Scientific Research Publishing

Open Journal of Social Sciences, 2024, 12, 109-136
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Developing a Large-Scale Language Model to Unveil and Alleviate Gender and Age Biases in Australian Job Ads

Ruochen Mao, Liming Tan, Rezza Moieni, Nicole Lee

Diversity Atlas, Melbourne, Australia
Email: ruochen.m@diversityatlas.io, limtan@diversityatlas.io, rezza.moieni@diversityatlas.io, nicole.lee@diversityatlas.io

How to cite this paper: Mao, R. C., Tan, L. M., Moieni, R., & Lee, N. (2024). Developing a Large-Scale Language Model to Unveil and Alleviate Gender and Age Biases in Australian Job Ads. *Open Journal of Social Sciences, 12,* 109-136.
https://doi.org/10.4236/jss.2024.126006

Received: October 28, 2023
Accepted: June 14, 2024
Published: June 17, 2024

Abstract

This study aims to explore the application of large-scale language models in detecting and reducing gender and age biases in job advertisements. To establish gender and age bias detectors, we trained and tested various large-scale language models, including RoBERTa, ALBERT, and GPT-2, and found that RoBERTa performed the best in detecting gender and age biases. Our analysis based on these models revealed significant male bias in job ads, particularly in the information and communication technology, manufacturing, transportation and logistics, and services industries. Similarly, research on age bias revealed a preference for younger applicants, with limited demand for older candidates in job ads. Furthermore, we explored the application of natural language generation using ChatGPT to mitigate gender bias in job advertisements. We generated two versions of job ads: one adhering to gender-neutral language principles and the other intentionally incorporating feminizing language. Through user research, we evaluated the effectiveness of these versions in attracting female candidates and reducing gender bias. The results demonstrated significant improvements in attracting female candidates and reducing gender bias for both versions. Overall, gender bias was reduced, and the appeal of job ads to female candidates was enhanced. The contributions of this study include an in-depth analysis of gender and age biases in job advertisements in Australia, the development of gender and age bias detectors utilizing large-scale language models, and the exploration of natural language generation methods based on ChatGPT to mitigate gender bias. By addressing these biases, we contribute to the creation of a more inclusive and equitable job market.

Keywords

Gender Bias, Age Bias, Natural Language Generation, Large Language

Models, Machine Learning, Natural Language Processing

1. Introduction

Gender bias and age bias hold significant importance in Australian job advertisements. These biases manifest through gendered and age-related language. Gendered language utilizes words and expressions with masculine or feminine tones, while age-related language incorporates words associated with age-related stereotypes. For example, words such as "glamorous" and "nurturing" often resonate with stereotypical feminine work environment. While words such as "young" and "fresh grad" could turn off the older candidates. The Victorian government has committed to promoting gender equality and inclusivity and has been using more gender-neutral and inclusive terms in communications over the years (Raichur, Lee, & Moieni, 2023). Despite advancements in governmental domain, male dominance persists in specific industries like STEM, and older job seekers encounter challenges in reemployment or reentering the workforce.

Gendered language in job advertisements influences applicants' perception of job descriptions and gender equality. Advertisements adopting masculine language tend to discourage qualified female applicants, exacerbating gender inequality in the workforce (Bem & Bem, 1973; Arceo-Gomez et al., 2022; Askehave & Zethsen, 2014). In a study of Gaucher, Friesen and Kay (2011), adult readers expected a greater number of men in the occupation when the job advertisements use masculine language. Likewise, age bias subtly embedded in job advertisements restricts employment opportunities for older job seekers (Burn et al., 2019; Burn et al., 2021). To address these issues, Australia and other countries have enacted anti-discrimination labour laws such as the Sex Discrimination Act 1984 (SDA) (Australian Government, 2014a) and the Age Discrimination Act 2003 (ADA) (Australian Government, 2014b). Unfortunately, discriminatory practices in job ads often remain implicit, as research suggests that companies utilizing age-biased language tend to favour younger applicants (Burn et al., 2019; Burn et al., 2021). Detecting and mitigating these biases is crucial for fostering fair hiring practices.

Multiple methods have been developed to tackle gender and age biases in job advertisements, including regular expression matching (Ningrum, Pansombut, & Ueranantasun, 2020), lexicon-based approaches (Hu et al., 2022; Tang et al., 2017), end-to-end methods (Cryan et al., 2020), and hybrid methods (Böhm et al., 2020; Frissen, Adebayo, & Nanda, 2022). These approaches employ language models and classifiers to analyze the degree and severity of biases present in job ads.

This research aims to collect job advertisement data from the Australian job search website and employs advanced large-scale language models such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and GPT-2 (Radford et al.,

2019) to construct gender bias detectors and age bias detectors. These detectors are deployed to detect and analyze gender and age biases across various industries in Australia. A natural language generation method, based on ChatGPT, is proposed to eliminate gender bias and enhance job advertisements' appeal to female candidates. Two approaches are implemented: one aligned with gender-neutral principles (adhering method) and the other intentionally incorporates language favouring underrepresented genders (steering method). The effectiveness of these methods is evaluated through user research, gathering feedback on job advertisements' attractiveness, and perceived levels of gender bias.

This research contributes the following:

- Valuable insights into the prevalence of gender and age biases in Australian job advertisements.
- We developed gender bias detectors and age bias detectors to detect gender bias and age bias in advertisement texts, which can also be applied to other text detection tasks. This study represents the first research in utilizing large-scale language models to detect age bias.
- Insights into the efficacy of ChatGPT-based natural language generation in addressing gender bias.

By addressing these biases, we contribute to creating a more inclusive and equitable job market, ensuring equal opportunities for all applicants irrespective of their gender or age.

2. Background

1) Discrimination and inequality in job ads

Age and gender biases in job advertisements can be either explicit or subtle. Explicit biases refer to direct, obvious, and overt discriminatory expressions, while subtle biases are conveyed through language in a subtle manner (Gaucher, Friesen, & Kay, 2011). These biases often manifest through gender-coded terms. Much research indicates that we use different words and expressions when describing men and women. Trix and Psenka (2003) found gender differences in recommendation letters for medical faculty, where recommendations for female applicants were confined to teaching roles, while male applicants were recommended for research and other professional roles. Guerin (1994) revealed underlying gender biases in word usage through participants' use of abstract words to describe their own gender behaviours.

Multiple studies have shown that gender-coded words have a significant impact on how potential applicants perceive job descriptions and gender equality in different industries. Job advertisements that contain more male-coded language tend to deter female applicants even when they are well qualified for the role (Bem & Bem, 1973). Another instance of gender stereotypes contributing to the gender pay gap in the labour market is where job advertisements associated with female traits offer lower salaries than those associated with male traits, indicating the influence of gender stereotypes on the gender pay gap (Arceo-Gomez

et al., 2022). Gender biases exist in job ads for all levels, including top management positions, with most of the wording aligning with traditional male attributes (Askehave & Zethsen, 2014). Gaucher et al. (2011) found that job advertisements in male-dominated industries tend to be biased toward male applicants by using words like "leader," "competitive," and "dominant," while job advertisements in female-dominated industries show bias towards female applicants. If job advertisements unknowingly adopt the male-coded language, they may continue to attract predominantly male applicants, perpetuating a vicious cycle and reinforcing a male-dominated culture.

Biases towards specific age groups are also a significant issue. Similar to gender biases, many studies suggest that potential age biases are embedded in language. Moon (2014) found that English adjectives used to describe young people are mostly positive, while those used to describe older people are often negative. Diaz et al. (2018) analysis of age-related sentiment on Twitter also revealed a more negative sentiment in tweets related to older individuals. Burn et al. (2019) collected words from job advertisements related to age stereotypes from the perspectives of health, personality, and skills, such as "cannot accept new technology." This study indicated that these age stereotypes unconsciously hinder older potential applicants. Age biases can lead to greater difficulties for older job seekers in finding employment opportunities. This bias has profound implications for economic productivity and social equality, particularly considering the aging populations in many societies, including Australia.

2) Detect gender bias in job ads

In the context of detecting gender bias in job advertisements, the regular expression matching method involves directly searching for gender-related keywords such as "male" and "female" (Ningrum, Pansombut, & Ueranantasun, 2020). Lexicon-based methods rely on establishing male-oriented and female-oriented word libraries and determine the presence of gendered language by analyzing the proportions of masculine and feminine words in the text (Cryan et al., 2020; Tang et al., 2017; Hu et al., 2022). On the other hand, end-to-end methods do not require the extraction of specific words from the text but directly input the text into a large-scale deep-learning model to obtain the probability of the text leaning towards a male or female stereotype.

Ningrum et al. (2020) utilized regular expressions and n-grams to match gender-discriminatory keywords (such as man, male, female, and women), demonstrating the effectiveness of this method in identifying gender bias in recruitment ads. However, the regular expression matching approach may not detect more implicit forms of gender bias, such as those related to personality traits, skills, and abilities.

Further research has been conducted using the other two methods. In terms of employing lexicon-based methods to detect the degree of gender bias. Gaucher, Friesen and Kay (2011) calculated the proportions of masculine and feminine words in job ads to highlight differences in language use between male-dominated

and female-dominated domains. Hu et al. (2022) introduced a novel method for measuring gender bias in job ads. This method involved using a pre-trained word embedding models like GloVe (Pennington, Socher, & Manning, 2014) to categorise words into strongly masculine, weakly masculine, strongly feminine, and weakly feminine categories. The study then employed two different bias assessment methods: word-based counting (similar to Gaucher's approach) and similarity-based calculations such as Relative Norm Distance (Garg et al., 2018) and the Word Embedding Association Test (Caliskan, Bryson, & Narayanan, 2017) to evaluate gender bias in job ads. Tang et al. (2017) allocated weights to gender-indicative words based on their implied gender levels and calculated the number of gendered words to determine the target gender group of the ads. And used the sigmoid function to obtain a gender target score, and the gendered tone considered the meaning and bias of the words, combining weighted gender-oriented words to obtain a total score. These two indicators can evaluate gender bias in job ads, with the word categories and weights organized through tools.

