Assessing and Mitigating Social Bias in Text and Natural Language Processing Systems

by

Lei Ding

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

This thesis presents a comprehensive exploration of social biases embedded within texts and Natural Language Processing (NLP) models. It develops innovative algorithms to evaluate and mitigate these biases, thereby enhancing the fairness and effectiveness of NLP applications. The initial phase of the research introduces a novel method for reducing gender bias in static word embeddings, meticulously designed to preserve maximum semantic integrity and explainability. This approach not only achieves state-of-the-art results in gender debiasing tasks but also enhances performance in word similarity evaluations and various downstream NLP tasks.

Expanding the scope, subsequent sections delve into broader evaluations of social biases. A new evaluation framework employing Masked Language Models is introduced, which quantitatively assesses social bias using validated inventories of social cues and words, enabling a systematic linguistic analysis. This framework was applied in a large-scale evaluation of the ChatGPT model in high-stakes environments such as the job market. Our findings reveal how the increasing use of generative AI by both employers and job seekers can reinforce gender and social disparities through biased language.

The final section proposes a statistical hypothesis-testing framework to detect biases in texts generated by MLMs. This unsupervised approach uses sentence perturbation techniques to facilitate effective bias testing across various linguistic contexts. Empirical validation confirms its ability to identify subtle biases, enhancing the framework's practical utility and effectiveness.

Together, these investigations provide a series of comprehensive, effective, and

efficient algorithms for studying social bias in textual contexts. They offer valuable insights and practical tools for future researchers and significantly advance the state of the art in NLP research. This thesis contributes to academic knowledge and represents a crucial step toward creating more equitable technological solutions in language processing.

Preface

The research conducted for this thesis is part of a series of collaborative projects under the supervision of Professor Linglong, Kong and Professor Bei, Jiang.

Chapter 2, is titled "Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, & Bei Jiang. (2022). Word embeddings via causal inference: Gender bias reducing and semantic information preserving" has been published in Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11864-11872). My contributions to this paper include the development of the general idea, writing, coding, and numerical experiments. Other authors contributed to the paper's discussion, formatting, and proofreading.

Chapter 3, titled "Lei Ding, Yang Hu, Nicole Denier, Enze Shi, Junxi Zhang, Qirui Hu, Karen D. Hughes, Linglong Kong, & Bei Jiang. Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach" was submitted to Neural Information Processing Systems 2024. My involvement in this paper consists of proposing the main algorithm and idea, writing the majority of the content, and developing all the coding and experiments. Yang Hu, Nicole Denier, and Karen D. Hughes helped with part of the related works on sociology and the social impact part. Enze Shi, Junxi Zhang, and Qirui Hu helped with the discussion by providing theoretical proofing.

Chapter 4, which will be submitted to ACL 2025, is titled "Lei Ding, Qirui Hu, Linglong Kong & Bei Jiang. A Statistical Testing Framework for Bias Word Detection with Masked Language Models." For this paper, I propose the idea of bias word detection with hypothesis testing. I and Qirui Hu designed the whole framework.

Additionally, I completed all coding and numerical experiments and wrote the majority of the paper. Qirui Hu aided in the discussion and the self-normalization part.

*To my family: My **wonderful** Mother and Father, my **cherished** GrandFathers and GrandMothers, my **dearest** eldest aunt, and second aunt, my uncles, and my brothers and sisters. Their unwavering love and support have been my cornerstone.*

To Haizhou: no need to say more.

Acknowledgements

I want to start by thanking my supervisors, Prof. Linglong Kong and Prof. Bei Jiang, for bringing me on the journey of academic life. Their continuous guidance, support, and motivation during my research. Their expertise, insightful advice, and feedback have been instrumental in shaping my thesis and ensuring its completion. They have been extraordinary mentors, inspiring me to strive for academic excellence.

I also extend my appreciation to my colleagues Yi Liu, Qirui Hu, Enze Shi, Xiaotian Chang, Yangdi Jiang, Dengdeng Yu, and Junxi Zhang. Their valuable contributions have greatly enriched both my academic endeavors and personal growth. A special thanks to all members of the BIAS project: Yang Hu, Nicole Denier, Alla Konnikov, Karen D. Hughes, Shenggang Hu, and Hongsheng Dai. Their knowledge, dedication, and innovative perspectives have profoundly enhanced our methodology and analysis.

Furthermore, my heartfelt thanks go to my high school classmates, my friends in Beijing, my friends in Minnesota, my friends in Edmonton, and my friends from around the world, including those I hold dear. Your supports mean the world to me.

Love until death and unchanging.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The rapid advancement of Artificial Intelligence (AI) is transforming society in profound ways. Among the most significant developments are in Natural Language Processing (NLP) and Large Language Models (LLMs), which have revolutionized how machines process, understand, and generate human-like text. These technologies are increasingly integrated into systems that influence our social interactions. Despite their success, there is a critical issue: these models can learn, perpetuate, and even amplify harmful social biases, potentially exacerbating inequalities and impacting users and society at large. This thesis addresses this pressing challenge by exploring methods to mitigate bias within NLP models, evaluate bias in textual settings, and develop a comprehensive framework for bias assessment.

At the foundation of all NLP and LLM systems is the word embedding—the numerical representation that encapsulates the meaning of word tokens. Given its fundamental role, it is crucial to examine and address biases at this initial stage. In the first paper of this thesis, we introduce a causal framework aimed at reducing bias within static word embeddings. Our comprehensive experiments demonstrate that this method not only achieves state-of-the-art results in gender debiasing tasks but also enhances performance in word similarity assessments and various downstream NLP applications.

We then shift our focus to evaluating biases in generative language models, such as

ChatGPT, which are gaining widespread use across multiple sectors. The potential of these models to propagate and amplify social biases, especially in high-stakes settings like the job market, presents a significant concern. Our research employs a novel experimental design to analyze social biases in ChatGPT-generated job applications in response to actual job postings. By simulating the job application process, we uncover language patterns and biases, introducing a novel evaluation framework that utilizes Masked Language Models (MLMs) to quantitatively assess social biases using validated social cue inventories. Our findings reveal how the increasing use of generative AI by both employers and job seekers could reinforce gender and social disparities through biased language.

Finally, we propose a novel statistical hypothesis-testing framework to detect biases in textual content generated by MLMs. This unsupervised approach uses sentence perturbation techniques to create robust datasets from individual text instances, facilitating practical and effective bias testing. The framework is versatile, incorporating various bias measurement and variance calculation methods to suit different linguistic contexts and research needs. Empirical validation with real-world data confirms the framework's ability to identify subtle biases, underscoring its utility and effectiveness.

Collectively, these studies enhance our understanding of and provide innovative solutions for addressing bias in NLP. They offer valuable insights and practical tools for data analysts and fairness researchers, contributing significantly to the fields of AI ethics and social responsibility. This thesis not only advances the state of the art in NLP bias mitigation but also equips developers and researchers with actionable methods to enhance the fairness and transparency of AI technologies.

Chapter 2

Debias Static Word Embedding

2.1 Abstract

With widening deployments of Natural Language Processing (NLP) in daily life, inherited social biases from NLP models have become more severe and problematic. Previous studies have shown that word embeddings trained on human-generated corpora have strong gender biases that can produce discriminative results in downstream tasks. Previous debiasing methods focus mainly on modeling bias and only implicitly consider semantic information while completely overlooking the complex underlying causal structure among bias and semantic components. To address these issues, we propose a novel methodology that leverages a causal inference framework to effectively remove gender bias. The proposed method allows us to construct and analyze the complex causal mechanisms facilitating gender information flow while retaining oracle semantic information within word embeddings. Our comprehensive experiments show that the proposed method achieves state-of-the-art results in gender-debiasing tasks. In addition, our methods yield better performance in word similarity evaluation and various extrinsic downstream NLP tasks.

2.2 Introduction

Word embeddings are dense vector representations of words trained from human-generated corpora [1, 2]. Word embeddings have become an essential part of natural

language processing (NLP). However, it has been shown that stereotypical bias can be passed from human-generated corpora to word embeddings [3–5].

With wide applications of NLP systems to real life, biased word embeddings have the potential to aggravate and possibly cause serious social problems. For example, translating 'He is a nurse' to Hungarian and back to English results in 'She is a nurse' [6]. In word analogy tasks appears in Bolukbasi *et al.* [7], wherein $\overrightarrow{she}$ is closer to $\overrightarrow{nurse}$ than $\overrightarrow{he}$ is to $\overrightarrow{doctor}$. Zhao *et al.* [8] shows that biased embeddings can lead to gender-biased identification outcomes in co-reference resolution systems.

Current studies on word embedding bias reductions can be divided into two camps: word vector learning methods [8] and post-processing algorithms [7, 9]. Word vector learning methods are time-consuming and suffer from the high computational cost required to train word embeddings from scratch. To overcome these limitations, post-processing algorithms have emerged as popular alternatives. Yang and Feng [10], for example, proposes a simple and efficient algorithm that projects embeddings into a space that is orthogonal to gender-specific words such as *mother* and *father* and is successful in reducing gender bias. However, the critical issue of using gender-specific word vectors remains: information on gender and semantics entangled within these words. For example, the gendered word pair *bride* and *bridegroom* exhibit gender information as well as semantic information pertaining to weddings. Therefore, eliminating gender information through pairs of gendered words such as *policeman* and *policewoman* also eliminates intrinsic semantic information: this is clearly not ideal.

As a solution, we propose utilizing the differences between vectors corresponding to paired gender-specific words to better eliminate gender bias while retaining important semantic information. These differences are between embedded vectors for male- and female-gendered words, such as $\overrightarrow{father} - \overrightarrow{mother}$ or $\overrightarrow{bridegroom} - \overrightarrow{bride}$.

As a motivating example[1], Table 2.1 demonstrates that this simple change from

[1]Please refer to the appendix for detail explanation

	Task 1 *Wedding*	Task 2 *Service*	Task 3 *Family*	Task 4 *Religion*
Oracle	11.22 (0.2)	9.96 (0.11)	13.51 (0.3)	20.27 (0.3)
DeSIP	7.01 (0.15)	6.67 (0.10)	10.69 (0.25)	13.59 (0.25)
HSR	4.34 (0.14)	5.61 (0.10)	8.90 (0.22)	9.85 (0.20)
Win	100.00%	99.00%	100.00 %	100.00%

Table 2.1: Semantic information preservation experiment.

gender-specific word vectors to the differences between word-pair vectors indeed retains more semantic information than the state-of-the-art post-processing framework [10].

In this paper, we propose novel causal frameworks for reducing bias in word embeddings while maximally preserving semantic and lexical information. Our contributions are summarized as follows.

- We develop two causal inference frameworks for reducing biases in word embeddings that improve upon existing state-of-the-art methods.
- We find an intuitive and effective way to better represent gender-related information that needs to be removed and use this approach to achieve oracle-like semantic and lexical information retention.
- We show that our methods outperform other *state-of-the-art* debiasing methods in various downstream NLP tasks.

The rest of this paper is organized as follows. We first present a thorough review of current studies on word embedding bias evaluation and debiasing algorithms. We then define two types of bias and propose frameworks for dealing with each. The comprehensive experimental results on a series of gender bias evaluation and semantic evaluation tasks demonstrate the effectiveness of our proposed methods.

2.3 Related Works

2.3.1 Quantifying Gender Bias

Numerous studies have demonstrated that word embeddings trained by human-generated corpora exhibit human stereotype bias. Caliskan *et al.* [3] develops the Word Embedding Association Test (WEAT) as an analogue to the Implicit Association Test used in psychology [11] to detect implicit stereotypes. WEAT measures the association between a word and an attribute using cosine similarity; the test compares two sets of target words against a pair of attribute sets.

Bolukbasi *et al.* [7] applies word analogy tests as a way to demonstrate bias. The task uses a word embedding to find an output to pair with a given input word, say, *doctor*, such that the (target, output) pair is in analogy to the gender pair (he, she). The word embedding passes the test if the output is stereotype-free, say, *physician* instead *nurse* for the input *doctor*. However, this task requires crowd-sourcing to set the benchmark and has been replaced by other evaluation methods in more recent works.

Another approach from Bolukbasi *et al.* [7] for evaluating gender bias involves computing projections onto a gender direction, the difference between vector embeddings of a pair of gender-specific words (e.g. he and she, as the most widely accepted definition). This debiasing metric is used in many other studies [12]. Such a method has failed to become the gold standard because a "true" gender direction if it exists, is used in the evaluation.

Gonen and Goldberg [13] later points out that direct projection does not eliminate gender bias from the geometry of the embedding and that biased words tend to cluster together even after debiasing. To account for this, the neighborhood bias metric was introduced to measure the bias of a word by counting the difference in the number of (socially) male- and female-biased neighbors among the word's K-nearest neighbors.

2.3.2 Prior Debiasing Methods

Current studies on word embedding bias reductions can be divided into two camps: word vector learning methods [8] and post-processing algorithms for instance [7] and [9] and many more. Word vector learning methods require retraining of the word embedding and can be time-consuming due to the retraining of the word embedding. Therefore, most of the works on debiasing word embeddings choose to remove the bias through post-processing, including algorithms like [7, 9, 10, 14–16].

From a technical perspective, we see that Bolukbasi *et al.* [7] formulates the core idea of detecting the subspace that contains the most information related to gender Based on the idea of removing gender subspace, other works have incorporated different strategies, e.g., maximizing the distance between masculine and feminine words [8], detecting gender direction using partial projection [14], or detecting and mitigating distortion in gender direction due to word frequency [15]. Various extensions of [7] are also developed, for instance removing bias with respect to multiclass attributes (like ethnic) [12] or debiasing multilingual word embeddings [17].

More recent works [10, 16] have considered the problem beyond just detecting and removing gender direction from gender-neutral word vectors. Shin *et al.* [16] models a word vector as a sum of two components, each containing latent gender information and semantic information respectively. An autoencoder is trained to disentangle these two components and gender-neutral words are debiased using a counterfactual copy of itself, i.e. a synthesized word vector with the same semantic component but biased in the other gender direction.

Similarly, Yang and Feng [10] approaches the problem using a causal framework in which it is assumed that latent gender information affects both gendered and gender-biased words. The model aims to recover gender-specific information in gender-biased words from the gendered words through a linear ridge regression. In comparison, the causal framework used in our approach not only distinguishes gender information from

semantic information but also takes into account the potential effect of the former on the latter through causal inference. This causal path from gender information to semantic information is overlooked by the causal model used in [10].

2.4 Methodology

2.4.1 Preliminary Definitions

We characterize two types of gender bias in the causal framework and propose algorithms for removing each type. Specifically, we use model intervention techniques to determine causal effects in a causal model. It is more manageable to apply the model intervention to proxy variables of the gender bias rather than the gender bias variables themselves (represented by the differences between gender-specific word pair vectors, such as $\overrightarrow{he}-\overrightarrow{she}$ or $\overrightarrow{male}-\overrightarrow{female}$), since the latter are generally regarded as inherited attributes for which interventions are often impossible in practice.

We consider five types of variables corresponding to five word-related matrices: an s_1-dimensional pure gender bias variable D with a corresponding matrix $\mathbf{D} \in \mathcal{R}^{N\times s_1}$ composed of pure gender bias vectors such as $\overrightarrow{he}-\overrightarrow{she}$ and $\overrightarrow{male}-\overrightarrow{female}$; an s_2-dimensional gender bias variable proxy P with a corresponding matrix $\mathbf{P} \in \mathcal{R}^{N\times s_2}$ composed of vectors that are directly influenced by D that should not affect the final prediction; an m-dimensional resolving, non-gender-specific word variable Z with a corresponding matrix $\mathbf{Z} \in \mathcal{R}^{N\times m}$ composed of vectors that are influenced by D in a manner that we accept as non-discriminatory; a d-dimensional, non-gender-specific word variable Y with a corresponding matrix $\mathbf{Y} \in \mathcal{R}^{N\times d}$ composed of word vectors potentially containing gender bias that needs to be removed, such as $\overrightarrow{nurse}$ and $\overrightarrow{engineer}$; and another p-dimensional, non-gender-specific word variable X with a corresponding matrix $\mathbf{X} \in \mathcal{R}^{N\times p}$ that may retain semantic information. Here N is the dimension of the word embedding vector, and s_1, s_2, m, d, and p are the sizes of the variables D, P, Z, Y and X, respectively.

It is clear that using the vectors in $\mathbf{D}$ can eliminate pure gender bias information contained in word embeddings. In this way, semantic information can be preserved. As shown in Figures 2.1 2.2 and 2.3 2.4, we generally allow influence along the pathway $D \to X \to Y$ in our framework. Motivated by Kilbertus *et al.* [18] and these conventions, we introduce the following definitions.

