Human Associations of Gender in Job Vacancies Compared to Gender Bias Reflected in AI

Vrentzou, Marina

Leiden University

Faculty of Humanities
Leiden University Center of Linguistics
MA Linguistics: Linguistics

Human Associations of Gender in Job Vacancies Compared to Gender Bias Reflected in AI

Marina Vrentzou
December,31 2022

Universiteit Leiden

Supervisors: Prof. Dr. S.A. Raaijmakers
Dr. J. Prokic
Second Reader: Dr. M. Westera

Acknowledgements

I would like to give special thanks to my supervisors, Prof. Dr. S.A. Raaijmakers and Dr. J. Prokic for giving me the opportunity to work on a topic on a quite new to me field, this of Natural Language Processing. The valuable contribution of Prof. Dr. S.A. Raaijmakers with the code as well as the immense knowledge and the insightful feedback of both of them have been really helpful throughout the entire process.

I would also like to thank the Data Science department of TNO and most importantly my mentor, Steven Vethman. His constant support, encouragement and guidance are deeply appreciated.

Abstract

Gender bias is an identified issue in recruitment which is often transmitted through linguistic means in job postings. Language models utilised in the hiring process appear to also reflect this type of bias resulting in unfair evaluations and therefore reinforcing gender inequality and discrimination, particularly against marginalised groups. Experimental research has demonstrated the effects gendered wording has on potential candidates, while extensive work in the field of Natural Language Processing (NLP) has investigated the predominance of gender bias in language models. By integrating ideas from the fields of Sociolinguistics and NLP, this thesis explores the level of agreement between human perceptions of gender bias in job postings and revealed gender bias in NLP models, specifically, BERT. Furthermore, it assesses whether context words trigger individuals' perceptions, similar to BERT. Following a mixed-methods approach, both humans and BERT are submitted to an experimental gender bias detection task, addressing shared job postings data. The results show a significant difference between BERT and men in the attribution of genders, while little overlap is reported in the indicative words of the female gender between BERT and women. Potential other reasons behind people's decisions are investigated through a thematic analysis.

Contents

List of Tables

List of Figures

1 Introduction

Gender equality and non-discrimination are fundamental principles of the United Nations Charter [107]. Gender employment discrimination is illegal under the EU Charter of Fundamental Rights (Article 21) in Europe [49] as well as under the Title VII of the Civil Rights Act of 1964 in the United States [1]. However, millions of women and LGBTQIA+ people around the world continue to experience discrimination in various areas of everyday life; one of them being the labour market.

In particular, according to the International Labour Organisation (ILO), the current global labour force participation rate for women is almost 47% compared to 72% for men [65], a percentage which represents the difficulty women face finding a job. In addition, according to a study by Williams Institute [19], around 45% of LGBTQIA+ individuals have experienced discrimination in the labour market, including not being hired or being harassed because of their gender identity.

Gender bias is defined as the systematic discrimination of people of a certain gender in favour of others of a different one [52]. This type of bias often finds its way into job advertisements, as a reflection of human bias. This leads to unfair associations in language representations based on gender that are derived from culture, norms and tradition. Language is a mirror reflecting gender bias not only in an explicit but also in a subtle, implicit way by encoding, for example, gender biases at the level of syntax [51]. Language, therefore, perpetuates gender bias which appears to also be present in Natural Language Processing (NLP) systems. These systems are often utilised in the hiring process to assist recruiters. Large-scale language models, such as BERT [39], underlie search engines used in recruitment, as is the case in InterviewBERT [100]. However, there are incidents which showed that these machines are not that fair after all [33].

A variety of studies has investigated binary gender bias in recruitment. Important empirical research in the area of gender bias was conducted by Gaucher et al.(2011) [54] that showed that job postings for male-dominated fields employ greater masculine wording than those in female-dominated areas. Similarly, extensive research has investigated the existence of this kind of bias in language models. Works, such as the ones by Bolukbasi et al. (2016) [15] and Kurita et al.(2019) [77], explored fairness by detecting and measuring gender bias in standard and contextual word embeddings, like BERT.

1.1 Motivation

Gender bias and stereotypes affect the perception of individuals and the way they are treated. Instead of judging people on their own merits, people are put into categories and they are judged based on pre-existing beliefs. This phenomenon often results in unfair evaluations and discrimination.

When applied to organisational context, candidates may experience gender bias and discrimination at all stages of employment. During decision-making processes, the bias of individuals

and gender stereotypes, play an important role in the evaluation of the candidates [59]. Exemplary for this is the fact that women are criticised when they display stereotypical masculine traits that do not align with the female role expectations and they are considered unsuitable for stereotypical male positions [59, 4]. Even when in a profession, the goal of gender balance has been achieved, it has been observed that individuals who claim that gender bias is not an issue anymore and think they are objective, are prone to evaluate others in a biased way [8, 106].

Apart from certain stages of recruitment, such as interviews, that are infamous of being susceptible to bias, other steps of the hiring process, such as job postings, have also been claimed to be gender biased [61, 3, 54]. Vacancy advertisements can be effectively thought as the first step in the communication process between a company and the labour force. According to the Signaling Theory, a job posting is a signal that conveys the organisational attributes to the job seekers [30]. The way it is written reflects what a company looks for and its values. At the same time, it affects whether a potential candidate will assess they are a great fit for the vacancy and they will apply. Hence, it is relevant to examine further whether individuals evaluate a series of job postings as biased.

Despite the fact that several studies have explored gender bias in the area of job postings, none of them has systematically compared the revealed gender bias of an algorithm to the bias of humans. In this thesis, emphasis is placed on exploring whether an NLP model, and in particular BERT, exhibits potential gender bias that aligns with the human bias, based on a shared dataset of job advertisements. In doing so, it can be assessed whether these algorithms are less biased than people and therefore should be utilised in recruitment support or if there is still more work that needs to be done. This study, thus, may nuance earlier analyses of BERT gender bias.

This research falls into the field of Natural Language Processing combining ideas from the field of Sociolinguistics, since in that way a deeper understanding of the issue of bias can be acquired [13]. By systematically comparing human and algorithmic gender bias to job advertisements, this thesis aims to bring together an empirical and an algorithmic perspective on gender bias. The motivation behind the study is to paint a fuller picture on the topic of gender bias in general and in recruitment in specific and to create an inclusive study that aligns with today's society. After all, exposing potential gendered wording may lead to less gender-biased job postings.

1.2 Research Questions

With the aforementioned central goals, the questions under discussion with their corresponding hypotheses are the following:

Research Question 1 : Does a significant difference exist in the attribution of genders to job vacancies between subgroups of the population as well as between them and BERT?
Hypothesis 1: *The gender of a person influences their perception of the preferred gender in a*

job posting.

Research Question 2 : Are people triggered by context words, similar to BERT, in the attribution of gender?
Hypothesis 2a: *Similar to BERT, humans are triggered in the attribution of gender by context words, reflecting bias in language models.*
Hypothesis 2b: *Gender and context words together are not sufficient for explaining assigned genders.*

To answer the research questions and to test the hypotheses, the study consists of different components. At first, an analysis of a shared dataset is conducted in order to detect gender denoting words and to prepare the dataset for the next steps of the research. Then, a user study explores people's bias in a systematic way and whether there are certain words or other reasons that trigger this bias. At the same time, BERT is utilised in an unmasking procedure so as to examine whether it exhibits gender bias. The results from these processes are compared by examining on a first level whether there is an agreement between people and BERT regarding the gender they assign to postings. On a second level, a comparison is made between the words that influenced participants' decisions and the words that are the most important for predicting a certain gendered wording in BERT. At the end, additional reasons that may have influenced people's perceptions are analysed.

1.3 Layout of the thesis

To set out the research, the thesis starts with some preliminaries (section 2) for easier understanding of the terminology that follows and the way gender is viewed in this study. Then, a theoretical framework is presented which is divided into three main sections. In section 3, gender stereotypes and bias are described. In section 4, emphasis is placed upon the role of language and the effects gendered wording has on potential candidates for a job position. In section 5, the focus is on gender bias in the field of NLP. This type of bias within NLP techniques and its measures as well as various detection tasks are described. Additionally, works that combine an algorithmic and a human perspective and gender diversity within the NLP world are discussed. In section 6, which is the methodology section, the dataset selection and its preprocessing are described. The recruitment of participants, the instrument of the user study as well as the unmasking procedure followed by BERT are presented. Then, the way the data were analysed is explained. In section 7, the results from the unmasking procedure and the empirical study are presented. In section 8, the findings are discussed and in section 9, a conclusion of this study is given. In section 10, the limitations of this research and avenues for further research are presented and in section 11, the ethical considerations are discussed.

2 Preliminaries

2.1 Terminology

In academia, various terms have been used over the years to describe gender identities that are not confined to the two fixed genders constructed by Western society. The two main umbrella terms that have been embraced the most by scholars are *genderqueer* and *non-binary*. However, there are some limitations with these.

The term *genderqueer* includes the word *queer* which in the past was considered a derogatory term. Although it has been reclaimed by a big part of the LGBTQIA+ community, there are still individuals who are triggered by it due to past experiences with that word being used against them [85]. Regarding the *non-binary* term, it indicates that there is the binary based on which the non-binary gender identity exists. The prefix "non" creates an *othering* environment, since the non-binary individuals are defined as not fitting in the binary [16].

In recent years, the term *gender diverse* has emerged as a more suitable term, as the word *diverse* may contribute to the conceptualisation of gender as multifaceted [101, 62]. Henceforth, this term is used throughout the paper to describe a range of different gender identities that do not conform to the traditional binary. It should be emphasized, though, that gender identity is based on the personal experiences of each individual and, therefore, it is a sensitive topic. Thus, by no means it should be assumed that all individuals falling into this umbrella term identify in the same way.

Additionally, the use of the terms *female* and *male* as nouns to describe people was avoided, since in this way they are often used to dehumanise and disrespect individuals [38]. Instead, throughout this thesis the nouns "woman" and "man" are used to refer to people with a cisgender identity.

2.2 Beyond the binary

Sex and gender are two terms that are used interchangeably even though they are not equivalent. The first time a distinction was made between these terms in the research community was back in 1945, when Bentley (1945) [10] suggested that gender refers to the social roles deriving from sex. Since then, authors [111, 28] pointed out the difference between these two notions. Sex is assigned at birth and refers to biological characteristics (e.g. chromosomes and hormones), whereas gender is a more challenging concept. At first, the latter was used as a reference to socially constructed characteristics of sex, since there was, and to a certain extent, there is still a held belief that gender stems from sex. This belief resulted in the view of gender as binary with individuals being divided in two discrete groups, male and female [88].

Yet, in many societies, the distinction between genders has always been more fluid. For example, *hijra* is considered an official third gender identity in countries such as India, Bangladesh and Pakistan [66, 89]. People who identify as hijras are, among others, individuals who are born intersex or were assigned the gender of male or female at birth but self-identify in a different

way.

The last few years in the Western world, the binary view of gender has been understood to be problematic and thus the definition of gender has been challenged a lot [97]. Gender, among others, expresses gender identity, namely, self-categorisation [87, 115] and for that reason it should be best described as a spectrum [64, 46]. On that note, scholars, such as Lips (2020) [79], suggested that gender should not be seen as a dichotomy of two opposites with clear boundaries. Similarly, others such as Butler (2004) [23], tried to deconstruct the concept of cisnormativity, namely the assumption that all individuals' gender corresponds to their sex. Besides, gender identity is not static and may change over time, as it has been observed through experimental studies [109].

