1. Introduction and Data

We plan to examine gender bias in online job screenings, plus bias in specific industries of job listings. We want to investigate the possible presence of gender bias in the online job screening process across different industries. We also want to explore how gender-coded language found in job listings and resumes may influence hiring decisions. As women in STEM fields, we are acutely aware of the barriers that gender bias, whether subtle or implicit, may impose in hiring, advancement, and representation. The STEM field is a rapidly growing employment sector and one where women remain underrepresented. Furthermore, as college students, we are beginning to experience the benefits and harms of online job screenings. As individuals who hope to have jobs one day, we are interested in understanding the systems in place today that can both enhance and hinder our work experiences. When firms use algorithmic systems for screening, such as automated resume parsers, coding tests, or AI-assisted tools, there is a risk that these systems amplify or encode social biases. Our project, therefore, examines not only whether gender bias exists in online job screening but also how contextual features such as gender-coded words in job descriptions and industry differences influence candidate outcomes. By combining resumes (Kaggle, HireItPeople), screening systems (HackerRank, CodeSignal, ChatGPT), and analysis of job description text, we aim to provide evidence that speaks directly to questions of fairness in hiring algorithms.

1.1 Research Questions

We propose the following research questions:

1. Do male and female candidates with similar qualifications have different job screening outcomes?

2. Does gender bias in job screening vary across industries (specifically gender-dominated fields)?

3. Are gender-coded words (e.g., “collaborative” vs. “assertive”) in job descriptions associated with differences in the gender distribution of job applicants?

1.2 Dataset Construction

The dataset is constructed through web scraping of resumes available on Hire IT People. Each resume entry contains text under sections such as Summary, Technical Skills, and Professional Experience. After scraping, we preprocess the HTML to plain text and parse these sections to extract key variables. Specifically, we identify the applicant’s years of experience, skills, number of past jobs, and industries worked in. As with the first dataset, we add a randomized “Gender” and “Pronoun” column to test our research question about bias.

1.3 Expected Findings

We hypothesize that there will not be a large bias based on gender alone, but algorithms' heavy reliance on other variables, such as gender-coded keywords in specific industries, may reflect implicit biases. We plan to use a regression model to evaluate whether hiring rates are different based on gender. A separate regression can evaluate industry-specific hiring decisions to see if industry-specific hiring preferences reflect gender bias.

1.4 Analysis of Data

a. Exploratory Data Analysis (EDA)

We began by summarizing the dataset using descriptive statistics and visualizations. This included examining distributions of years of experience, number of past jobs, and the frequency of technical skills across resumes. This step helped verify that the extracted features were consistent and allowed us to identify anomalies, missing information, or formatting issues before applying the scoring function.

b. Resume Scoring and Outcome Generation

Rather than using AI-based screening tools, we developed a transparent scoring function to simulate automated hiring assessments. Each resume was evaluated based on a weighted combination of three factors: semantic similarity between the resume and a job description, years of experience, and number of past jobs. These scores served as our outcome variable, representing how an automated system might rank applicants in a real-world screening process. The scoring system allowed us to systematically examine the relationship between resume content and assigned gender labels without relying on black-box AI models.

c. Feature Analysis and Gender-Coding

We analyzed how specific resume characteristics, particularly the presence of gender-coded language, influenced scoring outcomes. Using methods inspired by Gaucher, Friesen, and Kay (2011), we identified masculine- and feminine-coded words in resumes and examined whether resumes containing these terms systematically received higher or lower scores. This allowed us to test whether linguistic patterns associated with gender could create disparities in ranking outcomes, even when gender was not explicitly used in the scoring function.

d. Robustness and Fairness Checks

To assess fairness, we compared resumes with similar professional attributes (years of experience, technical skills, number of past jobs) but different assigned genders. This allowed us to isolate the effect of gender-coded language on the ranking outcomes. Differences in scores within these matched pairs highlighted whether gender-driven disparities emerged when other resume features were held constant.

1.5 Outcome and Key Variables

Our primary outcome variable is the resume score, which reflects whether a candidate would be ranked higher in a simulated automated screening process. Key explanatory variables include technical skills, years of experience, number of past jobs, assigned gender, and counts of masculine and feminine-coded words. These variables form the basis of our analysis on how gender and language interact to influence resume rankings.

1.6 Data Cleaning and Preparation

To clean the data, we separated the resume data into more specific column categories. First, we separated the data by “summary,” “skill level,” and “technical skills.” Then, we started extracting certain details such as number of past jobs and number of years of experience.

To prepare the data, we scraped 600 resumes from HireItPeople.com under the Java Developers/Architects Resumes page. Hire IT People provides thousands of resumes within industries for employers to view. We created the resume.csv file by opening each link that holds a resume and then added the text to the CSV. Once we had a CSV containing the URL link and the long resume text, we were able to begin separating the different variables into separate columns.

