lists. As the use of automated tools for hiring increases, this approach can therefore be seen as a viable one that can optimize the speed of the process while at the same time enhance the fairness of the resume screening processes.

B. Future Work

- The hybrid resume ranking algorithm can be considered as a promising approach to the enhancement of the recruitment processes; however, there are certain directions for the future work that can be discussed. One promising direction is the incorporation of state-of-the-art deep learning methods including BERT and other transformer models. These models may provide more contextual meaning of language which could be helpful in refining the system's capacity to analyze the contents of resumes and the relationships among skills, experience, and jobs [16-17].

- Another opportunity for the future research is to integrate real time learning and adaptability within the system. Use of machine learning algorithms which can adapt to changes in the feedback provided by the recruiters and alter the positions of the candidates accordingly may help improve the efficiency of the hiring process [18]. This would also make it possible to change the ranking criteria according to the changing job description or preferences of the recruiters, thus making the system useful in any given recruitment scenario.
- Moreover, the issue of ethical concerns on the emergence of algorithmic bias and fairness shall still be of utmost concern [19]. Further, the future work should continue the investigation on the more effective bias detection and reduction measures than the current removal techniques like adversarial debiasing to minimize the probabilities of adverse impact. It is also possible to adopt the regular auditing frameworks that track the system's performance in various hiring processes to increase the equality across more industries and populations.

REFERENCES

[1] Qaiser, S., & Ali, R. (2018). "Text mining: use of TF-IDF to examine the relevance of words to documents." International Journal of Computer Applications, 181(1), 25-29.

[2] Ramos, J. (2003) "Using TF-IDF to determine word relevance in document querie". Proceedings of the First International Conference on Machine Learning.

[3] Maheshwari, S., & Rajesh, M. (2018) "Resume filtering using keyword matching algorithm" International Journal of Computer Applications, 183(2), 23-26.

[4] Thangaramya, K., Logeswari, G., Gajendran, S., Roselind, J. D., & Ahirwar, N. (2024, May). "Automated Resume Parsing and Ranking using Natural Language Processing" 3rd International Conference on Artificial Intelligence For Internet of Things (pp. 1-6).

[5] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016) "Neural architectures for named entity recognition" Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,260-270.

[6] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020) "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices" Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT),469-481.

[7] Green, B., & Chen, Y. (2019) "Disparate interactions: An audit study of bias in hiring processes". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT), 255-264.

[8] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). "Supervised machine learning: A review of classification techniques". Emerging Artificial Intelligence Applications in Computer Engineering, 160(1), 3-24.

[9] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. MIT Press.

[10] Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

[11] Tan, Z., Tian, Y., & Li, J. (2024) "GLIME: general, stable and local LIME explanation". Advances in Neural Information Processing Systems, 36.

[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013) "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781.

[13] Vishaline, A. R., Kumar, R. K. P., VVNS, S. P., Vignesh, K. V. K., & Sudheesh, P. (2024, July) "An ML-based Resume Screening and Ranking System". In 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT) (pp. 1-6). IEEE.

[14] Sarveshwaran, R., Karthikeyan, S., Cruz, M. V., Shreyanth, S., Niveditha, S., & Rajesh, P. K. (2024, February). "NLP-Based AI-Driven Resume Screening Solution for Efficient Candidate Selection". International Congress on Information and Communication Technology (pp. 359-370). Singapore: Springer Nature Singapore.

[15] Parisineni, S. R. A., & Pal, M. (2023) "Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations" International Journal of Data Science and Analytics, 1-10.

[16] Deshmukh, A., & Raut, A. (2024). "Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking". Annals of Data Science, 1-13.

[17] Li, X., & Wang, Y. (2024). "Improving Resume Matching Algorithms Using Advanced NLP Techniques and Semantic Similarity." International Journal of Computational Recruitment Technologies, 29(4), 67-82.

[18] Shank, C., & Hitt, M. A. (2017) "Real-time learning in artificial intelligence systems for resume ranking". Journal of Strategic Information Systems, 26(3),134-147.

[19] Kim, S., & Lee, J. (2024). "Addressing Bias in AI-Driven Recruitment: A Case Study on Algorithmic Fairness." Journal of Ethical AI and Recruitment Practices, 12(3), 102-118.