To keep on claiming that gender is binary is pejorative for all the people who do not identify as such and encourages ongoing discrimination towards them in various aspects of their lives. Therefore, in this research gender is defined as a spectrum.

3 Gender stereotypes & bias

Stereotypes are held beliefs about people based on their membership in a certain social group. Gender stereotypes are generalisations about attributes that are assigned to individuals based on their gender and they are distinguished in two categories: prescriptive and descriptive [74]. Prescriptive stereotypes pertain to views regarding roles and behaviours with which each gender is expected to comply, whereas descriptive stereotypes are beliefs about the qualities that each gender possesses [22]. According to Social Role Theory, gender stereotypes are a result of the distribution of genders into social roles in different fields of life, such as the labour [43, 75]. A main characteristic of this type of stereotypes is their binary categorisation, based on which, male and female are viewed as polar opposite genders [48, 79]. For example, when a woman is not connected with her emotions, she is considered straight away masculine and the other way round. This leads to the reinforcement of the perception of genders as two total opposites and creates a bigger gender gap. This essentialist view has also contributed to the hierarchical privilege of one gender over the other, with men being viewed as having more power compared to women [73].

Deriving from both descriptive and prescriptive stereotypes, two categories have been created with regard to the stereotypical attributes of the masculine and feminine behaviour. In particular, men are associated with agentic traits, whereas women with communal ones. Qualities such as assertiveness, decisiveness, competence as well as goal-achievement are considered agentic. On the other hand, traits which are associated with the preservation of relationships and others such as trustworthiness, morality and benevolence are referred to as communal [2].

Many suggest that these stereotypical beliefs about the attributes of each gender are impervious to change [81, 57], an argument which is supported by phenomena such as illusory correlation and confirmation bias. However, experimental and meta-analysis studies [44, 41] as well as research utilising Machine Learning techniques [12] showed that the stereotypes about women have to a certain extent changed.

What makes gender stereotypes so persistent, though, is the fact they are built in individuals' minds from the early childhood in the same way as gender bias [103]. Gender bias is a mechanism that leads to favouritism for a gender over another. It can be a subtle mechanism and therefore people are not always aware of its existence. Gender bias in the form of benevolent sexist beliefs and compensatory and complementary stereotypical thoughts justifies and rationalises gender inequality and stratification [90] which contributes to the maintenance of stereotypical gender roles.

4 The role of language

4.1 Gendered wording

Modern English is considered a natural gender language, since it lacks a grammatical gender system [32]. This means that nouns are not classified based on gender, as is the case in other languages (e.g. French, German). There is, though, a referential and a lexical gender [27]. An example of a referential gender is the third-person pronouns (*she*, *he*, *they* and neopronouns such as *ze*, *xe* for individuals with a gender diverse identity), namely, linguistic expressions that refer to the extra-linguistic reality.

Although it has been observed that gender discrimination is more common in languages with grammatical gender [35], studies have shown that biased wording is still prevalent in job postings written in English [91, 54]. Gender inequalities and discrimination are transmitted through linguistic cues which lead to the reinforcement of gender imbalance in the workplace. Based on the ascribed attributes of the two traditional genders that were mentioned in section 3, the linguistic style of speech of people with a female and people with a male gender is characterised as communal and agentic respectively. Words such as *committed, creative and understanding* are perceived as communal and therefore feminine, whereas words such as *driven, influential and independent* as agentic and therefore masculine.

A few studies have focused on the detection of these words in vacancy postings and on whether the use of them has an impact on the appeal the postings have to each gender. In particular, methods such as surveys [54, 99] were utilised to assess people's perception towards the advertisements. In some studies, postings from various fields were included, while in others the focus was on a certain field [91, 113].

4.1.1 The effects of the gendered wording in job postings

One of the first studies to document the existence and the effect of gendered wording in job postings was published back in 1973 [9]. At that time, gender bias was explicit, since words such as pronouns (e.g. "he") and gendered nouns (e.g."man") were included. Bem & Bem[9] found that job ads which contain a great amount of wordings associated with a specific gender discourage the other gender from applying. In addition to that, it was observed that, when the masculine wording was replaced by feminine, more women were interested in these postings even if they were describing a male dominated job.

Since then, explicit gender bias has disappeared to a large extent. Nowadays, it is implicit and hence harder to track. Scholars have mainly focused on subtle linguistic cues that ascribe gender-based characteristics and as a result reinforce gender stereotypes. These linguistic cues are based on the two groups mentioned earlier; the communal and the agentic one. The existence of agentic words in job postings has been proven to be a factor that deters women from applying for a job because of negative evaluation of career advancement prospects and a lacking sense of belonging [61, 54, 91]. This aligns with the Attraction-Similarity Theory [24] adapted in

an organisational context, according to which, a job seeker finds appealing a job posting that portrays characteristics similar to their own. It is interesting, though, that this is not always the case for men. In many cases, the type of wording didn't affect men's perception of the job position [17, 61]. A reason behind this phenomenon possibly lies in the fact that individuals from minority groups are aware that the others perceive them as having a low status [17, 47]. Therefore, women being aware of the gender injustice in the workplace have a different reaction to the wording being used. On a similar note, Tang et al.(2017) [99] suggested that individuals' biased perception and not the biased wording is the main source of the problem. Cognitive bias patently affects the way people process information and perceive others. However, the way language is used is of equal importance. Language shapes the way individuals think [105] and subtle linguistic cues, in the present study words associated with a specific gender, can have a causal effect on people's behaviours [95, 50].

Changes have been observed in the stereotypes that are ascribed to the two fixed genders [41]. Furthermore, the focus has been only on the two fixed genders of the traditionally binary and, as Hentschel et al.(2021) [61] state, future research should put emphasis on the other genders, too.

5 Using technology to assess bias

5.1 Gender bias within NLP techniques

In Natural Language Processing (NLP), the issue of gender bias has also been identified. Gender bias is not confined to a single part of an NLP application. On the contrary, it can be found in almost all parts of it, ranging from the training datasets and the pretrained models to the algorithms themselves.

The last years, text classification has been done using feature extraction and language modeling techniques. Through these techniques, word embeddings, namely word representations in the form of numeric vectors, can be obtained. Models, such as GloVE [92] and Word2Vec [86], create static embeddings that are binded directly to dictionary words. In such way, these models do not take into account the context in which each word is used. Early studies [25, 15] examined fairness in non-contextualised language models or in other words explored whether such models have discriminatory outcomes based on features such as gender. These models reflect the stereotypes and the bias that are present on the training data [53] resulting in stereotypical links such as the famous *Man is to Computer Programmer as Woman is to Homemaker* [15].

Recently, progress in word embeddings has been attained with context dependent embeddings. For reference, contextualised word embeddings assign each word a different vector representation, depending on the linguistic context the word occurs [80]. One of the best known contextual language models is the Bidirectional Encoder Representation from Transformers (BERT) [39]. BERT is based on the encoder attention layer of the Transformer architecture and captures the semantics with the use of the Masked Language Modeling objective and the Next Sentence Prediction. In that way, it is bi-directional, namely it learns information from both left to right and right to left at once. One of its main characteristics is that it takes polysemy into consideration, as it is able to generate different embeddings for different meanings of a certain word. This is a major issue in numerous languages and BERT seems to give better results in such a task compared to previous models [112].

In a comparative analysis, Basta et al.(2019) [7] found that contextual word embeddings are less gender biased compared to the standard ones, even in the case that the latter are debiased. However, this does not mean that in the former indications of bias have not been detected. Investigations conducted with BERT showed that it encodes gender biases just like standard word embeddings and in particular tasks exhibits a male bias [6, 77]. Specifically, Bartl et al.(2020) [6], following the approach of Kurita et al. (2019) [77], showed that BERT displays gender bias which corresponds to the real-world workforce statistics about professions with the highest ratio of male and female individuals, respectively. In certain balanced professions, though, the model presented biases that stem from the language use in its training data.

Due to the fact it relies on co-occurrences, it can pick up on the stereotypes in the pretraining data, which is referred to as intrinsic bias in the language model [36, 55]. In contrast, the bias displayed in downstream tasks and in the fine-tuning process for a certain application is called

extrinsic bias.

5.1.1 Measures of gender bias

In an attempt to tackle these issues and move towards fairer frameworks, scholars have focused on various tasks including detecting, mitigating and measuring bias. Various measures have been proposed to estimate gender bias in language models and in downstream tasks. Methods such as the Word Embedding Association Test (WEAT) [25], which was developed based on the Implicit Association Test (IAT) [56], and a direct bias metric [15] were among the first that measured bias in standard word embeddings. Specifically, the WEAT test is a semantic association statistical test that assesses the level to which a model relates sets of target words (e.g. doctor, nurse) with sets of attributes (e.g. male, female). This association is examined by measuring the cosine similarity between the 2 vectors. For reference, cosine similarity measures the angle between two vectors by using the dot product and normalising them. The direct bias metric [15] works by determining a pair of gendered words, such as *he* and *she*, based on which a vector space is defined. Then, the relationship of a word embedding to a gender axis is explored by measuring the cosine similarity. The closer a word is to one of the gendered words the more related it is to this gendered word.

In contextualised word embeddings, various methods have been suggested as an additional way of estimating bias. A standalone method that has been proposed is the Sentence Encoder Association Test (SEAT) [84], which is an alternative to WEAT. Instead of sets of words, the SEAT test examines the degree of association between sets of targets and attributes with sentence templates. Other studies utilised templates [77, 6] and randomly sampled sentences from a corpus [119, 7]. These studies used methods that are not cosine-based, since it has been observed that cosine-based approaches produce inconsistent results on sentence embeddings [84]. For example, Kurita et al(2019)[77] proposed sentence templates to measure bias based on the Masked Language Modelling objective of BERT. This bias measure technique also takes into account the prior bias of the model, that is to say, the probability of the model unmasking, for example, the pronoun *he* based only on the structure of the sentence. In sentiment classification tasks, Jentzsch & Turan (2022) [67] presented a new bias measure in BERT models which is based on the valuation capacity of sentiment classifiers.

A few works have explored the relation between intrinsic and extrinsic bias based on the metrics that have been used over the years [36, 55]. What has been observed, though, is that the various metrics in language models are incongruent with each other and also dependent on each task [55].

5.2 Detection of gender bias by word analysis

Extensive algorithmic research has focused on gender bias by exploring the association of gender-coded words with occupations. Bolukbasi et al.(2016) [15] defined a gendered vector space based on gendered word pairs, such as *he* and *she*, and assessed the correlation between

occupational words and the gendered vector space. Building on this work, Matthews et al. (2021) [82] examined gender bias in nine different languages. Both studies showed that profession words display a gender bias.

Katsarou et al. (2022) [71] deployed the T5 model and showed that professions with a higher status are associated more with the male than the female gender. On that matter, Zhao et al.(2018) [118] discussed that resume filtering systems are biased because they tend to bind together certain occupations with a specific gender. In the context of sentiment analysis, Bhaskaran & Bhallamudi (2019) [11] explored whether three different models, with BERT being one of them, reflect occupational gender stereotypes. BERT displayed a lower predicted positive probability class for sentences with female nouns compared to the sentences with males nouns.