We might remove some filler information, such as repetitive language, and exclude resumes with the “professional summary” section (3 entries). We randomly assigned gender (male/female) to each entry. We also added columns for years of experience and number of previous jobs held, which were pulled from the resume descriptions. In addition, we added our own data for awards to relate to gender, specifically for females. We added awards that highlighted their role as a “woman in STEM.”

1.7 Gender-Coded Words

We used Gender Decoder, which was inspired by a research paper written by Danielle Gaucher, Justin Friesen, and Aaron C. Kay (2011). In this paper, *Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality*, researchers showed job adverts that included different kinds of gender-coded language to men and women and recorded how appealing the jobs seemed and how much the participants felt that they 'belonged' in that occupation.

2. Methods

Our resume scoring algorithm combined three quantitative indicators of job fit: (1) semantic similarity between each resume and the job description, (2) the applicant’s years of experience, and (3) the number of previous jobs listed. The three inputs were normalized and combined using a weighted linear formula, with semantic similarity contributing the most (70%), followed by years of experience (20%) and job count (10%). These weights were chosen to prioritize content relevance while still incorporating basic measures of experience.

Semantic similarity was calculated using a TFIDF Vectorizer, which measured if applicants' resumes matched against a set of keywords that suited the resume description. TF-IDF similarity was chosen because it is simple, transparent, and interpretable. It provides a baseline measure of how well the content of each resume aligns with the language and requirements of the job description.

Experience features were normalized using min-max scaling so that each feature contributed proportionally to the final score regardless of its original range. Years of experience and job count were selected as lightweight proxies for career depth and stability.

In the analysis, for each applicant, the algorithm computed:

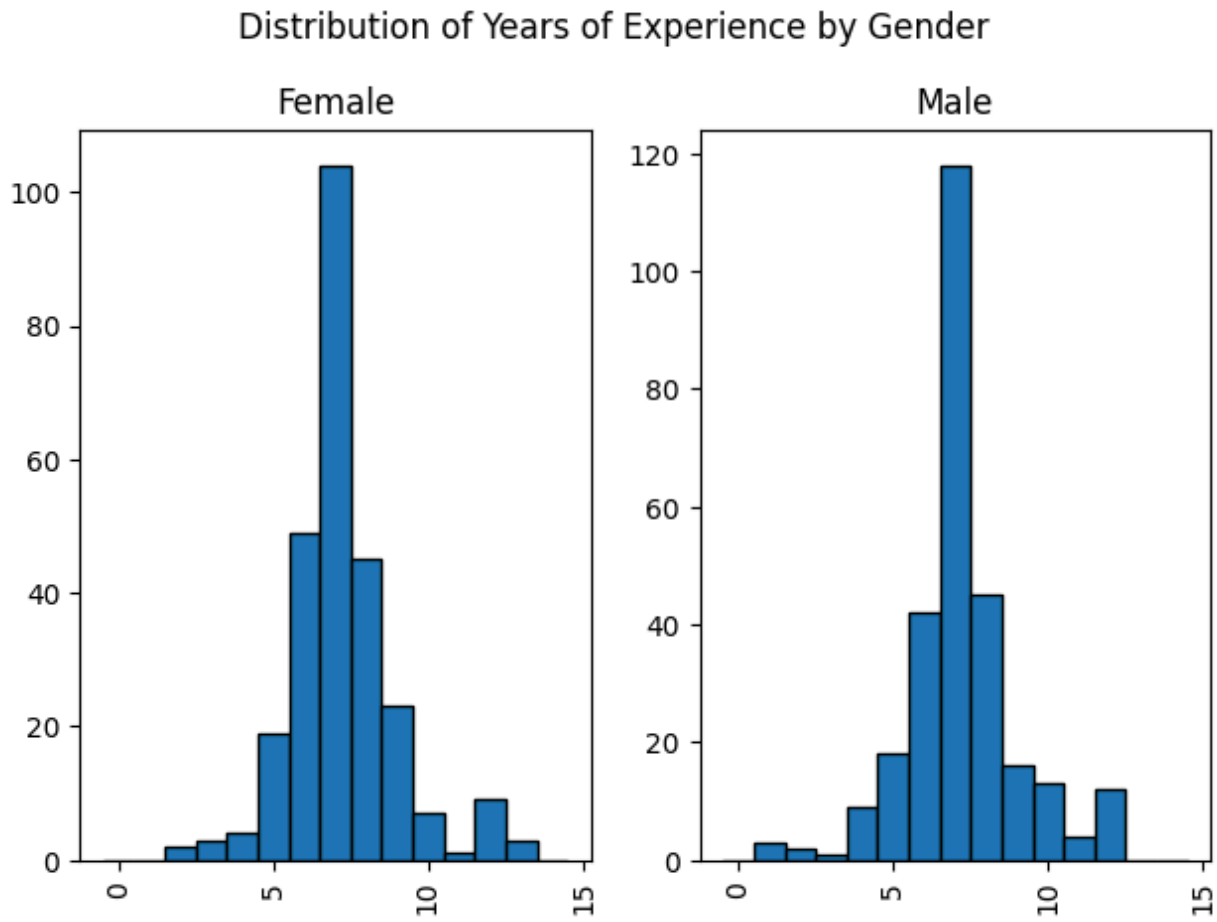
- score = (0.70 * semantic similarity) + (0.20 * years of experience) + (0.10 * number of past jobs)

