Machine Learning Bias in Resume Screening

Ishita Samadhiya*

March 25, 2024

Abstract

Machine Learning (ML) has revolutionized hiring processes, introducing new dynamics and challenges, especially in the realm of resume screening. This study delves into the challenges and implications of ML-driven resume screening, specifically addressing gender bias. We utilized a dataset containing resumes from different sectors in India to train models, including Random Forests and Multilayer Perceptron Classifier. These models were employed to categorize resumes into sectors such as technology, business, etc. Our results reveal gender biases, with men more frequently predicted as executives and women in technical roles, reflecting historical disparities. Furthermore, we identify limitations and ethical concerns surrounding such classifiers, emphasizing the need for responsible AI deployment in recruitment processes. By shedding light on the complexities of bias in ML-based recruitment, this research contributes to the ongoing discourse on ethical AI deployment, offering insights and recommendations for fostering fairness and accountability in hiring-based automated decision-making processes.

1 Introduction

In recent years, machine learning has gained increasing prominence in the domain of resume screening, revolutionizing the hiring process in numerous fields and organizations [Woo23]. This shift towards automation in candidate evaluation not only enhances efficiency but also introduces new dynamics and considerations that warrant careful analysis.

The traditional approach to resume screening often involved human recruiters manually reviewing and shortlisting candidates. While this method has served its purpose, it has its limitations. It can be time-consuming, prone to bias, and less effective in handling the overwhelming volume of applications in the age of overpopulation and digitization. Moreover, manual hiring was often characterized by inherent biases and unfairness when compared to seemingly unbiased algorithms [CM13].

*Advised by: Angelina Wang of Princeton University

As a result, machine learning algorithms or artificial intelligence (AI) tools are being adopted to streamline the resume screening process. These systems, driven by their ability to analyze vast amounts of data and identify patterns, offer a promising solution to the challenges of manual hiring practices [NFC21]. With the advent of Natural Language Processing (NLP) and other deep learning techniques, these models can sift through resumes more efficiently and consistently than their human counterparts.

Machine learning-based resume screening systems typically involve the automatic extraction of key information from resumes and the assessment of candidates' suitability for specific roles. These systems consider factors such as skills, qualifications, experience, and other relevant criteria. The use of keyword matching and classification tools has further enhanced the capabilities of these systems, allowing for more precise candidate selection.

While the potential benefits are evident, the widespread adoption of machine learning in resume screening raises concerns about biases and the fair treatment of all candidates [Ale21]. Our approach to understanding bias in machine-learning based resume screening involved replicating current machine-learning practices. We extracted key features from a large dataset of resumes and employed two types of machine-learning approaches that resume screeners might use. After conducting these analyses, we identified clear instances of bias within the model's predictions, which raised concerns about the potential adverse impact on both women and men.

One of the most pressing concerns regarding the use of machine learning in resume screening is its potential for bias [YW18]. Biased algorithms may inadvertently favor or disfavor certain groups of candidates, leading to unequal opportunities and reinforcing historical disparities due to biased data. This bias can have profound implications, particularly for marginalized groups, such as women in STEM who may already face hurdles in the job market.

In this paper, we examine the critical issue of bias in machine learning-based resume screening systems. We explore how these systems can perpetuate or even exacerbate existing biases, with a specific focus on gender bias. By shedding light on how these biases manifest, we aim to contribute to the ongoing discourse on ethical and responsible AI in recruitment. Our analysis also includes recommendations for mitigating bias in resume screening to ensure fair and equitable hiring practices for all candidates.

2 Related Works

In this subsection, we provide an overview of research related to the broader field of hiring processes. We explore studies that discuss the evolution of bias in hiring practices, the importance of mitigating bias, and notable trends or changes in recruitment methods.