Definition 2.1 *(Potential proxy bias.) A variable Y in a causal graph exhibits potential proxy bias if there exists a directed path from D to Y that is blocked by a proxy variable P and if Y itself is not a proxy.*

This definition indicates that potential proxy bias from P articulates a causal criterion that is in a sense dual to unresolved bias from Z.

Definition 2.2 *(Unresolved bias.) A variable Y in a causal graph exhibits unresolved bias if there exists a directed path from D to Y that is not blocked by a resolving variable Z and Y itself is non-resolving.*

This definition implies that all paths from a gender-bias variable D are problematic unless they are justified by a resolving variable Z.

2.4.2 Removing Potential Proxy Bias

We now develop a practical procedure for removing proxy bias in a linear structural equation model. For each $\mathbf{y} \in \mathcal{R}^N$, the column vector of $\mathbf{Y}$, it can be decomposed into two parts as $\mathbf{y} = \mathbf{y}_\Delta + \mathbf{y}_{\Delta^\perp}$, where $\mathbf{y}_\Delta$ and $\mathbf{y}_{\Delta^\perp}$ are the projections of $\mathbf{y}$ onto the mutually orthogonal spaces Δ and $\Delta^\perp$. In particular, let $\boldsymbol{\phi}_j \in \mathcal{R}^N$ denote the basis vectors for Δ and $\boldsymbol{\psi}_{j'} \in \mathcal{R}^N$ denote the basis vectors for $\Delta^\perp$. The whole space $\Omega = \Delta \cup \Delta^\perp$. We can write $\mathbf{y} = \sum_{j:\boldsymbol{\phi}_j \in \Delta} \xi_j \boldsymbol{\phi}_j + \sum_{j':\boldsymbol{\psi}_{j'} \in \Delta^\perp} \kappa_{j'} \boldsymbol{\psi}_{j'}$, where $\xi_j, \kappa_{j'} \in \mathcal{R}$. In this paper, we take $\Delta = \text{Span}(\mathbf{D})$, namely, the linear space spanned by the column vectors of $\mathbf{D}$. Consequently, $\Delta^\perp$ contains the semantic information not described by $\mathbf{D}$. As bias reduction is primarily concerned with reducing bias along paths starting from D, we do not remove information from $\mathbf{y}_{\Delta^\perp}$.

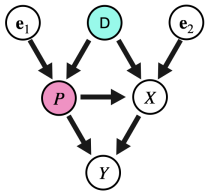

Figure 2.1: Proxy bias

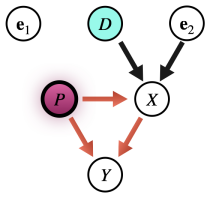

Figure 2.2: Intervention on proxy bias

We next propose an algorithm for debiasing non-gender-specific word vectors $\mathbf{y}$. As illustrated in Figure 2.1, 2.2 the corresponding linear structural equations are

$$\begin{aligned} \mathbf{P} &= \mathbf{D}\boldsymbol{\alpha}_0 + \boldsymbol{e}_1 \\ \mathbf{X} &= \mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{P}\boldsymbol{\alpha}_2 + \boldsymbol{e}_2 \\ \mathbf{Y} &= \mathbf{P}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2, \end{aligned} \tag{2.1}$$

where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are unobserved errors and $\boldsymbol{\alpha}_0 \in \mathcal{R}^{s_1 \times s_2}$, $\boldsymbol{\alpha}_1 \in \mathcal{R}^{s_1 \times p}$, $\boldsymbol{\alpha}_2 \in \mathcal{R}^{s_2 \times p}$, $\boldsymbol{\beta}_1 \in \mathcal{R}^{s_2 \times d}$ and $\boldsymbol{\beta}_2 \in \mathcal{R}^{p \times d}$ are parameters. Here, we note that the proxy matrix $\mathbf{P}$ contains vectors of words that are direct descendants of $\mathbf{D}$ and should not affect the prediction of $\mathbf{Y}$. In this paper, we pre-specify $\mathbf{P}$ using the gendered-word pairs listed in Zhao *et al.* [8]. We build predictors that remove proxy bias by intervening on P, that is, by setting $P = p'$, where p' is a random variable: this is similar to the approach in Kilbertus *et al.* [18]. In particular, we want to guarantee that P has no overall influence on the prediction of the non-gender-specific variable Y by adjusting the $P \to Y$ pathway to cancel the influence along $P \to X \to Y$. We do

Algorithm 1 (P-DeSIP) Removing potential proxy bias.

Input: $\mathbf{D}$, $\mathbf{P}$, $\mathbf{X}$, $\mathbf{Y}$.

1: Solve $\mathbf{X} = \mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{P}\boldsymbol{\alpha}_2$ by PLS to get $(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2)$

2: Solve $\mathbf{Y} = \mathbf{P}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2$ by PLS to get $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$

3: Compute $\widehat{\mathbf{Y}} = (\mathbf{X} - \mathbf{P}\widehat{\boldsymbol{\alpha}}_2)\widehat{\boldsymbol{\beta}}_2$

4: Compute $\widehat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$

5: Compute $\widehat{\mathbf{Y}}_{\text{P-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$

Output: $\widehat{\mathbf{Y}}_{\text{P-DeSIP}}$ as the debiased word matrix.

not generally prohibit the potential for the gender bias variable D to influence the non-gender-specific variable Y in this case: see Figure 2.1, 2.2. The non-gender-specific word matrix $\widehat{\mathbf{Y}}$ with potential proxy bias removed is[2]

$$\widehat{\mathbf{Y}} = (\mathbf{X} - \mathbf{P}\widehat{\boldsymbol{\alpha}}_2)\widehat{\boldsymbol{\beta}}_2, \tag{2.2}$$

where the parameters $\widehat{\boldsymbol{\alpha}}_2$ and $\widehat{\boldsymbol{\beta}}_2$ are estimated by partial least squares (PLS), a supervised dimension reduction method that works particularly well when variable dimensionality is very large [19] and becomes a popular tool in various scientific areas in recent years [20]. However, since the debiasing procedure above does not retain any information of $\mathbf{Y}_{\Delta^\perp}$ since $\widehat{\mathbf{Y}}$ is a descendant of $\mathbf{D}$, we must find a way to restore the information of $\mathbf{Y}_{\Delta^\perp}$.

In particular, we propose obtaining a least-squares estimate $\widehat{\mathbf{Y}}_\Delta$ of $\mathbf{Y}_\Delta$ through multivariate linear regression of $\mathbf{Y}$ on $\mathbf{D}$. We then use the residual $\widehat{\mathbf{Y}}_{\Delta^\perp}$ as an estimate of $\mathbf{Y}_{\Delta^\perp}$. Finally, we compute $\widehat{\mathbf{Y}}_{\text{P-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$ as the bias-reduced version of $\mathbf{Y}$. This post-processing algorithm is formally presented in Algorithm 1.

In practice, when the dimensionality of $\mathbf{X}$ is extremely high, the computational cost of this algorithm becomes a concern. With this in mind, we introduce a preliminary screening step to reduce ultrahigh dimensionality to a moderate level before conducting a refined analysis. Before conducting a simple screening procedure using correlation learning, each column of $\mathbf{X}$ and $\mathbf{Y}$ are standardized to a mean of zero and a standard deviation of one. Inspired by Fan and Lv [21] and Xie *et al.* [22], we propose the

[2]Please refer to appendix for detail derivation

following marginal screening utility to measure the dependence between $\mathbf{Y}$ and the columns $\mathbf{x}_k$ $(k = 1, \ldots, p)$ of $\mathbf{X}$: $\tau_k = \max_{j=1,\ldots,d} |\mathbf{x}_k^\top \mathbf{y}_j|/N$, where $\mathbf{y}_j$ $(j = 1, \ldots, d)$ denotes the j-th column of $\mathbf{Y}$. We propose ranking $\mathbf{x}_k$ by sorting τ_k from largest to smallest. We denote the reduced non-gender-specific word matrix by $\mathbf{X}_{\widehat{\mathcal{M}}}$, where $\widehat{\mathcal{M}} = \{k : \tau_k \geq \gamma_n\}$ and γ_n is a pre-specified threshold value.

2.4.3 Removing Unresolved Bias

We take a similar approach to remove unresolved bias when a proxy gender bias matrix $\mathbf{P}$ is not attainable. We consider the resolving non-gender-specific word matrix $\mathbf{Z} \in \mathcal{R}^{N \times m}$ that directly affects $\mathbf{X}$ instead of the proxy bias matrix $\mathbf{P}$: this is illustrated in Figure 2.3, 2.4.

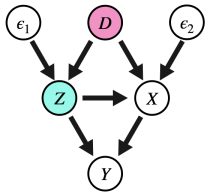

Figure 2.3: Unresolved bias

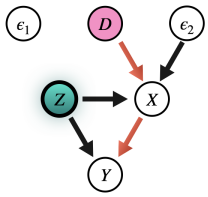

Figure 2.4: Intervention on unresolved bias

Resolving variables are influenced by $\mathbf{D}$ in a manner that we accept as non-discriminatory: therefore, $\mathbf{Z}$ is chosen to directly affect $\mathbf{X}$ and have some correlation with $\mathbf{D}$. In particular, we choose $\mathbf{Z}$ containing the adjectives and nouns correlated to $\mathbf{D}$ based on mean cosine similarity, while $\mathbf{X}$ includes the words that are otherwise

contained by $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{D}$. Since all adjectives in English have an adverb form, this ensures that the path from $\mathbf{Z}$ to $\mathbf{X}$ exists.

The causal dependencies in the corresponding linear structural equation model are equivalent to those in Figure 2.1, 2.2 for potential proxy bias:

$$\begin{aligned} \mathbf{Z} &= \mathbf{D}\boldsymbol{\gamma}_0 + \boldsymbol{\epsilon}_1 \\ \mathbf{X} &= \mathbf{D}\boldsymbol{\gamma}_1 + \mathbf{Z}\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}_2 \\ \mathbf{Y} &= \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are unobserved errors and $\boldsymbol{\gamma}_0 \in \mathcal{R}^{s_1 \times m}$, $\boldsymbol{\gamma}_1 \in \mathcal{R}^{s_1 \times p}$, $\boldsymbol{\gamma}_2 \in \mathcal{R}^{m \times p}$, $\boldsymbol{\theta}_1 \in \mathcal{R}^{m \times d}$, and $\boldsymbol{\theta}_2 \in \mathcal{R}^{p \times d}$ are parameters. We can proceed as before by intervening on Z, that is, by setting $Z = z'$. In this case, we want to cancel the remaining information from D to Y by intervening on Z: Figure 2.3, 2.4 illustrates this procedure. The non-gender-specific word matrix $\widehat{\mathbf{Y}}$ with unresolved bias removed is

$$\widehat{\mathbf{Y}} = \mathbf{Z}\widehat{\boldsymbol{\theta}}_1. \tag{2.4}$$

This debiasing procedure does not retain any information of $\mathbf{Y}_{\Delta^\perp}$. Therefore we restore the information from $\mathbf{Y}_{\Delta^\perp}$ by taking a similar way to the previous procedure.

2.5 Experiments

In this section, we compare the proposed methods against other debiasing algorithms in a set of comprehensive experiments. Our results show that the proposed methods not only reduce bias in various evaluation tasks, but also enhance the performance

Algorithm 2 (U-DeSIP) Removing unresolved bias.

Input: $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{Y}$.

1: Solve $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2$ by PLS to get $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$
2: Compute $\widehat{\mathbf{Y}} = \mathbf{Z}\widehat{\boldsymbol{\theta}}_1$
3: Compute $\widehat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$
4: Compute $\widehat{\mathbf{Y}}_{\text{U-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$

Output: $\widehat{\mathbf{Y}}_{\text{U-DeSIP}}$ as the debiased word matrix.

of word embeddings in semantic evaluation tasks. Our debiasing methods outperform in downstream part-of-speech (POS) tagging, POS chunking, and named-entity recognition tasks.

We apply the proposed debiasing methods to 300-dimensional GloVe embeddings pre-trained on English Wikipedia data with 322,636 unique words [2]. As baselines, we also compare our results against previous state-of-the-art debiasing methods, including the hard-debiasing method (Hard) [7], the gender-preserving debiasing method (GP) [9], word vector learning method (GN) [8], and the half-sibling regression debiasing method (HSR) [10]. For a fair comparison, we utilize the other authors' implementations. [3]

To separate the words in the following experiments, we manually pick 11 pairs of pure gender words such as (*he*, *she*) and (*him*, *her*)[4]. We form $\mathbf{D}$ using the differences between the vector embeddings corresponding to these word pairs. We similarly compute $\mathbf{P}$ using the gendered word pairs listed in Zhao *et al.* [8]. The words represented in $\mathbf{P}$ contain significant non-gender-related information and gender-related information, e.g., *bride* and *bridegroom*. We choose the 50,000 most frequent words in GloVe to form $\mathbf{Y}$, which contains the words to be debiased, following the evaluation procedure in Gonen and Goldberg [13]; $\mathbf{X}$ is formed using the remaining words. In all of the below experiments, we use a fixed screening parameter $\gamma_n = 0.92$ in P-DeSIP and $\gamma_n = 0.80$ in U-DeSIP.

2.5.1 Quantitative Evaluation for Bias Tasks

Throughout this section, the top N gender-biased words are chosen by evaluating dot products with the gender direction $\overrightarrow{he} - \overrightarrow{she}$ in the original word embedding (i.e. GloVe) and choosing the most positive and negative values as the most male- and female-biased words, respectively.

[3] https://github.com/Lei-Ding07/Word_Debias_DeSIR

[4] See the accompanying appendix for details of word list

Bias-by-projection Task.

Bias-by-projection uses the dot product between the gender direction $\overrightarrow{he} - \overrightarrow{she}$ and the word to be tested. We compute and average the absolute projection bias of the top 50,000 most frequent words.

The first column of Table 2.2 shows that our methods achieve very good results. Its performance is just below that of Hard-GloVe, which can be explained by the fact that Hard-Glove is trained by removing projections along the gender direction.

Sembias Analogy Task.

The SemBias test was first introduced in Zhao *et al.* [8] as a set of word analogy tests. The task is to find the word pair in best analogy to the pair (*he*, *she*) among four options: a gender-specific word pair, e.g., (*waiter*, *waitress*); a gender-stereotype word pair, e.g., (*doctor*, *nurse*); and two highly-similar, bias-free word pairs, e.g. (*dog*, *cat*). The dataset contains 440 instances, of which 40 instances, denoted by SemBias(subset), are not used during training. We report accuracy in identifying gender-specific word pairs.

The second and third columns of Table 2.2 quantify accuracy in identifying gender-specific word pairs. Our P-DeSIP methods achieve very good performance in both tasks. Specifically, in the subset test, P-DeSIP outperforms GloVe by almost 40%.

	Bias-by-projection	SemBias	SemBias (subset)
GloVe	0.0375	0.8023	0.5750
Hard	**0.0007**	0.8250	0.3250
GP	0.0366	0.8432	0.6500
GN	0.0555	**0.9773**	<u>0.7500</u>
HSR	0.0218	0.8591	0.1000
P-DeSIP	<u>0.0038</u>	<u>0.9523</u>	**0.9750**
U-DeSIP	<u>0.0038</u>	0.9090	0.5000

Table 2.2: Gender-direction-related task performance. In each column, the best and second-best results are boldfaced and underlined, respectively.

Clustering Male- and Female-biased Words.

As noted in Gonen and Goldberg [13], biased words tend to cluster together. Even some debiased embeddings were unable to escape from this phenomenon. Here we take the top 500 male-biased words and the top 500 female-biased words and partition them via K-means clustering (K=2) [23]. Accuracy in splitting the 1,000 words into male and female clusters is presented in Table 2.3. Our methods achieve the best performance among all other methods.

Correlation between Bias-by-projection and Bias-by Neighbors.

Taking again the top 50,000 most frequent words as targets, we compute the Pearson correlation coefficient between the bias-by-projection and bias-by-neighbor results. The latter is computed using the neighborhood metric, which counts the percentage of male- and female-biased words within the K-nearest neighbors of each target word [13, 15]. Here, we take $K = 100$. Referring to the second column of Table 2.3, our methods generally achieve the best performance.

Bias-by-neighbors for Profession Words.

In this task, we assess the effect of debiasing by calculating the correlation between bias-by-neighbor measures before and after debiasing. We use the neighborhood metric, as in the previous task, but we restrict our targets to the list of professional words in Bolukbasi *et al.* [7] and Zhao *et al.* [8]. Results, in the third column of Table 2.3, show that our methods outperform GloVe and are comparable to HSR-GloVe.