The final rankings were assigned by sorting applicants in descending order of their overall score. Their individual scores, based on semantic similarity alone, and demographic features were preserved for descriptive statistics, but were not considered in the final ranking/scoring.

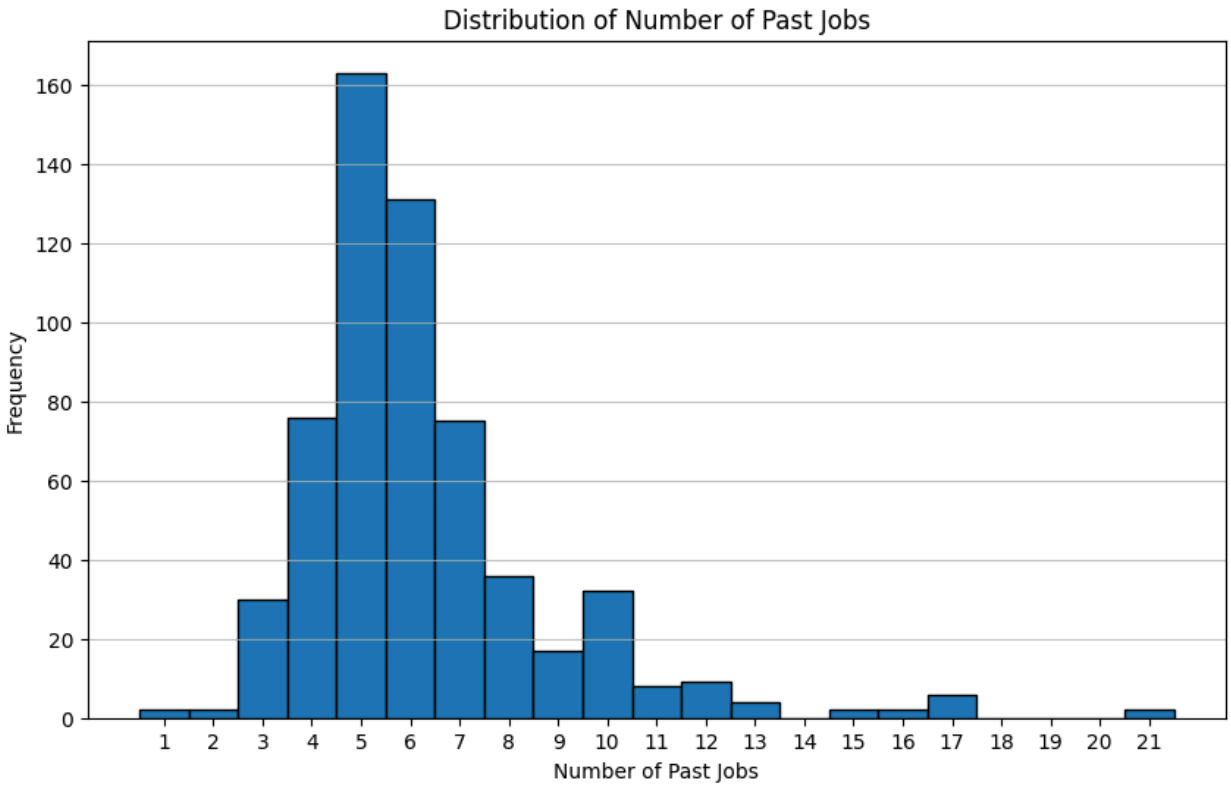
3. Results

3.1 Descriptive Profile of the Applicant Pool

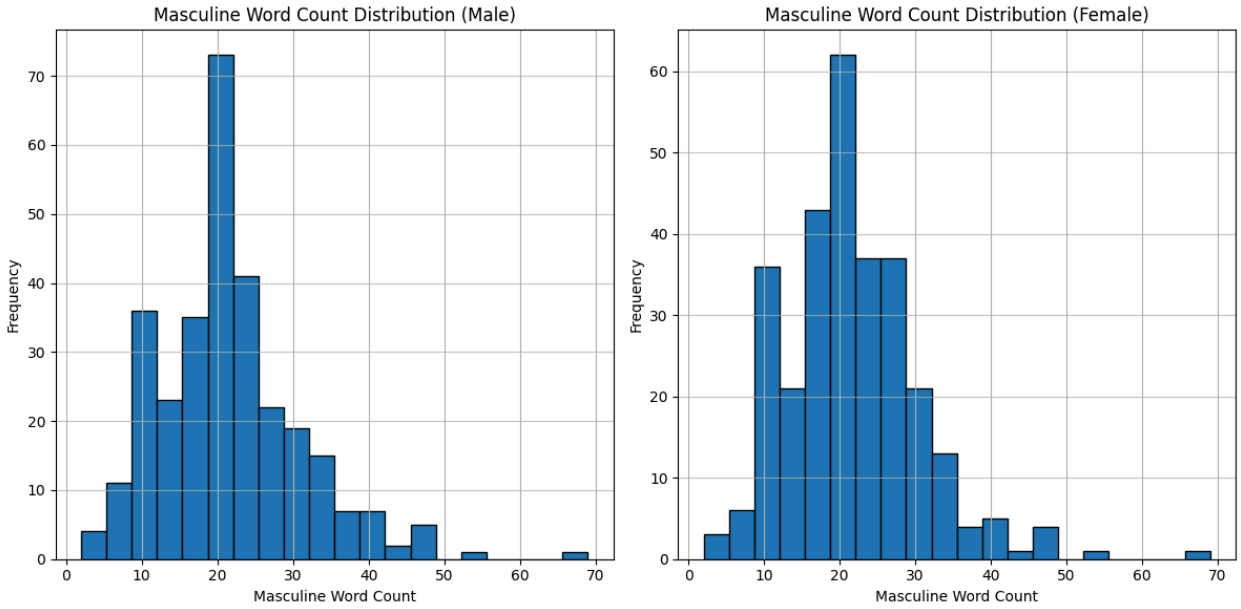
Across the entire dataset, applicants show a relatively consistent pattern of career experience and resume structure. Years of experience form an approximately normal distribution centered around seven years. Most applicants fall between four and ten years of experience, which is typical for mid-career hiring pools. The distribution appears slightly different