2.1 Gender Bias in Machine Learning

De-Arteaga et al. (2019) conducted a large-scale study of gender bias in occupation classification, a task with potential real-world consequences. They explored the impact of gender indicators in semantic representations of online biographies on occupation classification. The study revealed the potential allocation harms and proxy behavior related to gender bias, highlighting the challenges of debiasing machine learning models in practical applications [DARW$^+$19]. Gonen and Goldberg (2019) examined gender bias in word embeddings and critiqued existing debiasing methods. Their work focused on exposing the limitations of current techniques, suggesting that merely reducing bias is insufficient. They argued that true bias removal is challenging to achieve in word embeddings, and this bias can still be recovered from gender-neutralized embeddings [GG19]. Caliskan et al. (2017) delved into the ethical implications of applying machine learning to human language, highlighting the presence of human-like semantic biases in text corpora. Their research demonstrated that language models trained on textual data inherited and replicated various biases, including those related to race and gender. The study underscored the importance of identifying and addressing bias in culture and technology [CBN17].

Furthermore, "On the Dangers of Stochastic Parrots" emphasized the need for ac- countability and fairness in AI systems, urging a reevaluation of the continuous expansion of large language models. The paper recommended considering environmental and financial costs, dataset curation, stakeholder values, and research beyond bigger language models to mitigate the risks associated with this technology [BGMM21].

In the broader context of biases within machine learning, the work of Dietterich and Kong sheds light on the multifaceted nature of biases, encompassing statistical bias and variance. Their research explores the intricate relationships between machine learning bias, statistical bias, and the statistical variance of decision tree algorithms. While their primary focus is on general biases in learning algorithms, the principles they discuss provide valuable insights into the broader discussion on biases, including gender bias [DK95].

Turning specifically to gender bias in algorithms and embeddings, existing literature has examined biases encoded in language models, particularly word embeddings. Bolukbasi et al. (2016) have highlighted that word embeddings often encode gender stereotypes, leading to biased results in NLP tasks. Dietterich and Kong's exploration of statistical bias and variance contributes to the understanding of biases in machine learning algorithms, providing a broader perspective on the challenges and potential solutions in mitigating biases, including gender biases [BCZ$^+$16].

However, Leavy's work states that these studies overlook decades of research on the embedding of gender ideology in language. Her research underscores the significance of incorporating historical insights into approaches to machine learning from text. Her observations suggest that individuals who are potentially affected by bias are more inclined to recognize, comprehend, and endeavor to resolve it. Thus, achieving gender balance in the field of machine learning

is crucial to prevent algorithms from perpetuating gender ideologies that disadvantage women [Lea18].

On the other hand, Langenkamp's work critically assesses the legal and technological landscape of algorithmic hiring, proposing enhanced transparency as a solution. Advocating for algorithmic transparency reports, Langenkamp suggests that employers using automated hiring software should be mandated to publish detailed reports on their algorithms' operation and impact. Such transparency, he argues, would enable accountability and regulatory action, fostering fairer hiring practices [LCC20].

2.2 Bias in Hiring

Understanding the nuances of biased decision-making at the resume screening stage is critical for organizations leveraging machine learning in their hiring processes. By recognizing the broader implications of the model, organizations can take proactive measures to mitigate gender bias, fostering fair and equitable hiring practices. Khurana and Lee (2022) investigated gender bias in high-stakes pitching scenarios, specifically in the context of venture capital funding. They applied Natural Language Processing techniques to analyze the sentiment and reactions of judges during pitch evaluations. The study found differences in the reactions of the men and women judges and their impact on securing funding, shedding light on non-rational thinking and gender bias in early-stage investment [KL23]. Gender Bias impacts the ability of women entrepreneurs to get investors, showing the implications of bias even in non-corporate fields.

Another model introduced by [DR19], though centered on ethnic bias, offers valuable insights into gender bias in resume screening through machine learning. In situations where machine learning algorithms process resumes with limited job-related, personalized information and non-job-related stigmatizing cues, there exists a risk of biased decision-making against certain gender groups.