Classifying Previously Female- and Male-biased Words.

After selecting the top 2,500 biased words for each gender, for each baseline model we train a support vector machine (SVM) model using 1,000 randomly sampled words. This classifier is then applied to the remaining 4,000 words to predict gender bias direction. Prediction accuracy is shown in the last column of Table 2.3: a lower

accuracy indicates the trained model is unable to capture gender-related information from the original embedding and thus, that the debiasing method is superior. Again, both of our methods outperform the other methods.

	Clustering	Correlation	Profession	Classify
GloVe	1.0000	0.7727	0.8200	0.9980
Hard	0.8050	0.6884	0.7161	0.9068
GP	1.0000	0.7700	0.8102	0.9978
GN	0.8560	0.7336	0.7925	0.9815
HSR	0.9410	<u>0.6422</u>	**0.6804**	0.9055
P-DeSIP	**0.7910**	0.6431	0.7096	**0.8547**
U-DeSIP	<u>0.7920</u>	**0.6421**	<u>0.7060</u>	<u>0.8550</u>

Table 2.3: Gender bias word relation task performance. In each column, the best and second-best results are boldfaced and underlined, respectively.

Word Embedding Association Test (WEAT)

The WEAT test [3] is a permutation-based test that measures bias in word embeddings. We report effect sizes (d) and p-values (p) in our results. The effect size is a normalized measure of how separated two distributions are. A higher value indicates a larger bias between target words with respect to attribute words. The p-values denote whether the bias is significant or not.

We conduct three tests using the Pleasant & Unpleasant (Task 1), Career & Family (Task 2), and Science & Art (Task 3) word sets. We consider male and female names as attribute sets.[5]. As shown in Table 2.4, we achieve results comparable to those for other methods. In two out of three tasks, the p-value is not significant. We also achieve a reasonably small effect size in all three tasks.

[5]All word lists are from Caliskan *et al.* [3]. Because GloVe embeddings are uncased, we use lower case words.

	Task1		Task2		Task3	
	p	d	p	d	p	d
GloVe	0.090*	0.704	0.000	1.905	0.026	0.987
Hard	0.363*	**0.187**	0.000	1.688	0.583*	-0.104
GP	0.055*	0.832	0.000	1.909	0.025	0.997
GN	0.157*	0.541	0.074*	**0.753**	0.653*	-0.222
HSR	0.265*	0.340	0.000	1.555	0.410*	0.122
P-DeSIP	0.755*	-0.373	0.001	<u>1.459</u>	0.486*	<u>0.019</u>
U-DeSIP	0.732*	<u>-0.335</u>	0.001	1.462	0.491*	**0.012**

Table 2.4: WEAT test result. In each column of p-value, * indicates statistically **non**-significant compare with $\alpha = 0.05$; In each column of d, the best and second-best results are boldfaced and underlined, respectively.

2.5.2 Visualization

In order to visually illustrate that our proposed methods effectively reduce gender bias, we took the top 500 male- and female-biased embeddings and generated a t-SNE projection [24] for all of the baseline embeddings. In Figures 2.5, 2.6, 2.7, 2.8, 2.9 and 2.10 , the two colors in the graphs indicate male- and female-biased embeddings. We can see our two methods more effectively mix up the male- and female-biased embeddings.

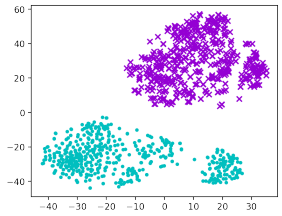

Figure 2.5: t-SNE visualization of GloVe

2.5.3 Word Similarity Tasks

Another important aspect of word embedding is its ability to encode words' semantic information. While bias removal is our main goal, it is unacceptable to disregard how semantic information is influenced by the debiasing process. We next implement several

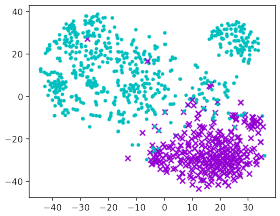

Figure 2.6: t-SNE visualization of Hard-debias

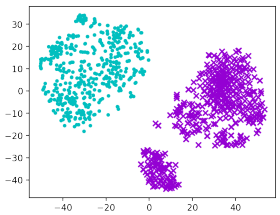

Figure 2.7: t-SNE visualization of GP-debias

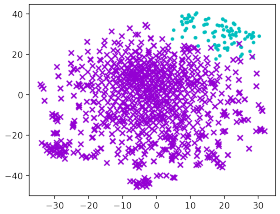

Figure 2.8: t-SNE visualization of HSR

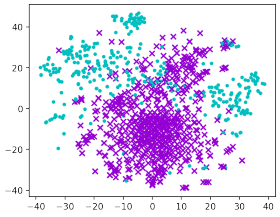

Figure 2.9: t-SNE visualization of P-DeSIP

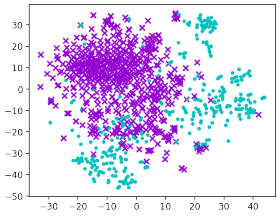

Figure 2.10: t-SNE visualization of U-DeSIP

Embedding Matrix Replacement						
	POS Tagging			POS Chunking		
	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall
Hard	-0.0776	-0.0736	-0.2079	-0.0653	-0.1500	-0.1009
GP	-0.1021	-0.1910	-0.2068	-0.0702	-0.1385	-0.1301
GN	-0.0987	-0.1001	-0.2554	-0.0702	-0.1269	-0.1401
HSR	-0.0666	-0.0589	-0.1820	-0.0377	-0.0753	-0.0689
P-DeSIP	<u>-0.0133</u>	<u>-0.0006</u>	<u>-0.0471</u>	**-0.0108**	**-0.0036**	<u>-0.0346</u>
U-DeSIP	**-0.0107**	**0.0033**	**-0.0405**	<u>-0.0110</u>	<u>-0.0073</u>	**-0.0324**

Table 2.5: Result of downstream tasks for POS Tagging and POS Chunking. Positive value means the task has better performance than using Original GloVe. In each column, the best and second-best results are boldfaced and underlined, respectively.

Embedding Matrix Replacement (continued) and Model Retraining						
	Named Entity Recognition			Model Retraining		
	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall
Hard	-0.0118	-0.0187	-0.0238	-0.0194	<u>0.0078</u>	-0.0741
GP	-0.0353	-0.0366	-0.0871	-0.0071	0.0011	-0.0264
GN	-0.0294	-0.0610	-0.0472	-0.0027	**0.0089**	-0.0174
HSR	-0.0055	-0.0068	-0.0128	-0.0055	-0.0009	-0.0192
P-DeSIP	<u>-0.0014</u>	**0.0002**	<u>-0.0052</u>	<u>-0.0018</u>	0.0002	<u>-0.0068</u>
U-DeSIP	**-0.0007**	<u>0.0013</u>	**-0.0035**	**-0.0010**	0.0000	**-0.0036**

Table 2.6: Result of downstream tasks for Named Entity Recognition and Model Retraining. A positive value means the task has better performance than using Original GloVe. In each column, the best and second-best results are boldfaced and underlined, respectively.

word similarity tests to evaluate our algorithms against existing baseline methods. We consider the following tasks: RG65 [25], WordSim-353 [26], Rarewords [27], MEN [28], MTurk-287 [29], and MTurk-771 [30]. *SimLex-999* [31], and *SimVerb-3500* [32]. These datasets associated with each task contain word pairs and a corresponding human-annotated similarity score.

As an evaluation measure, we compute Spearman's rank correlation coefficient between these two ranks. Results are shown in Table 2.7 and 2.8. We see that our

methods have the leading performance for most of the tasks.

	RG65	WS	RW	MEN
GloVe	0.7540	0.6199	0.3722	0.7216
Hard	0.7648	0.6207	0.3720	0.7212
GP	0.7546	0.6003	0.3450	0.6974
GN	0.7457	0.6286	**0.3989**	0.7446
HSR	0.7764	0.6554	0.3868	0.7353
P-DeSIP	**0.7794**	**0.6856**	<u>0.3970</u>	**0.7484**
U-DeSIP	<u>0.7734</u>	<u>0.6828</u>	0.3956	<u>0.7478</u>

Table 2.7: Word similarity task performance 1. In each column, the best and second-best results are boldfaced and underlined, respectively.

	MT-287	MT-771	SimLex	SimVerb
GloVe	<u>0.6480</u>	0.6486	0.3474	0.2038
Hard	0.6468	0.6504	0.3501	0.2034
GP	0.6418	0.6391	0.3389	0.1877
GN	**0.6617**	0.6619	0.3700	0.2219
HSR	0.6335	0.6652	**0.3971**	**0.2635**
P-DeSIP	0.6452	**0.6741**	<u>0.3765</u>	<u>0.2286</u>
U-DeSIP	0.6455	<u>0.6731</u>	0.3756	0.2273

Table 2.8: Word similarity task performance 2. In each column, the best and second-best results are boldfaced and underlined, respectively.

2.5.4 Downstream Task Utility Evaluation

In order to demonstrate that our de-biased word embeddings still retain good downstream utility and performance, we follow the CoNLL2003 shared task [33] and use POS tagging, POS chunking, and named-entity recognition(NER) as the evaluation tasks. Following Manzini *et al.* [12] we evaluate each task in two ways: embedding matrix replacement and model retraining.

In embedding matrix replacement, we first train the task model using the original biased GloVe vectors and then calculate test data performance differences when using the original biased GloVe embeddings versus other debiased embeddings. Table 2.5

suggests constant performance degradation for all debiasing methods relative to the original embedding. Despite this, our methods outperform all the other tasks (in the sense of minimizing degradation) by a large margin across all the tasks and evaluation metrics (i.e., F1 score, precision, and recall). Furthermore, we even achieve a small improvement in precision on the NER task.

In model retraining, we first train two task models, one using the original biased GloVe embeddings and the other using debiased embeddings. We then calculate differences in test performance. Table 2.6 again suggests that our methods have the closest performance to the model trained and tested using the original GloVe embeddings. Our method also displays the most consistent and comparable performance across the three tasks.

2.6 Conclusion

In this paper, we develop two causal inference methods for removing biases in word embeddings. We show that using the differences between vectors corresponding to paired gender-specific words can better represent and eliminate gender bias. We find an intuitive and effective way to better represent gender information that needs to be removed and use this approach to achieve oracle-like retention of semantic and lexical information. We also show that our methods outperform other debiasing methods in downstream NLP tasks. Furthermore, our methods easily accommodate situations where other kinds of bias exist, such as social, racial, or class biases.

There are several important directions for future work. First, we only consider the linear relationship among the proposed causal inference frameworks. Further investigation is warranted to extend these frameworks to incorporate the non-linear causal relationship [34]. Second, when P is not attainable, we select the resolving variables Z to contain the adjectives and nouns correlated to gender bias variables D. This selection method is rather heuristic. If prior knowledge about resolving variables was introduced, it would surely improve the performance of the unresolved

bias removal. Third, we introduce a residual block to restore the information not retained from the debiasing procedure. The construction of it is rather intuitive and requires more rigorous justification. Finally, although our methods facilitate easy accommodations for situations where other kinds of bias exist, how the proxy and resolving variables as well as the bias variables are properly pre-specified may require non-trivial efforts.

2.7 Appendix

2.7.1 Detail explanation of Table 1

For each of the four pre-determined words *Wedding*, *Service*, *Family*, and *Religion*, we identify the top 200 most cosine-correlated words. For each of the 200 words, we fit a ridge regression against gender-specific words defined in Yang and Feng [10] (HSR), and a linear regression against the differences between gender-specific word pairs from this paper (DeSIP). The fitted word vectors are used as reduced-bias word vectors. To quantify the semantic information preservation, the mean absolute dot product between the pre-determined words and their bias-reduced versions over the 200 most related words are presented, with standard errors in parentheses. Note that, the oracle preservation semantic information is achieved by using the original word vector instead of the fitted one. The last row shows the proportion of these 200 words for which DeSIP outperforms HSR with respect to semantic information preservation.

2.7.2 Pure gender word list of D

Male words: he, him, man, his, himself, son, father, guy, boy, male, men, sons, fathers, guys, boys, males, sir, gentleman, gentlemen, mr

Female words: she, her, woman, hers, herself, daughter, mother, gal, girl, female, women, daughters, mothers, gals, girls, females, madam, lady, ladies, mrs

D is formed by subtraction of each word in Male words with the corresponding word in Female words.

2.7.3 Detail derivation of equation (2) and (4)

We present the details about how to obtain the equations (2) and (4) here as follows:

- Intervene on $\mathbf{P}$ by removing all incoming arrows, see Figure 2.1, 2.2, and set $\mathbf{P} = p'$, where p' is a random variable. Then we obtain:

$$\mathbf{P} = p', \mathbf{X} = \mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{P}\boldsymbol{\alpha}_2 + \boldsymbol{e}_2, \mathbf{Y} = \mathbf{P}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2.$$

- Integrate the first and second equations into the third equation from their structural equations.

$$\mathbf{Y} = p'(\boldsymbol{\beta}_1 + \boldsymbol{\alpha}_2\boldsymbol{\beta}_2) + (\mathbf{D}\boldsymbol{\alpha}_1 + \boldsymbol{e}_2)\boldsymbol{\beta}_2.$$

- Require the distribution of $\mathbf{Y}$ to be independent of p', i.e. for all p_1 and p_2, $\Pr\{p_1(\boldsymbol{\beta}_1+\boldsymbol{\alpha}_2\boldsymbol{\beta}_2)+(\mathbf{D}\boldsymbol{\alpha}_1+\boldsymbol{e}_2)\boldsymbol{\beta}_2\} = \Pr\{p_2(\boldsymbol{\beta}_1+\boldsymbol{\alpha}_2\boldsymbol{\beta}_2)+(\mathbf{D}\boldsymbol{\alpha}_1+\boldsymbol{e}_2)\boldsymbol{\beta}_2\}$, which simply yields $\boldsymbol{\beta}_1 = -\boldsymbol{\alpha}_2\boldsymbol{\beta}_2$. Hence $\mathbf{Y} = (\mathbf{X} - \mathbf{P}\boldsymbol{\alpha}_2)\beta_2$.

- Given the dataset, we estimate the parameters $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_2$ by partial least squares method, denoted the estimators as $\widehat{\boldsymbol{\alpha}}_2$ and $\widehat{\boldsymbol{\beta}}_2$. Then, the equation (2) can be obtained.

Similar to equation (2), we can get equation (4).

- Intervene on $\mathbf{Z}$ by removing all incoming arrows, see Figure 2, and set $\mathbf{Z} = z'$, where p' is a random variable. Then we obtain:

$$\mathbf{Z} = z', \mathbf{X} = \mathbf{D}\boldsymbol{\gamma}_1 + \mathbf{Z}\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}_2, \mathbf{Y} = \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2.$$

- Integrate the first and second equations into the third equation from their structural equations.

$$\mathbf{Y} = z'(\boldsymbol{\theta}_1 + \boldsymbol{\gamma}_2\boldsymbol{\theta}_2) + \mathbf{D}\boldsymbol{\gamma}_1\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}_2\boldsymbol{\theta}_2.$$

- Require the distribution of $\mathbf{Y}$ to be invariant under interventions $\mathbf{D}$, i.e. for all d_1 and d_2, $\Pr\{z'(\boldsymbol{\theta}_1 + \boldsymbol{\gamma}_2\boldsymbol{\theta}_2) + d_1\boldsymbol{\gamma}_1\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}_2\boldsymbol{\theta}_2\} = \Pr\{z'(\boldsymbol{\theta}_1 + \boldsymbol{\gamma}_2\boldsymbol{\theta}_2) + d_2\boldsymbol{\gamma}_1\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}_2\boldsymbol{\theta}_2\}$, which simply yields $\boldsymbol{\theta}_2 = 0$. Hence $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta}_1$.

- Given the dataset, we estimate the parameter $\boldsymbol{\theta}_1$ by partial least squares method, denoted the estimator as $\widehat{\boldsymbol{\theta}}_1$. Then, equation (4) can be obtained.

Chapter 3

Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach

3.1 Abstract

As generative large language models (LLMs) such as ChatGPT gain widespread adoption in various domains, their potential to propagate and amplify social biases, particularly in high-stakes areas such as the labor market, has become a pressing concern. AI algorithms are not only widely used in the selection of job applicants, individual job seekers may also make use of generative LLMs to help develop their job application materials. Against this backdrop, this research builds on a novel experimental design to examine social biases within ChatGPT-generated job applications in response to real job advertisements. By simulating the process of job application creation, we examine the language patterns and biases that emerge when the model is prompted with diverse job postings. Notably, we present a novel bias evaluation framework based on Masked Language Models to quantitatively assess social bias based on validated inventories of social cues/words, enabling a systematic analysis of the language used. Our findings show that the increasing adoption of generative AI, not only by employers but also increasingly by individual job seekers, can reinforce

and exacerbate gender and social inequalities in the labor market through the use of biased and gendered language.