across genders when those genders are assigned randomly: women show a modest left skew, while men show a slight right skew. This difference, however, is small and does not indicate any meaningful variation in qualifications at this stage of analysis.

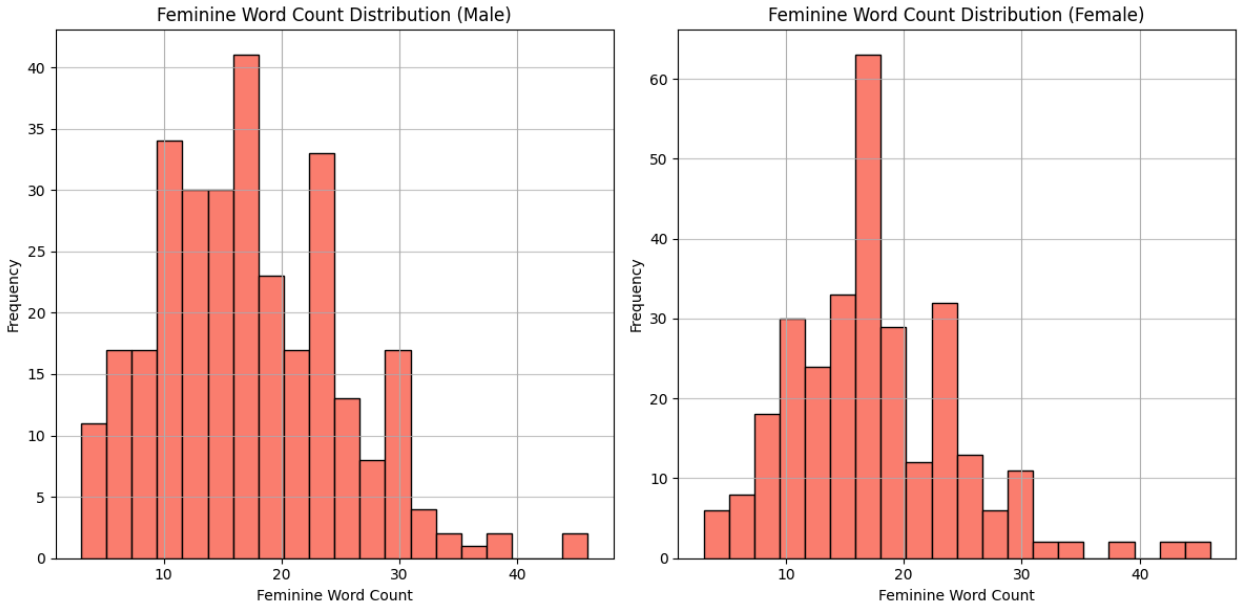


The number of past jobs follows a right-skewed distribution, with most candidates having held approximately five positions. Very few candidates list more than ten prior roles. This pattern is consistent across genders: regardless of assignment method, men and women display nearly identical distributions in job mobility. This suggests that the dataset does not contain systematic gender differences in employment stability or career transitions. **(need to clean Number of Past Jobs column)**