Another study highlights a new facet of bias against women in hiring: casual work. The prevalence of women in casual work puts them on the fringes during crucial times when they should be climbing mainstream career ladders. This observation aligns with the challenges faced by women in casual employment, potentially affecting their visibility and opportunities for career advancement [W$^+$13].

Addressing these challenges requires proactive government policies supported by financial incentives. For instance, deferring government pensions to the age of 67 and encouraging women to continue working can be strategies to retain women in the workforce. However, the paper also calls attention to the issue of casual workers lacking superannuation, which remains a significant challenge for public policy because of the prevalence of women in casual work [W$^+$13]. Therefore, not only does bias prevent women from getting hired, but it also pushes them into highly unstable work environments such as casual work.

3 Model Training

Here we will talk about the data, models, and model architectures used for the resume classification task.

3.1 Data

The dataset employed for resume classification, obtained from an open source dataset Kaggle [Dat18], is a diverse collection of Indian resumes spanning multiple sectors. It serves as the dataset for our classification task, encompassing a wide range of resume content to deeply analyze bias within the classification models. While the dataset also provided information such as companies worked at, college name, degree, years of experience, etc., we specifically chose to prioritize skills in predicting designations or sectors. This decision is rooted in the understanding that skills are often more indicative of a candidate's suitability for a particular role or sector than other factors. By focusing on skills and predicted designations, which reflect the real-world implications of these predictions, we aim to enhance the precision and relevance of our analysis. Predicted designations signify which sector an applicant did or did not get a job in, providing valuable insights into the practical outcomes of the classification process and the bias within it.

3.1.1 Data Extraction

For our classification task, relevant information including "skills" and "designations" was extracted from each resume sample. Each resume had various skills defining competencies or qualities that the candidate has, such as Python, collaboration skills, etc. Designations define which sector each candidate is currently employed in, such as tech, business, etc. These extracted features served as our input (X), and the associated designations were used as labels (y) for our classification models.

3.1.2 Data Partitioning

To assess the performance of our models, we partitioned the multi-sector resume dataset into training and testing subsets. Specifically, we allocated 80% of the data to the training set and the remaining 20% to the testing set. This division is crucial as it allows us to gauge how well our models generalize to unseen resumes.

Furthermore, within the training set, we employed an additional split. This time, we divided the data into two subsets: 80% for training the model and 20% for validation. This validation set aids in fine-tuning the model's hyperparameters, such as the depth of decision trees in the random forest classifier or the number of neurons in the hidden layers of the MLP classifier.

3.2 Model Architecture

We chose two different models to ensure we could deeply analyze bias in multiple algorithms and ensure the accuracy of our results.

3.2.1 Random Forest Classifier

In our resume classification task, we utilized a machine-learning model known as the random forest classifier for its versatility and efficacy. Despite its name, this model is not a literal forest but rather a collection of individual decision trees. Each tree independently makes predictions based on different subsets of the data. To ensure that these decision trees remain interpretable and avoid overfitting—where they become overly complex and capture noise rather than genuine patterns—we employed a parameter called "maximum depth." This parameter restricts the depth of each decision tree within the forest, effectively controlling the level of detail it can incorporate into its predictions. By imposing this constraint, we encourage the trees to focus on the most relevant features of the data, thereby enhancing the model's overall performance.

3.2.2 MLP Classifier

In our resume classification task, we also employed the MLP Classifier, which stands for Multi-Layer Perceptron Classifier. This is a type of neural network model. Neural networks are computational models inspired by the structure and function of the human brain. The MLP Classifier consists of multiple layers of artificial neurons, organized in a specific way. It has an input layer, which receives the data, two hidden layers where computations are performed, and an output layer, which produces the final prediction. Each neuron in the hidden layers processes information and passes it along to the next layer. Within each hidden layer, we included 50 neurons, which are essentially computational units that perform calculations on the input data. To allow these neurons to learn and make predictions effectively, we used an activation function called ReLU, which stands for Rectified Linear Unit. This function is a mathematical operation that helps the neural network understand and learn patterns within the data more efficiently. It works by turning negative values into zeros and leaving positive values unchanged, thereby introducing non-linearity into the model's calculations, which is crucial for learning complex relationships in the data.