Drawing attention to reference letters from the finance market and the way they are written, Eberhardt et al.(2022) [45] focused on the terms and traits that are related to each gender in order to assess the unconscious bias, by applying the LASSO technique [102] and then by building dictionary of words. They found that women tend to be described with fewer ability traits and more "grindstone" terms [104], namely terms used to describe one who works hard and persistently. One of these terms is *hardworking* which contradicts, though, the leadership style idea of a male [93], according to which, the *hardworking* characteristic is associated with a male and not a female leader.

5.2.1 Detection of gender bias in job postings

In the topic of job postings, in particular, research underscores the problematic nature of the way they are written with regard to the existence of gender denoting words. This is achieved not only through empirical studies, which has already been mentioned in section 4.1.1, but also through the utilisation of algorithms.

Although, a longitudinal analysis [99] on online job postings showed that there is, to a certain extent, a decrease of gender bias, discrimination in the form of this type of bias is still a current issue [108]. A study [68], deploying word vectors and a smaller pretrained version of BERT, focused on the detection of additional gender denoting words and the identification of gender bias at the grammatical/ sentence level. By utilising BERT-small, that is a BERT model with 4 layers of encoder blocks with a hidden size of 512 and 8 attention heads, it was observed that gender bias does not just arise due to the existence of stereotypical words. Context plays an important role on whether such a word would be indicative of gender bias. In addition to that, a preliminary study on German job postings [14], suggested that it is important not only the number of feminine words to be increased but also the amount of masculines words to be diminished.

A few years back, an academic gender decoder [83] evaluated Facebook's job postings, since the company had announced that it would increase the number of employees from minority groups. The results showed that there was an actual change in the word choice. The postings included way more feminine words than masculine which validated the company's announcement.

In recent years, automated tools (e.g. Textio [1], Gender Decoder [2]) have been designed which enable the detection and mitigation of gendered wording in job advertisements. These web services are built based on the list of masculine and feminine wording provided by Gaucher et al. (2011) [54]. Although, this is an important step towards a more inclusive and gender fair way of writing job descriptions, the list of words they are based on may be incomplete and not up to date.

5.3 Comparison of the algorithmic and people's perspective on bias

So far, studies that utilised either an empirical or an algorithmic approach were presented. However, there are some works that compared the revealed gender bias of an algorithm to the human bias on the same task in order to get a better understanding and validate their results.

A study conducted by Caliskan et al. (2017) [25] revealed that word embeddings encode human-like biases which replicate the results of the IAT, namely the main trends of the unconscious bias of humans. To evaluate certain automatic metrics of social biases, Dhamala et al. (2021) [40] conducted a series of surveys. These surveys showed there is a strong correlation between annotators' labels and the automatic metrics they presented. Similarly, in a face recognition task, Dooley et al. [42] found that even though algorithms showed a higher accuracy on both verification and identification tasks, they exhibited similar biases with the participants from the surveys. Finally, in a comparative analysis, an effort was made to explore further the relationship between human and algorithmic bias by showing similarities and suggesting future research to focus more on such a comparison [69].

5.4 Gender diversity in the NLP world

Even though a priority area of work in language technologies has been towards promoting non-biased frameworks, the studies that treat gender as a variable, by considering it a spectrum of various gender identities, are limited [38].

This is a serious issue, since current language models can cause harms on individuals and marginalised groups. These harms, which are a result of the algorithmic bias, are distinguished in two types; allocational and representational harms [5]. Harms of allocation arise when a system distributes opportunities, such as a job vacancy, unfairly to different groups of people. Harms of representation occur when a system doesn't recognise the existence of certain groups at all or presents them in a subordinate way.

Specifically, misgendering and erasure are two examples of harms towards people with a gender diverse identity. Sample size disparities, tainted data and limited features are a few of the factors that contribute to the creation of these harms [37]. Annotators may not pay attention to gender diverse identities and gender diverse data may be deemed outliers by the system. It is, therefore, of uttermost importance gender diverse individuals to be included in

[1] `https://textio.com/`

[2] `https://gender-decoder.katmatfield.com/`

the dataset creation and analysis, as it has already been stated by Kuhlman et al.(2020) [76]. Only recently, a study was presented about the creation of a taxonomy of gendered language inclusive to people with various gender identities [58]. This taxonomy was then used for the annotation of a dataset.

Representational harms of erasure have been observed in both GloVE and BERT with the latter, though, displaying better results [37]. In GloVe representations, the neopronouns *ze* and *xe*, which a lot of people with a gender diverse identity use, do not have semantically meaningful associations and their close neighbours are acronyms and Polish words. In addition, words such as *genderqueer* are associated with negative adjectives, such as *dishonest* and *arrogant*. This stands in stark contrast with the nearest neighbours of *male* and *female* which are, among others, *good* and *loving*. In BERT, the neopronouns *xe* and *ze* are not part of the word-piece vocabulary. Furthermore, although this model is able to distinguish *they* used as a singular form from *they* as a plural, its accuracy is not high (67,7 %).

Other works available that do not neglect the existence of gender diverse individuals are focused on tasks such as Coreference Resolution and Automatic Gender Recognition. Studies on these tasks showed that the way systems operate leads to harms which amplify the normative view that individuals with various gender identities do not matter in the greater scheme of things [27, 72].

Many scholars who explored gender bias acknowledge the existence of more than two genders but they stick to a binary classification for various reasons, the main one being convenience [6, 36, 26, 98, 34]. They rather suggest that future work should focus on a non-binary classification of gender. This is exactly one of the issues that Hovy & Spruit (2016) [63] address; the exclusion of certain groups of the population in NLP which leads to demographic bias.

6 Methodology

6.1 Research Design

To investigate the two research questions (RQ1, RQ2) and to test the corresponding hypotheses, a convergent mixed-methods approach [31] was followed. Based on previous literature focusing on job postings, which showed that gendered wording influences people's perception [54, 14], it was important to create a dataset consisting of a high number of gender denoting words. Therefore, the first step was to create a dataset as to examine whether the detected gender wording still influences individuals' views.

By conducting a series of surveys, it was possible to examine people's perception on certain job advertisements. Surveys were conducted, since it is a tool that has been utilised in the past for similar tasks [91, 54, 113, 99]. In particular, online surveys offer the opportunity to researchers to reach a big number of people, even individuals who would be difficult to be approached through other means, in a timely manner [116].

In order to explore the algorithmic perspective on bias, it was important to utilise a model that is able to understand the various senses of a word depending on the linguistic environment. For example, a non-contextualised model would consider the same the words "lead" (i.e. the chemical element) and "lead" (i.e. to be in charge of). Various models have been developed that encompass this feature including XLNet [117] and T5 [96]. Albeit their systematic and accurate results in many tasks, their great number of parameters makes them computationally expensive. Therefore, the pre-trained BERT base model (uncased)[39] was utilised by exploring the prediction of masked tokens. This model was originally trained on English Wikipedia and BookCorpus and the uncased version was selected, since in English the capitalisation of a word does not indicate a change in its meaning.

Regarding the analysis of the data, a series of statistical approaches was followed in combination with a qualitative analysis. Specifically, contingency tables were created to assess the agreement in the assignment of genders between groups and then the McNemar- Bowker Chi-square test [18] as well as Cohen's Kappa [29] were utilised. Then, the TF-IDF model alongside the Rank Biased Overlap measure [110] were employed. Finally, a thematic analysis was conducted so as to analyse the qualitative data and identify patterns.

6.2 Dataset & Preprocessing

At first, three shared datasets with job postings (2 from *data.world*[1] and 1 from *kaggle*[2]) were selected as potential materials for the study. Since each one consisted of a large number of entries ($N = 30.000$), a random selection of 1000 entries per dataset without a preselected seed was made. Two of them were queried with the website's built-in SQL, while the other with Python (in a Pandas Dataframe representation). Each dataset was parsed in order to examine the frequency

[1] `https://data.world/`

[2] `https://www.kaggle.com/`

of stereotypically masculine and feminine words based on previous literature [54, 99, 61, 93, 114] and therefore to assess which one the most suitable for the purpose of the current research was. In the Appendix A, the list of the gendered wording is provided. The dataset with the largest amount of gendered wording included 3359 typically feminine and 2793 typically masculine words. For reference, the second one contained 3423 and 1940 typically feminine and typically masculine respectively, while the third one 2553 and 1160 respectively. Thus, from the first dataset, 200 job postings were randomly selected as the material of the surveys and the algorithmic assessment. The number 200 was chosen, since it is high enough to get a clear understanding of the model's potential bias, but also not too high so that it is feasible all the job postings to be analysed by people.

As the main focus of this research was the evaluation of people's and BERT's perspective on gender bias, the pronouns of each job description were detected in order to be masked in a similar way as in BERT, since they are gendered words. In that way, this process provided a great basis for comparison. To circumvent the fact that most of the job advertisements were naturally lengthy and this would maybe cause issues later in the surveys, they were shortened. In order, though, to still retain the meaning of the posting, one or more of the following three steps were taken: 1) Each description was shrunk to three sentences with the second sentence including the pronoun. As this process was not definitive, in some cases there were more than one pronouns in a row of sentences in the refined job description. 2) In cases where there were not any pronouns available at all, a pronoun was inserted in a sentence without changing the intended meaning of the original job posting. 3) Finally, in the refined job description, when a sentence was too long, it was shortened, if no gendered denoting words were included at the end of it. For each of these cases, an example is presented in Table 1 below.

Table 1: Exemplary cases in each step of the preprocessing

	Examples
1) Pronoun in the second sentence	The successful candidate will need to be adaptive and be comfortable working in a dynamic team. **You** will be the first point of contact for our patients and health professionals, therefore a high level of maturity, professionalism, and sensitivity is required. ESSENTIAL CRITERIA: Previous experience working within a healthcare setting.
2) Addition of a pronoun	Have a passion and commitment to aged care and an affinity for the aged. **You** must have excellent customer service, interpersonal and communication skills. Have previous experience with an electronic care management system.

Continued on next page

Table 1 – continued from previous page

	Examples
3) Shortened sentence	This individual must have advanced skills in desktop and server administration, support and implementation, effective decision-making and problem-solving, excellent communication abilities, and be highly collaborative in nature. You will be required to be very technical with a proven ability to assist in system administration, project delivery, and troubleshooting both remotely and in a hands-on capacity. This is a dynamic role and requires the applicant to be able to identify issues, effectively communicate with the customer and accurately capture their requirements, present well, employ excellent oral and written skills **[...]**.

After that, each job description was processed aiming at investigating how many stereotypically feminine and masculine words it included. This process was conducted manually, since, as noted by Kanij et al. (2022) [70], text analysis tools that detect gendered language can not make a distinction between a word that is considered biased but in certain cases is used as a technical term. For example, words such as *analytical, analyse* and *analysis* derive from the same root which is considered stereotypically masculine. However, the phrase *Data analytics* is an established term that is considered biased using the aforementioned tools, although there is no other way of phrasing it. By detecting the gendered wording, each job posting was annotated as either mainly feminine, mainly masculine or neither feminine nor masculine. In total, there were 94 mainly feminine, 55 mainly masculine and 51 job descriptions that were neither masculine nor feminine. All of them were distributed equally at a later step between each of the five surveys.

Then, due to privacy concerns, in case the name of the company and the location were mentioned, they were either not included, if that was possible, or the name of the company was replaced with the phrase "the company" or the pronoun "this".