3.2 Introduction

The rapid advancements in generative Large Language Models (LLM) like ChatGPT [35], mark a significant technological shift. These models have not only propelled the field of Natural Language Processing (NLP) but have also found widespread application across numerous sectors [36, 37]. However, as these models are incorporated into social and economic practices, they bring to the fore critical ethical concerns, especially regarding their potential to propagate and amplify existing social biases and attendant inequalities, particularly within high-stakes domains such as the labor market [38].

Recognizing the growing potential of generative AI use in employment practices, our research primarily aims to identify and understand the impact of biases in the application of generative LLM within the labor market. We focus particularly on ChatGPT, investigating how this widely used LLM influences the propagation of biases in job advertising and application processes.

The complexity of *automating* bias evaluation in textual content poses significant challenges. Traditional approaches in social sciences, such as content analysis, often rely on manual word counts from static lists [39], which may miss the subtleties and unlisted language cues that advanced NLP technologies can detect. In addition, by considering words individually, these traditional approaches often fail to capture the contextual meanings that emerge from the interplay of words within entire sentences. To address this limitation and build toward a more solid bias evaluation method, we develop a novel bias evaluation algorithm called **PRISM**: **P**robability **R**anking b**I**as **S**core via **M**asked language model. PRISM involves masking words sequentially within texts and using the Masked Language Models (MLM) [40, 41] to predict the likelihood of alternative tokens, thus allowing us to assess bias with a ranking-based approach that leverages established word lists from social science research to provide

contextual sensitivity, enabling a systematic and detailed analysis of language use.

Additionally, the inherently opaque nature of LLMs like ChatGPT, which function as black boxes without transparent access to their internal structures or parameters, adds another layer of complexity. We propose a method of probing these biases by simulating and analyzing how job seekers use ChatGPT to craft applications (output texts) in response to real job postings (input texts), as illustrated in Figure 3.1. This simulation reveals insights into the biases embedded within ChatGPT's training data and their potential impacts on real-world human resource practices.

Utilizing our PRISM algorithm in tandem with job posting and application text pairs, we explore the correlation between generated content and bias propagation. This comprehensive and novel simulation offers a distinctive lens through which to view how biases might influence the job application process.

In essence, this paper seeks to bridge the gap between rapid technological advancements and the ethical considerations raised by the use of generative LLMs. Through our research, we emphasize the importance of ensuring that AI use promotes core social values of fairness and equality in the labor market as these technologies become increasingly integral to our daily lives.

Our key contributions include:

- We propose **PRISM**, a brand new paradigm for bias evaluation combines with validated word lists capturing directional cues (based on social science research) with MLM to assess biases in texts. It advances existing methods in terms of efficiency, flexibility, robustness as well as theoretical properties.

- We draw on a novel experimental design to probe the black-box of social biases in ChatGPT models to understand both the biases inherent in their training data and their implications for real-world job application scenarios.

- Analysis of bias across four different social dimensions demonstrates inherent biases in job postings are likely reproduced in ChatGPT-generated job applications,

with a tendency for the model to exacerbate and reinforce these biases.

This paper is structured as follows: we first review the current landscape of bias evaluation in NLP and social sciences. Following this, we introduce our bias scoring algorithm and provide experimental evidence supporting our methodology. We conclude with an analysis of job postings and applications mediated by ChatGPT, evaluating our approach's broader applicability and discussing the social implications of our empirical findings.

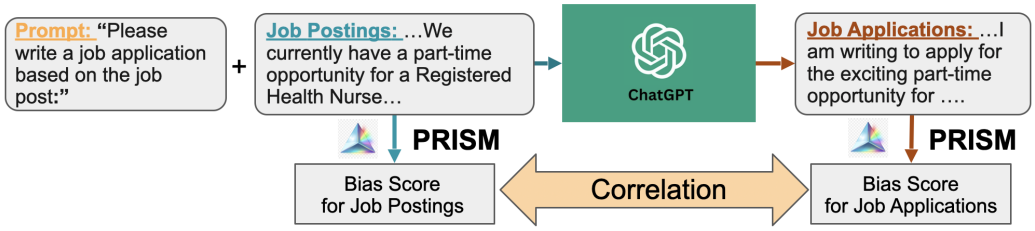

Figure 3.1: Overview of the paradigm for bias probing experimental design.

3.3 Background and Related Works

Bias Evaluation in NLP The evaluation of bias within natural language processing (NLP) presents complex challenges, as methodologies vary significantly across studies [42, 43]. Traditional approaches range from analyzing cosine similarity in word embeddings [44] to diverse methods such as correlation, clustering, classification, and visualization [13, 45, 46]. Recent works have focused on detecting bias in language models that rely on manual sentence templates [47] or creating benchmarks that require high-cost crowd-workers [48, 49] and across various NLP downstream tasks including text classification [50, 51], coreference resolution [52], natural language inference [53], and machine translation [54].

Bias Evaluation for Text The domain of text bias evaluation is notably more challenging than evaluating the NLP models, often requiring extensive human expert intervention or resorting to simplistic and heuristic methodologies. Many existing

approaches are also limited to specific types of bias, making them difficult to adapt to other contexts. Dhamala *et al.* [55] measure bias by computing the cosine similarity of word embeddings [2, 56] with respect to the gender direction ($\overrightarrow{he} - \overrightarrow{she}$) [44] and averaging over sentences. Cryan *et al.* [57] compare a lexicon-based approach and a fine-tuned BERT model with a Crowdsourced label dataset. Spinde *et al.* [58] developed a media bias dataset through costly expert annotation, a process not easily generalizable to other domains. Raza *et al.* [59] explore the use of named entity recognition for detecting biased words within texts. Yet this approach also requires the creation of costly labeled training data for each task and model training.

Labor Market Bias Evaluation in Social Sciences A substantial body of research has documented prevalent gender stereotypes and their role in (re)producing inequalities – gender segregation [60, 61], gender wage/promotion gaps [62], motherhood penalties [63], and fatherhood premiums [64] – in the labor market. Further research shows that gendered language plays a crucial role in maintaining and reproducing gender stereotypes [65]. Psychological studies also show that women and men, given their gender socialization, tend to use and be attracted to different gendered languages and linguistic styles [39]. For example, women tend to employ and identify with a more communal language style, including the use of words related to social and emotional contexts [66]. In contrast, masculine language is typically characterized by a style that highlights agentic traits. Gendered language is found across a wide range of contexts, and in the labor market, it features prominently in job advertisements, the language used in job applications and interviews, as well as performance management processes [67]. While existing research has often focused on gendered language from the labor demand side in terms of, for example, employers' wording of job advertisements [68], far less attention has been paid to the language used by job candidates in response to job advertisements in order to secure a job, despite an increase in individual job seekers' use of ChatGPT. This study thus fills

this important gap by assessing gendered languages from both the labor demand (job advertising) and supply (job application) sides. In doing so, it highlights the relational use of language in the job application process as a quintessential example of social interactions in action. It aims to explore and reveal the extent to which gender biases are present and indeed circulated and exacerbated through the interplay between languages used in job advertisements and job applications.

3.4 Bias Evaluation Algorithm for Text

3.4.1 Motivations

When assessing social bias in textual content, previous methodologies often begin with a straightforward approach: selecting keywords for simple frequency counts. For instance, this might involve comparing the total word count of feminine and masculine words. This technique is prevalent in psychological and sociological studies as described in Section 3.3. More contemporary methods have advanced to include the use of static word embeddings to measure semantic similarities among words, although these approaches still treat each word individually. To go further, researchers need to acquire expensive, labeled training data for specific tasks and do the model training.

In contrast, our objective is to refine and further advance these existing approaches to measuring textual bias with three useful and practical settings:

- Beyond merely analyzing each word individually, the algorithm should aim to consider the contextual meanings of entire sentences, allowing for a more nuanced and comprehensive view of the text.

- The algorithm does not require costly human-labeled training data and circumvents the process of model training or fine-tuning. This aspect is particularly valuable in scenarios where the necessary labeled data is not readily available, allowing for more flexible and scalable applications.

- The algorithm should incorporate established and rigorous word inventories from social science research to guide the bias calculation in a contextually embedded and domain-specific manner (e.g., accounting for specificities of the labor market context). This incorporation of domain knowledge ensures that the assessments are both empirically grounded and contextually salient.

3.4.2 Problem Setup and Algorithm Implementation

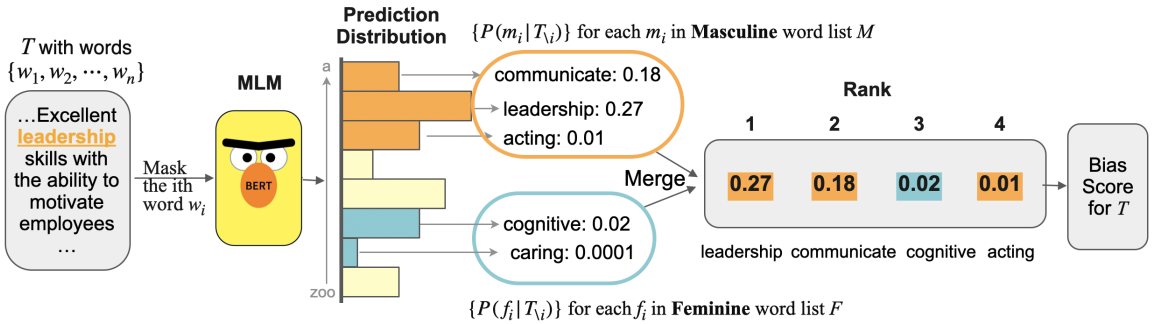

Figure 3.2: An illustration of the paradigm for **PRISM** that uses word lists for directional cues with MLM to compute bias score for text.

In this section, we detail our algorithm under the settings introduced above. Given a text T comprising n words $T = \{w_1, w_2, \ldots, w_n\}$, we iteratively mask each word w_i and input the modified masked text $T_{\setminus i} = \{\ldots, w_{i-1}, [\text{MASK}], w_{i+1}, \ldots\}$ into an MLM, which outputs the probability distribution over the vocabulary for the masked position i, denoted as $P(\cdot \mid T_{\setminus i})$.

Then, to obtain the direction signal for score calculation, we require two predefined word lists representing different contexts—such as gender with a feminine word list $F = \{f_1, \ldots, f_{|F|}\}$ and a masculine word list $M = \{m_1, \ldots, m_{|M|}\}$. For each word in F and M, we obtain the probability from the distribution $P(\cdot|T_{\setminus i})$. This yields two sets of probabilities: $P_F = \{P(f|T_{\setminus i})\}_{f \in F}$ and $P_M = \{P(m|T_{\setminus i})\}_{m \in M}$.

Next, we merge P_F and P_M and filter the probabilities by taking the top α percent of the probabilities, as lower probabilities represent less likely predictions by the MLM and thus contribute minimally to our analysis of bias. This step allows us to focus

on the most influential predictions which significantly determine the context of the sentence.

Finally, we calculate the rank of each probability within this merged list. The rank of probability for each word w_i in the text T using the word lists F and M is denoted as $R_{F\cup M}(P(f|T_{\setminus i}))$ and $R_{F\cup M}(P(m|T_{\setminus i}))$, respectively. And the lower the rank indicates the higher the probability. The bias score for each word w_i is computed as the difference between the mean ranks of the two word lists:

$$S(w_i) = \frac{1}{|F|}\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i})) - \frac{1}{|M|}\sum_{m\in M} R_{F\cup M}(P(m|T_{\setminus i}))$$

A positive score indicates a bias toward a masculine orientation, while a negative score suggests a bias toward a feminine orientation. This differential allows us to detect the direction of the bias, providing deeper insights into how gender nuances are embedded within the language.

Finally, the overall bias score for the text T is the mean of the scores for all words in the text:

$$B(T) = \frac{1}{n}\sum_{i=1}^{n} S(w_i)$$

This score quantifies the bias present in T. By analyzing these scores across various texts, we can assess the extent and direction of linguistic bias present, providing insights into the underlying gender biases conveyed through language. The overall algorithm is illustrated in Figure 3.2, and is detailed in Algorithm 3.

3.4.3 Methodological Benefits of PRISM

Efficiency Our algorithm eliminates the need for costly data labeling and model training. By leveraging predefined word lists developed by existing sociological research, our method avoids the resource-intensive processes associated with supervised learning, such as gathering expert annotations and training models from scratch. This approach not only expedites deployment but also ensures that the algorithm can be scaled

Algorithm 3 PRISM: Probability Ranking bIas Score via Masked language model

Input: Text T with n words $\{w_1, w_2, \ldots, w_n\}$, Word lists F and M
Ouput: Bias score $B(T)$

for each word w_i in T **do**
    Create $T_{\backslash i}$ by masking w_i in T
    Predict distribution $P(\cdot \mid T_{\backslash i})$ using MLM
    Initialize $P_{merged} = [\,]$
    for all words w in $F \cup M$ **do**
        Append $\{w, P(w|T_{\backslash i})\}$ to P_{merged}
    end for
    Sort and filter P_{merged} to retain top $\alpha\%$ of entries
    Calculate ranks for $R_{F\cup M}(P(f|T_{\backslash i}))$ and $R_{F\cup M}(P(m|T_{\backslash i}))$ in the filtered list
    $S(w_i) = \frac{1}{|F|}\sum_{f\in F} R_{F\cup M}(P(f|T_{\backslash i})) - \frac{1}{|M|}\sum_{m\in M} R_{F\cup M}(P(m|T_{\backslash i}))$
end for
$B(T) = \frac{1}{n}\sum_{i=1}^{n} S(w_i)$
return $B(T)$

and adapted swiftly and economically, making it highly practical for researchers and practitioners needing quick and reliable bias assessments in various settings.

Computational Flexibility The inherent flexibility of our method allows for the evaluation of bias across various dimensions simply by altering the word list cues. This adaptability means that different types of bias can be assessed without the need to relabel data or retrain models, significantly reducing the time and resources required for analysis. Whether exploring gender, race, age, or any other form of bias, our algorithm can adjust to new research questions with minimal adjustments. This also allows for the incorporation of substantively meaningful domain-specific word inventories from social science disciplines such as sociology, management studies, psychology, etc.

Robustness Robustness in our method is two-fold. Firstly, we utilize ordinal measurements of word probabilities, focusing on relative positions (ranking) rather than absolute values. This method effectively mitigates issues arising from the predominance of low probabilities within a large pool of candidate words, which can

lead to nonsensical outcomes. Secondly, our approach ensures robust results across different MLMs. Unlike other scoring methods using raw probabilities for calculation, our rank-based bias score method remains consistent even when different MLMs produce varying output probabilities. This dual approach minimizes the influence of outliers and maintains reliability across various computational models.

Theoretical Properties Moreover, we can test whether MLM's predictions have the same distribution on two word lists (M and F). Consider two groups of probabilities, $\{P(f|T_{\backslash i})\}_{f\in F}$ and $\{P(m|T_{\backslash i})\}_{m\in M}$, representing the probability distributions P_F and P_M. The rank sums, denoted by $\sum_{f\in F} R_{F\cup M}(P(f|T_{\backslash i}))$ and $\sum_{m\in M} R_{F\cup M}(P(m|T_{\backslash i}))$ respectively, allow us to test the hypotheses $H_{i0} : P_F = P_M$ vs. $H_{i1} : P_F \neq P_M$. The null hypothesis holds if there is no statistically significant bias toward masculine or feminine language in a particular word w_i. The following theorem provides a rigorous formulation of the test statistic and its asymptotic result.

Theorem 3.1 *When $|F|$ and $|M|$ are large, for each $i \in [n]$, under H_{i0}:*

$$\sum_{m\in M} R_M(P(m|T_{\backslash i})) \sim N(\frac{|M|(|F|+|M|+1)}{2}, \frac{|F||M|(|F|+|M|+1)}{12})$$

If further we have $|M| = |F| = K$, for each $i \in [n]$, under H_{i0}:

$$S(w_i) \sim N(0, \frac{2K+1}{3})$$

3.4.4 Algorithm Validation

To demonstrate the reliability of our scoring algorithm in identifying social biases within texts, we validate our method on two different tasks:

Human Experts Validation This validation involved collaboration with six experienced professionals from the fields of sociology and management science. Each coder manually labeled a randomly selected subsample of job advertisements. Leveraging

their extensive domain knowledge, these experts meticulously classified the advertisements, assessing them for levels of perceived gender bias. These categorical labels were then transformed into ordinal variables, enabling a detailed statistical comparison with the results produced by our scoring algorithm. This rigorous, expert-driven coding process ensured the reliability of our evaluation methodology.