3.3 Hyperparameter Selection

To fine-tune our models' hyperparameters and optimize their performance, we leveraged a validation set within our training data. This validation set, distinct from the testing data, served as a benchmark for evaluating different hyperparameter configurations. Through a combination of domain expertise and grid search methodology, we embarked on an extensive exploration of hyperparameter space. This involved systematically varying hyperparameters and evaluating model performance on the validation set. Our grid search approach

	param_alpha	param_hidden_layer_sizes	param_learning_rate_init	mean_test_score	rank_test_score
0	0.01	(20, 20)	0.0001	0.21875	81
1	0.01	(20, 20)	0.001	0.31250	48
2	0.01	(20, 20)	0.01	0.34375	31
3	0.01	(20, 20)	0.1	0.40625	8
4	0.01	(20, 20)	1	0.40625	8
...	...	...	...	...	...
91	10	(25, 25)	0.001	0.37500	24
92	10	(25, 25)	0.01	0.37500	24
93	10	(25, 25)	0.1	0.31250	48
94	10	(25, 25)	1	0.43750	5
95	10	(25, 25)	10	0.40625	8

96 rows × 5 columns

Figure 1: Example grid demonstrating hyperparameter values and corresponding performance scores on the validation set for the MLP CLassifier.

meticulously examined numerous hyperparameter combinations, ranking them based on their efficacy in classifying resumes across sectors. As shown in Figure 1, an example grid illustrates various hyperparameter values and their corresponding performance scores on the validation set. From this array of configurations, we identified commonalities among the top-performing ones. By analyzing these commonalities, we discerned the hyperparameters that consistently demonstrated optimal performance across various combinations. These refined hyperparameters, selected based on their consistent performance on the validation set, ensured that our models were finely tuned for effective resume classification.

3.3.1 Random Forest Classifier (Model: 'forest')

The choice of hyperparameters significantly influences model performance. For the Random Forest Classifier, we meticulously selected the following hyperparameters:

- **Random State**: Set to 0 for consistent results.
- **Max Depth** (defined in section 3.2): A depth of 9 was chosen to regulate the depth of individual decision trees within the forest.
- **Number of Estimators**: A total of 13 estimators were used to construct the forest. In a random forest classifier, an estimator typically refers to an individual decision tree within the ensemble, collectively used to make predictions by aggregating the outputs of each tree.

3.3.2 MLP Classifier (Model: 'model')

Hyperparameter selection for the MLP Classifier was equally meticulous, with the following parameters fine-tuned:

- **Random State**: Set to 0 for consistent results.
- **Alpha (L2 Regularization Term)**: An alpha value of 10 was selected to penalize large weights in the model to prevent overfitting during training.
- **Hidden Layer Sizes**: The neural network architecture was configured as (50, 50) for the two hidden layers. Hidden layer size refers to the number of neurons (or units) in each hidden layer. It determines the complexity and capacity of the neural network to learn from the data, with larger sizes allowing for more intricate representations but potentially increasing the risk of overfitting.
- **Initial Learning Rate**: An initial learning rate of 0.01 was employed to set the rate at which the model updates its parameters (weights and biases) during training using optimization algorithms like gradient descent. It influences the speed and convergence of the training process, with higher rates potentially causing faster initial progress but risking instability or overshooting optimal solutions.

4 Gender Classification on Names

To perform gender classification, we gathered a dataset of Indian names with associated gender labels. These labels were obtained from publicly available data sources, official records, or self-reported gender information. The dataset was carefully curated to ensure diversity and accuracy. We chose a dataset of Indian names to better predict the genders of the Indian applicants in our dataset.

4.1 Data Preprocessing

Names themselves served as the primary features for our gender classification task. To ensure uniformity, we processed the names by standardizing letter cases and removing special characters or spaces, facilitating consistency in the classification process. We also limited the names to first names to achieve a higher accuracy after trying full names and getting better results from first names.