Finally, we computed the number of masculine- and feminine-coded words in each résumé. Masculine-coded words were more common across the dataset as a whole, and this pattern remained stable regardless of gender assignment method. Even among applicants randomly labeled as female, masculine-coded phrasing appeared frequently, indicating that linguistic style is not evenly distributed across the applicant pool.



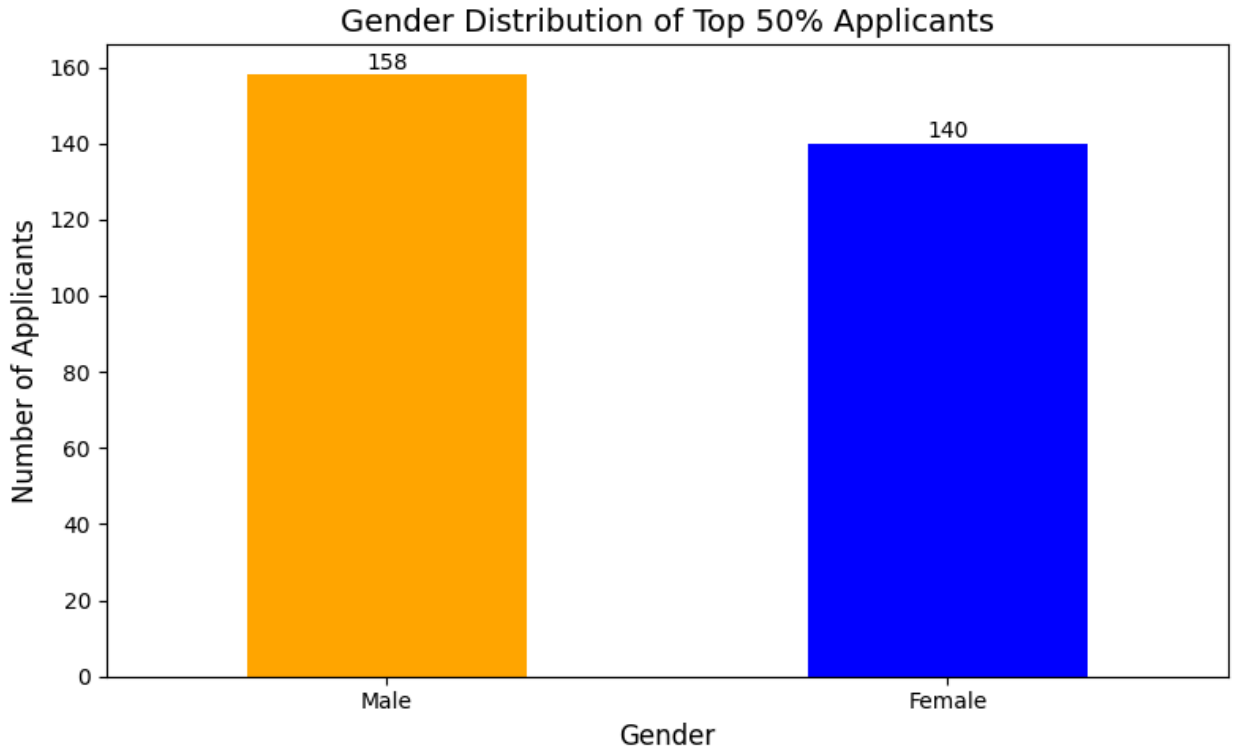


3.2 Algorithmic Rankings Under Random Gender Assignment

To evaluate whether the resume-scoring algorithm produced different outcomes for men and women when gender was not correlated with résumé content, we first randomly assigned gender to all applicants without an existing label, fixing the percentage of male and female to be roughly 50% to have a balanced dataset.

When the algorithm ranked applicants by the weighted score combining semantic similarity, years of experience, and number of past jobs, the top 50% of applicants contained proportions of men and women that were similar to the overall distribution. In other words, when gender was independent of resume content, the algorithm did not disproportionately favor one

group.