6.2.1 Masking Procedure

For the final step of the dataset preprocessing, which is the masking of the pronouns, the part-of-speech (POS) tagging from the *spaCy* library was utilised. POS tagging is the process of marking a token to a specific part-of-speech, that is to say a syntactic category.

For transparency reasons, details of the code implementation are provided. At first, the modules *sys* and *re* as well as *spacy* and *sent_tokenize* from the *nltk.tokenize* were imported. Additionally, one of *spaCy's* trained pipelines, *en_core_web_sm*, was pip installed to ensure that all the necessary dependencies were in place. Then, the *nltk.download(punkt)* was run once in order for the model to tokenize efficiently the text at hand.

In the main function, *en_core_web_sm* from the spaCy library was loaded into an object, named *nlp*. The *sent_tokenize*, then, enabled the division of the text at a sentence level. Each sentence was then cleaned, meaning trailing whitespace characters were removed. By using a *control* statement, the program evaluated whether a word was a pronoun and only in the case it was, it replaced the pronoun with an "[X]". At the end, the console output was saved in a CSV file.

6.3 User study

6.3.1 Participants

The initial number of participants was 157 (75 with a female identity, 20 with a gender diverse, 57 with a male identity, 4 non-disclosed and 1 who preferred not to say) and the age ranged from 20 to 39. Eighty of them (37 with a female identity, 16 Gender Diverse, 22 with a male identity, 4 non-disclosed and 1 who chose not to answer) were eliminated right away, because they didn't complete the whole survey. The aim was a stratified balanced sample in order to assess the effect of the independent variable (gender of participants) on the dependent variable (the assigned gender to each posting). Therefore, there was another round of eliminations where 7 participants with a female, 4 with a gender diverse and 5 with a male identity were not included in the data analysis. Out of them, two participants (1 with a female and 1 with a male identity) were eliminated as outliers due to their age (36 and 39 respectively) and the rest were eliminated randomly. Although the goal was to include participants with various gender identities, unfortunately, the amount of respondents who identified with a gender diverse identity was low and unbalanced so as to be a part of the analysis. The age of the final participants ranged from 23 to 31 and their nationalities were, among others, American, British, Chinese, Dutch, Greek, Iranian, Irish and Italian.

The final sample used for the analysis, after the aforementioned rounds of elimination, included 60 people in total, 12 per survey, as Table 2 depicts. More specifically, each of the five surveys consisted of a balanced distribution of people who self-identified as female (n=6) and male (n=6). The mean age of the participants was 27.75 in Survey 1 and Survey 5, 26 in Survey 2, 26.16 in Survey 3 and 26.58 in Survey 4.

6.3.1.1 Recruitment procedure

The participants were recruited through the online distribution of the surveys. In particular, a series of different approaches was followed in order to avoid selection bias. Emails were sent to various groups of many universities in the Netherlands, such as the ones responsible for student matters and the LGBTQIA+ groups as well as to the department of Gender Studies of Leiden University. In addition, since the aim was to include people with a wide range of gender identities, messages were sent to various LGBTQIA+ groups online via *Facebook*, *Instagram* and *Reddit* and numerous NGOs were approached. Finally, a snowball procedure enabled the recruitment of more participants. In other words, participants spread the surveys further to

Table 2: Participants' distribution per survey

	Survey 1	Survey 2	Survey 3	Survey 4	Survey 5
Initial number of participants	34	41	21	38	23
First round eliminations	16	28	6	24	6
Second round eliminations	6	1	3	2	5
Participants per survey	12	12	12	12	12
Total			60		

other individuals. Information regarding the ethical considerations of the study is presented in section 11.

6.3.2 Instrument

To investigate the potential bias of individuals and the hypothesis that a person's own gender influences their perception of the preferred gender in a job advertisement, five online surveys were conducted. The dataset was split into five surveys so that each survey takes maximum 30 minutes and the survey software *Qualtrics* [1] was utilised.

All five surveys had the exact same format with the only difference being that each one contained different job descriptions from the aforementioned dataset. In particular, each survey consisted of 40 job postings plus the demographic questions. Prior to filling out the surveys, participants had to give active consent. Thus, they were informed about the voluntary nature of their participation and the confidentiality of the data. In addition to that, an introductory note was presented with all the relevant information about the study without stating, though, the actual goal of the research. In this way, potential biased responses could be avoided.

After giving their consent, participants were directed to the demographic questions. Individuals were asked about their gender, their age and their nationality so as to get a better view of the diversity of the respondents. With respect to the gender question, it has been found that individuals with a gender diverse identity prefer an open-ended question, a fill-in bar in order to avoid misgendering, and also the option not to answer at all [94]. Furthermore, as Broussard et al. (2018) [21] suggest, the best option to ask about gender is a combination of a multiple choice and an open-ended question in order to avoid methodological issues. Thus, in each survey the people who chose the gender diverse identity had a write-in option to specify. In addition, some gender diverse individuals recommend to not place the cisgender options of male and female first in the list of the answers [94]. For this reason, the available answers were placed in an

[1] `www.qualtrics.com`

alphabetical order (i.e. *Female, Gender Diverse, Male, Prefer not to answer*).

Throughout the whole survey participants were asked whether they think that the author of each job description favours a certain gender. By putting the focus on the author instead of asking the participants straight forward whether they think each description shows a preference to a specific gender, the possibility of people trying to be politically correct could be minimised. This question was multiple-choice and the available answers were: *-Yes, the preferred gender is female, - Yes, the preferred gender is gender diverse, - Yes, the preferred gender is male, - No, there is no preferred gender, - Not sure.* The last option was available in case individuals were not able to decide in a short period of time.

To get a better understanding of the reason behind the answers, participants had to highlight words that may have influenced their decision and/ or give extra motivation for their answer. The latter was in an open text format giving, in that way, the freedom to participants to elaborate as much as they wanted. In Appendix B, a fragment of a survey is provided. It was, thus, possible to explore whether certain linguistic cues, either the ones already available in the literature or new ones, triggered the bias or whether the preconceived notion of the job type was the main reason behind people's judgements. Therefore, this section provided also relevant information for the hypothesis 2a; that humans are triggered by context words in the attribution of a gender to a posting.

At the end of each survey, there was a fill-in box for participants to provide their feedback and/or write their email in case they wanted to be updated on the results. Since each survey took around 30 minutes for an individual to fill out, participants had the option to finish taking the survey at a later point and not in one go.

Prior to the actual research, a pilot study was conducted in order to ensure that the questions were clear and therefore examine the validity of the investigation. Six people (3 men and 3 women) between the age of 22 to 26 participated in it. Through this preliminary small- scale study, it was ensured that the survey could be completed in the expected time frame. In addition, feedback was received and therefore small changes were made. For instance, instead of asking respondents about the signifying wording in an open text format, a hightlight question type was created. This ensured an easier process for the participants and a better way of collecting data for the analysis.

6.4 Unmasking procedure

The BERT model is trained to unmask masked tokens in the form [MASK] considering the linguistic context of the sentence. By unmasking the pronouns which, as it was mentioned in section 6.2.1, were masked during the dataset processing, it was possible to explore whether the model gives highest weights on words that denote a certain gender.

This process can be thought as a two steps procedure; an exploratory step and a conclusive one [2]. At first, the [X] that replaced the pronouns during the masking procedure was replaced

[2]The scripts can be found at https://github.com/MarinaVrentzou/thesis

with [MASK] for easier manipulation. Then, a pandas dataframe named *input_df* was created by reading the CSV file described in the masking procedure (6.2.1).

As far as the exploratory design is concerned, a function was created, named *unmask_sentence*, which predicted the masked tokens. In short, the function took as an input a masked sentence and produced as an output the top_k = 5 predictions accompanied by their weights. The predictions were evaluated by the *BertForMaskedLM* function from the *transformers* package and the final choices were estimated based on the softmax function in the *torch* package. For reference, the softmax function converts values to a normalised range of [0,1], which can be thought as probabilities. Afterwards, a function named *get_output_from_text* was defined which used the aforementioned *unmask_sentence* to split each job posting in distinct rows for each masked token. To ensure that no punctuation elements were predicted for a [MASK] at the end of a sentence, a small check was added. In case of multiple masked tokens in a posting, in order to assist the user which token had already been predicted in previous iterations, the [MASK] was replaced with "Unmasked".The dataframe was then exported (*export_function*) to a CSV file named *UnmaskedJobPostings*. This paved the way for an exploratory analysis of the top_k = 5 results in the *UnmaskedJobPostings* file. The analysis showed that apart from the nouns *he, she, they*, the nouns *candidates* and *applicants* were among the top results of the model. These non-gendered terms are strongly associated with the task at hand and thus proved to be an ideal cornerstone for the next step. Therefore, the list *target_words* was created consisting of the three pronouns and the two nouns previously mentioned.

As for the conclusive step, a function *unmask_sentence_desired_words* was defined which greatly resembled the *unmask_sentence* but instead of looking at the top_k = 5 predictions, it searched among the top_k = 1000 predictions for the words in *target_words*. The choice of 1000 words prediction was based on the fact that most target words would fetch an actual value and not N/A. The rest of the process was similar to the first step, namely, the function *get_output_from_text* was used and then the final dataframe was exported to a CSV file named *TargetWordsUnmaskedJobPostings*. Figure 1 illustrates the pipeline of the unmasking procedure.

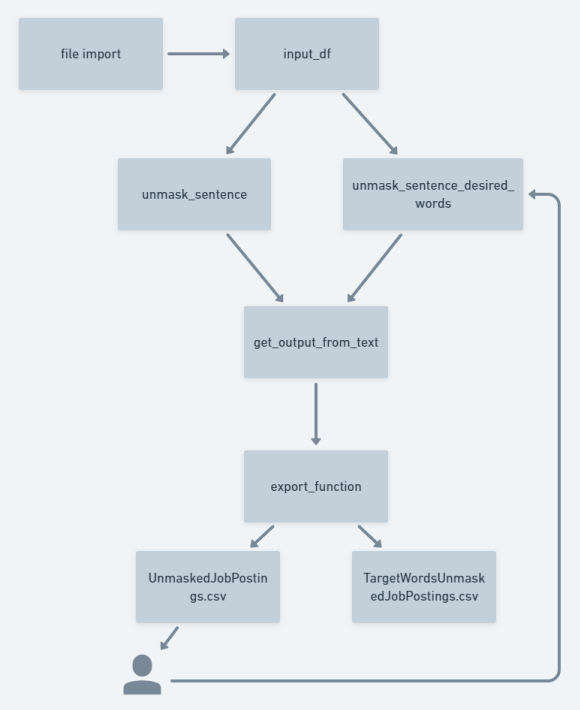


Figure 1: Pipeline of the unmasking procedure

6.5 Data Analysis

6.5.1 Assignment of genders to postings

The first step of the analysis process was assigning a gender (Male, Female, No Preferred gender, Gender Diverse[3]) to each posting based on, on one hand, the results attained from the unmasking procedure and, on the other hand, the results attained from each survey. In order to assess the gender that BERT assigned to each posting, at first it was observed whether each posting contained more than one masked tokens. In case it didn't, then the highest bearing unmasked gendered word determined the gender. For example, if the pronoun "she" had the highest weight, then the female gender was assigned to the posting. The same applied in the case there were more than two masked tokens and there was alignment between the unmasked results. In case, though, there was no agreement between the top weights of the masked tokens in a posting, the ratio between the weights was measured and then the weight with the highest ratio determined the gender. For example, in the case of a job advertisement consisting of two masked tokens each one with a different highest weight, the following procedure was followed:

1st Mask	he weight: 0.2082 they weight: 0.2080	Ratio 1: $\frac{0.2082}{0.2080} = 1.0009$
2nd Mask	they weight: 0.6742 he weight: 0.0190	Ratio 2: $\frac{0.6742}{0.0190} = 35.484$

In this case, the highest weight of the first masked token is *he*, whereas the second highest weight is *they*. Likewise, the highest weight of the second masked token is *they* and the second highest weight is *he*. Since *Ratio* 2 > *Ratio* 1 and in 2nd Mask *they weight* > *he weight*, the No Preferred Gender ("they") was assigned to the posting. Then, a file was created consisting of a column of all the postings and a column of the assigned genders by BERT.