Cryan et al. (2020), in their study, established a gender bias word library and the first manually annotated text corpus of gender bias. They compared the performance of lexicon-based methods and end-to-end-based methods in identifying gender bias, with the results indicating the superiority of end-to-end-based methods over lexicon-based methods. Some studies have also focused on collecting and identifying gender-biased words, such as masculinised and feminised words. For example, the Multi-Dimensional Gender Bias Classification dataset (Dinan et al., 2020) is built on a generic framework that decomposes gender bias in the text along practical and semantic dimensions, including gender bias towards mentioned individuals, gender bias towards said individuals, gender bias towards the target of speech, and gender bias towards the speaker. The Gender-Bias-Datasets-Lexicons (Doughman & Khreich, 2022) publicly provides labelled datasets and comprehensive lexicons by collecting, annotating, and augmenting relevant sentences. These gender bias datasets provide strong support for end-to-end-based methods.

In terms of employing a mixed-method approach, Böhm et al. (2020) focused on gender bias in IT job advertisements in the German employment market and utilized job descriptions as the dataset and employed deep learning techniques to establish a gender recognition model. Through manual annotation using keywords, they constructed a lexicon containing stereotypical terms associated with men and women. A custom vectorisation algorithm was employed to perform clustering analysis on these keywords. The study revealed a higher correlation between male stereotypical terms and gender bias space. Based on these analyses, three distinct lexicons were created for replacing potentially biased terms, providing an encouraging vocabulary for female applicants, and offering language that attracts female candidates. After building the lexicons, the researchers proposed a method for calculating gender bias in job recruitment using a single

score to convey information about the number of biased terms associated with "push" and "pull" biases. Frissen, Adebayo and Nanda (2022), on the other hand, investigated the identification and classification of bias and discriminatory language in job advertisements. They employed Machine Learning techniques and constructed a lexicon comprising unique biased and discriminatory terms based on relevant literature in behavioural studies. Subsequently, they used a word-based approach to annotate job ads from the publicly available dataset, resulting in an annotated corpus. The annotated corpus was utilized to train state-of-the-art Machine Learning classifiers, combining linguistic features with the latest word embedding representations to capture the natural language semantics of the biased language.

3) Detect age bias in job ads

Research on age bias in ads has been far less extensive than research on gender bias in job ads. Currently, most detection methods involve regular expression keyword matching and lexicon-based methods using age-discrimination word datasets.

Ningrum, Pansombut and Ueranantasun (2020) directly used regular expression matching and n-gram methods to detect age discrimination in job ads in Indonesia. They searched for word phrases related to "age" in the ad text, such as "between age" and "underage", and found that more than 50% of job ads contained age discrimination. Pillar, Poelmans and Larson (2022) also employed a regular expression matching approach to detect age bias in job ads, but she pointed out that while regular expression matching can identify discrimination in recruitment ads, it cannot understand semantic information. For example, a sentence containing the word "age" or "male" may not necessarily convey discriminatory information. Therefore, the regular expression matching method has certain limitations.

Burn et al. (2019) introduced a method that combines Machine Learning analysis and experimental measurement to investigate whether the use of age-related stereotypes in job ads by certain companies leads to fewer older applicants applying for these positions. They collected age stereotyping words from psychological literature and analyzed their semantic relevance to word phrases used in job ads. The results showed that job ads using these age stereotyping words made older applicants less willing to apply for the positions. This study provides substantial support for subsequent research on age discrimination in job advertisements. In subsequent research, Burn et al. (2021) explored ageist stereotypes in recruitment ads by measuring the semantic similarity between job ad texts and age stereotyping words. They investigated whether discriminatory stereotypes against older people exist in recruitment ads and whether these languages are perceived as biased against older individuals. The study demonstrated that Machine Learning methods can sensitively detect the presence of stereotyping language. In a user survey, the experiments also revealed that sentences classified as age discrimination-related by the Machine Learning algorithm were generally perceived as ageist by people.

4) Gender debiasing method in job ads

The main approaches to mitigating gender bias in textual content and recruitment advertisements involve the removal or substitution of gender-biased terms, the generation of new text, and the redesign of recruitment processes.

Sczesny, Formanowicz and Moser (2016) research highlights the significant impact of linguistic forms on gender equality. Using gender-fair language (GFL) can alter people's mental representations and promote gender equality. Hu et al., (2022) devised a debiasing strategy and algorithm by modeling the frequencies of each word group and sampling the word composition. Böhm et al. (2020) developed a tool called "betterads" that detects and prompts gender-biased words in advertising texts to reduce gender discrimination.

Strengers et al. (2020) introduced the contributions of natural language generation (NLG) to gender equity in society. They outlined three approaches to treating gender under NLG: adhering, steering, and queering, and applied them to job advertisements. In the adhering approach, the treatment involves neutralizing the text or removing any gender-specific or sensitive attributes from the advertisement to promote equality. The steering approach intentionally uses language that favors a specific gender or gender group in job advertisements to encourage applications from the desired gender or gender group.

Kanij et al. (2022) conducted a study using the GenderMag method to assess gender bias in software engineering job advertisements and found that cognitive walkthroughs effectively identified potential biases. They plan to expand the investigation and apply the InclusiveMag method in further research.

3. Data Collection

We obtained job advertisement data from the Australian job website. These job advertisements are publicly available and do not require a registered account to view. All online job advertisement websites have their own classification for different types of jobs and the number of job advertisements in each category. Based on this information, we selected seven popular industry categories with a high number of job advertisements: Advertising, Arts & Media; Education & Training; Healthcare & Medical; Hospitality & Tourism; Information & Communication Technology; Manufacturing, Transport & Logistics; and Trades & Services.

We developed a program using Python's BeautifulSoup library to scrape job advertisements, successfully retrieving a total of 21,683 data entries. **Figure 1** illustrates the distribution of data across different industries, while **Figure 2** showcases the distribution of job types. Each job advertisement includes the following information: job title, company name, location, category, job description, job URL, and job type (full-time, casual/vacation, contract/temp, and part-time).

4. Methodology

Our objective is to examine gender bias and age bias in job advertisements

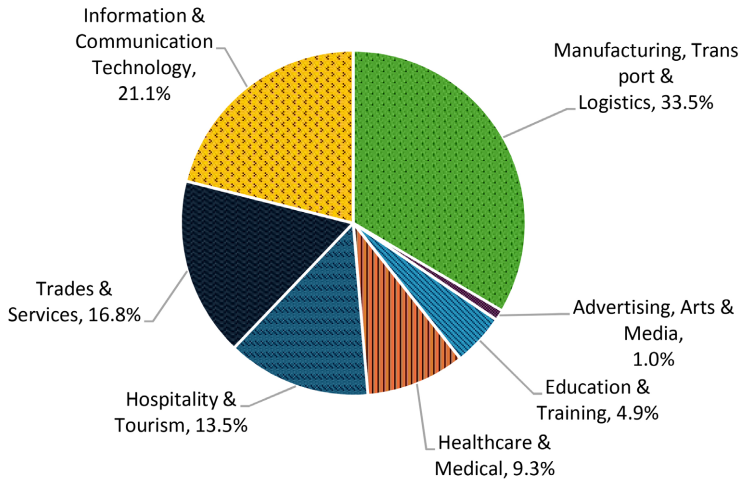

Figure 1. Percentage of job categories.

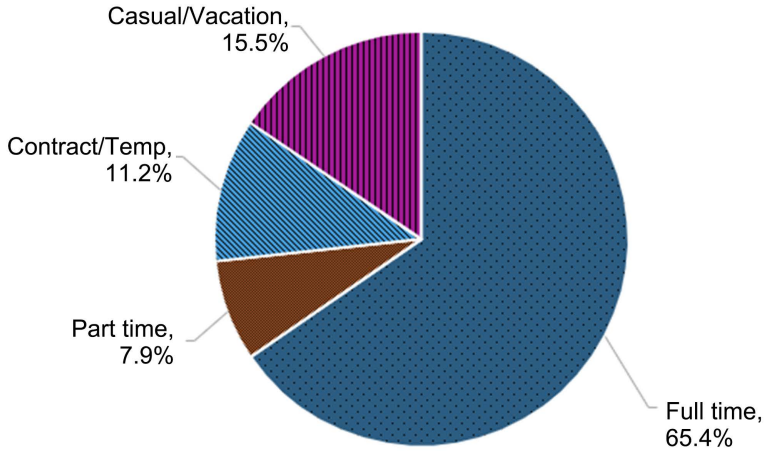

Figure 2. Percentage of job types.

within popular industries in Australia and explore methods to mitigate gender bias in these ads. To accomplish this, our task is divided into two stages.

In the first stage, we develop detectors to accurately identify gender bias and age bias independently, as depicted in **Figure 3** and **Figure 4**. The gender bias detector utilizes an end-to-end approach, comparing the performance of three large language models: RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and GPT-2 (Radford et al., 2019). Similarly, for the age bias detector, we adopt a hybrid approach that combines end-to-end and lexicon-based methods, evaluating the performance of the three models. These three models have demonstrated excellent performance on various NLP tasks.

- RoBERTa: It stands for "Robustly optimized Bidirectional Encoder Representations from Transformers approach", which is an optimized model based on BERT architecture (Devlin et al., 2019). It is pretrained on 160 GB of unlabeled text data and improved training methods to enhance performance. It is capable of understanding context in text and generating human-like and longer sentences.