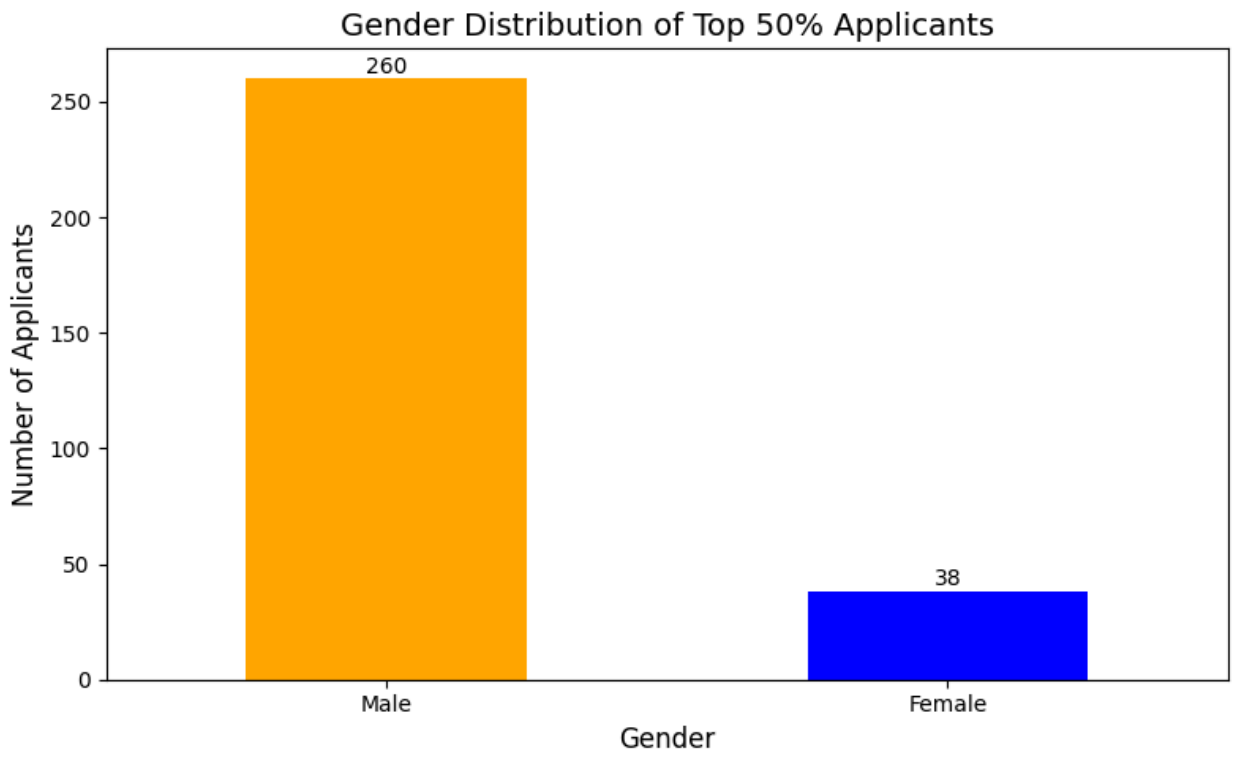
This result provides an important baseline: if gender has no structural relationship to the features the algorithm rewards, then the ranking outcomes appear unbiased.

3.3 Algorithmic Rankings Under Text-Based Gender Assignment

The interpretation changes substantially when gender is inferred from linguistic signatures, specifically, the relative frequency of masculine- and feminine-coded words. Under this condition, the dataset becomes uneven. A noticeably larger share of resumes are classified as male-coded due to their higher counts of masculine language. Feminine-coded résumés appear less frequently, suggesting that applicants in this dataset, regardless of actual gender identity, tend to describe their skills and accomplishments using more stereotypically masculine phrasing.

This imbalance has downstream consequences. Because masculine-coded resumes tend to be longer, contain more technical verbs, and score slightly higher in TF-IDF similarity to the job description, their semantic similarity scores are, on average, higher. Moreover, these resumes often list more detailed tasks and responsibilities, inflating their number of past jobs and boosting their overall score. Thus, although gender labels themselves never enter the scoring formula, features correlated with masculine-coded language do.

When we examine the top 50% of algorithmic rankings under this condition, the difference becomes pronounced. Male-coded resueés occupy a disproportionately larger share of the top-ranking applicants, while female-coded resumes appear far less frequently. This shift is not due to algorithmic bias toward gender per se, but rather reflects the way linguistic style interacts with the scoring features, particularly semantic similarity, which was weighted most heavily.



3.4 Comparison Between the Two Gender Conditions

Juxtaposing the two gender-assignment methods highlights a crucial insight: **the perceived fairness of the resume-screening algorithm depends strongly on how gender is defined in the dataset**.