4.2 Model Architecture

For the gender classification task based on names, we employed a neural network model with the following architecture:

- Embedding Layer: An embedding layer that maps input names to 128-dimensional embeddings.
- LSTM Layer: A Long Short-Term Memory (LSTM) layer with 128 input features and 256 hidden units. This layer captures sequential dependencies within names.

- Linear Layers: Two linear layers with 256 neurons each, separated by a ReLU activation function. The final linear layer produces output logits.

This architecture was chosen for its ability to capture meaningful patterns and relationships in names for gender classification. The model was trained using a suitable loss function and optimizer, as described earlier.

4.3 Adaptation for Indian Names

Given the dataset's inclusion of Indian names, we recognized the need to adapt our model to handle this subset effectively. To achieve this, we incorporated a separate dataset of Indian names, split into training (80%) and testing (20%) subsets. By training our model with this specific Indian name dataset, we aimed to enhance its accuracy when classifying Indian names. This approach yielded an impressive 93% accuracy rate on unseen or test data, specifically for Indian names within the dataset.

4.4 Benefits and Weaknesses of Using Such a Classifier

Benefits:

- Simplicity: Gender classification based on names is a straightforward and understandable approach.
- Efficiency: The classifier is computationally efficient and suitable for real-time or large-scale applications.
- Data Availability: Name data is widely accessible from various sources.
- Adaptation for Indian Names: Our model achieved high accuracy (93%) for categorizing Indian names within the dataset.

Weaknesses:

- Stereotyping: Name-based gender classification may reinforce stereotypes and lead to misclassification.
- Inaccuracy: The classifier may be affected by unisex names, cultural variations, and evolving naming trends.
- Bias and Fairness: The classifier can exhibit bias if the training data is imbalanced or biased.
- Binary Nature: Our model typically assumes a binary gender model, not accounting for non-binary or gender-diverse identities.

In conclusion, our gender classification approach based on names, while simple and efficient, has limitations related to stereotyping, accuracy, privacy, bias, and its binary nature. We adapted it to handle Indian names effectively and

achieved high accuracy in this context. This classifier should be used thoughtfully, especially in contexts valuing individual identities and gender diversity. In our case, we used this model to draw meaningful patterns between resume screening by machine learning models and gender, which wouldn't have been possible without this model.

5 Results

In this section, we present the results of our gender bias analysis and test hypotheses related to occupational markers, gender bias, and historical trends.

5.1 Hypothesis 1: Occupational Marker Bias

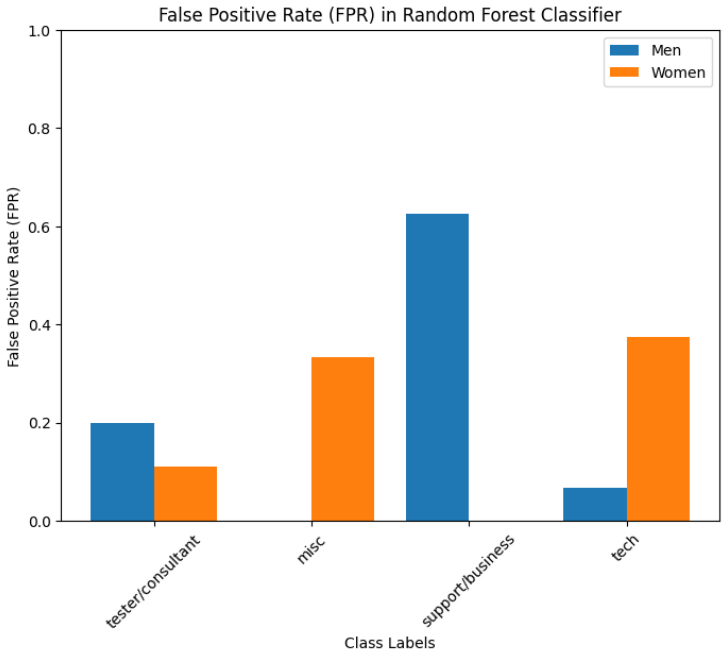

Figure 2: Random Forest Classifier F1 Score Analysis for Men vs. Women. Men were more often incorrectly predicted in testing or business. Women were more often incorrectly predicted into the miscellaneous and tech designation.