With regard to the perceived favoured gender of each posting based on the surveys' results, in case there was not absolute agreement between the participants in each gender group, the reply that the majority of respondents gave determined the assigned gender of the group. However, in the case there was not one answer which was selected by most of the participants, it was presumed that there was no preferred gender. This was based on the fact that there was not a specific gender which stood out the most. This procedure was followed for each of the five surveys. At the end, the results from the surveys were merged into two files. One file included all the postings and the gender that women ascribed to each job advertisement and the other all the postings accompanied by the ascribed gender by men.

[3]The BERT model in this study does not assign a Gender Diverse class to the postings. More information is provided in section 7.1.1.

6.5.2 Quantitative analysis

6.5.2.1 Testing Hypothesis 1

To investigate the first research question and to test the hypothesis that the gender of a person influences their perception of the preferred gender in a job posting, a series of quantitative approaches in Python were followed. For every statistical test, an alpha level of α=0.05 was used.

At first, a cross-tabulation was used to analyse the relationship between the categorical variables and, in particular, the relationship between the genders assigned by BERT and the genders assigned by women and men. In such a way, it was possible to examine quantitatively the amount of times BERT, women and men agreed on the assignment of a certain gender. Then, a series of non-parametric tests was run, since the underlying distributions were unknown.

The contingency tables were at first computed along with the McNemar-Bowker Chi-square test [18], which is an extension of the McNemar test that allows for multiple classes. The null hypothesis is that the marginal distribution is the same for the two classification groups that are compared, or in other words $P_{i.} = P_{.i}$.

In addition to that, the Cohen's Kappa metric [29] evaluated the level of agreement between the annotators, i.e. between BERT and what the majority of women and men answered respectively as well as between the latter two. The null hypothesis of this metric is the opposite of the one of McNemar-Bowker and states that κ is 0. This κ metric falls in the range [0,1], where 0 indicates an agreement equivalent to chance, and 1 indicates a perfect agreement between the annotators.

By utilising these statistical tools, it was possible to determine the overall agreement between the three groups with regard to the assigned gender. So as not to make any assumptions regarding the agreement or disagreement of the classification groups, both tests were utilised since they have contrasting null hypotheses.

6.5.2.2 Testing Hypothesis 2a

To test hypothesis 2a, namely, to assess whether, similar to BERT, individuals were triggered by context words, a term weighting scheme, the Term Frequency- Inverse Document Frequency (TF-IDF) was employed. By utilising TF-IDF, the top 50 most important words were detected from a corpus comprising postings attributed to the same gender by each group. The TF-IDF model designated how relevant a word is to a posting considering the corpus this posting was a part of. In other words, this model detected which words were indicative of a specific attributed gender.

To examine, more specifically, whether there was an agreement between the three groups with regard to the words that signified each assigned gender, the overlap-based tool, Rank-Biased Overlap (RBO) [110] based on TF-IDF, was implemented. This was achieved by comparing per case two files, each one containing all the posting that were assigned the same gender, but from different groups. This tool measured the overlap between the two documents and returned a

value varying from 0 to 1. An RBO value of 0 indicated that the two ranked documents were disjoint, while a value of 1 meant that they were identical.

This measure was chosen for the following reasons. Firstly, the two files per case were not conjoint, namely they did not necessarily include the same postings. Secondly, the first words from the TF-IDF results were more important compared to the bottom ones and finally, the decision to restrict the list to the 50 top words was arbitrary. These qualities conform to the notions of *non-conjointness*, *top-weightedness* and *indefiniteness*, all of which RBO takes into account [110].

6.5.3 Qualitative analysis

To explore further the second research question and whether words or other factors have influenced people's perceptions (i.e. hypothesis 2b), a qualitative analysis was conducted.

As it was mentioned in the Instrument section (6.3.2), as a part of each survey, participants had to highlight the words which influenced their decision as well as to elaborate on the reasons behind their decision in a fill-in box. By following Thematic Analysis and specifically the guide provided by Braun and Clarke (2012) [20], common themes were distilled from the data in an inductive way. These themes reflected the reasons behind participants' decision and the type of words that determined those decisions.

After the collection of all the data, the first step was to analyse each survey on its own. At first, each data item was transferred to a separate file, while at the same time notes were made for each case in order to get a general idea and understanding of the dataset's content.

The next step was to generate codes, namely labels to organise the data. This phase was focused on creating codes and discarding data that were of no importance, such as comments similar to "I do not see any gendered language in this text". Some codes were created by just describing participants' comments, while other by interpreting the comments. An example of the former case is the code "certain job positions typically linked to a specific gender" which was created for the comment "Banking makes me think of male so I make the assumption before I even read on any further". An example of an interpretative code is "manifestation of the oppression created by the patriarchy" for the comment "submissive qualities expected shrouded in a happy vocabulary sounds like this job application has a preference for the female gender". The same types of codes were also created for the highlighted words. For example, words that denoted a job position, such as "labourers" and "operators", and were attributed to a specific gender received the descriptive code "words attributed to a certain gender based on the profession", while words such as the adjective "positive" used in phrases such as "positive manner", "positive attitude" were interpretatively coded as "words linked to emotional bias towards females".

After generating the codes, patterns were identified and codes were sorted for a draft creation of themes. In case there was an overlap between certain codes, one theme was created with sub-themes consisting of these codes so to describe a coherent pattern. For the current study, at

first separate themes were created based on the data women and men provided respectively.

The next phase was refining the draft themes and exploring whether each theme included coherent codes. In case there were codes that didn't fit in the theme, they were either discarded or were used for the creation of another theme. At the same time, it was examined whether there were enough data that supported each theme or whether certain themes were thin and therefore not so meaningful. To finalise the analysis of this phase, the dataset was explored again to ensure that all the themes were representative of it.

Next, the themes were defined by determining the essence of each one. An additional step in the current study was to examine the final themes from all the surveys with the purpose of extracting the overall main themes. During this procedure, in case there were identical themes between all the surveys, one theme was created including a large amount of data. Additionally, the most interesting extracts per theme were selected for the final presentation.

The final step was the analysis and presentation of the themes in a concise and meaningful way.

7 Results

7.1 Quantitative analysis

7.1.1 Exploration of agreement between the three groups

To answer the first research question and to examine the hypothesis that the gender of a person influences their perception of the preferred gender in a job posting, contingency tables were created and the McNemar-Bowker test as well as Cohen's Kappa were used.

In Table 3, the three cross-tabulation tables of the groups are presented. In each of the three tables, the amount of times there was an agreement between the groups with regard to the gender they assigned to the postings as well as the number of times that each group assigned different genders compared to their pair are presented. In the case of BERT, the Gender Diverse (GD) identity group is absent, due to the fact that people with gender diverse identities self-identify with various pronouns and therefore the model couldn't make a distinction for this class. Note that F stands for Female, GD for Gender Diverse, M for Male and NP for No Preferred gender.

Table 3: Contingency tables

(a) Assigned Genders by BERT and Women

BERT \ Women	F	GD	M	NP	Total
F	2	0	0	3	5
M	2	0	7	29	38
NP	4	1	14	138	157
Total	8	1	21	170	200

(b) Assigned Genders by BERT and Men

BERT \ Men	F	GD	M	NP	Total
F	1	0	0	4	5
M	0	0	9	29	38
NP	19	0	18	120	157
Total	20	0	27	153	200

(c) Assigned Genders By Men and Women

Men \ Women	F	GD	M	NP	Total
F	2	1	0	17	20
GD	0	0	0	0	0
M	0	0	15	12	27
NP	6	0	6	141	153
Total	8	1	21	170	200

The highest agreement between the groups was observed in the assignment of No Preferred gender. On the contrary, the highest disagreement between the groups was observed in the attribution of the female gender. Specifically, there were 20 postings to which men assigned the female gender, out of which only 2 and 1 got ascribed the same gender by women and BERT,

respectively.

The McNemar-Bowker test showed a significant difference between the distributions of BERT and men with $p = .006$, whereas contrary to the hypothesis, in the cases of BERT and women and women and men, there was no significant difference with $p = .060$ and $p = .064$, respectively. Similarly, the Cohen's Kappa was $\kappa = .056$ between BERT and men and $\kappa = .14$ between BERT and women, values which show a slight agreement. On the contrary, the same metric for women and men was $\kappa = .36$ which signals a fair agreement among the two groups based on the interpretation guidelines by Landis & Koch (1977) [78]. Based on these statistical tests the aforementioned hypothesis is not supported, since there is no significant difference in the attribution of gender to job postings between the two subgroups of the population, women and men. This may be a result of insufficient evidence, since as it was observed, although there was a high agreement in the attribution of the No Preferred gender, there was a low one in the attribution of the Female and Male class.

By exploring empirically the level of agreement between the groups, there were only four occasions where all three groups assigned the gender of male to the same posting and only one occasion, where all of them attributed the gender of female to the same job description. In the four postings which were perceived as favouring the male identity, participants highlighted phrases and words such as "manual tasks", "Motor Mechanic", "Artilleryman" and "craft beers" as indicative of favouritism. With regard to the feminine perceived job description, participants highlighted the phrases "meaningful relationships" and "social, emotional and physical needs". More information with regard to the wording is provided in the next sections. In addition, there were 114 postings that all three groups considered as showing no preference towards a specific gender.

7.1.2 Signifying wording

To examine hypothesis 2a, according to which, similar to BERT, humans are triggered in the attribution of gender by context words, reflecting bias in language models, at first the TF-IDF technique was implemented by extracting the 50 most indicative words per case. In Table 4, the top 5 highest weighted words per case are presented.

When the assigned gender was female, the common words between the three groups were only "team" and "skills", whereas men and women shared in addition to the aforementioned, words including "care", "understanding", "flexible", "positive" and "excellent".

When the assigned gender was male, 15 out of the 50 words were common among all the groups, some of which were "work", "team", "role", "experience", "seeking", "responsible" and "business". Women and men shared 11 additional words in common including "provided", "require" and "duties".

As for the case of No Preferred gender, the three groups had in common the highest number of words, namely 42. A few of these words were "team", "work", "experience", "position" and "provide". The full list of words for each group can be found in Appendix C.

Table 4: TF-IDF top 5 highest ranking words

	Female As Assigned Gender	Male As Assigned Gender	No Preferred Gender
BERT	role, successful, professional, team,organisation	team, role, experience, customer, work	team, work, skills, ability, experience
Women	people, care, management, skills, recruitment	work, ability, able, team, skills	team, work, skills, role, experience
Men	team, skills, work, care, communication	work, role, ability, team, provide	team, work, experience, skills, role

By exploring this matter also empirically based on the results from the surveys, the wording indeed influenced participants. Certain words polarised the two groups of participants, although in theory, they were indicative of a specific gender. For instance, the verb "commit" and its derivatives were perceived as mainly masculine by men, but mainly feminine by women. However, words such as "progressive", "resilient", "autonomous" and "independent" had no strong gender connotations, although they were included in the list of gendered wording (Appendix A).

