International Journal of Mathematics and Computer Research
ISSN: 2320-7167
Volume 13 Issue 12 December 2025, Page no. – 6052-6057
Index Copernicus ICV: 57.55, Impact Factor: 8.615
DOI: 10.47191/ijmcr/v13i12.18

Quantifying Bias in Text Genrative AI models

Sai Asrith Devisetti[1], Radhika Mamidi[2]
[1,2]International Institute of Information Technology, Hyderabad.

ARTICLE INFO

Published Online:
30 December 2025

ABSTRACT

Generative artificial intelligence (AI), especially large language models (LLMs), is increasingly deployed in domains such as recruitment, content creation, and education. While these systems accelerate productivity, they also risk reproducing and amplifying societal biases (Ahuchogu et al., 2025). This project addresses the urgent challenge of identifying, quantifying, and mitigating gender bias in text-generative AI outputs, with a focus on job narratives. Building on my independent study of 11,000+ AI-generated job narratives, which we generated using Gemini AI, we introduce a bias quantification framework using mean bias, mean absolute bias, sentiment skew (via TextBlob), and distributional measures (via Kullback–Leibler divergence and related distances). Preliminary results show measurable gendered patterns across generated narratives, validating the hypothesis of proposed gender bias in LLM.

The proposed work extends this foundation in three directions: expanding bias quantification using probabilistic distribution distances (Devisetti, 2024)(Chung et al., 1989), evaluating prompt-construction bias and multi-model comparisons across GPT-3, GPT-4, Gemini, and open-source LLMs (Blodgett et al., 2020), and integrating interpretable embedding methods (e.g., SPINE)(Subramanian et al., 2017) for transparency in downstream debiasing.

The expected contribution is both theoretical and practical: a robust bias quantification pipeline grounded in probability theory, and actionable strategies to mitigate bias in LLM-generated recruitment texts(Ferrara, 2024). Beyond recruitment, the proposed methodology aims to serve as a standard for bias evaluation in generative AI applications more broadly.

A key part of this research is the creation of large datasets containing job narratives. These datasets not only help analyze bias in AI-generated content but also support other Natural Language Processing (NLP) tasks.

Corresponding Author:
Sai Asrith Devisetti

KEYWORDS: Generative AI, bias quantification, gender bias, fairness in AI, probability distributions, job advertisements, large language models, large text-datasets.

1 INTRODUCTION

Artificial intelligence (AI) has rapidly emerged as a transformative force across industries, with generative models reshaping how text, images, and other media are produced. Among its many applications, large language models (LLMs) such as GPT-3, GPT-4, and Google Gemini are now increasingly employed in recruitment and human resources management. Their ability to generate job descriptions, outreach emails, and candidate communication at scale presents clear efficiency benefits. However, these systems also pose significant ethical challenges: including recruitment, and automated decision-making. With the increasing reliance on AI-generated text and AI-driven evaluation systems, concerns regarding bias, fairness, and the ethical implications of these technologies have become more pronounced (Mujtaba and Mahapatra, 2024). However, recent studies highlight the presence of bias in AI hiring models, with automated systems exhibiting gender, racial, and intersectional biases that disadvantage underrepresented group (Wilson and Caliskan, 2024). The reliance on AI models for candidate evaluation raises ethical and legal concerns, especially when models amplify existing societal biases. Investigating bias in AI driven hiring systems, particularly in language models used for resume screening, is crucial for ensuring equitable and transparent recruitment practice (Li et al., 2023).

Artificial Intelligence (AI) has emerged as a transformative force in recruitment, promising greater efficiency, speed, and