The first hypothesis tests whether occupational markers are better for men, while women are more likely to be put in different sectors, harming the quality of service. Figure 2 illustrates the F1 score analysis for the random forest classifier. Our findings indicate a significant discrepancy in the misclassification patterns between genders, with men more frequently mispredicted in testing

or business roles, and women often misclassified into miscellaneous and tech designations. Moreover, the analysis reveals that a higher proportion of men experienced greater false positive rates, suggesting a potential bias toward favoring candidates who are men in certain industries.

5.2 Hypothesis 2: Gender Bias in Executive Predictions

The second hypothesis suggests that men are more likely to be predicted as executives, with no women in the executives sector in the data, affecting the visibility of women in that sector.

5.2.1 Percentage of Men vs. Women as Executives

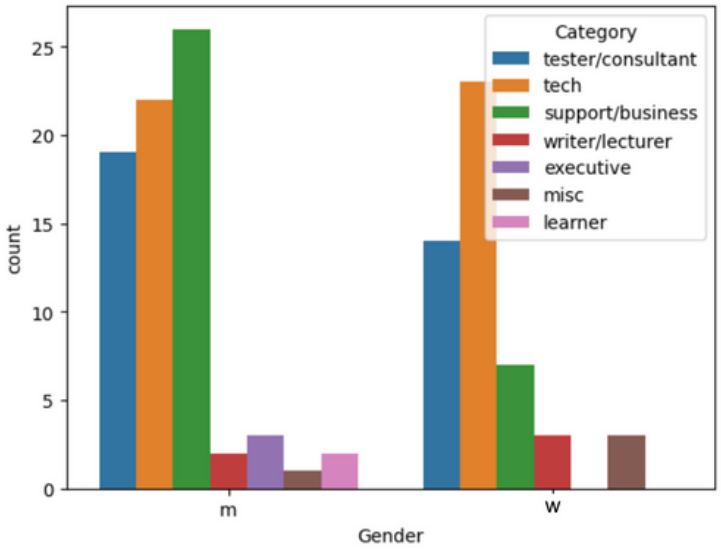

Figure 3: The distribution of Men vs. Women by Sector. The graph displays men and women having nearly equal amounts of samples in tech fields. Men have more business, consultant, executive, and learner samples, while women have more lecturers and miscellaneous samples.

Figure 3 provides an insightful representation of the percentage of men and women predicted as executives in our dataset. This analysis unveils a notable observation with implications for gender representation in executive roles.

Our findings suggest that men are more likely to be predicted as executives. It is important to note that women in the executive sector are absent from our dataset, and this absence has a significant impact on the model's predictions. With no women in the executive sector, the predictive model tends to assign executive roles more frequently to men. This discovery is of great importance, as it directly relates to gender equity and representation in leadership positions.

The absence of women in the executive sector in our dataset highlights a limitation in the available training data, which may be reflected in current industry models, resulting in an imbalance in predictions. This observation raises concerns about the visibility and opportunities for women applicants in executive roles. It underscores the importance of addressing data limitations and model biases to ensure that predictive systems yield fair outcomes for candidates of all genders.

5.3 Hypothesis 3: Historical Trend Reinforcement

The third hypothesis examines whether historical trends are reinforced in the data, with more men in business and more women in tech.