We compute the Spearman rank correlation between the bias scores generated by our algorithm and the results from the manual labeling process. A Spearman correlation coefficient of 0.85[1] was obtained (Figure 3.4), indicating a strong positive association between our algorithm's scores and the human experts' assessments. This result validates the algorithm's capacity to accurately reflect human judgments of bias, confirming its effectiveness as a tool for social bias detection.

Benchmark Validation Further validation was conducted using the BIOS dataset [50], which comprises personal biographies categorized by gender and various occupations. We employed gender-specific word lists from [69], such as {man, his, he ...} versus {woman, her, she...}, as binary directional cues and designated gender as the ground truth label. Our algorithm demonstrated high performance, achieving an AUC of 0.97 in classifying gender, as illustrated in Figure 3.5. The AUC, or Area Under the ROC Curve, measures the ability of our model to distinguish between classes — here, gender categories. This performance surpasses that of three baseline methods in [55] that rely on unigram or word embeddings, highlighting the effectiveness and potential applicability of our bias detection approach in broader NLP tasks.

[1]This correlation is notably higher compared to those typically observed in non-experimental social sciences.

3.5 Probing Methodology and Job Application Data Generation

Probing Methodology To explore the social biases inherent in ChatGPT, particularly in the context of the labor market, our study simulates the typical use case where job seekers employ ChatGPT to assist in drafting job applications. This approach allows us to investigate not only the biases that may emanate from ChatGPT's training data but also to understand how these biases could potentially influence real-world job application/hiring processes.

Probing the social biases within ChatGPT presents several challenges. Firstly, ChatGPT's model operates as a 'black box,' making it difficult to discern the internal processes that contribute to bias propagation. Secondly, the lack of access to the model's architecture or parameters further complicates direct examination. Therefore, our analysis adopts an indirect method, employing our known bias evaluation algorithm to detect and quantify the biases exhibited by ChatGPT, thereby illuminating how these biases might manifest in practical applications.

Job Application Data Generation Our dataset comprises over 33,000 job postings collected from LinkedIn, reflecting a diverse range of industries and job types. To simulate realistic job application processes, we utilize the OpenAI API to prompt ChatGPT with these job advertisements, instructing it to generate corresponding job applications for each job posting.

This method does more than replicate real-world scenarios where individuals respond to job postings—it also facilitates a comprehensive analysis of the generated texts across various sectors. By using job advertisements as standardized prompts, we ensure that any observed deviations from neutrality in the generated texts are attributable to the model's ingrained biases, rather than the content of the advertisements themselves. This setup is crucial for isolating the effects of ChatGPT's biases, allowing for an accurate assessment of bias presence and intensity using the quantifiable metrics

provided by our bias score calculation method.

3.6 Analysing the Bias inside ChatGPT

3.6.1 Dimensions of Gender Bias

We begin by introducing the four gender dimensions, each defined by a distinct set of gender-related word lists, which will form the basis of our analysis. In recent social science research, understanding gender bias involves not just recognizing the existence of biases but also evaluating their impacts in various contexts. Building on the framework proposed by Gaucher *et al.* [39], Bem [66], and Konnikov *et al.* [69], we utilize specialized word lists to apply our social bias analysis across four different dimensions. Each dimension not only helps identify specific instances of bias but also offers insights into the broader social and psychological dynamics at play.

Psychological Cues: The psychological dimension assesses language context leaning towards communal attributes (e.g., "caring," "sympathetic," "attentive") commonly associated with femininity, or agentic attributes (e.g., "authoritative," "active," "confident") typically linked to masculinity.

Role Description: We evaluate job descriptions and roles using word lists that categorize terms associated with "soft" and "social" skills for feminine orientation, and "time-compressed" and "stressful" tasks, such as "multitasking," "pressure," "speed," for masculine orientation.

Work–Family Characteristics(WFC): This dimension examines employer policies and cultural expectations affecting gendered labor force participation, scrutinizing terms like "parental leave" and "flexible work" for feminine orientation versus "irregular and long work hours" and "weekend work" for masculine orientation.

Social Characteristics: We also analyze explicit gender references such as gendered pronouns and identity markers ("she," "he," "his," "her," "man").

3.6.2 Correlation Analysis

We first analyze the correlation of job postings and job applications across each dimension of gender bias. Our findings indicate a consistent positive linear correlation between the bias scores of job postings and the ChatGPT-generated job applications. This trend suggests that the biases inherent in job postings are likely to be reproduced in job applications by generative AI, reinforcing and possibly amplifying the initial biases. This correlation is visually captured in Figure 3.3, illustrating the potential for cyclical reinforcement of biases through the use of generative AI in job application practices.

Figure 3.3 presents the statistical parameters for each analyzed dimension of social bias. The strongest correlation is observed in the Social Characteristics dimension with a correlation coefficient of 0.777, indicating a very strong positive relationship. This is followed by the Role Description dimension, which shows a correlation coefficient of 0.708. Both of these correlations suggest significant potential for the biases in job postings to be reproduced by AI in job applications in these dimensions.

The Psychological Cues and WFC dimensions exhibit lower but still substantial correlation coefficients of 0.644 and 0.451, respectively. The slopes of these relationships indicate the rate at which the bias scores from job postings predict those in job applications, with steeper slopes observed in the Social Characteristics dimension. This analysis clearly supports the hypothesis that inherent biases in job postings are likely reproduced in ChatGPT-generated job applications.

3.6.3 Statistical Testing for Analysis

In this section, we delve deeper into how ChatGPT influences bias reproduction within the job application process. Let $\{X_i\}$ and $\{Y_i\}$, for $i = 1, \ldots, n$, represent the bias scores of original job postings and those of the job applications generated by ChatGPT, respectively. A score close to zero indicates minimal bias (i.e., gender neutrality that is neither feminine nor masculine), a higher positive score signifies a

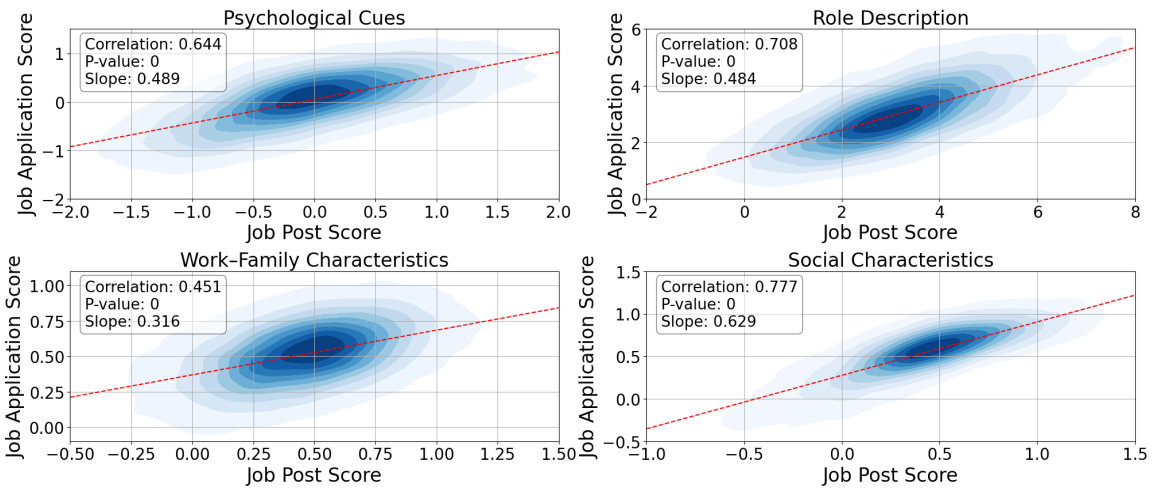

Figure 3.3: Result scatter density plot, for each of the bias dimensions where the x-axis is the job posting bias score and the y-axis is the job applications bias score. Where the darker color means there are more dots. The p-value is the significance of the correlation coefficient.

bias towards masculine language, and a lower negative score indicates a bias towards feminine language. The aim is to evaluate how ChatGPT may exacerbate or mitigate these biases. We denote the population mean and variance of X as μ_X and σ_X^2. The histogram and summary statistics of the bias scores are in Appendix 3.8.3.

Shift in Mean We propose the following hypothesis tests to assess shifts in mean:

$$H_0 : \mu_X \geq \mu_Y \quad \text{vs.} \quad H_1 : \mu_X < \mu_Y$$

Using the Wilcoxon signed-rank test, we determine whether there is a significant change in the mean bias score from the job postings to the applications.

Shift in Magnitude For the magnitude of bias, we assess:

$$H_0 : |\mu_X| \geq |\mu_Y| \quad \text{vs.} \quad H_1 : |\mu_X| < |\mu_Y|$$

This test measures the central tendency of bias scores, examining if the absolute values (regardless of bias direction) decrease. The less the magnitude(i.e. closer to zero) the less bias it has.

Change in Variance We also explore the variability in bias scores:

$$H_0 : \sigma_X^2 \leq \sigma_Y^2 \quad \text{vs.} \quad H_1 : \sigma_X^2 > \sigma_Y^2$$

This variance test, employing Levene's test [70] for equality of variances, explores whether ChatGPT produces job applications with more uniform bias expressions compared to the job postings. It helps determine if there is a reduction in variance, which would suggest that ChatGPT standardizes the use of gendered language cues. Such standardization could potentially reinforce specific gender biases more consistently.

Dimensions	**Mean**	**Magnitude**	**Variance**
Psychological Cues	↑	↓	↓
Role Description	↓	↓	↓
Work–Family Characteristics	↑	↑	↓
Social Characteristics	↑	↑	↓

Table 3.1: Statistical testing results for each dimension. The mean result indicates whether the overall bias score is shifting toward the masculine (↑) or feminine (↓) direction. The magnitude result reveals whether the bias is moving toward zero (↓) or away from zero (↑). The variance assesses whether job application bias scores exhibit greater (↑) or lesser (↓) variance compared to the job postings. Please refer to Table 3.2, 3.3,3.4, 3.5 and 3.6 in Appendix for detail statistics.

Shift in Mean The testing for the mean shift in Table 3.1 reveals significant findings across several dimensions. Except for Role Description, all other dimensions exhibit statistically significant shifts toward more masculine language. This indicates a predominant inclination for ChatGPT to amplify the use of masculine language in simulated job applications over and above the original job postings, possibly due to its training on historically biased data. This shift raises concerns about the consolidation and exacerbation of masculine language. Such biases in AI-generated content could perpetuate gender disparities in professional settings, emphasizing the need for interventions in AI training processes to address and correct historical biases. In contrast, the Role Description dimension shows a mean shift toward a less masculine

direction, but the bias in job postings has already been shown to be skewed toward a very masculine direction. In this case, ChatGPT seems to help mitigate this extreme masculine bias.

Magnitude of Bias The magnitude of bias, assessed through the mean of the absolute bias scores, varies across the dimensions. The Psychological Cues and Role Description dimensions suggest that the overall intensity of bias—regardless of direction—does not increase. This could imply that while the direction of bias towards masculinity is pronounced, the degree of bias embedded within job applications does not intensify. Conversely, the WFC and Social Characteristics dimensions exhibit an increase in bias magnitude, indicating not only a shift towards masculine language but also an overall increase in the strength of biased expressions. This finding is particularly troubling as it suggests that AI-generated job applications in these areas may become more polarized, further entrenching gender-specific expectations in roles traditionally associated with work-life balance and social interactions.

Variability in Bias Expression The variance results across all dimensions reveal a consistent decrease in job applications compared to job postings. This decrease in variance suggests that the language used by ChatGPT is more uniform across different applications, potentially indicating a standardization of language that leans towards masculine expressions. Such uniformity in language use could narrow the range of expressions and perspectives presented in job applications, limiting diversity and potentially skewing hiring decisions in favor of male candidates.

3.6.4 Implications and Extended Analysis

Our statistical results underscore a critical issue: biases in job postings are not merely replicated but are amplified in job applications created by generative AI in response to the postings. This phenomenon can be explained by the reinforcement of initial biases through the language processing and text generation capabilities of AI tools

like ChatGPT, which tend to replicate and often intensify the language patterns they are trained on.

Societal and Labor Market Implications: The amplification of gender biases in AI-generated job applications has profound societal and labor market implications, suggesting that not only are stereotypical roles perpetuated through biased language, but they are also strengthened when individuals use AI tools like ChatGPT to assist with drafting job applications. This use of generative AI plays a crucial role in circulating and amplifying biases, which reinforces, rather than challenges, the gender biases underpinning persistent gender inequalities in the workplace. Such biases can compound, influencing job satisfaction, employee retention, and career advancement. The misallocation of human resources due to biased AI could reduce economic efficiency and innovation, potentially causing sectors to overlook qualified candidates. Furthermore, these persistent inequalities may spur regulatory and legal challenges, especially in countries with robust equal employment opportunity laws, with significant implications for social ethics, justice, and economic equality.

Recommendations for Intervention: To mitigate the reproduction of gender biases through LLMs, it is recommended that employers and AI developers implement more rigorous bias monitoring and mitigation strategies. This could include the use of debiased language models, regular audits of AI-generated content by independent third-party organizations, and the development of enhanced AI training datasets that reflect the diversity of the global job market. Additionally, public awareness and education initiatives should be promoted to increase understanding of AI's role in job application and its potential impacts, fostering a critical approach to AI tool usage in professional settings.

3.7 Conclusion

Our paper – including a novel experiment, new algorithm development, and empirical application and findings – contributes to the ongoing debates and developments

in the ethical use of AI in labor market processes and practices. By identifying underlying biases in AI-driven text generation, this paper proposes novel strategies and methods for detecting and mitigating such biases. Through our **PRISM** algorithm and empirical application, we show that these strategies are not just theoretical but are intended as actionable steps toward ensuring that the integration of AI in the labor market supports equitable and fair employment opportunities for both employers and job seekers.

3.8 Appendix & Supplemental Material

3.8.1 Proof of Theorem 3.1

Since each word from two word lists M and F are selected independently. Therefore, the first result in Theorem 3.1 is implied directly from the Wilcoxon rank sum test [71].

If $|M| = |F| = K$, the bias score $S(\omega_i)$ can be rewritten as

$$\begin{aligned} S(\omega_i) &= \frac{1}{K}\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i})) - \frac{1}{K}\sum_{m\in M} R_{F\cup M}(P(m|T_{\setminus i})) \\ &= \frac{1}{K}\left[\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i})) - \left(\frac{2K(2K+1)}{2} - \sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i}))\right)\right] \\ &= \frac{2\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i}))}{K} - (2K+1)\,, \end{aligned} \tag{3.1}$$

where the second equality follows from the fact:

$$\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i})) + \sum_{m\in M} R_{F\cup M}(P(m|T_{\setminus i})) = \frac{2K(2K+1)}{2}.$$

From the first result in Theorem 3.1, we have

$$\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i})) \sim N(\frac{K(2K+1)}{2}, \frac{K^2(2K+1)}{12})$$

for each $i \in [n]$, under H_{i0}. Therefore, under H_{i0}, the result $S(w_i) \sim N(0, \frac{2K+1}{3})$ follows directly from the relationship (3.1) and the distribution of $\sum_{f\in F} R_{F\cup M}(P(f|T_{\setminus i}))$.