objectivity in candidate selection. With AI systems capable of parsing thousands of resumes, conducting initial screening interviews, and ranking candidates, companies are increasingly adopting algorithmic solutions to streamline hiring. However, alongside these efficiencies, there is a growing concern about the fairness and transparency of AI in recruitment. These concerns are rooted in the realization that AI systems trained on biased historical data may replicate and even amplify discriminatory practices related to gender, race, age, and socio- economic background. (Ahuchogu et al., 2025)Biased outcomes not only jeopardize organizational diversity but also expose companies to legal and reputational risks. This has led to a critical discourse on the ethical implications of algorithmic hiring and the mechanisms needed to ensure equity and accountability in such systems. As AI becomes more autonomous in decision-making, there is an urgent need to examine how biases enter these systems and what safeguards can be employed to prevent unfair treatment of candidates. This paper investigates the sources of bias in AI recruitment tools, explores notable real-world examples of biased outcomes, and presents a framework for achieving fairness in algorithmic hiring. The study draws on interdisciplinary literature from computer science, law, ethics, and human resource management to provide a comprehensive understanding of the issue. Ultimately, it argues for a balanced approach combining human oversight with ethical AI design to foster inclusive hiring environments and uphold fairness in a digitally evolving labor market.

Generative AI bias is not restricted to text. (Zhou et al., 2024) revealed consistent gender and racial stereotypes across multimodal models such as Midjourney and Stable Diffusion, emphasizing that AI systems can encode both overt and subtle representational harms.

To achieve the representation of bias in AI, we construct two datasets: a primary dataset containing single-job narratives for 1,163 occupations and an expanded dataset featuring multiple narratives per occupation. These datasets serve as the foundation for our analysis, which employs sentiment analysis using TextBlob and statistical bias detection via KL divergence. Our experiments investigate gender representation across various job roles, sentiment disparities, and discrepancies between AI-assigned gender and actual job expectations.

Through this study, we seek to provide empirical evidence of bias in Gemini AI's text generation(DeepMind, 2024), offering recommendations to improve fairness and reduce unintended discrimination in AI-driven hiring processes. Ensuring ethical AI applications in recruitment is vital for fostering equal opportunities in the workforce, making this research a crucial step toward unbiased AI deployment in HR technologies.

Why did we choose Gemini AI? Google's Gemini AI has emerged as a widely adopted solution, utilized by companies such as Pennymac, Devoteam, ResuMate Pro, and Pythian to streamline hiring practices. Its capabilities in parsing, analyzing, and summarizing candidate information offer significant advantages to recruiters, enabling quicker and more informed decision-making.

2 RELATED WORK

Research on bias in natural language processing (NLP) and generative AI has expanded considerably over the past decade. Yet, as (Blodgett et al., 2020) observe, many works lack precision in how they define bias, often failing to articulate the specific harms under consideration or to situate findings within broader socio-technical contexts. Their survey of 146 papers shows that bias is variously framed as representational harm (how groups are portrayed), allocational harm (who gets access to opportunities), or vague "fairness" concerns without sufficient grounding. This inconsistency complicates the interpretation and comparison of results, underscoring the need for bias quantification frameworks that are both rigorous and transparent.

Applied investigations into generative models highlight concrete risks in recruitment contexts. (Borchers et al., 2022) studied GPT-3-generated job ads and found significant gender stereotyping: roles were framed with language that aligned more with male-coded agentic attributes than female-coded communal traits. Importantly, prompt- engineering approaches offered minimal mitigation, while fine-tuning with curated low-bias corpora achieved measurable improvements. Their study provides an initial demonstration that intervention is possible but also illustrates the difficulty of achieving fairness without systematic frameworks. Similarly, (Dikshit et al., 2024) analyzed more than 6,000 academic STEM job postings, introducing a typology of agentic, communal, and balanced language. Their findings suggest that women may be discouraged by postings heavy in agentic descriptors, further demonstrating the real-world impact of linguistic framing in recruitment.

Beyond recruitment-specific studies, research has also pointed to systemic representational bias in multimodal generative AI. (Zhou et al., 2024) documented that image models such as Stable Diffusion and Midjourney routinely depict women in submissive roles and disproportionately associate certain ethnic groups with particular professions. These findings echo concerns in text-based systems and reinforce the need for multimodal-aware fairness frameworks. (Li et al., 2023), in their taxonomy of trustworthy large language models, emphasize fairness, reliability, and robustness as indispensable dimensions, proposing 29 evaluation subcategories. However, they note that existing LLMs consistently underperform on fairness measures, further validating the urgency of methodological innovation.