As an additional step, to investigate the level of agreement based on the TF-IDF results, the Rank-Biased Overlap measure was employed and its scores are reported in Table 5. In the cases where the attributed genders were female and male, the highest overlap was reported in the informative words between women and men (.4332 and .6854 respectively), while the lowest between BERT and women (.0561 and .4437 respectively). However, the overlap score was the highest between BERT and women (.9154) in respect of the wording signifying a non-biased posting. This is in alignment with the fact that these two groups agreed the most in the attribution of the No Preferred gender, as it is depicted in the Table 3a. In this case, the lowest score was reported between BERT and men (.8825). Since only the group of women suggested that one posting was favouring individuals with a gender diverse identity, it was not possible to make comparisons with regard to this identity as an attributed gender.

Table 5: Rank-Biased Overlap values

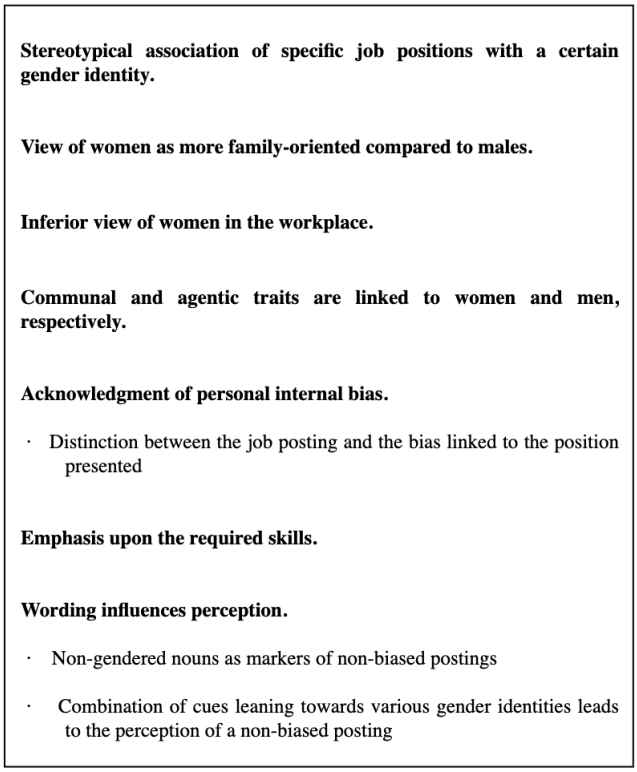
	Female As Assigned Gender	Male As Assigned Gender	No Preferred Gender
BERT- Women	.0561	.4437	.9154
BERT- Men	.1921	.4653	.8825
Women- Men	.4332	.6854	.9134

Overall, all three groups were influenced by context words and therefore the hypothesis 2a is supported. However, differences between the influential words were observed, especially in the case of female as the attributed gender.

7.2 Qualitative analysis: Thematic Analysis

To get insights into potential other reasons behind respondents' answers in the surveys and to explore the wording that influenced people's perceptions specifically in each job posting, a thematic analysis was conducted. In that way, hypothesis 2b, according to which gender and context words together are not sufficient for explaining assigned genders, was explored. Seven overarching themes for both genders were created that reflect individuals' motivation and the influential wording. In Figure 2 the main themes with their sub-themes are presented.

Figure 2: List of themes and sub-themes



7.2.1 Stereotypical association of specific job positions with a certain gender identity

Participants of both gender identities expressed in various cases that the job position itself and the field of work was what influenced their decision. Job positions such as "Motor Mechanic", "Technicians", "Labourers" and "Truck Drivers" were highlighted by the majority of respondents as mainly masculine, whereas the field of IT was perceived mainly masculine mostly by men. On the contrary, positions including "care workers" and "nurses" were viewed as mainly feminine, with the latter being perceived as such mainly by men. Quotes such as "A *she* truck driver is more rare than water in desert" expressed by a female participant as well as "Sounds like a nurse and sadly nurses are still thought of as a female dominated field" expressed by a male respondent reflect that. Words and phrases, which were highlighted, were among others "welding", "operate

machinery", "stoop", "crouch" and "technical".

There were, though, a few "reverse psychology" cases, where male participants suggested that a posting favours the female gender because it describes a position in a male-dominated field. Exemplary of this is the following quote: "Since computer programming jobs tend to be male dominated, I can see that this company would favor a female applicant. ". This is in accordance with the expression of male participants that these stereotypical beliefs are "changing year by year".

7.2.2 View of women as more family-oriented compared to men

Another reason behind people's perceptions was the association of the female identity with family. There were numerous occasions where participants of both gender identities suggested that certain job vacancies were discriminatory towards women because of that. Quotes such as "I think there may be the underlying bias that women are more family oriented and thus less flexible", "Flexibility and traveling made me think of men, without having the responsibility of family" and "In a society where women are often assumed to be responsible for raising their children, a female hire may be assumed by the author to be incompatible with these working hours" support this. Therefore, words such as "weekends", "flexible" and "overtime" were underscored.

7.2.3 Inferior view of women in the workplace

A part of only male participants explicitly stated that senior positions are linked to the male gender and that women have an inferior position in the workplace. Quotes such as "The seniority of this role in the financial field makes me think of a 'man'." and "This job application feels like it has a preference towards the female gender. Contradictingly expecting *submissive* loyalty while granting autonomy." support this. These beliefs were accompanied by the perception of women as being less professional: "I would like to point out that in general men are more professional in their work than women. It seems natural to me that this job posting is male.". Phrases including "great attitude will teach you great skills" and words such as "professionalism" and "loyal" were highlighted to underpin these claims.

7.2.4 Communal and agentic traits are linked to women and men, respectively

Words pertaining to communal and agentic traits were highly underscored by respondents. Female participants perceived communal words as either feminine or non-gendered, whereas male participants viewed them mainly as feminine. In particular, male participants were highly influenced by the phrases "interpersonal skills" and "communication skills". However, phrases which included the words "relationship" and "care" were perceived as feminine by participants of both genders.

With regard to the agentic words, the adverb "competently" was linked to a masculine trait only by women, whereas adjectives including "assertive" and "driven" were viewed as mainly

masculine by both groups of respondents.

7.2.5 Acknowledgment of personal internal bias

Although gender stereotypes are built deeply in people's minds without people being always aware of that, participants, mainly the ones with a male identity, expressed implicitly this realisation. Evidence of this are the following comments: "I connect the highlighted words to 'men' *unfortunately*" and "...it's hard for me to notice if something is more masculine oriented because that's what I am familiar with". The latter showed that this participant being always exposed to male bias made him not realise whether a posting is biased against another gender. Only one female participant explicitly stated that the implicit bias was the source of her decision: "I don't necessarily think the author has a preferred gender in mind, but the "precision welding skills" sounds like a job opening that is traditionally more masculine. However, this is more my internal bias than what the posting is presenting.".

7.2.5.1 Distinction between the job posting and the bias linked to the position presented

The acknowledgement of the internal bias resulted in many cases in men to not let their biased perceptions influence their evaluation of a posting. In particular, male participants assigned the No Preferred gender to certain advertisements by commenting, though: "The role makes me think of a woman, even if there is no preferred gender." and "No bias - However, I do think of a man when I think of an 'accountant' ".

7.2.6 Emphasis upon the required skills

According to the respondents, the focus mainly on skills was an indication of the job advertisement not showing a preference towards a certain gender as it is implied in the quote: "I feel like this application is very straight to the point and doesn't favor a particular gender.". In majority of cases, the requirement of generic personality traits not really relevant to the required skills was what triggered people the most which is apparent in the following quote: "...but I don't see how this word is relevant in a job offer. Why is someone supposed to be recruited according its 'positive' personality?".

7.2.7 Wording influences perception

So far, attention was focused on the internal gender bias which influenced people's perceptions. Another factor, though, that triggered the decision of participants was the particular choice of wording.

The choice of certain words, the context and their frequency were crucial for the way job vacancies were perceived. Exemplary of this is the case of the word "strong" being used as an intensifier. This resulted in many cases the vacancy to be considered biased towards men by women and this word being highlighted as indicative of male bias. Additionally, when the word

"lead" and its derivatives were included more than once in a posting, the latter was perceived as favouring men by female participants. The plural form of the noun "leaders" prompted a male participant to think of: " 'old boys network' stories where the same type of men keep choosing similar people for their jobs.". Regarding the frequency of words, a female respondent suggested that the repetition of the auxiliary verb "must" made a posting "too direct" and therefore exuding a male bias, whereas another commented "if the word support wasn't used that many times then I would say it's neutral".

With regard to the selection of specific wording, a male respondent commented "Partnering instead of communication, and more words strongly implying leadership and autonomy that seem more preferred with the male gender." and "This feels like the male counterpart of 'being able to be well organized', but now it has the words 'responsible' (leadership), and 'logistics' instead of planning.". Additionally, certain words such as the noun "workmanship" were criticised by many. In particular, a female participant commented: "The use of the word workmanship just sounds so masculine". On the contrary, words linked to positive behaviour were perceived mainly feminine, especially by male participants.

Finally, a few vacancies emphasised the fact they value diversity. The majority of participants from both genders highlighted the relevant words like "diversity" and "inclusive" as indicative of a non-biased or gender diverse posting.

7.2.7.1 Non-gendered nouns as markers of non-biased postings

The inclusion of non-gendered nouns such as "candidate", "incumbent", "applicant" and "person" was mentioned a few times as an indication of a non-biased or gender diverse posting. Some participants expressed that the inclusion of these nouns was enough for them to determine that the posting was bias-free, even if they acknowledged the existence of gendered wording in it. Others highlighted them alongside other non-biased words as it is evident from the following quote: "The use of the incumbent to refer to the individual as well as the neutral adjectives that describe the role as highlighted make this sound gender diverse...".

7.2.7.2 Combination of cues leaning towards various gender identities leads to the perception of a non-biased posting

There were numerous occasions where individuals of both gender identities claimed that postings did not favour a certain gender because words with different gender connotations were included and there was, thus, a balance. For example, a female participant highlighted in a vacancy the words "fun", "love" as mainly feminine and the phrase "physically fit" as mainly masculine to support her decision.

On a similar note, postings referring to a position typically associated with a certain gender, but including words skewed towards another gender were perceived in many cases as non-biased or made participants question their initial stereotypical thought. For example, a participant commented "The phrase 'dedicated team' makes me think of a woman, the position though of a man. That's why I think it's neutral", while another replied: "Although the application sounds

like it comes from ICT, which is probably more geared towards the male gender in general, the organizing and multitasking factors could also hint a preference towards the female gender. For me, I am not sure if one is actually preferred or not.".

8 Discussion

Through this research, two questions were investigated. It was firstly assessed whether there is a significant difference in the attribution of genders to job vacancies between subgroups of the population and between each group and a BERT model. Therefore, the hypothesis that the gender of an individual influences their perception of the preferred gender in a job posting was tested (hypothesis 1). Secondly, it was examined whether similar to BERT, people are triggered by context words in the attribution of gender. For this task, two hypotheses were tested. The first one stated that similar to BERT, humans are triggered in the attribution of gender by context words, reflecting bias in language models (hypothesis 2a). The second one stated that gender and context words together are not sufficient for explaining assigned genders (hypothesis 2b).