3.8.2 Algorithm Validation Result

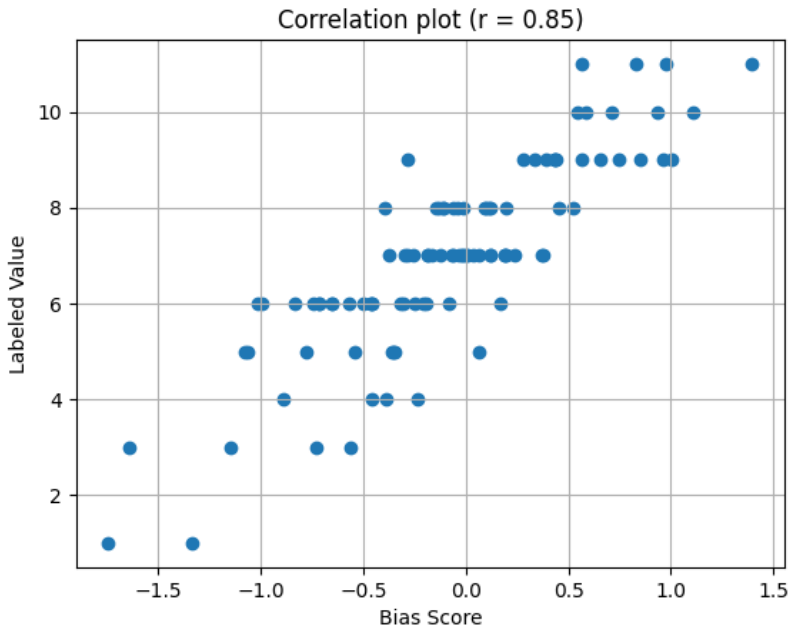

Figure 3.4: Human Expert Validation

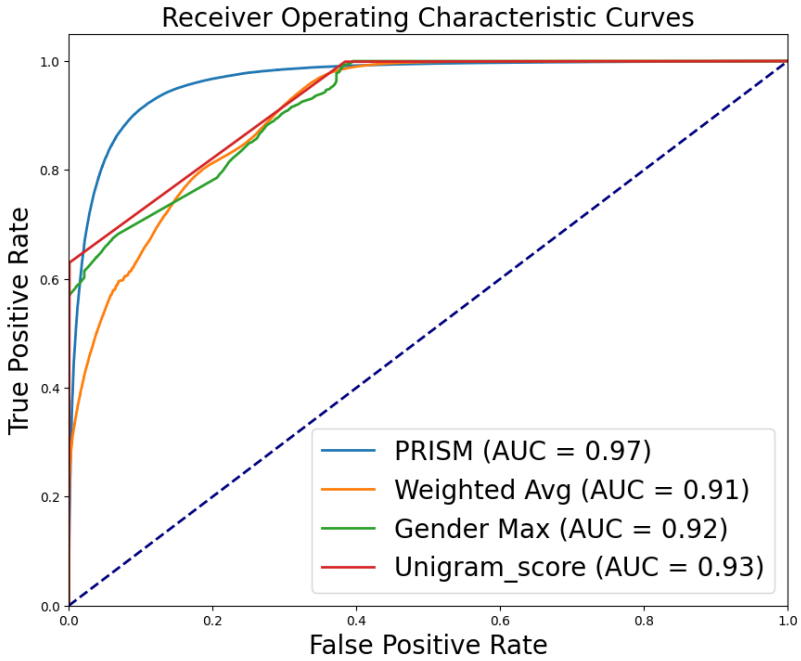

Figure 3.5: Benchmark Validation

In Figure 3.4, A labeled value of 7 signifies neutrality, while values less than 7

suggest femininity, and values greater than 7 imply masculinity. A Bias Score close to zero indicates neutrality, a positive value suggests masculinity and a negative value denotes femininity.

In Figure 3.5, We follow three gender metrics for evaluating gender bias in texts in [55]. The first metric, *unigram matching*, counts gender-specific tokens like 'he', 'him', 'she', 'her' etc., and labels texts with more male tokens as male, more female tokens as female, and texts with equal counts as neutral. The second metric assesses words indirectly related to gender via a normalized projection of word vectors in the gender direction, defined by $\vec{she} - \vec{he}$, using a Word2Vec embedding. Word-level gender scores are calculated as $b_i = \frac{\vec{w}_i \cdot \vec{g}}{\|\vec{w}_i\|\|\vec{g}\|}$. These are aggregated either by a weighted average (*Gender-Wavg*):

$$\text{Gender-Wavg} = \frac{\sum_{i=1}^{n} \text{sgn}(b_i) b_i^2}{\sum_{i=1}^{n} |b_i|}$$

or by taking the score from the most gender-polar word (*Gender-Max*):

$$i^* = \arg\max_i(|b_i|), \quad \text{Gender-Max} = \text{sgn}(b_i^*)|b_i^*|$$

Texts are classified as male if the score is less than -0.25 and as female if the score is greater than 0.25.

3.8.3 Histogram of Bias Scores

In Figure 3.6, we present the histogram of bias scores for Job Postings and Job Applications on different dimensions.

3.8.4 Statistical Tests Results

3.8.5 Experiment Setting & Computational Resources

For our analysis, we utilize the 'bert-base-uncased' model from the Hugging Face library with a selection threshold, α, set to 20 for choosing the top probability percentage. The temperature parameter for the ChatGPT API is set to its default value of 7. For

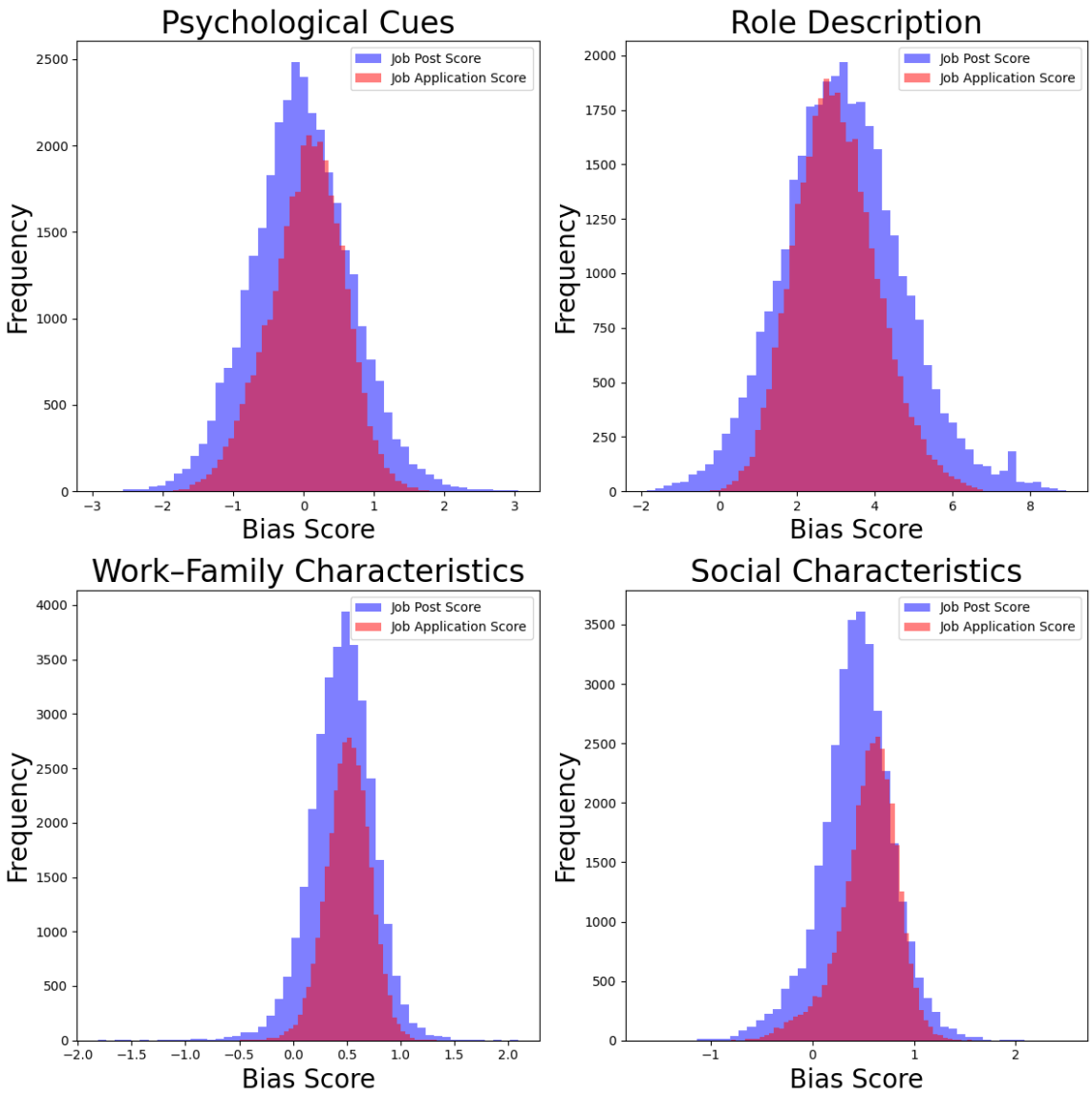

Figure 3.6: Result histogram, for each of the bias dimensions, we use different colors to distinguish the Job Postings and Job Applications

Dimension	**Mean**	**Magnitude**	**Std**
Psychological Cues	-0.030	0.552	0.703
Role Description	3.188	3.205	1.550
Work–Family Characteristics	0.455	0.476	0.290
Social Characteristics	0.435	0.484	0.355

Table 3.2: Mean, Magnitude, and Standard Deviation for job postings across different dimensions.

Dimension	**Mean**	**Magnitude**	**Std**
Psychological Cues	0.039	0.427	0.533
Role Description	3.020	3.020	1.060
Work–Family Characteristics	0.513	0.515	0.203
Social Characteristics	0.550	0.569	0.288

Table 3.3: Mean, Magnitude, and Standard Deviation for job applications across different dimensions.

Dimension	**Statistic**	p **- value**	H_1
Psychological Cues	229792649.0	1.06×10^{-136}	$\mu_X < \mu_Y$
Role Description	316042300.0	0.0	$\mu_X > \mu_Y$
Work–Family Characteristics	207756451.0	0.0	$\mu_X < \mu_Y$
Social Characteristics	119987669.0	0.0	$\mu_X < \mu_Y$

Table 3.4: Wilcoxon Test Results for Mean Shift

Dimension	**Statistic**	p **- value**	H_1
Psychological Cues	353433005.0	0.0	$\lvert\mu_X\rvert > \lvert\mu_Y\rvert$
Role Description	319570996.0	0.0	$\lvert\mu_X\rvert > \lvert\mu_Y\rvert$
Work-Family Characteristics	219288122.0	0	$\lvert\mu_X\rvert < \lvert\mu_Y\rvert$
Social Characteristics	146778531.0	0.0	$\lvert\mu_X\rvert < \lvert\mu_Y\rvert$

Table 3.5: Wilcoxon Test Results on Absolute Values

Dimension	**Statistic**	p **- value**	H_1
Psychological Cues	1825.094	0.0	$\sigma_X^2 > \sigma_Y^2$
Role Description	3410.619	0.0	$\sigma_X^2 > \sigma_Y^2$
Work–Family Characteristics	2491.084	0.0	$\sigma_X^2 > \sigma_Y^2$
Social Characteristics	922.186	1.80×10^{-201}	$\sigma_X^2 > \sigma_Y^2$

Table 3.6: Levene's test results for variance between job and application data across different dimensions, analyzed with a one-sided interpretation. These results indicate significant differences in variance, with the job postings consistently showing greater variance compared to the job applications.

the preprocessing, when we iterate through the text, we skip some of the function words like Articles ('a', 'an', 'the'), Prepositions ('on', 'by'), Conjunctions('and', 'but', 'if'), etc. As our algorithm solely requires a forward pass and no training, this enhances computational efficiency. To further optimize performance, we employ an Nvidia RTX A5000 GPU. All experiments are conducted on an Ubuntu server equipped with an AMD Ryzen Threadripper 3990X 64-Core Processor and 256 GB of RAM.

Limitations & Future Works

All our results are based on an English dataset; however, additional complexities may arise with other languages due to more intricate word splitting or tokenization challenges. The Masked Language Models (MLMs) utilized in our study are sourced from public, open-source pre-trained models. The inherent biases of these pre-trained models might impact our results, although we have attempted to mitigate this issue through a robust rank-based method that reduces sensitivity to changes in probability distributions. Currently, our job application generation relies on basic prompts; exploring the effects of varied prompts to capture a broader spectrum of biases constitutes part of our future work. Moreover, while our framework is initially designed to use pairs of word lists, it possesses the flexibility to accommodate single or multiple word lists with minimal adjustments, an extension we also plan to explore in future research endeavors.

Chapter 4

A Statistical Testing Framework for Bias Word Detection with Masked Language Models

4.1 Abstract

We introduce a novel statistical hypothesis-testing framework designed to detect biases in textual content by Masked Language Models (MLMs). Our unsupervised approach leverages sentence perturbation techniques to construct robust datasets from singular text instances, enabling practical and effective statistical testing of bias. The framework is adaptable, incorporating various methods for measuring bias and calculating variance, thereby accommodating different linguistic contexts and analytical needs. Through rigorous empirical validation using real-world data, we demonstrate the framework's capacity to identify subtle biases, highlighting its utility and effectiveness. Our contributions are significant, offering a tool that enhances the transparency and fairness of AI technologies. This work not only advances the field of AI and ethics but also provides actionable insights for developers and researchers striving to mitigate bias in AI-generated texts.

4.2 Introduction

As artificial intelligence (AI) continues to evolve, with Natural Language Processing (NLP) and Large Language Models (LLMs) becoming increasingly prevalent, their influence permeates every facet of social and economic life. The rapid deployment and integration of these technologies into daily interactions, decision-making processes, and broader societal functions spotlight their potential to reshape numerous aspects of human activity. However, alongside the technological advancements, there is a growing concern about the embedded social biases and ethical issues that these technologies may harbor. Such concerns underscore the urgent need for rigorous scrutiny and methodologies capable of addressing and mitigating bias within AI systems.

Recognizing the profound impact generative AI has on society, this paper focuses on identifying and understanding the biases manifested in generated textual content. To address these challenges, we are the first to develop a statistical hypothesis testing framework designed specifically for bias detection in texts produced by Masked Language Models (MLMs). This framework is not only innovative but also adaptable, accommodating various bias measurement and variance calculation methods.

The crux of our research is the establishment of a statistical testing framework that measures biased words within textual data. This framework is distinctive in its flexibility, allowing for the integration of diverse bias measurement techniques. Moreover, our method stands on solid theoretical ground, possessing several desirable statistical properties that enhance its reliability and applicability in bias detection.

One of the highlights of our framework is its unsupervised nature, which is particularly advantageous given the typical scarcity and expense of labeled data. This is especially relevant in domains where bias-labeled data are rare or non-existent. Furthermore, our approach addresses the real-world challenge of data scarcity in statistical testing scenarios. Typically, when a sentence is identified as a candidate for bias testing, the available data may be limited to that single instance rather than a

larger sample. Our method creatively employs sentence perturbation techniques to generate a sufficient sample of data points, thus enabling the robust estimation of distributions necessary for effective statistical testing.

In essence, this paper seeks to bridge the gap between advanced computational techniques and the critical need for fairness and transparency in AI. Through our innovative approach, we aim to contribute significantly to the ongoing dialogue and efforts in AI ethics, particularly in mitigating bias in language models.

Our key contributions are threefold:

- We propose a novel statistical testing framework specifically designed for detecting bias in textual contexts, which is both robust and flexible.
- Our algorithm is distinguished by several aspects of methodological soundness, including theoretical validity and the ability to handle diverse data scenarios without the need for extensive labeled datasets.
- We validate our framework with real-world data, demonstrating its effectiveness and practical utility in identifying and understanding biases within large language models.

This paper is structured as follows: We first discuss the related work about quantifying social bias in text and NLP. Then we talk about the problem setting and motivation. For the main algorithm, we start by introducing the sentence perturbation followed by the statistical hypothesis testing framework. Finally, we present the experiment result.

4.3 Related work

Evaluating bias in textual contexts remains a formidable challenge within the field of natural language processing. Despite significant advancements in machine learning and AI technologies, effectively detecting and quantifying biases in text not only

requires sophisticated algorithms but also a nuanced understanding of language and context. This has spurred ongoing research efforts aimed at developing more robust and adaptable methodologies to address this critical issue.

The measurement methods for evaluating bias in pre-trained word embeddings and language models can be broadly divided into two categories: Intrinsic and Extrinsic evaluations. Intrinsic bias evaluations probe the bias within pre-trained word embeddings and language models. Common methods include measuring the geometry in embedding space, such as the Word Embedding Association Test (WEAT; [45]) and Sentence Encoder Association Test (SEAT; [72]). Additionally, [47–49] propose metrics using the likelihood score. Furthermore, research suggests that some debiasing methods may only hide bias, and thus additional measurement approaches are needed [13].

The extrinsic bias is specific to certain downstream tasks. In the text classification task, De-Arteaga *et al.* [50] and Blodgett *et al.* [51] proposed two benchmark datasets and used the equal opportunity measure from fairness literature. Zhao *et al.* [52] proposed the WinoBias benchmark for Coreference resolution. As well as other benchmarks, such as Bias-NLI [53] and in machine translation [54]. However, recent research has indicated that intrinsic bias in embeddings or models typically does not have a strong correlation with bias in downstream tasks[73, 74]. Kaneko *et al.* [75] found out that the debiased models re-learn the bias from the fine-tuning datasets, showing that only debiasing upstream models may not be enough to eliminate bias in downstream tasks.