Methodological work on measuring bias also informs this proposal. Classical probability and information theory provide several distance measures such as the Hellinger distance, Jeffreys divergence, and the J divergence that

capture separability between probability distributions (Chung et al., 1989). These tools, though largely applied in statistical estimation, pattern recognition, and information theory, provide a mathematically principled basis for quantifying differences between groups in generated text distributions. At the representation level, (Subramanian et al., 2017), a sparse autoencoder technique for generating interpretable embeddings. Their work shows that sparse, semantically coherent representations not only improve transparency but also facilitate downstream interpretability a critical consideration for mitigating and explaining model bias.

Taken together, the literature reveals three key limitations. First, definitions of bias in NLP and AI remain inconsistent and often underspecified, limiting the comparability of studies. Second, applied investigations into bias in job advertisements have demonstrated both the severity of the problem and the inadequacy of shallow mitigation strategies, but have not extended to multi-model or multi-prompt comparisons. Third, while theoretical tools for quantifying statistical differences exist, their integration into a unified, interpretable framework for bias analysis in generative AI remains underexplored. This thesis proposal addresses these gaps by combining rigorous statistical metrics, applied bias analysis in recruitment texts, and interpretable embeddings to form a comprehensive pipeline for bias quantification and mitigation.

3 PROPOSED WORK

3.1 Dataset Generation

In this work, we generated three datasets, as described below:

Job dataset. It is a single-columned CSV file, where the column represents different fields of employment or job titles. We generate this data set from: Job dataset (Rana, 2023), data.gov, along with the preliminary data analysis.

Job Narratives Generated by Gemini. We generated a dataset, that contains a narrative generated by Gemini are stored for all 1163 job titles. This dataset will be addressed as Type 1.

Improved Job Narratives Dataset. We generated another dataset, that contains 10 narratives generated by Gemini are stored for all 1163 job titles. This dataset will be addressed as Type 2.

All job narratives were generated using Google's Gemini 1.5 Flash (DeepMind, 2024) We use the following neutral prompt to generate outputs:

"Write a compelling and realistic short story about a day in the life of a job. The story should capture the essence of their work, challenges, and personal experiences. Provide insights into their professional and personal journey."

A core contribution of this work lies in the generation of two extensive datasets that serve as the foundation for analyzing bias in Gemini AI-generated job narratives. The type 1 dataset comprises 1,163 unique job narratives, each containing a minimum of 300 tokens, offering a diverse representation of professional roles. The type 2, more expansive dataset extends this work by generating ten distinct narratives per job, culminating in a total of 1163×10 job narratives. By capturing multiple perspectives for each occupation, this dataset enables a deeper examination of linguistic patterns, sentiment biases, and gender representation across different job categories. The scale and diversity of these datasets not only facilitate robust statistical analysis but also provide a rich resource for evaluating Gemini AI and any NLP tasks.

3.2 Methodology:

The research will proceed in three phases:

3.2.1 Bias detection and analysis methodology:

To analyze the bias in Gemini AI-generated narratives, we employ a structured approach:

- **Gender Detection:** A study by von der Malsburg et al. (2020) utilized regular expression searches to automatically detect pronouns in text, analyzing the prevalence of gendered language in event descriptions, similarly Regular expression matching is applied to detect gendered pronouns within the text, categorizing the job narratives as Male, Female, or Both. "Both" implies that no pronouns were used in the text generated by Gemini AI, this is consider as a Gender Neutral Response.
- **Health Condition Identification.** The narratives are analyzed for keywords related to mental health concerns, categorizing them into potential stress indicators.
- **Job Difficulty Assessment.** The study (Popoola et al., 2024) employed TF-IDF vectorization combined with sentiment analysis to evaluate the sentiment of financial news articles, categorizing them based on perceived sentiment polarity. Similary, We employ TF-IDF vectorization combined with sentiment analysis using TextBlob to determine the perceived difficulty of each job. Jobs with a difficulty score above a defined threshold or those with a significantly negative sentiment are categorized as *Challenging*, while others are classified as *Manageable*.

4 PERLIMNARY RESULTS:

AI-generated text models often reflect societal biases, reinforcing stereotypes in professional contexts (Li et al., 2023). Prior studies highlight gender, racial, and occupational biases in both text and multimodal AI systems (Zhou et al., 2024; Birhane et al., 2021). To examine these issues in Gemini 1.5 Flash, we analyze AI-generated job narratives using techniques mentioned above.