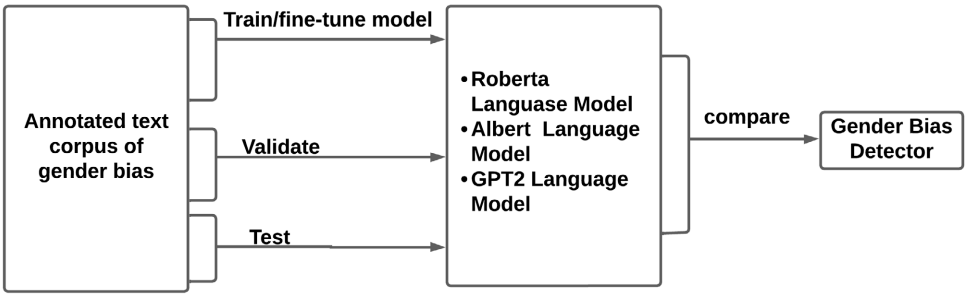

Figure 3. Pipeline for building gender-bias detector.

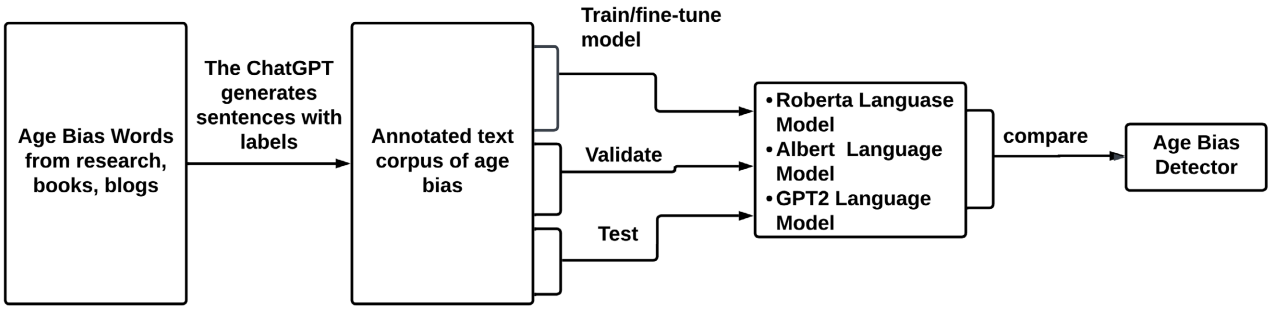

Figure 4. Pipeline for building age-bias detector.

- ALBERT: It stands for "A Lite BERT", which is a model based on BERT architecture with fewer parameters. It employs techniques such as parameter sharing and parameter factorization to reduce the number of parameters while maintaining high performance. It is pretrained on 16 GB of text data, and is capable of generating human-like sentences.
- GPT-2: It stands for "Generative Pre-trained Transformer 2", which is an optimised model of GPT. It is based on the Transformer architecture (Vaswani et al., 2017) and trained on a large-scale unsupervised dataset of 8 million web pages, and capable of predicting text from the input text.

The second stage focuses on exploring methods to reduce gender bias in job advertisements, with a specific emphasis on the IT industry, as shown in **Figure 5**. We gathered the top 40 job advertisements in the IT industry that have been identified with the highest gender bias by our gender bias detector. Subsequently, we utilize ChatGPT (Open AI, 2023) as a tool to generate new job advertisement texts based on two feminist-oriented natural language generation approaches: "adhering" and "steering" (Strengers et al., 2020). We then test whether the gender bias level has been reduced by inputting the generated texts into our gender bias detector. In the subsequent section, we present the outcomes of these experiments and proceed with a confirmatory user study.

5. Bias Detector Development

1) Data

There are various methods for annotating datasets, including crowd annotation (Cryan et al., 2020), where workers on crowdsourcing platforms annotate

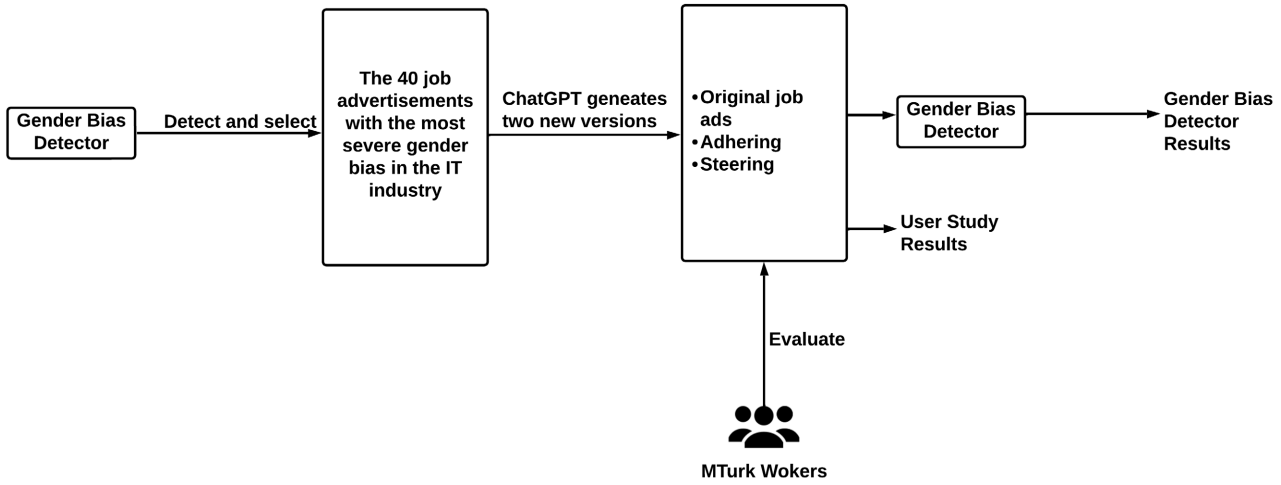

Figure 5. Pipeline for building gender-bias detector and generate results from the 40 job advertisements in IT industry.

data according to specified criteria and submit the results; rule-based annotation (Frissen, Adebayo, & Nanda, 2022), where data is automatically annotated based on predefined rules and models; semi-supervised learning (Böhm et al., 2020; Doughman & Khreich, 2022), which combines labeled and unlabeled data for annotation. By the end of 2022, with the widespread popularity of ChatGPT worldwide, its emergence has brought significant changes to many industries. Some studies have started exploring the potential of ChatGPT as an annotation tool. According to Gilardi, Alizadeh, and Kubli's findings (2023), they manually annotated 2382 tweets and compared the classification performance of ChatGPT and MTurk on the same tasks. They found that ChatGPT outperformed MTurk in all four tasks in terms of accuracy. Furthermore, ChatGPT's accuracy demonstrated a certain level across different tasks and category sizes, especially considering that these annotations were zero-shot. Gilardi's study demonstrated the remarkable performance of ChatGPT on complex tasks and provided a potential solution for zero-shot annotation. Kuzman, Mozetič and Ljubešić (2023) research revealed that ChatGPT performed better than the fine-tuned X-GENRE classifier in the task of automatic genre identification. Even in under-resourced languages like Slovenian, ChatGPT's performance was on par with English. However, prompting the model in under-resourced languages resulted in decreased performance. These results highlight the potential of ChatGPT in reducing manual annotation efforts.

a) *Gender bias dataset*: For training the gender bias detector, we utilized the "Annotated Text Corpus of Gender Bias" dataset (Cryan et al., 2020). This dataset was annotated through manual labeling, where different texts describing males and females were categorized into four labels: male-consistent (m_consis), male-contradictory (m_contra), female-consistent (f_consis), and female-contradictory (f_contra). The dataset comprises 4550 labeled instances, with 1138 instances labeled as m_consis, 1130 as f_consis, 1140 as m_contra, and 1142 as f_contra.

b) *Age bias dataset*: When it comes to age bias, there is a lack of pre-existing annotated datasets similar to the annotated text corpus of gender bias. To address this, we initially gathered age-related stereotypical terms for both young and old individuals from relevant research papers and articles, as well as blog posts (Gendron et al., 2016; Reissmann et al., 2021; Australian Human Rights Commission, 2013; Devlin, 2006). Based on studies exploring the possibility of labeling with ChatGPT, we utilized ChatGPT to generate stereotypical sentences corresponding to these age-related terms, along with their respective labels. Similar to the annotated text corpus of gender bias, the labels for age bias are categorized into four types: old-consistent (o_consis), old-contradictory (o_contra), young-consistent (y_consis), and young-contradictory (y_contra). The dataset consists of 448 instances related to young individuals, with 229 instances labeled as y_consis and 219 instances as y_contra. Similarly, there are 443 instances related to old individuals, with 225 instances labeled as o_contra and 218 instances as o_consis.

2) Creating models

We aim to develop a method that can detect subtle gender and age biases in textual descriptions and visually present these differences. To achieve this goal, it seems essential to establish gender bias detectors and age bias detectors.

After obtaining the relevant dataset, we proceed to detect the level of gender bias in a given text by building two classifiers: a male language classifier and a female language classifier. The male language classifier outputs two probabilities: the probability of alignment with male stereotypes and the probability of contradiction. Similarly, the female language classifier provides similar outputs. These probabilities range from 0 to 1. Based on previous similar studies (Cryan et al., 2020; Tang et al., 2017; Hu et al., 2022), we define the gender bias score of a text as the probability of alignment with female stereotypes v_f minus the probability of alignment with male stereotypes v_m. In other words, the Gender Bias Value (GBV) ranges from −1 to 1, with a value closer to 1 indicating a language bias towards femininity, closer to −1 indicating a bias towards masculinity, and a value of 0 representing neutral language, as shown in Formula (1).

$$\mathrm{GBV} = v_f - v_m \tag{1}$$

Similarly, for age bias detection, we establish two classifiers: an older people language classifier and a younger people language classifier. The final Age Bias Value (ABV) is calculated as the probability of alignment with younger people stereotypes v_y minus the probability of alignment with older people stereotypes v_o, as shown in Formula (2). The result falls between −1 and 1. A value closer to −1 indicates a language bias toward older people, a value closer to 1 indicates a bias toward younger people, and a value closer to 0 represents language with no age bias. In this way, we have established a standardized measurement for detecting gender bias and age bias.

$$\mathrm{ABV} = v_y - v_o \tag{2}$$

With this standardized measurement established, we select three large language models: RoBERTa, ALBERT, and GPT-2, to build corresponding classifiers for male language, female language, older people language, and younger people language. We partitioned the gender bias and age bias datasets into training, validation, and testing sets in a 6:2:2 ratio. The RoBERTa-Large (huggingface.co., n.d. a), GPT-2 (huggingface.co., n.d. b), and ALBERT-Base-v2 (huggingface.co., n.d. c) models from Hugging Face were utilized as pre-trained models in this study. These models were fine-tuned by incorporating them into downstream tasks. Table 1 shows the hyperparameter information of each model. And the performance of these three models on various classification tasks is presented in Table 2 and Table 3.