5.3.1 Prediction Percentage of Men vs. Women in Different Sectors

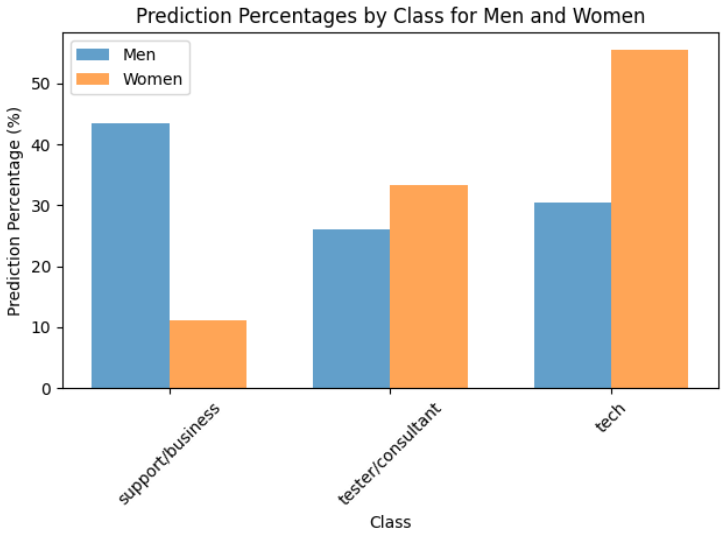

Figure 4: Men vs. Women Prediction Percentages by Sector. While men were predicted to be more likely to be in business, women were predicted more likely to be in tech or testing.

Figure 4 provides a visual representation of the prediction percentage of men and women in various job sectors. Our analysis uncovers intriguing patterns in the distribution of genders across these sectors, shedding light on the influence of historical trends and bias in job market dynamics.

Our findings reveal a notable trend where men are more frequently predicted in business roles. This aligns with historical trends where men have traditionally dominated leadership positions and executive roles. The predictive model tends

to favor men in these sectors, reflecting a historical bias that may perpetuate existing disparities.

Conversely, our analysis shows that women are more frequently predicted in technical roles, such as STEM fields. The predictive model, influenced by historical data and biases, tends to predict women in these roles more frequently, potentially exacerbating the gender gap in these sectors.

The observed patterns in sector predictions highlight the significant impact of historical biases on the distribution of genders in the job market. These biases not only influence the model's predictions but also reinforce historical trends. This reinforcement can have far-reaching consequences, including the perpetuation of gender disparities and the limited representation of women in leadership roles and men in technical sectors.

5.3.2 False Positive Rates

Figure 2 displays the false positive rates for men's sector predictions. Our findings indicate the rates at which the models incorrectly predict sectors, revealing interesting patterns that align with historical trends in the representation of genders in different roles.

Our analysis reveals that false positive rates are notably higher for men being falsely predicted into business roles, while women are more likely to be falsely predicted into technical roles. This observation reflects and reinforces historical trends in the job market. Historically, business roles have been predominantly occupied by men, while women have faced underrepresentation in business or management sectors.

The higher false positive rates for men in business roles may be attributed to the model's bias, which tends to overpredict men in roles traditionally associated with leadership positions. On the other hand, the elevated false positive rates for women(shown in figure 2) in technical roles may stem from the underrepresentation of women in these fields, leading the model to make frequent mistakes by predicting women in technical roles.

These findings emphasize the importance of addressing bias and promoting fair representation in gender predictions, as they can perpetuate and reinforce historical trends. To ensure equitable opportunities in the job market, mitigating such biases and improving the accuracy of predictions are essential steps forward.

6 Discussion

Our research sheds light on the intricate dynamics of machine learning bias in resume screening, uncovering several critical insights and implications for both the industry and academia. In this section, we delve deeper into our key findings and discuss their broader implications.

The pervasive presence of biases in machine learning-driven resume screening systems necessitates the implementation of robust mitigation strategies. Our

study emphasizes the imperative for organizations to adopt comprehensive bias detection and mitigation frameworks. By integrating fairness-aware machine learning techniques and conducting routine bias audits, organizations can proactively identify and rectify biases in their recruitment processes. This approach aims to cultivate a more inclusive and equitable environment for all candidates, mitigating the potential impact of biased decision-making.