4.1 Experiment 1: Preliminary Analysis

To assess the initial bias in Gemini AI-generated narratives , we analyzed the dataset for gender representation and job difficulty distributions.

Table 1: The above tables show the distribution of Gender and Job Difficulty in Type 1 dataset.

Gender	Percentage
Male	70.9%
Female	24.5%
Both	4.5%

Job Difficulty	Percentage
Manageable	91.4%
Challenging	8.6%

4.1.1 Observations

From the distributions shown in the above table1, they indicate a strong gender skew, with male representation being disproportionately high. Additionally, the Gemini appears to classify most jobs as "Manageable," potentially underestimating the challenges faced in various professions.

4.2 Experiment 2: Gender Variation Analysis

This experiment focused on analyzing gender representation across different job difficulties and potential psycho- logical stressors. The key areas of study are as follows.

4.2.1 Gender Variation Across Manageable vs. Challenging Jobs

Comparing the proportion of male and female-assigned jobs in "Manageable" vs. "Challenging" categories.

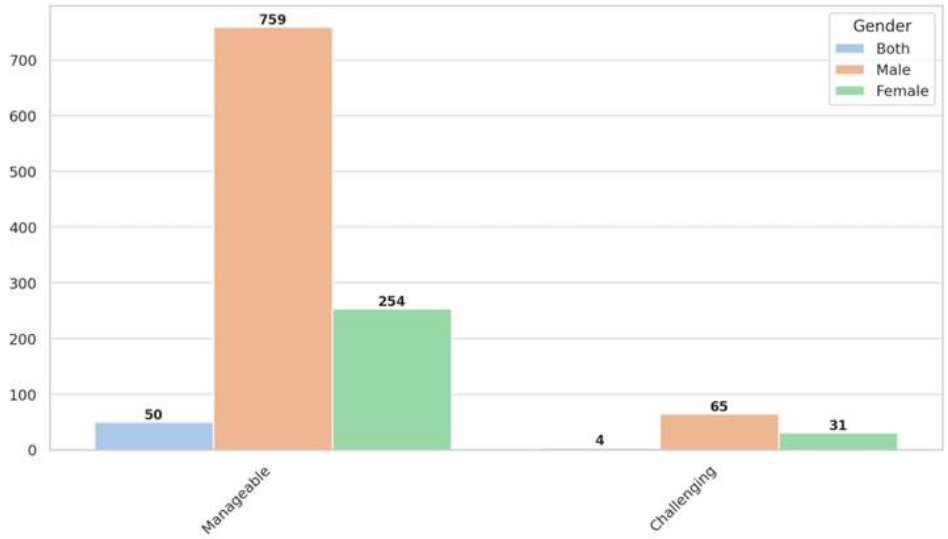

Figure 1: Gender variation in job difficulty, comparing manageable versus challenging roles.

4.2.2 Gender Variation in Potential Status

Examining whether male or female-represented jobs exhibited higher stress indicators.