For gender bias detection, the RoBERTa model achieved an accuracy of 0.7912 and an AUC (Area Under Curve) of 0.8660 in identifying female bias, and an accuracy of 0.7982 and an AUC of 0.8785 in identifying male bias, demonstrating excellent performance in gender bias detection. Data analysis reveals that RoBERTa exhibited the best performance in identifying gender bias, followed by GPT-2, while the ALBERT model performed slightly lower in terms of performance.

For age bias detection, the RoBERTa model achieved a high accuracy of 0.9889 and an AUC of 0.9965 in detecting bias towards young individuals, and an accuracy of 0.9101 and an AUC of 0.9454 in detecting bias towards elderly individuals, indicating outstanding performance in age bias detection. In the domain of age bias identification, RoBERTa displayed remarkably high accuracy and AUC, followed by GPT-2, while the ALBERT model showed slightly lower performance. Overall, RoBERTa demonstrated the best performance in gender bias and age bias detection tasks, with GPT-2 and ALBERT relatively falling behind in terms of performance. Based on these results, we ultimately selected the RoBERTa model as our gender bias and age bias detectors, with hyperparameters set at 2e−6 and 6 epochs.

6. Analysis of Model Results

1) Gender bias analysis

After using bias detector models for our test data, we obtained the bias value of each job advertisement, and then we will analyze the detection effect of the model.

Table 1. Model hyperparameters.

Model name	Base model name	Learning rate	Train batch size	Optimizer
RoBERTa	RoBERTa-Large	2e−6	6	AdamW with betas = (0.9, 0.999) and epsilon = 1e−8
ALBERT	albert-base-v2	1e−5	2	AdamW with betas = (0.9, 0.999) and epsilon = 1e−8
GPT-2	gpt2	2e−5	1	AdamW with betas = (0.9, 0.999) and epsilon = 1e−8

Table 2. Performance of models in detecting gender bias.

Model	Acc (F)	AUC (F)	Acc (M)	AUC (M)
RoBERTa	0.7912	0.8660	0.7982	0.8785
GPT-2	0.7780	0.8369	0.8162	0.8649
ALBERT	0.7682	0.8047	0.7763	0.7945

Table 3. Performance of models in detecting age bias.

Model	Accuracy (Young)	AUC (Young)	Accuracy (Old)	AUC (Old)
RoBERTa	0.9889	0.9965	0.9101	0.9454
GPT-2	0.9222	0.9816	0.9551	0.9959
ALBERT	0.9380	0.9654	0.9541	0.9784

Firstly, we performed macro statistics on the data, here a simple weighting algorithm was used to calculate the overall value of gender bias for all job ad data for both genders, for example overall bias for males value = Formula (3), where v denotes the bias value for each text given by the model, n denotes the number of this bias value, and N denotes the size of different bias values in the data set.

$$\sum_{i=1}^{N} v_i \cdot n \tag{3}$$

The results of the calculation show that the overall bias value for men is 0.64 and for women is 0.51. Therefore, the degree of bias for men is slightly greater than that for women in the job advertisements, i.e., these job advertisements show more masculine words. Moreover, we use GBV (Formula (1)) to measure the bias degree in recruitment. As shown in Figure 6, the texts with absolute bias values over 0.6 points occupy 60.9% of all data, and the bias value increases significantly after 0.6, which indicates that the bias felt by most people in the job advertisements is relatively strong.

In addition to looking at the overall data, we also analyzed the gender bias in different fields, and the results presented an interesting data distribution as shown in Table 4. We found that male bias is significantly greater than female bias in Information & Communication Technology, Manufacturing, Transport & Logistics, and Trades & Services. Most of these fields are in the technology sector or require some degree of physical labor. Companies in these fields prefer to hire employees with certain masculine traits, while Advertising, Arts & Media, Education & Training, Healthcare & Medical, Hospitality & Tourism prefer female employees. The distribution of bias value given by the model is consistent with our intuitive impressions about these occupations in our daily life, which indicates that our model is valid.

Another perspective for the analysis of the experimental data can be developed based on job types, our data contains different job types (Full time, Casual/Vacation, Part-time, Contract/Temp). In using job types as a new perspective to analyze gender bias, we found some new characteristics of data distribution.

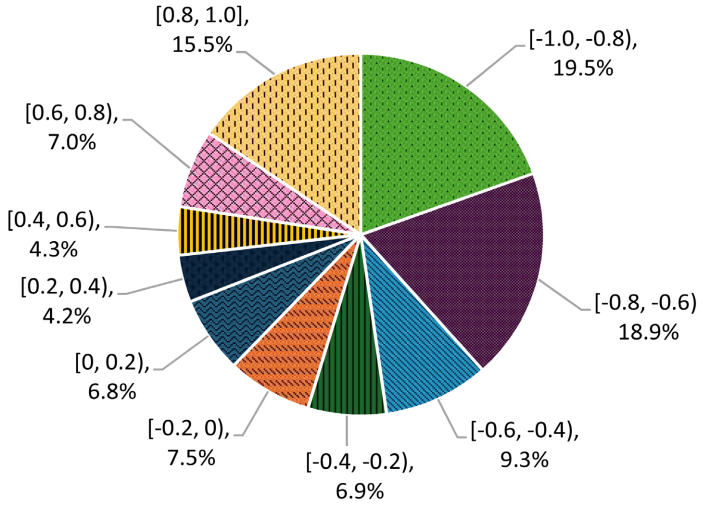

Figure 6. Gender bias distribution.

Table 4. Gender bias distribution over categories.

Category	Mean (GBV)	Std (GBV)	Mean Male Bias	Mean Female Bias
Advertising, Arts & Media	0.30	0.49	0.28	0.58
Education & Training	0.61	0.45	0.17	0.78
Healthcare & Medical	0.62	0.40	0.22	0.84
Hospitality & Tourism	0.55	0.43	0.32	0.88
Information & Communication Technology	−0.51	0.44	0.78	0.27
Manufacturing, Transport & Logistics	−0.39	0.48	0.84	0.45
Trades & Services	−0.45	0.51	0.82	0.36

Firstly, we divided the different job types into 4 levels, which are based on experience, i.e. from Part-time to Contract gradually requires more experience. From Figure 7, we can see that the degree of gender bias decreases as the level of job advertisement increases, i.e., the required experience increases. This may be due to the fact that as the skills required for the job become more complex, the gender factor tends to equalize in the job advertisement, i.e., the hiring company no longer has implicit expectations about the gender of the employee. It is also worth noting that with the increase in experience, companies prefer to hire men, while the preference for women in job postings is declining. According to the data in Figure 8, this trend is evident and shows a similar pattern across industries.

2) Age bias analysis

In general, our model gives data indicating that there is a much greater preference for younger than older job seekers in job ads across all industries. Here we define an age bias value, ABV = young bias - old bias, similar to GBV, which ranges from −1 to 1. If the ABV value of a job ad is larger, it indicates that the

company wants to hire younger candidates. After the statistics, as shown in Figure 9 below, more than 77.1% of the job advertisements have an ABV value greater than 0.6, showing a higher implicit age requirement for job seekers (requiring young people). About 17.9% of these job postings show a very strong age bias (ABV larger than 0.8), and only a very small number of companies show interest in older people in their job postings (less than 3%). This situation does not change significantly even for contract-type jobs that require more experience. As shown in Figure 10, almost all types of job advertisements want to hire young people.

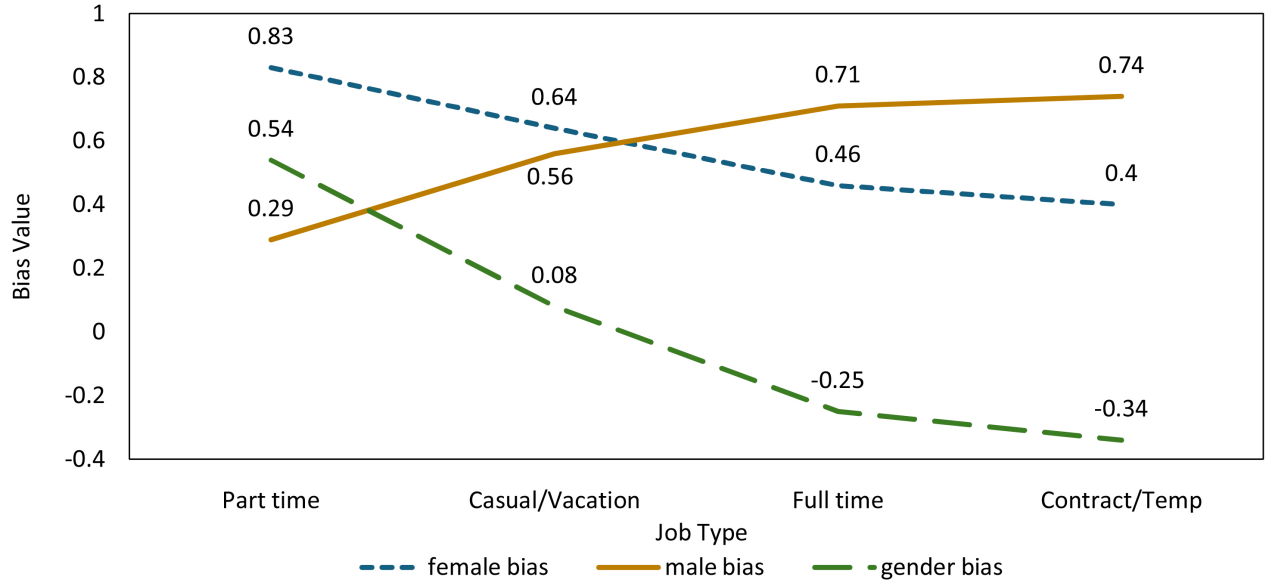

Figure 7. Gender bias values over job types.