Under random assignment, the rankings appear balanced and gender-neutral. Under text-based assignment, they appear skewed, with male-coded resumes overrepresented among top performers. Both results are technically “correct” given the assumptions but they tell very different stories.

This contrast suggests that disparities emerge not because the algorithm directly encodes gender bias, but because certain linguistic patterns, more frequently associated with masculine-coded writing, align more strongly with the algorithm’s scoring dimensions. The fact that masculine-coded resumes tend to be longer, more technical, and more detailed means they naturally accumulate higher similarity scores and, in turn, higher final rankings.

4. Discussion

Our study examined whether gender bias emerges in automated resume screening systems and how the appearance of bias changes depending on the way gender is defined in the dataset. Across our analyses, several key findings emerged that directly address our research questions and connect to broader discussions in the hiring and algorithmic fairness literature.

4.1 Gender-Neutral Baseline

First, when gender was assigned randomly, the resume-scoring algorithm produced nearly identical proportions of male and female applicants in the top half of ranking outcomes. This suggests that the scoring method, which relied on semantic similarity, years of experience, and number of past jobs, does not inherently privilege either gender when no structural relationships exist between gender and resume features. Under this condition, the algorithm behaved as a relatively neutral baseline model. This serves as an important point of comparison: in a hypothetical world where linguistic and experiential features are evenly distributed, our scoring mechanism does not generate gender disparities on its own.

4.2 Gender-Coded Language and Emergent Bias

However, the results changed substantially when gender was inferred through gender-coded language, following methodologies grounded in Gaucher, Friesen, & Kay (2011), who showed that masculine- and feminine-coded wording shapes perceptions of belonging and job appeal. When this approach was applied to our resume dataset, masculine-coded resumes appeared far more frequently than feminine-coded ones. This reflects a well-documented pattern in technical industries—such as software development—where resume phrasing tends to emphasize assertiveness, leadership, autonomy, and technical mastery, traits commonly categorized as masculine-coded. Prior work has suggested that these linguistic patterns arise from broader occupational cultures rather than individual identity, and our data support this interpretation.

4.3 Consequences for Algorithmic Outcomes

Importantly, the shift in the gender distribution of the applicant pool had measurable consequences for algorithmic outcomes. Masculine-coded resumes tended to be longer, more detailed, and richer in technical vocabulary. Because semantic similarity contributed the greatest weight to our scoring formula, these resumes systematically received higher scores. As a result, male-coded resumes occupied a disproportionately large share of the top-ranked applicants, creating an apparent but not explicit gender disparity.

This finding illustrates a central theme in the algorithmic fairness literature: algorithmic systems can reproduce or amplify historical patterns in data even without using protected attributes directly. Our results parallel work such as Bolukbasi et al. (2016) and Kleinberg et al. (2018), which demonstrate that seemingly neutral computational models can encode social inequities when trained on or applied to skewed datasets. Similarly, our analysis highlights that linguistic style—particularly differences in how applicants describe their experience—acts as a latent variable correlated with gender-coded classifications. Because the resume-scoring algorithm rewards features that tend to appear more frequently in masculine-coded resumes, these resumes are systematically advantaged.

4.4 Methodological Implications

The fact that both gender-assignment methods tell plausible, yet contrasting stories underscores an important methodological implication: how researchers define and operationalize gender fundamentally shapes the observed outcomes of fairness analyses. When gender is independent of resume content, the model appears fair. When gender is tied to linguistic patterns that reflect socialized communication styles, disparities emerge, yet their source lies in the data rather than the scoring rule itself. This reinforces the argument, common in fairness scholarship, that algorithmic bias often originates from structural or linguistic patterns embedded in inputs, not from overtly discriminatory design.

4.5 Relevance to Broader Literature

Overall, our findings strengthen existing literature showing that gender-coded language has meaningful consequences for hiring evaluations, even when decisions are mediated by automated systems rather than humans. At the same time, our results contribute a novel perspective: disparities can arise even in simplified, transparent, hand-designed scoring systems, suggesting that the risk is even more substantial in commercial black-box models such as proprietary resume parsers or AI hiring platforms. By showing how a dataset's linguistic composition can shift the perceived fairness of an algorithm, our work highlights the need for practitioners to consider not only which variables an automated system uses, but also how underlying language conventions and occupational cultures shape the features on which such systems rely.

4.6 Conclusions

In conclusion, our study demonstrates that apparent algorithmic gender bias depends heavily on contextual factors—particularly the relationship between gender and linguistic patterns in applicant data. When gender is unrelated to resume content, rankings appear balanced. When gender-coded wording is used as a proxy, masculine-coded resumes consistently rise to the top. These findings suggest that efforts to reduce bias in automated hiring systems cannot focus solely on removing explicit gender indicators. Instead, they require a deeper understanding of the subtle, socially constructed features that influence algorithmic decision-making and shape who benefits from automated screening tools.