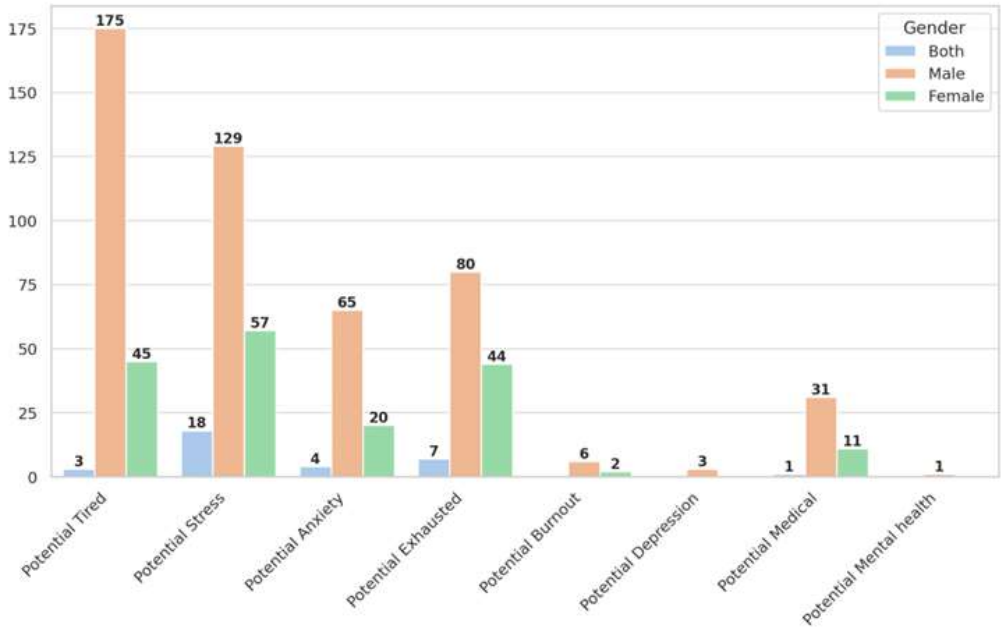

Figure 2: Gender variation in perceived potential status, including stress and exhaustion.

4.2.3 Comparison of the Gender Given by Gemini vs. the Preferred One

Evaluating whether Gemini AI aligns with common gender expectations in job roles. Refer figure 3 for the visual representation.

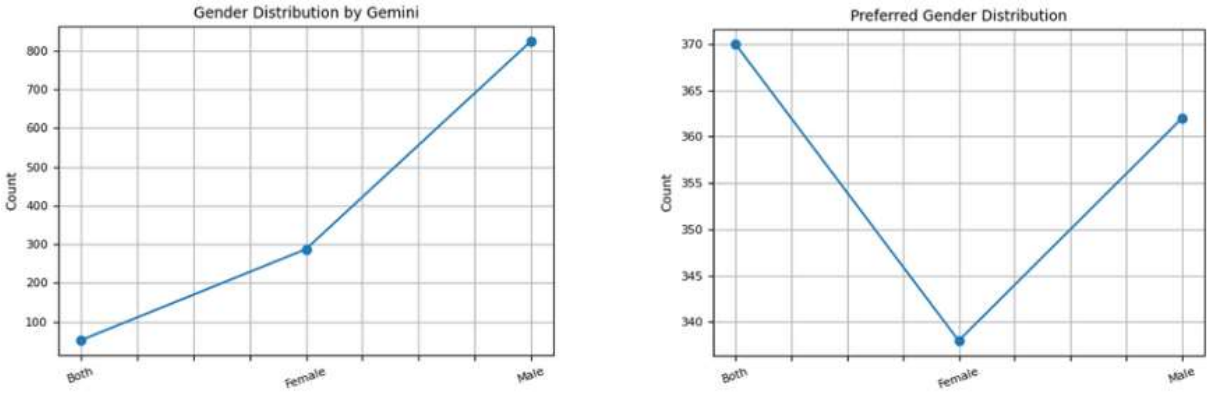

Figure 3: Comparison between Gender Generated by Gemini vs the preferred one

4.2.4 Observations

The results indicate a clear gender bias in AI-generated job narratives. Figure 1 shows that "Challenging" jobs were mostly assigned male pronouns (76%), reinforcing stereotypes of men in demanding roles. Figure 2 highlights that jobs linked to high stress were more often assigned male pronouns, suggesting a bias in how emotional burden is distributed. Figure 3 further reveals mismatches between Gemini AI-assigned and expected gender roles.

4.3 Experiment 3: KL Divergence for Bias Measurement

This experiment aims to analyze gender representation in textual narratives by computing the probability distributions of male and female references across multiple narratives . We quantify the divergence of these distributions from a uniform reference distribution using Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951).

4.3.1 Improvised Dataset

A new dataset consisting of ten narratives for each job is generated for this experiment.The gender analysis of all narratives have been made.

The significance of having ten narratives per job instead of a single narrative is that it allows for a more comprehensive and reliable measurement of gender representation. Instead of analyzing gender probability based on a single narrative, aggregating multiple narratives per job reduces variance and mitigates the impact of outliers. This structure provides a broader context for gender distributions, enabling a more robust analysis of overall trends in Gemini AI.