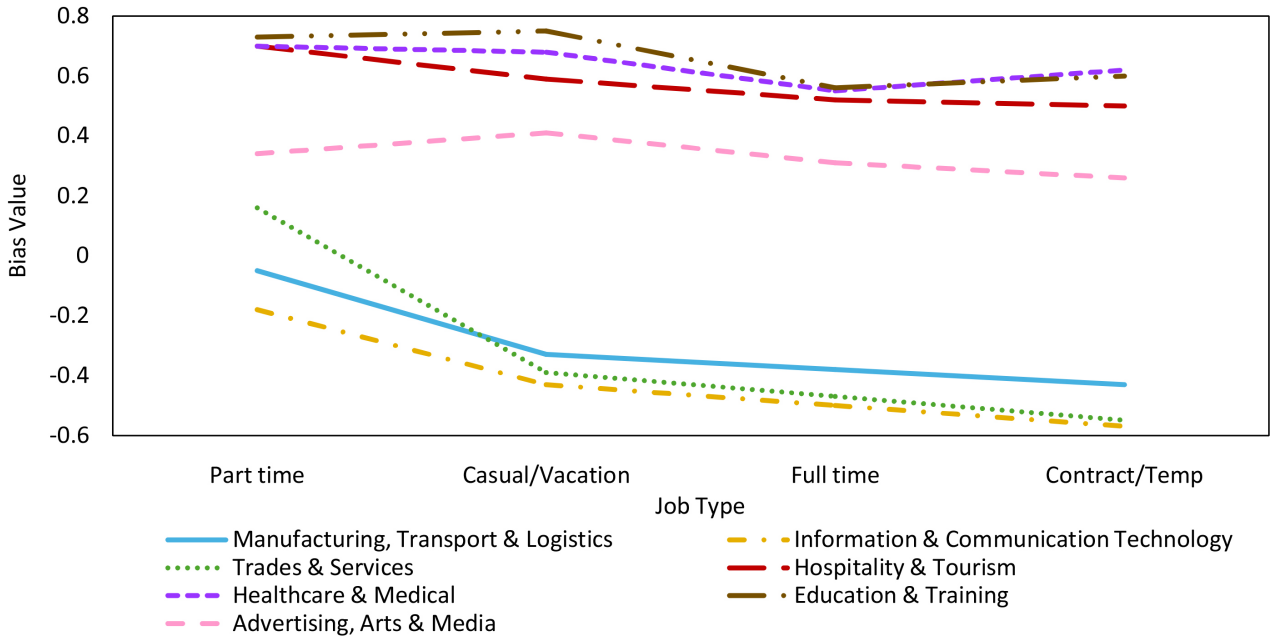

Figure 8. Gender bias distribution over job types and industries.

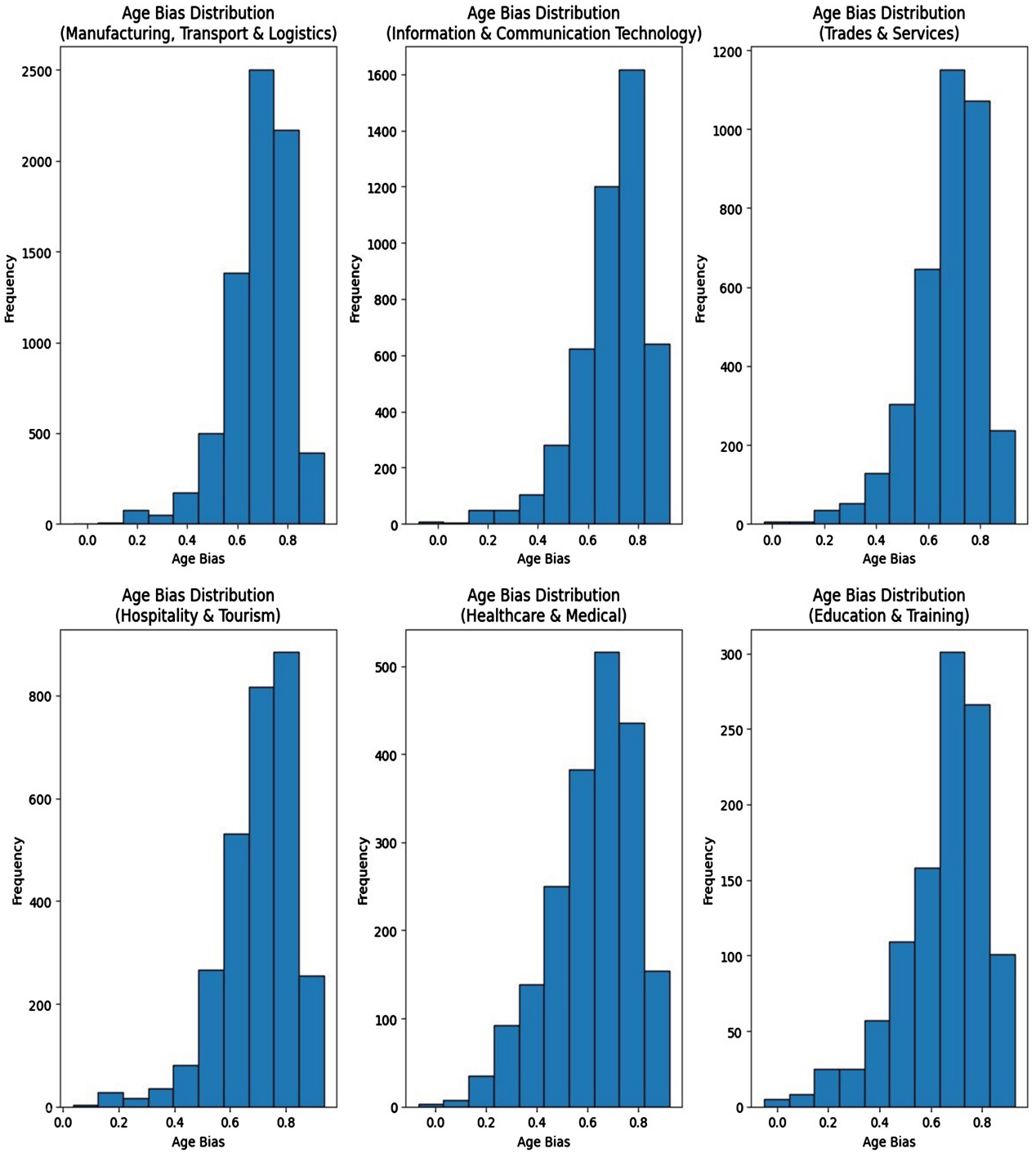

Figure 9. Age bias distribution over industries.

To summarize, in this section, we found some patterns of data distribution through macro analysis of data, analysis of data sub-domains and changes of data with job type level. The pattern of data distribution and the results of our analysis are consistent with our daily intuition and have overall and local validity, which proves to some extent the effectiveness of the model for predicting the bias value of the data, which is also reflected in our follow-up survey.

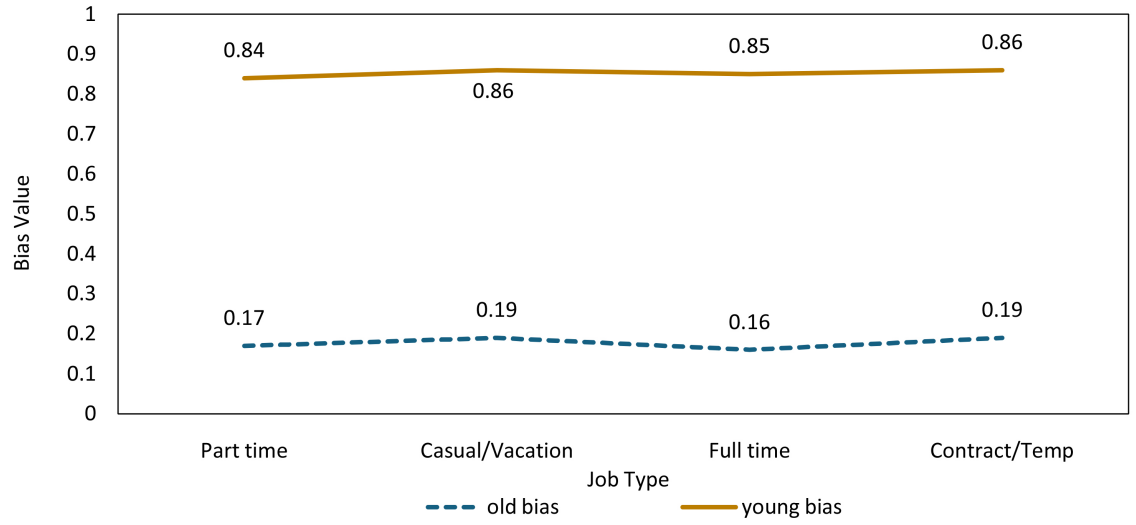

Figure 10. Age bias distribution.