4.7 Resume Scoring Function Limitations

The resume scoring/ranking function looked at a weighted combination of three factors: the semantic similarity between the applicant's resume text and the job description, the number of years of experience, and how many of the past jobs the applicant had. While this approach was easy to understand and implement, it had a few key limitations.

First, the semantic similarity between the resume text and job description was calculated using a TFIDF Vectorizer, which measures if keywords overlap. However, this scoring approach doesn't capture the semantic context or meaning of the words, so a resume that repeats the same words as the job description may score high even if the candidate's experience is weak.

At the same time, someone who describes the right skills using different wording may score lower.

Second, the years-of-experience and job-count features are very crude. They assume that “more years” and “more jobs” always mean “more qualified,” which is not always true. These features can also indirectly favor older candidates or penalize people with career gaps or fewer roles, even if they are strong fits for the job.

Third, the weights used for the scoring function are arbitrary (0.70, 0.20, and 0.10). They were not learned from data or validated in any way, and instead represented a programming decision, as we assumed the most influential aspect of an application would be its match to the job description, with years of experience and number of past jobs being less significant deciding factors, but still considered.

Finally, because everything is normalized across the current dataset, individual scores depend on who else is in the applicant pool. Adding or removing resumes would change the min-max scores and thereby change everyone’s scores, making the results less consistent. Overall, this method works as a simple baseline for ranking resumes, but it should not be used as a hiring tool because it misses the true semantic understanding of keyword phrases, relies on oversimplified numeric features, and lacks validation. Improving it would require better semantic models, explicit skill extraction, and fairness checks.

4.8 Data Reliability and Validity

There are multiple things to consider pertaining to the reliability and validity of our data and analytical approach. Scraping resume information from Hire IT People was done as systematically as we thought possible. While the resumes all have the same or similar information, each resume varies in length, format, detail, and level of description. We scraped the entirety of each resume and then weeded out information. While this made sense and was feasible for us to do, it does result in some reliability issues when extracting the data. The validity of our data should also be heavily considered. Due to the fact that we randomly assigned gender labels, our data is not directly related to real-world scenarios or examples. While we were able to use this variable to help analyze results, we can also only take our results with so much truth. As a result, our analysis looks at how an automated system responds to a labeled gender signal, not necessarily how actual employers may infer gender from names, writing style, or career trajectories. Overall, our methods allow us to examine potential mechanisms of algorithmic bias, but our analysis must be interpreted cautiously given the structured nature of our dataset and the limitations of our data.

4.9 Limitations for Generalization

Furthermore, the data we used limits any inferential conclusions or predictions for new data we could make. First, since our dataset was scraped from a single public online repository (Hire IT People), it is not fully representative of the broader applicant population. We took a handful of resumes under Java Developers/Architects. This means we looked at only a small

subset of resumes that are out there to analyze. In addition, many of these resumes are from high-achieving, highly technical professionals. This restricts our ability to make inferences about other groups of resumes under different job descriptions. We cannot make predictions for other fields, like non-STEM fields or entry-level positions. Second, the fact that we made gender random means that we created dummy data for us to use. This means that we would not really be able to make meaningful inferences about how the genders are treated. For example, gender may correlate with differences in experience patterns, skill domains, award types, or industry representation in actual labor markets, which our data does not take into account. Because of these limitations, our data and results should be interpreted as the potential presence of bias, rather than as predictive tools or conclusions that can be applied to new or real-world applicant data.

4.10 Ethical Considerations and Societal Implications

Our project raises ethical concerns and larger societal implications because it involves scraping resume data and simulating algorithmic hiring decisions. The fact that we used resumes from real applicants and personal information brings up privacy and data-use concerns. It is unknown whether or not the individuals who are connected to the resumes would want their personal information to be used for data research, such as our project. Ethical risks are also present when analyzing gender bias using randomly assigned gender labels, as there is a possibility of reinforcing stereotypes or oversimplifying the complexities of real gender identity. Our project does not accurately replicate what the real-world hiring process is like and how hiring algorithms or companies view or infer gender. One societal implication of our project is that if hiring algorithms do in fact have a gender bias, they could unfairly limit job opportunities for women. This means that biased algorithms could contribute to long-term inequalities in the job market. While our project has larger issues and implications, it also yields positive outcomes that help reduce bias in the hiring process, particularly for women, who continue to face challenges in this area.

5. Citations

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems, 29. https://papers.nips.cc/paper/6228

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. Journal of Personality and Social Psychology, 101(1), 109–128. https://doi.org/10.1037/a0022530

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2018). Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science (ITCS 2017), 43:1–43:23. https://arxiv.org/abs/1609.05807