The outcomes showed that there is a significant difference in the attribution of genders to job vacancies between men and BERT but not between women and BERT and women and men. Similar to BERT, people are triggered in the attribution of gender by context words, reflecting bias in language models. However, the findings showed that the gender of an individual and context words alone do not explain all results. Self-reported factors from the thematic analysis reveal there is more at stake.

In particular, there is, to a noticeable extent, an agreement between the BERT version that was used in this study, women and men regarding the postings which all of them perceived as showing no preference to a specific gender. However, there is a low agreement in respect of the postings to which each group assigned a female and a male gender, respectively. Context words influence to a great extent all three groups. There is, though, little overlap in the case that the assigned gender is female, which can be explained by the disagreement in the attribution of this gender. Apart from the wording, norm beliefs trigger individuals' decisions. Interestingly, though, traces of change are observed as people, specifically, men appear to be more aware of their internal biases.

The fact that there was no significant difference between BERT and women and there was overall a high agreement in the assignment of the No Preferred gender between all groups implies that the model follows to an extent similar patterns to people. However, the fact that there was a significant difference between BERT and men and a low agreement regarding the assignment of the female gender between women and men contradicts previous studies which suggest that BERT displays in general human-like biases [77]. The low agreement can be explained by the thematic analysis which showed that women associate in numerous cases stereotypically feminine wording as neither feminine nor masculine, whereas men consider it highly feminine.

Gender stereotypes played a major role in the perception of job vacancies by people, a finding which aligns with earlier work on the matter [99]. Stereotypes emanating from the patriarchal view of the society still hold strong, especially among male participants, even in young generations. However, participants and in particular men, acknowledged to an extent their prejudicial attitudes. A potential explanation may be that more conversations are held

around the topic of equality and dismantling patriarchy. Being exposed to such stimuli may have contributed to young men realising their privilege in society. However, this is up to various interpretations.

Since previous studies mentioned in section 4.1.1 mainly examined how gendered wording in job postings affects the job appeal, no concrete comparisons can be made. It is apparent, though, that the results complement previous studies on the overall impact wording has on individuals [61, 54]. In various cases, women indicated typically feminine words as neither feminine nor masculine, which is in line with past investigations suggesting that feminine-themed words are a sign of gender inclusivity [61]. Interestingly, though, feminine wording was underscored more times by men as being feminine and masculine wording was underscored by both groups which implies, in contrast to past studies [17, 61], that both genders are influenced by the wording and not only females. Furthermore, although previous studies have provided lists including gendered wording, not all these gendered words were perceived as biased by participants. Additionally, certain words such as "flexible" were considered mainly masculine, although previous studies suggested otherwise. Therefore, these lists on which many AI bias detection tools are based should be occasionally updated. This is crucial as these tools are often utilised for the creation of non-biased job advertisements.

Wording linked to personality traits had a high impact on people's perceptions. For example, words pertaining to positive emotions were mainly associated with the female gender which can be explained by the alleged female sensitivity to the concept of face. However, as it was observed, a moderate use of such gender denoting words and in combination with gendered wording linked to another gender may lead to the perception of bias-free postings. On this matter, although nouns such as "workmanship" are considered non-biased forms, in practice they were not perceived as such. Androcentric words used as a default perpetuate an underlying cognitive tendency of men to be considered the most suitable for a job vacancy.

Overall, participants of both groups got triggered both by stereotypes and the wording. However, there were differences with regard to the particular norm beliefs and words that influenced their perceptions. This phenomenon was also depicted by the low agreement in the assignment of the female gender to postings. This in combination with the fact that there was a significant difference between BERT and men suggests that people should cautiously utilise AI bias evaluation tools, which assess human-like biases. Since there were instances that, contrary to people, the BERT model exhibited biased results, recruiters should be careful when they utilise AI algorithms in the hiring process.

9 Conclusion

The aim of the study was twofold. At first, to examine whether a language model and subgroups of the population, based on the attribute of gender as a spectrum, view job postings as favouring the same gender or not. Secondly, to investigate whether specific wording influences such perceptions and whether there are additional other factors affecting those perceptions. Previous experimental investigations on the topic suggest that gendered wording in job listings affects the feeling of belongingness, especially on women, and therefore reinforces inequalities. Additionally, research in the field of NLP demonstrates that language models reveal human-like biases which is a result of several factors, such as the data they are trained on.

Although several studies have focused on each issue separately, there are limited works that systematically align the human perspective on gender bias with the algorithmic one on a similar task to the one under investigation. By creating a new dataset (Appendix E) and following a mixed-methods approach, the results of the current study showed that there is no significant difference between BERT and women and women and men in the attribution of genders. However, a significant difference was reported in the attribution of genders between BERT and men. Important differences were also empirically observed between women and men in the attribution of the female gender and between BERT and both groups of participants in the assignment of both genders.

Furthermore, all groups were influenced by context words. There was a high agreement between the groups with regard to the context words signifying a non-biased posting, although there was a low agreement regarding the informative words about a posting showing a preference towards the female gender. The qualitative analysis demonstrated that these preferences are based on the combination of specific words in job advertisements as well as on personal stereotypical beliefs.

This study suggests that the available lists of stereotypical feminine and masculine wording should be regularly updated, especially when they are used as a basis for AI systems detecting bias. In addition, this research contributes fresher insights on the form of information source on the topic of gendered language in job vacancies and in comparison to the bias reflected in language models. This investigation comes in contrast to previous works suggesting that BERT models depict in general human-like biases and, thus, pinpoints the importance of carefully using AI bias assessment tools.

10 Limitations & Future Directions

Although an effort was made to recruit a high number of participants, with various gender identities and in different life stages, this was not achieved, which raises concerns about external validity. A possible explanation behind this may be the length of the surveys in combination with the participation without compensation. Since the majority of respondents were from western, educated, industrialised, rich and democratic, the so-called WEIRD [60] societies, a Western bias may be exhibited. Furthermore, emphasis was placed upon the dimension of gender without considering interwoven identities by which it is formed. That being the case, the results can not be generalised to the whole population.

Since participants in general, and to a larger extent women, were polarised regarding the perceived gender of certain postings, further research should consider assessing the level of intra-agreement between the participants of each group. It is further interesting to explore whether BERT may predict different masked tokens after fine-tuning on the results from the surveys.

As a high disagreement was observed between women and men in the attribution of the female gender and a significant difference between BERT and men in the attribution of genders in job descriptions overall, more research should be undertaken concerning the evaluation of bias detection by AI with user experiments. Research is needed to report the perceptions of people with a gender diverse identity on similar tasks and to assess the level of agreement between them and algorithms on the topic of gender bias. There is abundant room for progress on the topic of fairness in AI and in order to talk about inclusive, fair systems, more effort should be made towards the inclusion of all demographics.

11 Ethical Considerations

To ensure that this study met the requirements for ethically responsible conduct, as a first step, the privacy officer of Leiden University was approached. In that way, it was assessed whether there are certain rules that needed to be followed for the consent and data management plan. Each survey was distributed through an "anonymous" link, that is to say, a link that does not collect personal information such as the name of participants. Additionally, the "Anonymize Responses" setting from *Qualtrics* website was enabled so as to prevent the collection of respondents' IP addresses.

People did not receive any compensation for their participation in the surveys. Prior to filling out the survey, they gave informed consent (Appendix D). They were informed that the participation was absolutely voluntary and that they were free to withdraw at any point. Furthermore, they were assured about the confidentiality of the study, namely the protection of private information.

References

[1] The United States Department of Justice. Laws enforced by the employment litigation section. title vii of the civil rights act of 1964. `https://www.justice.gov/crt/laws-enforced-employment-litigation-section`. Accessed: 2022-05-1.

[2] Andrea E Abele and Bogdan Wojciszke. Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology*, volume 50, pages 195–255. Elsevier, 2014.

[3] Inger Askehave and Karen Korning Zethsen. Gendered constructions of leadership in d anish job advertisements. *Gender, Work & Organization*, 21(6):531–545, 2014.

[4] Leanne E Atwater, Joan F Brett, David Waldman, Lesley DiMare, and Mary Virginia Hayden. Men's and women's perceptions of the gender typing of management subroles. *Sex roles*, 50(3):191–199, 2004.

[5] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. *In Proceedings of SIGCIS, Philadelphia, PA.*, 2017.

[6] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[7] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3805.

[8] Christopher T Begeny, Michelle K Ryan, Corinne A Moss-Racusin, and Gudrun Ravetz. In some professions, women have become well represented, yet gender bias persists—perpetuated by those who think it is not happening. *Science Advances*, 6(26): eaba7814, 2020.

[9] Sandra L Bem and Daryl J Bem. Does sex-biased job advertising "aid and abet" sex discrimination? 1. *Journal of Applied Social Psychology*, 3(1):6–18, 1973.

[10] Madison Bentley. Sanity and hazard in childhood. *The American Journal of Psychology*, 58(2):212–246, 1945.

[11] Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on*

Gender Bias in Natural Language Processing, pages 62–68, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3809.

[12] Nazlı Bhatia and Sudeep Bhatia. Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, 45(1):106–125, 2021.

[13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485.

[14] Stephan Böhm, Olena Linnyk, Jens Kohl, Tim Weber, Ingolf Teetz, Katarzyna Bandurka, and Martin Kersting. Analysing gender bias in it job postings: A pre-study based on samples from the german job market. In *Proceedings of the 2020 on Computers and People Research Conference*, pages 72–80, 2020.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[16] Kristina Boréus. Discursive discrimination: A typology. *European Journal of Social Theory*, 9(3):405–424, 2006.

[17] Marise Ph Born and Toon W Taris. The impact of the wording of employment advertisements on students' inclination to apply for a job. *The Journal of social psychology*, 150 (5):485–502, 2010.

[18] Albert H Bowker. A test for symmetry in contingency tables. *Journal of the american statistical association*, 43(244):572–574, 1948.

[19] Brad Sears, Christy Mallory, Andrew R. Flores and Kerith J. Conron. Lgbt people's experiences of workplace discrimination and harassment. `https://williamsinstitute.law.ucla.edu/publications/lgbt-workplace-discrimination/`. Accessed: 2022-04-24.

[20] Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012.

[21] Kristin A Broussard, Ruth H Warner, and Anna RD Pope. Too many boxes, or not enough? preferences for how we ask about gender in cisgender, lgb, and gender-diverse samples. *Sex Roles*, 78(9):606–624, 2018.

[22] Diana Burgess and Eugene Borgida. Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, public policy, and law*, 5(3):665, 1999.

[23] Judith Butler. *Undoing gender*. Routledge, 2004.

[24] Dorm Byrne, Charles Gouaux, William Griffitt, John Lamberth, NBPM Murakawa, M Prasad, Atma Prasad, and M Ramirez III. The ubiquitous relationship: Attitude similarity and attraction: A cross-cultural study. *Human Relations*, 24(3):201–207, 1971.

[25] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[26] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. *arXiv preprint arXiv:2206.03390*, 2022.

[27] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.418.

[28] Jenny Cheshire. 17 sex and gender in variationist research. *The handbook of language variation and change*, 11:423, 2002.

[29] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[30] Brian L Connelly, S Trevis Certo, R Duane Ireland, and Christopher R Reutzel. Signaling theory: A review and assessment. *Journal of management*, 37(1):39–67, 2011.