7. Gender Debiasing Method Design

1) Data

According to our analysis, the IT industry exhibits significant gender bias. To study gender debiasing methods, we selected 40 job advertisements with the most pronounced gender bias from this industry as our samples. **Figure 11** showcases the distribution of these job advertisements. The average of the bias scores for these advertisements is −0.9184, with a variance of 0.0004.

2) Methodology

We aim to explore the use of generated text to reduce gender bias in job advertisements. The process of the debasing method experiment is illustrated in **Figure 5**. Strengers et al. (2020) proposed two methods, namely adhering and steering, to mitigate gender discrimination in job advertisements using natural language generation techniques.

- **Adhering** involves generating new text by adhering to guidelines that remove expressions containing sensitivity and gender bias.
- **Steering** intentionally incorporates language that appeals to underrepresented gender groups in specific industries, aiming to attract more diverse practitioners.

The implementation of the gender debasing method involves two steps. Firstly, in this experiment, we used ChatGPT as the tool for generating new text. We replaced the original company names and address information in the job advertisements, replacing all company names with "companyA" and all location information with "locationA". We then generated new text for the 40 job advertisements with the highest gender bias in the IT industry, based on the adhering and steering guidelines, using a gender bias detector we built. Secondly, we evaluated the gender bias scores of the two versions of newly generated texts by inputting them into the gender bias detector. The purpose was to determine whether the gender bias scores decreased for the texts generated by ChatGPT.

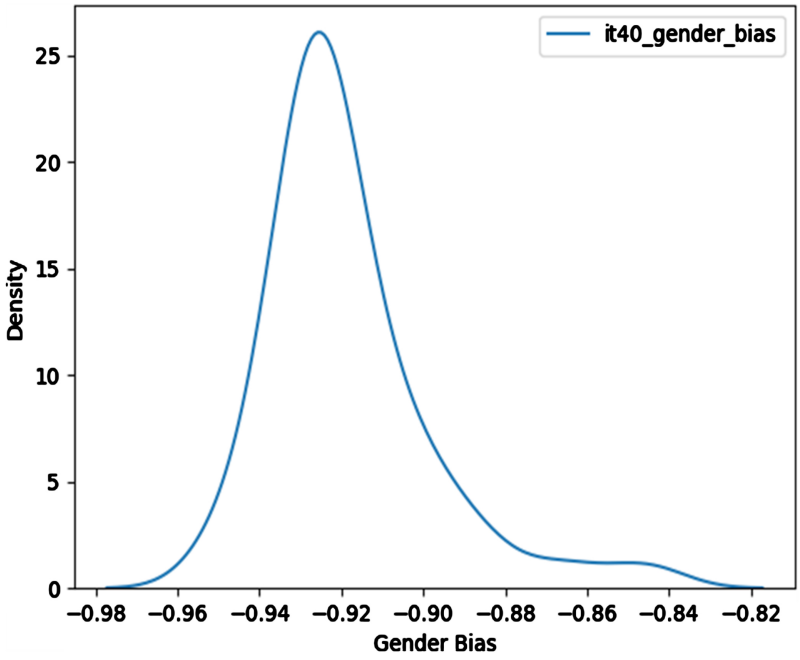

Figure 11. Distribution of selected job advertisements.

Table 5 presents the prompts for the steering and adhering approaches, with a temperature value of 0.5 employed to balance text generation diversity and determinism. Afterward, we conducted a user study to explore the effectiveness of the adherence and steering methods.

3) Participants

We recruited 410 distinct workers from Amazon Mechanical Turk (MTurk) as our survey respondents. Each worker who successfully completed the task received a payment of $1. To ensure the quality of the responses, we restricted the workers' geographical location to English-speaking countries, including Australia, New Zealand, the United States, Canada, and the United Kingdom, as the gender biases in the job advertisements were based on English detection. Additionally, considering that the job advertisements varied in length and workers were required to rate three versions of job advertisements in one session, we set a minimum completion time of 300 seconds. We deemed answers provided less than this time inaccurate and rejected responses from workers who completed the survey in less than 300 seconds. Calculating the time it takes could help control quality and valid data, in which a longer completion time may indicate that respondents are engaging with the advertisements and considering their answers, rather than rushing through it or providing inaccurate information. Ultimately, only 151 responses had completion times greater than 300 seconds.

As we selected 40 job advertisements and their modified versions for the user study, having 151 completed responses implies that each job advertisement was rated by at least three workers. This reduces the randomness of the answers and increases their reference value. However, considering that different job advertisements may have varying degrees of gender bias scores, a more significant

Table 5. Prompts for generating new advertisements with ChatGPT.

Natural language generation techniques	Prompts
Steering	Rewrite the following advertisement using language biased toward women to attract more female professionals to apply for this job without changing the original meaning
Adhering	Rewrite the following advertisement using neutral language without changing the original meaning

number of responses for a particular job advertisement could have a greater impact on the overall average. Therefore, for job advertisements with more than three respondents, we randomly retained the responses from three workers. This ensured that each job advertisement was rated by three workers, and the weight of each job advertisement's influence on the results was equal. In the end, we accepted responses from 120 distinct workers, with 57.98% (69 individuals) being female and 42.02% (50 individuals) being male.

4) User study design

For the user study, we selected the top 40 job advertisements in the IT industry that displayed the highest gender bias according to the gender bias detector. We used Qualtrics to create the questionnaire and published it on MTurk. Instead of simply replacing gender-biased words as done in previous studies, we employed the natural language generation method based on ChatGPT to create two versions of each advertisement: one with steering (using language biased towards femininity) and one with adhering (using more neutral language). For each version of the job advertisement, we asked three questions:

- Q1: Assuming you meet all the requirements, please rate the overall attractiveness of this advertisement on a scale of 1 to 10, with 1 being extremely unappealing and 10 being highly appealing.
- Q2: Please rate the effectiveness of this advertisement in attracting female professionals on a scale of 1 to 10, with 1 being extremely unappealing and 10 being highly appealing.
- Q3: Please rate the level of gender bias in this advertisement on a scale of 1 to 10, with 1 being not biased at all and 10 being highly biased.

We collected the gender, IP address location, response duration, and response date of the MTurk workers who participated in the study.

8. Gender Debiasing Method Results

1) Data reliability

Ensuring the consistency and reliability of the designed questions in the user study is of paramount importance in user research. Understanding the consistency of the questions helps determine the reliability and validity of the measurement tool, thereby providing more valuable data and insights. When there is higher consistency among the questions, we can have greater confidence in using them to evaluate concepts such as advertisement attractiveness, gender ap-

peal, and gender bias, thereby enhancing support for decision-making and practices. We employed a questionnaire as the measurement tool, consisting of a total of nine questions, including three versions (Q1, Q2, Q3) addressing advertisement attractiveness, gender appeal, and gender bias. We selected Cronbach's alpha as the indicator for consistency testing, as it is the most commonly used method to assess the internal consistency of a measurement tool. Cronbach's alpha evaluates the level of consistency by quantifying the variances between the questions and the overall variance. Our aim was to determine the internal consistency and reliability of the set of questions by calculating the reliability score. The formula for Cronbach's alpha is provided below as Formula (4).

$$\alpha = \frac{k}{k-1}\left(\frac{1-\sum_{j=1}^{k}\sigma_j^2}{\sigma_x^2}\right) \tag{4}$$

Here, $k = 9$. Based on our analysis, the reliability score of our set of questions was 0.8209. This indicates that in a single measurement, our set of questions exhibited good internal consistency and reliability. It provides an effective tool for measuring concepts related to advertisement attractiveness and gender.

2) User study results

After conducting internal consistency checks, we sought to examine the differences among different versions of job advertisements (original, steering version, and adhering version) in terms of job attractiveness, attractiveness to female job seekers, and degree of gender bias. To achieve this, we employed Analysis of Variance (ANOVA) to compare the significant differences among the versions. ANOVA results typically include F-statistic and *p*-value. F-statistic represents the ratio of between-group differences to within-group differences, with a larger F-statistic indicating more significant differences. The *p*-value is a probability value used to measure whether the observed differences are due to random factors, with a *p*-value less than 0.05 considered significant.

Initially, considering the responses from all participants, as shown in **Table 6**, concerning the attractiveness to female job seekers, there was a significant difference among different versions (F-statistic = 3.0931, *p*-value = 0.0466). Similarly, there was a significant difference among versions in terms of gender bias (F-statistic = 5.7382, *p*-value = 0.0035). However, there was no significant difference among versions regarding job attractiveness (F-statistic = 0.9824, *p*-value = 0.3754). **Figure 12** illustrates the average scores of different versions across various questions. Although there was no significant difference in job attractiveness, the steering version (7.25) and adhering version (7.43) showed percentage increases of 1.39% and 3.91%, respectively, compared to the original version (7.15). As for attractiveness to female job seekers, the steering version (7.63) showed a 7.31% increase compared to the original version (7.11), whereas the adhering version (7.20) increased by 1.27%. The most significant difference was observed in gender bias, as the adhering version (6.16) reduced gender bias by 13.60% compared to the original advertisement (7.13), while the steering version (6.61) also reduced it by 7.29%.