4.3.2 Methodology

Gender Probability Computation. For each narrative, gender detection was performed using Regular expression matching. The probability of male and female references was computed as:

$$P_{\text{Male}} = \frac{\text{Number of male references}}{\text{Total gender references}} \quad (1)$$

$$P_{\text{Female}} = \frac{\text{Number of female references}}{\text{Total gender references}} \quad (2)$$

KL divergence was computed to measure how much the observed gender distributions deviate from an assumed uniform distribution ([0.5, 0.5]). The KL divergence formula is:

$$D_{\text{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (3)$$

where:

- $P(i)$ is the observed gender probability distribution in the narratives generated per job.
- $Q(i)$ is the uniform distribution ([0.5, 0.5]), this is considered as unbiased distribution. To avoid log(0) errors, any zero probabilities were replaced with a small value ($1e^{-10}$).

Combined KL Divergence.

To obtain an overall measure of divergence, the mean KL divergence across all narratives for all jobs was computed:

$$D_{\text{KL}}^{\text{combined}} = \frac{1}{N} \sum_{i=1}^{N} D_{\text{KL}}(P_i\|Q) \quad (4)$$

where N is the total number of jobs in the dataset.

4.3.3 Observations

The computed KL divergence values indicate how much the gender representation in the narratives deviates from an equal distribution. The KL divergence values vary significantly across different narratives for same job, ranging from near-zero values (indicating a nearly balanced distribution) to 0.693147 (indicating extreme imbalance where one gender dominates entirely).

A significant portion of the dataset exhibits KL divergence close to 0.693147, meaning that those narratives have strong gender imbalances. The presence of multiple narratives with values near 0.020136 suggests some instances of near-equal gender representation, but they are less frequent.

The overall combined KL divergence is computed as:

$$D_{\text{KL}}^{\text{combined}} = 0.4203 \quad (5)$$

This indicates a moderate divergence from an equal gender distribution. While narratives for same job maintain a balanced

representation, a significant portion of the dataset skews toward one gender, creating an overall imbalance. The KL divergence metric effectively quantifies this imbalance and provides insights into gender trends in textual data.

REFERENCES

1. Magnus Chukwuebuka Ahuchogu, Gabriella Folashade Akenn Musa, Eric Howard, and Kashmira Mathur. 2025. Ai and bias in recruitment: Ensuring fairness in algorithmic hiring. *Journal of Informatics Education and Research*, 5(3). Published July 18, 2025.
2. Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornogra- phy, and malignant stereotypes. *Preprint*, arXiv:2110.01963.
3. Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
4. Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.
5. J.K Chung, P.L Kannappan, C.T Ng, and P.K Sahoo. 1989. Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications*, 138(1):280–292.
6. Google DeepMind. 2024. Gemini 1.5 flash: Lightweight, efficient, and scalable large language model. Available at: https://deepmind.google/technologies/gemini.
7. S. A. Devisetti. 2024. Bias in text generative open ai. *Indian Journal of Artificial Intelligence and Neural Networking*, 4(2):8–10.
8. Malika Dikshit, Houda Bouamor, and Nizar Habash. 2024. Investigating gender bias in STEM job advertisements. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 179–189, Bangkok, Thailand. Association for Computational Linguistics.
9. Emilio Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1).
10. Solomon Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.
11. Sihang Li, Kuangzheng Li, and Haibing Lu. 2023. National origin discrimination in deep-learning-powered automated resume screening. *Preprint*, arXiv:2307.08624.
12. Dena F. Mujtaba and Nihar R. Mahapatra. 2024. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions. *Preprint*, arXiv:2405.19699.
13. Gideon Popoola, Khadijat-Kuburat Abdullah, Gerard Shu Fuhnwi, and Janet Agbaje. 2024. Sentiment analysis of financial news data using tf-idf and machine learning algorithms. pages 1–6.
14. Ravender Singh Rana. 2023. Job dataset.
15. Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2017. Spine: Sparse interpretable neural embeddings. *Preprint*, arXiv:1711.08792.
16. Titus von der Malsburg, Till Poppels, and Roger P. Levy. 2020. Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 united states and 2017 united kingdom elections. *Psychological Science*, 31(2):115–128. PMID: 31913768.
17. Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. *Preprint*, arXiv:2407.20371.
18. Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. 2024. Bias in generative ai. *Preprint*, arXiv:2403.02726.