In this paper, we focus on the specific downstream task of bias evaluation of the generated text. Many current methods are only applicable to particular kinds of bias, making them less flexible for other scenarios. Dhamala *et al.* [55] calculate bias by determining the cosine similarity between word embeddings [2, 56] along the gender axis ($\overrightarrow{he} - \overrightarrow{she}$) [44], and then averaging these values across sentences. Cryan *et al.* [57] analyzes both a lexicon-based method and a BERT model that has been

fine-tuned using a dataset with labels gathered via crowdsourcing. Spinde *et al.* [58] has produced a dataset on media bias using expensive expert annotations, a technique that is not readily applicable to different fields. Raza *et al.* [59] investigate the application of named entity recognition to identify biased words in texts, a method that also necessitates the generation of expensive labeled training data for each specific task and model. Sociological and psychological studies correlate word usage with conventional gender roles and personality characteristics, pinpointing words typically linked to masculine or feminine traits (such as caregiving and assertiveness; agentic versus communal) [76]. The Bem Sex-Role Inventory serves as an example of a list of gender-related words created through participant studies, underscoring characteristics deemed desirable for different genders [39, 66].

4.4 The Problem Setting & Motivation

4.4.1 Motivation

In this section, we formally outline the objectives of our study. The primary aim is to develop a universal framework capable of estimating biases for individual words within a sentence using an unsupervised approach. This framework employs statistical testing to assess word biases and provides a corresponding p-value for each word within a text. Its flexibility is a key feature, allowing for various calculation methods, which we will elaborate on subsequently.

Our methodology operates within an unsupervised, or minimally supervised, framework. Lacking labeled data, we do not train a large model to predict word biases. Instead, we leverage a Masked Language Model (MLM) for extracting relevant information. Additionally, we utilize sets of keyword pairs that serve as directional cues in our analysis.

4.4.2 Notation

Given a text T comprising n words $T = \{w_1, w_2, \ldots, w_n\}$, we iteratively mask each word w_i and input the modified masked text $T_{\setminus i} = \{\ldots, w_{i-1}, [\text{MASK}], w_{i+1}, \ldots\}$ into an MLM, which outputs the probability distribution over the vocabulary for the masked position i, denoted as $P(\cdot \mid T_{\setminus i})$.

Then, to obtain the direction signal for score calculation, we require two predefined word lists representing different contexts—such as gender with a feminine word list $F = \{f_1, \ldots, f_{|F|}\}$ and a masculine word list $M = \{m_1, \ldots, m_{|M|}\}$. For each word in F and M, we obtain the probability from the distribution $P(\cdot|T_{\setminus i})$. This yields two sets of probabilities: $P_F = \{P(f|T_{\setminus i})\}$ for each word $f \in F$ and $P_M = \{P(m|T_{\setminus i})\}$ for each word $m \in M$.

4.4.3 Problem Setting

Our task is to test the biases of each word inside a sentence. From the viewpoint of statistical testing, we first start with the null hypothesis that the word is not biased. In order to do this, we first need to quantify the bias and denote that as a Bias Score. We will discuss more details and options for the Bias Score in the section below. In addition, after obtaining the Bias Score as test statistics, in order to estimate the distribution of the test statistics and conduct statistical inference, we need to have multiple scores for the distribution estimation, and this is done through the sentence perturbation method, and it will also discuss in detail in the following section.

4.5 Sentence Perturbation

To conduct accurate statistical hypothesis testing and estimate the distribution of test statistics, it is essential to have multiple samples. Yet, the primary constraint is often the availability of only the candidate text. To overcome this limitation, we propose several ways to generate data and text perturbation.

Sentence Transformation In our study, we explore sentence transformation. We employ a pre-trained language model to perturb text while ensuring that these transformations do not alter the original content's meaning, focusing solely on word changes. In detail, we utilize the Text-to-Text Transfer Transformer (T5) model [77], which excels in generating multiple sentences that preserve the semantic integrity of the original text. T5 is uniquely skilled at transforming text-based tasks into a text-to-text format, notably adept at producing paraphrases that maintain semantic consistency while varying lexical and syntactic elements.

The effectiveness of T5 stems from its comprehensive pre-training on a diverse corpus, enabling it to develop a nuanced understanding of language semantics and syntax. This pre-training allows T5 to generate semantically consistent perturbations, crucial for our methodology. In our approach, we first mask each word w_i in a text T with a masked token for the Masked Language Model (MLM) to predict. Additionally, we select several indices randomly and replace them with the `extra_id` token specifically designed for T5, resulting in a modified sentence fed into the T5 model, which then predicts and fills the gaps, creating randomly perturbed sentences. We provide a detailed illustration graph in Figure 4.1.

Redundant Prefix Another method for data perturbation involves appending a redundant prefix to the start of the original data, preserving the literal meaning of the original sentence. We modify the text style by introducing a redundant prefix at the beginning of the original text. These prefixes include "It is important to acknowledge that:", and "It is worth noting that:", which lend a nuanced and deliberate tone to the content. To create diverse samples, our approach requires utilizing various prefixes and suffixes. We employ ChatGPT to generate a predefined set of these elements, ensuring each modifies the original content differently. This strategy not only changes the tone but also adds a layer of careful consideration, emphasizing the message being conveyed.

Punctuation Marks Substitution Another straightforward yet effective perturbation technique we employ involves substituting the punctuation marks within a sentence. This method subtly alters the tone of the text without changing its factual content. However, the scope of this transformation is somewhat limited due to the restricted number of punctuation marks and their placement within the text.

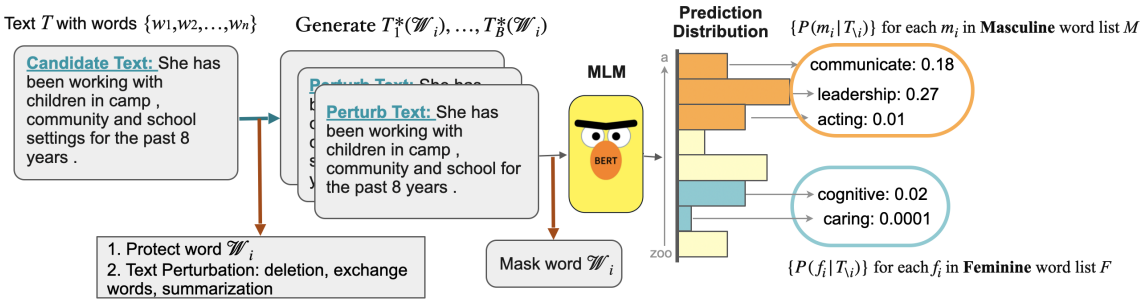

Figure 4.1: An illustration of the paradigm for text perturbation and use MLM for obtaining the score.

4.6 Word Level Bias Evaluation Framework

In this section, we first outline the full testing framework followed by the detail of each necessary component for our testing framework.

4.6.1 Algorithm for Testing Word Bias

To systematically determine whether a given word $\mathcal{W}_i$ in a sentence T exhibits bias, we devise an algorithm that employs natural language processing and statistical analysis techniques. The method involves perturbing the sentence context as we mentioned above, utilizing a Masked Language Model for contextual understanding, and performing a hypothesis test based on the generated data.

In our framework, the algorithm begins by accepting specific input parameters: a critical value Δ, a significance level α, a target word $\mathcal{W}$, and the sentence T that includes $\mathcal{W}$. The core of the algorithm involves a sentence perturbation process using the T5 model, which manipulates the words in T while keeping $\mathcal{W}$ static,

thus generating a series of perturbed sentences $T_1^*(\mathcal{W}), \ldots, T_B^*(\mathcal{W})$. Each perturbed sentence is then processed to mask $\mathcal{W}$, and a Masked Language Model (MLM) such as BERT is employed to predict the probability distribution of potential replacements at the masked position.

The MLM predictions are used to compute two specific sets of probabilities: P_F and P_M. P_F encompasses probabilities associated with female-biased words, while P_M covers male-biased words, both derived from a predefined list of gender-associated words.

With these probabilities, we can use them to form various metrics for quantifying social bias. Allow for the calculation of a Bias Score for each perturbed sentence, quantifying the level of gender bias that $\mathcal{W}$ introduces into T.

Following the bias score calculations, the variance of these scores is estimated to support the construction of a test statistic $\mathcal{V}$. This statistic is crucial for evaluating the significance of the observed bias, as it is compared against a standard or theoretically derived distribution to ascertain if the bias is statistically significant, culminating in the computation of a p-value. This p-value is instrumental in determining whether the bias associated with $\mathcal{W}$ is significant at the level α, thereby providing a robust measure of word bias in the context of the sentence.

This methodology leverages advanced machine learning techniques and statistical analysis to provide a systematic and reproducible approach to detecting and quantifying word bias in textual data, exemplified by processing a sentence like "The nurse attended the conference" where "nurse" might be perturbed to observe gender biases based on associative probabilities in different contexts. The detailed paradigm is in Figure 4.2 and the Algorithm step is in Algorithm 4.

This structured algorithmic approach integrates advanced machine learning techniques with robust statistical methods, facilitating precise and replicable bias detection in textual data.

In the following, we will provide details about each component of the framework

Algorithm 4 Bias Word Detection Algorithm

Input: Sentence $T(\mathcal{W})$, Target word $\mathcal{W}$, Critical Value Δ, Significance Level α, Number of Perturbation Samples B

Output: p-value indicating the bias of Word $\mathcal{W}$

Initialize Perturbation Samples: Generate $T_1^*(\mathcal{W}), \ldots, T_B^*(\mathcal{W})$ using T5 model by perturbing words in T, excluding $\mathcal{W}$
for $b = 1$ to B **do**
    Mask $\mathcal{W}$ in $T_b^*(\mathcal{W})$
    Use MLM to predict the probability distribution $P(\cdot|T_b^*(\mathcal{W})$ with $\mathcal{W}$ masked)
    Calculate probabilities P_F and P_M for female and male bias-related words respectively
    Compute Bias Score for $T_b^*(\mathcal{W})$ using P_F and P_M
end for
Compute the mean $\overline{\text{Bias Score}}$ and variance of Bias Scores across all B samples
Construct the test statistic $\mathcal{V} = (\overline{\text{Bias Score}} - \Delta)/\sqrt{\text{Var}}$
Determine the p-value by comparing $\mathcal{V}$ to a corresponding distribution
return p-value

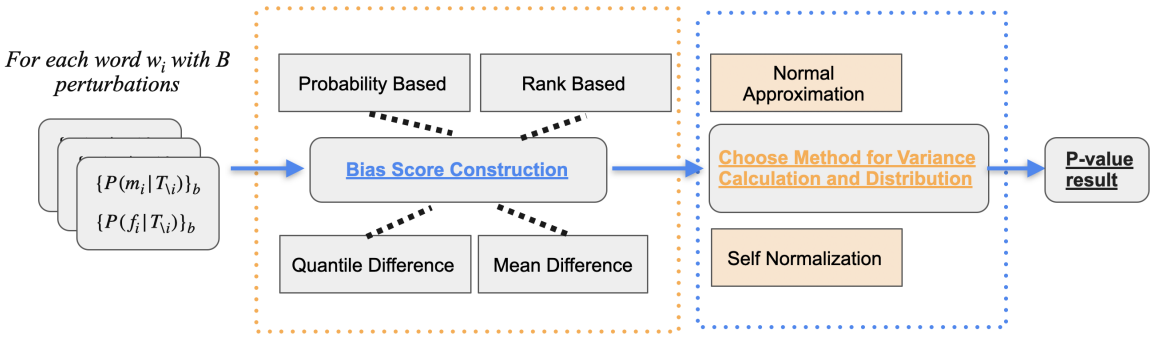

Figure 4.2: An illustration of the paradigm for the testing framework

and provide examples for illustration and recommendations.

4.6.2 Social Bias Quantification

Previous Methods

Here arises a key question on how to quantitatively represent the bias, many of the previous research works on evaluating bias in the context of word embedding, language models, and different downstream tasks(as discussed in the related work section). In this study, we concentrate on the assessment of unstructured text, which is inherently more difficult than evaluating natural language processing models. This often demands significant input from human experts or the use of basic and heuristic techniques.

In our previous work, we introduced two kinds of methods for bias calculation, the first is a static word embedding-based method for calculating the bias using word list words, the main idea is that for each word in the text we want to analyze, we use the word list pairs to convert the words into two sets of embedding vectors, and then we use the candidate word and also convert it to the embedding word vectors and then we calculate the cosine similarity of the candidate word vector with the two-word vectors sets. We can view the cosine similarity to be how biased this word is to male or female word sets. As for our framework, static embedding-based methods will yield exactly the same score for every perturbation of the sentence, since it looks at words individually. Thus failing to capture the nuanced changes induced by the perturbations.

The second metric is based on the masked language model, the key advantage is that the analysis will be context-aware, which means when we are looking at the candidate word, we are also considering the information of the whole sentence, not each word individually. The key idea is to mask the candidate word and use the Masked Language Model to output the prediction probabilities. We use this output and the probabilities of the two-word lists to construct the bias scores.

Bias Score Calculation Methods

We present various methods to quantify bias from the outputs of a Masked Language Model (MLM), specifically analyzing the probability distributions P_F and P_M for each target word $\mathcal{W}$. The bias score is fundamentally derived as a function:

$$\text{Bias Score} = \mathcal{F}(P_F, P_M)$$

where $\mathcal{F}$ represents a calculation method designed to measure the disparity between two sets of probabilities associated with gender-biased words.

To illustrate the versatility and adaptability of our framework, we present several exemplary methods for calculating the bias score, each tailored to meet different analytical needs and performance criteria within diverse linguistic research contexts:

1. **Average Probability Difference**: This method calculates the bias score as the difference between the average probabilities of female-biased and male-biased word sets:

$$\text{Bias Score} = \frac{1}{|F|} \sum_{f \in F} P(f|T_{\backslash i}) - \frac{1}{|M|} \sum_{m \in M} P(m|T_{\backslash i}).$$

2. **Quantile-Based Difference**: An alternative approach utilizes the η-quantiles of the probability distributions, providing a bias score that reflects more extreme probability values:

$$\text{Bias Score} = \frac{1}{|F|} \sum_{f \in F} Q_\eta(P(f|T_{\backslash i})) - \frac{1}{|M|} \sum_{m \in M} Q_\eta(P(m|T_{\backslash i})),$$

   where Q_η represents the η-quantile, with η being a predefined threshold.

3. **Rank-Based Evaluation**: This method assesses bias by comparing the average ranks within each set of probability distributions, where here we define function R as the rank of the probability with respect to the whole dictionary:

$$\text{Bias Score} = \frac{1}{|F|} \sum_{f \in F} R(P(f|T_{\backslash i})) - \frac{1}{|M|} \sum_{m \in M} R(P(m|T_{\backslash i})).$$

4. **Quantile of Ranks Difference**: Building on rank evaluations, this method calculates the bias score using the quantiles of the ranks, potentially highlighting disparities that are not evident from average ranks alone:

$$\text{Bias Score} = \frac{1}{|F|}\sum_{f\in F} Q_\eta(R(P(f|T_{\setminus i}))) - \frac{1}{|M|}\sum_{m\in M} Q_\eta(R(P(m|T_{\setminus i})).$$

5. **Merged Rank Differences**: A derivative of our previous methods, this approach merges the two probability sets before calculating the rank differences, here the function R' denotes the rank based on the union of two probability sets:

$$\text{Bias Score} = \frac{1}{|F|}\sum_{f\in F} R'(P(f|T_{\setminus i})) - \frac{1}{|M|}\sum_{m\in M} R'(P(m|T_{\setminus i})).$$

These methods provide a comprehensive toolkit for researchers to adapt and apply according to their specific datasets and hypothesis-testing needs, illustrating the flexibility and potential of our framework in exploring and quantifying biases in textual data. Each method highlights different aspects of bias, from subtle to overt, ensuring that researchers can choose the most appropriate approach for their specific study.

4.6.3 p-value calculation and the Corresponding Distributions.

With the bias score in hand, we can proceed with the calculation of the distribution and the corresponding p-value.

Normal Approximation

The most intuitive method for estimating variance in the context of bias score analysis is the ordinary variance calculation:

$$\text{Var} = (B-1)^{-1}\sum_{k=1}^{B}\left\{\text{Bias Score}_k - \overline{\text{Bias Score}}\right\}^2$$

Here, B represents the total number of perturbation samples, Bias Score$_k$ denotes the bias score from the k-th sample, and $\overline{\text{Bias Score}}$ is the mean bias score across all perturbation samples. This variance estimation is crucial as it quantifies the dispersion of bias scores around their mean, which is fundamental to the normal approximation.