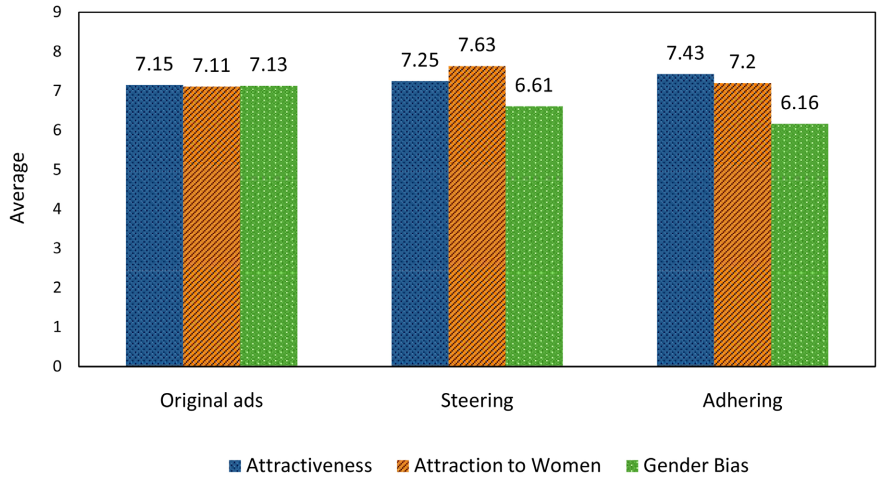

Figure 12. Gender bias and job attractiveness: all respondents' average.

Table 6. Comparison analysis of gender bias and attraction by all respondents.

Value	Attractiveness	Attraction to Women	Gender Bias
p_value	0.3754	0.0466	0.0035
F_statistic	0.9824	3.0931	5.7382

In conclusion, both the steering version and adhering version demonstrated significant improvements in attracting more female job seekers and reducing gender bias. Among them, the steering version (7.63) performed the best in attracting female job seekers, while the adhering version (6.61) had the lowest gender bias score. Although there was no significant difference in job attractiveness, the average job attractiveness scores of the steering version (7.25) and adhering version (7.43) were still higher than that of the original advertisement (7.15).

Next, we aimed to further analyze the differences in responses between male and female respondents. For female respondents, as shown in Table 7, there was a significant difference among versions regarding gender bias (F-statistic = 3.1075, p-value = 0.0468). From the perspective of averages, as shown in Figure 13, the gender bias differences followed the same trend as observed among all respondents, with the adhering version having the lowest score (6.26) and reducing gender bias by 12.69% compared to the original version (7.17), followed by the steering version, which reduced it by 5.85%. In terms of attractiveness to female job seekers, the steering version (7.86) had the highest score, showing a 6.08% increase compared to the original version (7.41), while the adhering version (7.38) slightly decreased by 0.41%. However, in terms of job attractiveness, both the adhering version and the steering version increased the attractiveness to female respondents. The adhering version (7.65) increased it by 5.95% compared to the original version (7.22), and the steering version (7.46) increased it by 3.32%. In summary, for female respondents, the adhering version (6.26) significantly reduced gender bias by 12.69%, consistent with the results from all

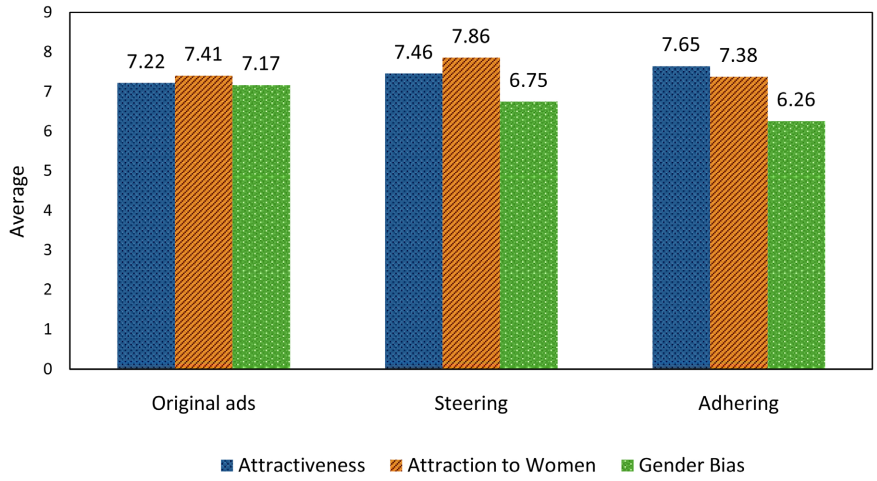

Figure 13. Gender bias and job attractiveness: female respondents' average.

Table 7. Comparison analysis of gender bias and attraction by female respondents.

Value	Attractiveness	Attraction to Women	Gender Bias
p_value	0.2016	0.1301	0.0468
F_statistic	1.6142	2.0597	3.1075

respondents. In terms of attractiveness to female job seekers, the steering version (7.86) performed the best, although the adhering version slightly decreased by 0.41%. However, both the adhering and steering versions exhibited improvements over the original advertisement in terms of job attractiveness to female job seekers.

For male respondents, as indicated in Table 8, the *p*-values for all three aspects did not reach the significance level (0.05), suggesting no significant differences among the versions. However, the *p*-value for gender bias (0.0735) was close to the significance level, indicating a certain trend of differences. Figure 14 presents the average values of different versions and questions. From the perspective of gender bias, the adhering version (6.02) reduced it by 17.28% compared to the original version (7.06), followed by the steering version (6.40) with a reduction of 10.32%. In terms of attractiveness to female practitioners, the steering version (7.32) showed a 9.25% increase compared to the original version (6.70), followed by the adhering version (6.96) with a 3.88% increase. However, in terms of job attractiveness, the adhering version (7.12) increased by 0.85% compared to the original version (7.06), while the steering version (6.96) decreased by 1.44%.

In conclusion, our analysis results indicate significant improvements in attracting female job seekers and reducing gender bias with the steering and adhering versions. The steering version performed the best in attracting female job seekers, while the adhering version exhibited the best performance in reducing gender bias. Although there was no significant difference in job attractiveness, the adhering version (7.43) and the steering version (7.25) both showed higher

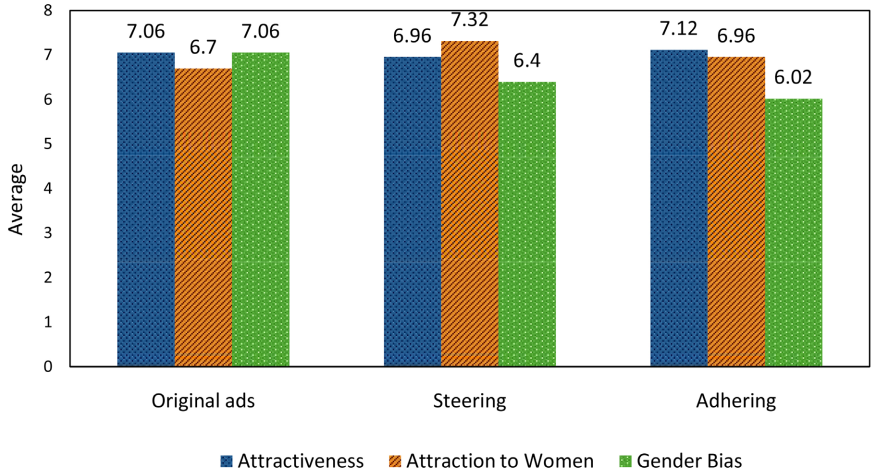

Figure 14. Gender bias and job attractiveness: male respondents' average.

Table 8. Comparison analysis of gender bias and attraction by male respondents.

Value	Attractiveness	Attraction to Women	Gender Bias
p_value	0.8906	0.2661	0.0735
F_statistic	0.1160	1.3359	2.6580

job attractiveness scores than the original version (7.15). Finally, we found differences in the effects of different versions on male and female respondents. Our research findings align with previous studies [24], suggesting that neutral language is beneficial in reducing gender bias and attracting both male and female job seekers, while language leaning towards femininity increases attractiveness to female practitioners. Therefore, the adhering and steering methods based on ChatGPT prove effective in this aspect.

3) Gender bias detector results

With the assistance of our gender bias detector, we evaluated the gender bias scores of three versions (original, steering, and adhering) of the 40 most gender-biased job advertisements in the IT industry. Through Figure 15, we can visually observe significant differences among the three versions.

To further analyze these differences, we conducted an ANOVA analysis, resulting in an F-statistic of 18.822 and an extremely small p-value (8.185e−08), indicating significant differences among the versions. Table 9 presents the average scores and distributions for each version. We can see that the steering version has the lowest gender bias score (−0.7352) and a more dispersed data distribution (variance of 0.0483). The adhering version has a reduced average score (−0.8887) compared to the original version (−0.9184), and the data distribution (0.0114) is also more dispersed than the original version (0.0004). This confirms the alignment of our gender bias detector with the user study results, further validating the effectiveness of the ChatGPT-based natural language generation methods guided by adhering and steering approaches in reducing gender bias in job advertisements.

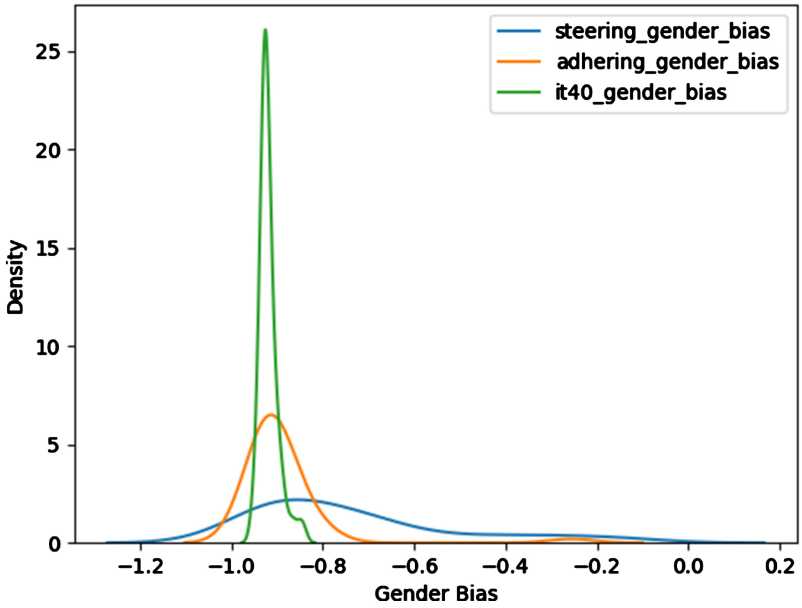

Figure 15. Distribution of gender bias scores across different versions.