The ethical ramifications of biased resume screening algorithms extend well beyond the boundaries of recruitment. Given the expanding influence of these algorithms across diverse sectors such as finance, healthcare, and law enforcement, unchecked biases pose heightened risks. Consequently, organizations and policymakers must prioritize ethical considerations in the development and deployment of AI technologies, particularly Natural Language Processors. This entails upholding principles of transparency and accountability to mitigate potential harm and uphold societal values.

While our research provides valuable insights into the challenges and implications of bias in machine learning-based resume screening, numerous avenues for future exploration beckon. Subsequent studies could focus on pioneering algorithmic approaches designed specifically to mitigate bias, employing advanced techniques like causal inference and counterfactual reasoning. Furthermore, investigating the intersectionality of multiple demographic factors, such as race, age, and socioeconomic status, promises a more nuanced understanding of bias dynamics. Such insights can guide the development of more equitable AI systems. As we navigate the evolving landscape of AI ethics and governance, the continuous exchange of ideas, ongoing research endeavors, and innovative solutions are indispensable. These efforts collectively ensure that AI functions as a force for societal good, propelling progress and fostering equitable opportunities for all.

7 Conclusion

The pervasive influence of machine learning in modern recruitment processes, as detailed in this paper, brings forth both opportunities and challenges. Our research underscores the dual nature of machine learning in recruitment, presenting both detriments and advantages. While automation offers efficiency and scalability, it also introduces the risk of perpetuating biases inherent in training data. Our focus on gender biases in machine learning-driven resume screening reveals that, despite their speed, algorithms can inherit historical imbalances and societal stereotypes.

Our findings, exemplified by the underrepresentation of women in executive roles, highlight the limitations of current AI systems and broader challenges faced by marginalized groups. Comprehensive, representative datasets and continuous model evaluation are crucial to ensuring fairness and inclusivity.

In deploying machine learning in recruitment, a balanced approach is imperative. While technology advances innovation, ethical considerations, transparency, and a commitment to equity must guide implementation. Recognizing

the limitations and biases in AI systems is crucial. By addressing these challenges collaboratively, we can harness machine learning's potential for efficiency while working towards a fairer and more just society. The significance lies in preventing the replication of biases present in human hiring practices, making AI a tool for fairer and more objective decision-making.

References

[Ale21] Salem Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 2021.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *ArXiv.org*, 2016.

[BGMM21] Emily M. Bender, Timnit Gebru, and Angelina McMillan-Major. On the dangers of stochastic parrots. *ACM Conferences*, 2021.

[CBN17] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 2017.

[CM13] Becca Carnahan and Christopher Moore. Actively addressing unconscious bias in recruiting. *Harvard Business School*, 2013.

[DARW+19] Maria De-Arteaga, Alex Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios. *ArXiv (Cornell University)*, 2019.

[Dat18] DataTurks. Resume entities for ner. *Kaggle*, 2018.

[DK95] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. *Citeseer*, 1995.

[DR19] Eva Derous and Ann Marie Ryan. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, 2019.

[GG19] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *ArXiv*, 2019.

[KL23] Indu Khurana and Daniel J. Lee. Gender bias in high stakes pitching: an nlp approach. *Small Business Economics*, 2023.

[LCC20] Max Langenkamp, Allan Costa, and Chris Cheung. Hiring fairly in the age of algorithms. *arXiv preprint arXiv:2004.07132*, 2020.

[Lea18] Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. *Proceedings of the 1st international workshop on gender equality in software engineering*, 2018.

[NFC21] Sean M Noble, Lori L Foster, and S Bartholomew Craig. The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment*, 2021.

[W+13] Jennifer Whelan et al. The barriers to equality of opportunity in the workforce: The role of unconscious bias. *Understanding the gender gap*, 2013.

[Woo23] Julie Woodworth. The rise of ai in resumes: How algorithms are changing hiring. *Linkedin*, 2023.

[YW18] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. *Hawaii International Conference on System Sciences*, 2018.