The test statistics can be constructed as:

$$\mathcal{V} = \frac{\overline{\text{Bias Score}} - \Delta}{\sqrt{\text{Var}}} \sim \mathcal{N}(0, 1)$$

The p-value for the word w can then be calculated using the following:

$$p_w = 1 - \sup\{\beta : \mathcal{V} > Z_\beta\}$$

where Z_β is the β-th quantile of the standard normal distribution.

Self Normalization

In typical statistical applications, the sample variance estimator presumes that observations are independent and identically distributed (iid). However, this assumption is problematic in context, where we deal with sentence perturbations—subtle alterations to syntax or diction that preserve the original meaning. These manipulations mean that sentences share inherent similarities and are not statistically independent.

Moreover, estimating variance in such dependent samples is notoriously challenging. To address this, under specific conditions, [78] introduced the self-normalization technique, which proposes a new estimator with variance proportional to that of the original data, referred to as the normalizer. Consequently, the ratio of two such statistics effectively eliminates the asymptotic variance, yielding a deterministic and known limiting distribution. Following the methodology outlined in [78], we apply a similar normalizer. Assuming certain technical conditions, this ratio is expected to weakly converge to the distribution given by $\frac{W(1)}{\sqrt{\int_0^1 (W(t) - tW(1))^2 dt}}$, where $W(t)$ denotes Brownian motion. Our variance estimator is defined as follows:

$$\text{Var} = B^{-2} \sum_{k=1}^{B} \left\{ \sum_{j=1}^{k} (\text{Bias Score}_j) - \overline{\text{Bias Score}} \right\}^2$$

This framework not only adheres to robust statistical theory but also ensures that our estimations are suitable for the complex and dependent nature of our framework.

4.7 Experiments

To analyze the performance of the framework, we conduct some real-world text analysis on various method combinations we have. To showcase it, we demonstrate an example text and analyze its result:

> 'She has been working with children in camp , community and school settings for the past 8 years . She believes in the importance of cultivating self - love and awareness in black children at a very young age and is excited to be apart of Black Lives Matter Toronto 's Freedom School !'

Using the text above, we deploy our framework using four different word lists with detail in the Appendix 4.9.1. Each of the word lists measures a different dimension of gender bias. For the result, when we are looking at the gender-social characteristic dimension, which only contains words with gendered pronouns, our framework is not only able to detect all the pronouns that it is meant to detect but also detects some words that are gender bias in another dimension, such as children, cultivation,.. etc. In addition, the other dimensions such as Psychological Cues, are able to detect the pronouns words such as she.

4.8 Conclusion

In this study, we introduced a novel statistical hypothesis-testing framework for detecting biases in textual content using Masked Language Models (MLMs). Our methodology stands out due to its unsupervised nature, leveraging sentence perturbation techniques to create robust datasets from individual text instances. This approach enables practical and effective statistical testing of biases, facilitating the identification of subtle biases in text.

Unique_Words	psy_diff	role_diff	wfc_diff	gsc_diff
she	0.036800	0.001200	0.709800	0.000000
working	0.017600	0.060200	0.242400	0.092600
children	0.004400	0.034400	0.009200	0.156200
camp	0.000400	0.243600	0.028800	0.132000
community	0.000000	0.005000	0.061200	0.110600
school	0.004800	0.459200	0.022800	0.061600
years	0.001200	0.005000	0.320400	0.213200
she	0.784200	0.085800	0.372400	0.013200
believes	0.203000	0.210800	0.013200	0.000000
importance	0.001000	0.067200	0.102200	0.236400
cultivating	0.265400	0.000000	0.000000	0.013000
self	0.024400	0.264400	0.000000	0.148800
love	0.036200	0.000000	0.005800	0.230600
awareness	0.632800	0.354600	0.000000	0.025200
black	0.058600	0.004600	0.000000	0.000000
children	0.479800	0.616400	0.058200	0.018200
young	0.137400	0.001400	0.002200	0.222400
age	0.112800	0.000000	0.000000	0.502200
excited	0.049800	0.001400	0.014800	0.074600
black	0.004800	0.000400	0.028000	0.047200
lives	0.118200	0.447800	0.168600	0.728800
matter	0.002000	0.018000	0.059800	0.082200
toronto	0.002000	0.001400	0.077600	0.425000
freedom	0.725200	0.000000	0.066200	0.006200
school	0.240800	0.000000	0.140800	0.592400

Table 4.1: Example text bias analysis

Our framework is highly adaptable, and capable of incorporating various methods for measuring bias and calculating variance. This flexibility allows it to accommodate different linguistic contexts and analytical needs, making it a versatile tool for bias detection across diverse settings. Through rigorous empirical validation of real-world data, we have demonstrated the framework's capacity to reliably identify biases, thereby underscoring its utility and effectiveness.

The contributions of this work are significant, enhancing the transparency and fairness of AI technologies. By providing a methodological advancement in the field of AI and ethics, this research not only propels the scientific community forward but also offers actionable insights for developers and researchers striving to mitigate bias in AI-generated texts. Looking forward, we envision this framework being applied more broadly to assess and improve the ethical considerations of AI systems, ultimately leading to more equitable AI applications. This work lays a solid foundation for future research aimed at refining bias detection techniques and developing more nuanced approaches to understanding and eliminating bias in automated systems.

Through continued innovation and collaboration, we can enhance the capabilities of AI systems, ensuring they serve society ethically and justly. The potential applications of our framework extend beyond the academic sphere into practical implementations, where developers can integrate these methodologies to audit and refine AI outputs, promoting fairness in automated decision-making processes.

4.9 Appendix & Supplemental Material

4.9.1 Dimensions of Gender Bias

We begin by introducing the four gender dimensions, each defined by a distinct set of gender-related word lists, which will form the basis of our analysis. In recent social science research, understanding gender bias involves not just recognizing the existence of biases but also evaluating their impacts in various contexts. Building on

the framework proposed by Gaucher *et al.* [39], Bem [66], and Konnikov *et al.* [69], we utilize specialized word lists to apply our social bias analysis across four different dimensions. Each dimension not only helps identify specific instances of bias but also offers insights into the broader social and psychological dynamics at play.

Psychological Cues: The psychological dimension assesses language context leaning towards communal attributes (e.g., "caring," "sympathetic," "attentive") commonly associated with femininity, or agentic attributes (e.g., "authoritative," "active," "confident") typically linked to masculinity.

Role Description: We evaluate job descriptions and roles using word lists that categorize terms associated with "soft" and "social" skills for feminine orientation, and "time-compressed" and "stressful" tasks, such as "multitasking," "pressure," "speed," for masculine orientation.

Work–Family Characteristics(WFC): This dimension examines employer policies and cultural expectations affecting gendered labor force participation, scrutinizing terms like "parental leave" and "flexible work" for feminine orientation versus "irregular and long work hours" and "weekend work" for masculine orientation.

Social Characteristics: We also analyze explicit gender references such as gendered pronouns and identity markers ("she," "he," "his," "her," "man").

Chapter 5

Conclusions and Future Work

5.1 Conclusion

This thesis has addressed the critical issue of social biases embedded in Natural Language Processing (NLP) models and textual data, offering innovative solutions for their evaluation and mitigation. Throughout this research, we have successfully developed and implemented algorithms that not only enhance the fairness and effectiveness of NLP applications but also push the boundaries of current methodologies in gender debiasing and bias evaluation. Our initial research phase introduced a novel method for reducing gender bias in static word embeddings, which preserved semantic integrity while achieving state-of-the-art results in debiasing tasks and improving performance in both word similarity evaluations and downstream NLP tasks. Subsequent studies expanded this approach to broader social biases, introducing a rigorous framework using Masked Language Models for quantitative bias assessment. This was particularly demonstrated in our large-scale evaluation of ChatGPT's performance in high-stakes environments like the job market, illuminating how generative AI can perpetuate social disparities. The final segment of our research introduced a groundbreaking statistical hypothesis-testing framework, using sentence perturbation techniques to detect subtle biases in MLM-generated texts, validated through empirical studies.

5.2 Future Work

Despite these advancements, the journey towards completely unbiased NLP models is ongoing, and several areas warrant further exploration:

- **Evaluation Benchmarks:** There is a significant scarcity of robust evaluation benchmarks for bias in textual contexts. Future research should focus on creating and refining these benchmarks to provide more reliable and standardized methods for bias evaluation.

- **Datasets:** The availability of training, testing, and validation datasets remains limited across various domains of bias evaluation. There is a pressing need to expand and diversify these datasets to enhance the comprehensiveness and reliability of bias assessments.

- **Interdisciplinary Approaches:** Integrating insights from psychology, sociology, and ethics can enrich bias mitigation strategies. Interdisciplinary research could lead to more holistic and effective solutions by incorporating diverse perspectives on what constitutes bias and fairness.

- **Broader Linguistic and Cultural Contexts:** Future research should consider applying the developed frameworks to a wider array of languages and cultural contexts. While most existing bias mitigation strategies focus predominantly on English, expanding this research to include other languages could help in understanding and reducing biases in a globally applicable manner.

- **Impact Assessments:** Future studies should also focus on the long-term impacts of debiased models in practical applications. It is crucial to evaluate whether changes at the algorithmic level translate into tangible improvements in fairness and equality in real-world scenarios.

By addressing these areas, future researchers can build on the foundation laid by this thesis, further enhancing the fairness, accuracy, and utility of AI and NLP technologies in diverse societal applications.

Bibliography

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[2] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[3] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[4] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, E3635–E3644, 2018.

[5] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.03310*, 2019.

[6] L. Douglas, "Ai is not just learning our biases; it is amplifying them," *Medium, December*, vol. 5, 2017.

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in Neural Information Processing Systems*, vol. 29, pp. 4349–4357, 2016.

[8] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *arXiv preprint arXiv:1809.01496*, 2018.

[9] M. Kaneko and D. Bollegala, "Gender-preserving debiasing for pre-trained word embeddings," *arXiv preprint arXiv:1906.00742*, 2019.

[10] Z. Yang and J. Feng, "A causal inference method for reducing gender bias in word embedding relations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9434–9441.

[11] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, "Measuring individual differences in implicit cognition: The implicit association test.," *Journal of personality and social psychology*, vol. 74, no. 6, p. 1464, 1998.

[12] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings," *NAACL*, 2019.

[13] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," *NAACL-HLT*, 2019.

[14] S. Dev and J. Phillips, "Attenuating bias in word vectors," in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 879–887.

[15] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong, "Double-hard debias: Tailoring word embeddings for gender bias mitigation," *arXiv preprint arXiv:2005.00965*, 2020.

[16] S. Shin, K. Song, J. Jang, H. Kim, W. Joo, and I.-C. Moon, "Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation," *arXiv preprint arXiv:2004.03133*, 2020.

[17] S. Bansal, V. Garimella, A. Suhane, and A. Mukherjee, "Debiasing multilingual word embeddings: A case study of three indian languages," in *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 2021, pp. 27–34.

[18] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *arXiv preprint arXiv:1706.02744*, 2017.

[19] V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang, "Handbook of partial least squares: Concepts, methods and applications," *Springer*, 2010.

[20] D. Yu, L. Kong, and I. Mizera, "Partial functional linear quantile regression for neuroimaging data analysis," *Neurocomputing*, vol. 195, pp. 74–87, 2016.

[21] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[22] J. Xie, Y. Lin, X. Yan, and N. Tang, "Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data," *Journal of the American Statistical Association*, vol. 115, no. 530, pp. 747–760, 2020.

[23] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[24] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, Citeseer, vol. 15, 2002, pp. 833–840.

[25] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[26] L. Finkelstein *et al.*, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 406–414.

[27] M.-T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology," in *Proceedings of the seventeenth conference on computational natural language learning*, 2013, pp. 104–113.

[28] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of artificial intelligence research*, vol. 49, pp. 1–47, 2014.

[29] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 337–346.

[30] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, "Large-scale learning of word relatedness with constraints," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1406–1414.

[31] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[32] D. Gerz, I. Vulić, F. Hill, R. Reichart, and A. Korhonen, "Simverb-3500: A large-scale evaluation set of verb similarity," *arXiv preprint arXiv:1608.00869*, 2016.

[33] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[34] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf, *et al.*, "Nonlinear causal discovery with additive noise models.," in *NIPS*, Citeseer, vol. 21, 2008, pp. 689–696.

[35] OpenAI, *Chatgpt*, Software, 2023. [Online]. Available: https://www.openai.com/chatgpt.

[36] W. X. Zhao *et al.*, *A survey of large language models*, 2023. arXiv: 2303.18223 [cs.CL].

[37] J. Yang *et al.*, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2024.

[38] P. Liang *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[39] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality.," *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[41] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[42] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López, "A survey on bias in deep nlp," *Applied Sciences*, vol. 11, no. 7, p. 3184, 2021.

[43] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp," *arXiv preprint arXiv:2005.14050*, 2020.

[44] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[45] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[46] L. Ding *et al.*, "Word embeddings via causal inference: Gender bias reducing and semantic information preserving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11 864–11 872.

[47] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 166–172.

[48] N. Nangia, C. Vania, R. Bhalerao, and S. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1953–1967.

[49] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5356–5371.

[50] M. De-Arteaga *et al.*, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 120–128.

[51] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case study of african-american english," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1119–1130.

[52] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 15–20.

[53] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, "On measuring and mitigating biased inferences of word embeddings," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 7659–7666.

[54] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1679–1684.

[55] J. Dhamala *et al.*, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 862–872.

[56] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[57] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, "Detecting gender stereotypes: Lexicon vs. supervised learning methods," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–11.

[58] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa, "Neural media bias detection using distant supervision with BABE - bias annotations by experts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1166–1177. DOI: 10.18653/v1/2021.findings-emnlp.101. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.101.

[59] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, "Nbias: A natural language processing framework for bias identification in text," *Expert Systems with Applications*, vol. 237, p. 121 542, 2024.

[60] R. Kjeldstad and E. H. Nymoen, "Underemployment in a gender-segregated labour market," *Economic and Industrial Democracy*, vol. 33, no. 2, pp. 207–224, 2012.

[61] P. England, "The gender revolution: Uneven and stalled," *Gender & society*, vol. 24, no. 2, pp. 149–166, 2010.

[62] F. D. Blau and L. M. Kahn, "The gender pay gap: Have women gone as far as they can?" In *Inequality in the United States*, Routledge, 2020, pp. 345–362.

[63] R. Glauber, "Trends in the motherhood wage penalty and fatherhood wage premium for low, middle, and high earners," *Demography*, vol. 55, no. 5, pp. 1663–1680, 2018.

[64] A. Killewald, "A reconsideration of the fatherhood premium: Marriage, coresidence, biology, and fathers' wages," *American sociological review*, vol. 78, no. 1, pp. 96–116, 2013.

[65] M. J. González, C. Cortina, and J. Rodríguez, "The role of gender stereotypes in hiring: A field experiment," *European Sociological Review*, vol. 35, no. 2, pp. 187–204, 2019.

[66] S. L. Bem, "The measurement of psychological androgyny.," *Journal of consulting and clinical psychology*, vol. 42, no. 2, p. 155, 1974.

[67] S. Hu *et al.*, "Balancing gender bias in job advertisements with text-level bias mitigation," *Frontiers in big Data*, vol. 5, p. 805 713, 2022.

[68] Y. Hu *et al.*, "Gendered stem workforce in the united kingdom: The role of gender bias in job advertising," 2022.

[69] A. Konnikov *et al.*, "Bias word inventory for work and employment diversity,(in) equality and inclusivity (version 1.0)," *SocArXiv*, 2022.

[70] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American statistical association*, vol. 69, no. 346, pp. 364–367, 1974.

[71] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, ISSN: 00994987. [Online]. Available: http://www.jstor.org/stable/3001968.

[72] C. May, A. Wang, S. Bordia, S. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 622–628.

[73] S. Goldfarb-Tarrant, R. Marchant, R. M. Sánchez, M. Pandya, and A. Lopez, "Intrinsic bias metrics do not correlate with application bias," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1926–1940.

[74] Y. Cao *et al.*, "On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 561–570.

[75] M. Kaneko, D. Bollegala, and N. Okazaki, "Debiasing isn't enough!–on the effectiveness of debiasing mlms and their social biases in downstream tasks," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 1299–1310.

[76] K. Carley, "Formalizing the social expert's knowledge," *Sociological Methods & Research*, vol. 17, no. 2, pp. 165–232, 1988.

[77] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[78] X. Shao, "A self-normalized approach to confidence interval construction in time series," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 3, pp. 343–366, 2010.