Table 9. Mean and variance of gender bias scores across different versions.

Version	Mean	Variance
Steering	−0.7352	0.0483
Adhering	−0.8887	0.0114
Original	−0.9184	0.0004

9. Conclusion

This study aims to analyze gender and age biases in job advertisements in Australia and explore the application of large-scale language models in detecting and reducing these biases. By collecting and analyzing job advertisement data from the Australian job-seeking website, Seek, this research reveals gender and age biases in different industries. The results show that male bias is significantly greater than female bias in Information & Communication Technology, Manufacturing, Transport & Logistics, and Trades & Services. Most of these fields are in the technology sector or involve some degree of physical labor. Companies in these fields tend to prefer employees with certain masculine traits, while industries such as Advertising, Arts & Media, Education & Training, Healthcare & Medical, and Hospitality & Tourism prefer female employees. Regarding age bias, only a very small number of companies demonstrate an interest in older individuals in their job postings, with the majority of advertisements leaning toward younger candidates.

Furthermore, to build gender bias detectors and age bias detectors, this study employs large-scale language models such as RoBERTa, ALBERT, and GPT-2 for training and testing. The results indicate that RoBERTa performs the best in gender bias and age bias detection tasks.

Additionally, by using ChatGPT as a natural language generation tool, this study explores two methods for generating new job advertisement texts: adhering (using neutral language in advertisements) and steering (using more feminine language in advertisements). User research results suggest that these methods, particularly adhering, can reduce gender bias and increase the attractiveness to female candidates.

10. Future Work

Our research also has some limitations that we hope to address in future work. Firstly, our age bias dataset consisted of a total of 443 labeled sentences generated from collected age bias words. The limited size of the age bias dataset may restrict the model's ability to learn comprehensive information about age bias during training. Future work could involve improving the establishment of age bias datasets to enhance the accuracy and comprehensiveness of age bias detection in text. Secondly, in reducing gender bias in job advertisements, we utilized text generation methods based on ChatGPT. However, there are numerous other text generation models that perform well in this domain, such as T5 (Raffel et al., 2019) and BART (Lewis et al., 2019). Exploring the performance of different text generation models and how they can be effectively combined with our downstream tasks would help to better address the reduction of gender bias in the textual content. Lastly, this study primarily focused on exploring methods to reduce gender bias in textual content. It would be intriguing to investigate approaches for mitigating age bias in future research, expanding the exploration of bias reduction methods beyond gender. These future research directions would contribute to addressing the limitations of our study and further advance the understanding and mitigation of bias in text generation and detection.

Acknowledgements

This research was supported by Diversity Atlas and their provision of data has been instrumental in shaping the findings of this study. We would like to express our special thanks to the University of Melbourne for supporting the internship program that allowed the authors to conduct this research endeavor.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Arceo-Gomez, E. O., Campos-Vazquez, R. M., Badillo, R. Y., & Lopez-Araiza, S. (2022). Gender Stereotypes in Job Advertisements: What Do They Imply for the Gender Salary Gap? *Journal of Labor Research, 43,* 65-102. https://doi.org/10.1007/s12122-022-09331-4

Askehave, I., & Zethsen, K. K. (2014). Gendered Constructions of Leadership in Danish Job Advertisements. *Gender, Work & Organization, 21,* 531-545.

https://doi.org/10.1111/gwao.12053

Australian Government (2014a). *Sex Discrimination Act 1984.* https://www.legislation.gov.au/Details/C2014C00002

Australian Government (2014b). *Age Discrimination Act 2004.* https://www.legislation.gov.au/Details/C2020C00283

Australian Human Rights Commission (2013). *Fact or Fiction? Stereotypes of Older Australians.* https://humanrights.gov.au/sites/default/files/document/publication/Fact%20or%20Fiction_2013_WebVersion_FINAL_0.pdf

Bem, S. L. & Bem, D. J. (1973). Does Sex-Biased Job Advertising 'Aid and Abet' Sex Discrimination? *Journal of Applied Social Psychology, 3,* 6-18. https://doi.org/10.1111/j.1559-1816.1973.tb01290.x

Böhm, S., Linnyk, O., Kohl, J., Weber, T., Teetz, I., Bandurka, K., & Kersting, M. (2020). Analysing Gender Bias in IT Job Postings. In *Proceedings of the 2020 on Computers and People Research Conference* (pp. 72-80). Association for Computing Machinery. https://doi.org/10.1145/3378539.3393862

Burn, I., Button, P., Corella, L. F., & Neumark, D. (2019). *Older Workers Need Not Apply? Ageist Language in Job Ads and Age Discrimination in Hiring.* NBER Working Paper No. 26552. https://doi.org/10.3386/w26552

Burn, I., Firoozi, D., Ladd, D., & Neumark, D. (2021). *Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment.* NBER Working Paper No. 28328. https://doi.org/10.3386/w28328

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science, 356,* 183-186. https://doi.org/10.1126/science.aal4230

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2020). Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-11). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376488

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 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)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Devlin, M. (2006). *Inequality and the Stereotyping of Young People.* https://mural.maynoothuniversity.ie/1185/1/Inequality.pdf

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery. https://doi.org/10.1145/3173574.3173986

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 314-331). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.23

Doughman, J., & Khreich, W. (2022). *Gender Bias in Text: Labeled Datasets and Lexicons.* arXiv: 2201.08675.

Frissen, R., Adebayo, K. J., & Nanda, R. (2022). A Machine Learning Approach to Recog-

nize Bias and Discrimination in Job Advertisements. *AI & SOCIETY, 38,* 1025-1038. https://doi.org/10.1007/s00146-022-01574-0

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences of the United States of America, 115,* E3635-E3644. https://doi.org/10.1073/pnas.1720347115

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology, 101,* 109-128. https://doi.org/10.1037/a0022530

Gendron, T. L., Welleford, E. A., Inker, J., & White, J. T. (2016). The Language of Ageism: Why We Need to Use Words Carefully. *The Gerontologist, 56,* 997-1006. https://doi.org/10.1093/geront/gnv066

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences of the United States of America, 120,* e2305016120. https://doi.org/10.1073/pnas.2305016120

Guerin, B. (1994). Gender Bias in the Abstractness of Verbs and Adjectives. *The Journal of Social Psychology, 134,* 421-428. https://doi.org/10.1080/00224545.1994.9712192

Hu, S., Al-Ani, J. A., Hughes, K. D., Denier, N., Konnikov, A., Ding, L., Xie, J., Hu, Y., Tarafdar, M., Jiang, B., Kong, L., & Dai, H. (2022). Balancing Gender Bias in Job Advertisements with Text-Level Bias Mitigation. *Frontiers in Big Data, 5,* Article 805713. https://doi.org/10.3389/fdata.2022.805713

huggingface.co. (n.d. a). *Roberta-Large · Hugging Face.* https://huggingface.co/roberta-large

huggingface.co. (n.d. b). *gpt2 · Hugging Face.* https://huggingface.co/gpt2

huggingface.co. (n.d. c). *Albert-base-v2 · Hugging Face.* https://huggingface.co/albert-base-v2

Kanij, T., Grundy, J., McIntosh, J., Sarma, A., & Aniruddha, G. (2022). A New Approach towards Ensuring Gender Inclusive SE Job Advertisements. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society* (pp. 1-11). IEEE. https://doi.org/10.1109/ICSE-SEIS55304.2022.9793874

Kuzman, T., Mozetic, I., & Ljubesic, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.* arXiv: 2303.03953.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations.* arXiv: 1909.11942.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension.* arXiv: 1910.13461.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M. S., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv: 1907.11692.

Moon, R. (2014). From Gorgeous to Grumpy. *Gender and Language, 8,* 5-41. https://doi.org/10.1558/genl.v8i1.5

Ningrum, P. K., Pansombut, T., & Ueranantasun, A. (2020). Text Mining of Online Job Advertisements to Identify Direct Discrimination during Job Hunting Process: A Case Study in Indonesia. *PLOS ONE, 15,* e0233746.

https://doi.org/10.1371/journal.pone.0233746

Open AI. (2023). *ChatGPT.* https://chat.openai.com/chat

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pillar, A., Poelmans, K., & Larson, M. (2022). Regex in a Time of Deep Learning: The Role of an Old Technology in Age Discrimination Detection in Job Advertisements. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 13-18). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1.2

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI Blog, 1,* 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* arXiv: 1910.10683.

Raichur, A., Lee, N., & Moieni, R. (2023). A Natural Language Processing Approach to Promote Gender Equality: Analysing the Progress of Gender-Inclusive Language on the Victorian Government Website. *Open Journal of Social Sciences, 11,* 513-529. https://doi.org/10.4236/jss.2023.119033

Reissmann, M., Geithner, L., Storms, A., & Woopen, C. (2021). Stereotypes about Very Old People and Perceived Societal Appreciation in Very Old Age. *Zeitschrift fur Gerontologie und Geriatrie, 54,* 93-100. https://doi.org/10.1007/s00391-021-01971-y

Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology, 7,* Article 25. https://doi.org/10.3389/fpsyg.2016.00025

Strengers, Y., Qu, L., Xu, Q., & Knibbe, J. (2020). Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376315

Tang, S., Zhang, X., Cryan, J., Metzger, M. J., Zheng, H., & Zhao, B. Y. (2017). Gender Bias in the Job Market. *Proceedings of the ACM on Human-Computer Interaction, 1,* 1-19. https://doi.org/10.1145/3134734

Trix, F., & Psenka, C. (2003). Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society, 14,* 191-220. https://doi.org/10.1177/0957926503014002277

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need.* arXiv: 1706.03762.