[31] John W Creswell. *Research design : Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, California : SAGE Publications, Inc., 5th edition, 2018.

[32] Anne Curzan. *Gender shifts in the history of English*. Cambridge University Press, 2003.

[33] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`, 2018. Accessed: 2022-05-01.

[34] Erenay Dayanık and Sebastian Padó. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, 2021.

[35] David DeFranza, Himanshu Mishra, and Arul Mishra. How language shapes prejudice against women: An examination across 45 world languages. *Journal of personality and social psychology*, 119(1):7, 2020.

[36] Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.122.

[37] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.150.

[38] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[40] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872, 2021.

[41] Kristin Donnelly and Jean M Twenge. Masculine and feminine traits on the bem sex-role inventory, 1993–2012: A cross-temporal meta-analysis. *Sex roles*, 76(9):556–565, 2017.

[42] Samuel Dooley, Ryan Downing, George Wei, Nathan Shankar, Bradon Thymes, Gudrun Thorkelsdottir, Tiye Kurtz-Miott, Rachel Mattson, Olufemi Obiwumi, Valeriia Cherepanova, et al. Comparing human and machine bias in face recognition. *arXiv preprint arXiv:2110.08396*, 2021.

[43] Alice H Eagly and Wendy Wood. Social role theory. *Handbook of theories of social psychology*, 2, 2012.

[44] Alice H Eagly, Christa Nater, David I Miller, Michèle Kaufmann, and Sabine Sczesny. Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American psychologist*, 75(3):301, 2020.

[45] Markus Eberhardt, Giovanni Facchini, and Valeria Rueda. Women are "hardworking", men are "brilliant": Stereotyping in the economics job market. https://cepr.org/voxeu/columns/women-are-hardworking-men-are-brilliant-stereotyping-economics-job-market, 2022.

[46] Susan K Egan and David G Perry. Gender identity: a multidimensional analysis with implications for psychosocial adjustment. *Developmental psychology*, 37(4):451, 2001.

[47] Naomi Ellemers, Cathy van Dyck, Steve Hinkle, and Annelieke Jacobs. Intergroup differentiation in social context: Identity needs versus audience constraints. *Social Psychology Quarterly*, pages 60–74, 2000.

[48] Naomi Ellemers et al. Gender stereotypes. *Annual review of psychology*, 69:275–298, 2018.

[49] European Union Law. Document 32000l0078-council directive 2000/78/ec of 27 november 2000 establishing a general framework for equal treatment in employment and occupation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078. Accessed: 2022-05-01.

[50] Gráinne M Fitzsimons and Aaron C Kay. Language and interpersonal cognition: Causal effects of variations in pronoun usage on perceptions of closeness. *Personality and Social Psychology Bulletin*, 30(5):547–557, 2004.

[51] Magdalena Formanowicz and Karolina Hansen. Subtle linguistic cues affecting gender in (equality). *Journal of Language and Social Psychology*, 41(2):127–147, 2022.

[52] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

[53] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

[54] Danielle Gaucher, Justin Friesen, and Aaron C. Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology vol. 101 iss. 1*, 101, 2011. doi: 10.1037/a0022530.

[55] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam 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)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150.

[56] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[57] Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. The times they are a-changing... or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363, 2016.

[58] Lucy Havens, Beatrice Alex, Benjamin Bach, and Melissa Terras. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, 2022.

[59] Madeline E Heilman and Suzette Caleo. Gender discrimination in the workplace. 2018.

[60] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[61] Tanja Hentschel, Susanne Braun, Claudia Peus, and Dieter Frey. Sounds like a fit! wording in recruitment advertisements and recruiter gender affect women's pursuit of career development programs via anticipated belongingness. *Human Resource Management*, 60 (4):581–602, 2021.

[62] Sally Hines and Matthew Taylor. *Is gender fluid?: a primer for the 21st century.* Thames & Hudson, 2018.

[63] Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.

[64] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, 74(2):171, 2019.

[65] International Labour Organization. The gender gap in employment: What's holding women back? `https://www.ilo.org/infostories/en-GB/Stories/Employment/barriers-women#intro`. Accessed:2022-05-1.

[66] Humaira Jami and Anila Kamal. Measuring attitudes toward hijras in pakistan: Gender and religiosity in perspective. *Pakistan Journal of Psychological Research*, pages 151–187, 2015.

[67] Sophie Jentzsch and Cigdem Turan. Gender bias in bert-measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, 2022.

[68] Mengmeng Ji. Understanding gender-coded wording in job postings with word-vectors and bert. `https://web.stanford.edu/class/cs224n/reports/final_reports/report009.pdf`.

[69] Gabbrielle M Johnson. Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198(10):9941–9961, 2021.

[70] Tanjila Kanij, John Grundy, Jennifer McIntosh, Anita Sarma, and Gayatri Aniruddha. A new approach towards ensuring gender inclusive se job advertisements. In *2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 1–11. IEEE, 2022.

[71] Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. Measuring gender bias in contextualized embeddings. In *Computer Sciences and Mathematics Forum*, volume 3, page 3. MDPI, 2022.

[72] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

[73] David Knights and Deborah Kerfoot. Between representations and subjectivity: Gender binaries and the politics of organizational transformation. *Gender, Work & Organization*, 11(4):430–454, 2004.

[74] Anne M Koenig. Comparing prescriptive and descriptive gender stereotypes about children, adults, and the elderly. *Frontiers in psychology*, 9:1086, 2018.

[75] Anne M Koenig and Alice H Eagly. Evidence for the social role theory of stereotype content: observations of groups' roles shape stereotypes. *Journal of personality and social psychology*, 107(3):371, 2014.

[76] Caitlin Kuhlman, Latifa Jackson, and Rumi Chunara. No computation without representation: Avoiding data and algorithm biases through diversity. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 3593, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3411074.

[77] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823.

[78] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[79] Hilary M Lips. *Sex and gender: An introduction*. Waveland Press, 2020.

[80] Qi Liu, Matt J Kusner, and Phil Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

[81] Lloyd B Lueptow, Lori Garovich, and Margaret B Lueptow. The persistence of gender stereotypes in the face of changing sex roles: Evidence contrary to the sociocultural model. *Ethology and Sociobiology*, 16(6):509–530, 1995.

[82] Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, 2021.

[83] James Mattone. Facebook is coding feminine language into job postings to achieve diversity goals. `https://www.businessofbusiness.com/articles/facebook-diversity-in-job-openings-gender-decoder/`, 2019.

[84] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel 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)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063.

[85] Mary McMahon. Is "queer" a derogatory word? `https://www.languagehumanities.org/is-queer-a-derogatory-word.htm`, 2022. Accessed:2022-6-13.

[86] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[87] Thekla Morgenroth and Michelle K Ryan. The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, 16(6):1113–1142, 2021.

[88] Thekla Morgenroth, M Gustafsson Sendén, Anna Lindqvist, Emma A Renström, Michelle K Ryan, and Thomas A Morton. Defending the sex/gender binary: The role of gender identification and need for closure. *Social Psychological and Personality Science*, 12(5):731–740, 2021.

[89] Serena Nanda. The hijras of india: Cultural and individual dimensions of an institutionalized third gender role. *Journal of homosexuality*, 11(3-4):35–54, 1986.

[90] Jaime L Napier, Hulda Thorisdottir, and John T Jost. The joy of sexism? a multinational investigation of hostile and benevolent justifications for gender inequality and their relations to subjective well-being. *Sex roles*, 62(7):405–419, 2010.

[91] Erin Oldford and John Fiset. Decoding bias: Gendered language in finance internship job postings. *Journal of Behavioral and Experimental Finance*, 31:100544, 2021.

[92] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[93] Helen Peterson. From "goal-orientated, strong and decisive leader" to "collaborative and communicative listener". gendered shifts in vice-chancellor ideals, 1990–2018. *Education Sciences*, 8(2):90, 2018.

[94] Jae A Puckett, Nina C Brown, Terra Dunn, Brian Mustanski, and Michael E Newcomb. Perspectives from transgender and gender diverse people on how to ask about gender. *LGBT health*, 7(6):305–311, 2020.

[95] Margreet Reitsma-van Rooijen, Gün R Semin, and Esther Van Leeuwen. The effect of linguistic abstraction on interpersonal distance. *European Journal of Social Psychology*, 37(5):817–823, 2007.

[96] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.

[97] Susan Stryker, Paisley Currah, and Lisa Jean Moore. Introduction: Trans-, trans, or transgender? *Women's Studies Quarterly*, pages 11–22, 2008.

[98] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159.

[99] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J Metzger, Haitao Zheng, and Ben Y Zhao. Gender bias in the job market: A longitudinal analysis. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017.

[100] TEAM PREDICTIVEHIRE. Introducing interviewbert: A world-first algorithm for better interviews. `https://sapia.ai/resources/blog/introducing-interviewbert-a-world-first-algorithm-for-better-interviews//`. Accessed: 2022-12-01.

[101] Nat Thorne, Andrew Kam-Tuck Yip, Walter Pierre Bouman, Ellen Marshall, and Jon Arcelus. The terminology of identities between, outside and beyond the gender binary–a systematic review. *International Journal of Transgenderism*, 20(2-3):138–154, 2019.

[102] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[103] Desiree D Tobin, Meenakshi Menon, Madhavi Menon, Brooke C Spatta, Ernest VE Hodges, and David G Perry. The intrapsychics of gender: a model of self-socialization. *Psychological review*, 117(2):601, 2010.

[104] Frances Trix and Carolyn Psenka. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14(2):191–220, 2003.

[105] Adam Tuszynski. How does language shape the way we think. 2010.

[106] Eric Luis Uhlmann and Geoffrey L Cohen. "i think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223, 2007.

[107] United Nations. Ohchr and women's human rights and gender equality. `https://www.ohchr.org/en/women`. Accessed: 2022-04-24.

[108] Steven Vethman, Ajaya Adhikari, Maike Ht de Boer, Joost Agm van Genabeek, and Cor J Veenman. Context-aware discrimination detection in job vacancies using computational language models. *arXiv preprint arXiv:2202.03907*, 2022.

[109] Afiah Vijlbrief, Sawitri Saharso, and Halleh Ghorashi. Transcending the gender binary: Gender non-binary young adults in amsterdam. *Journal of LGBT Youth*, 17(1):89–106, 2020.

[110] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[111] Candace West and Don H Zimmerman. Doing gender. *Gender & society*, 1(2):125–151, 1987.

[112] Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

[113] Lien Wille and Eva Derous. When job ads turn you down: how requirements in job ads may stop instead of attract highly qualified women. *Sex roles*, 79(7):464–475, 2018.

[114] John E Williams and Susan M Bennett. The definition of sex stereotypes via the adjective check list. *Sex roles*, 1(4):327–337, 1975.

[115] Wendy Wood and Alice H Eagly. Two traditions of research on gender identity. *Sex Roles*, 73(11):461–473, 2015.

[116] Kevin B Wright. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of computer-mediated communication*, 10(3):JCMC1034, 2005.

[117] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[118] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521.

[119] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. 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)*, pages 629–634. Association for Computational Linguistics, June 2019. doi: 10.18653/v1/N19-1064.

12 Appendix A

Table 6: List of stereotypical masculine and feminine words based on previous literature [54, 61, 99, 93, 114]

Mainly masculine words	Mainly feminine words
Active	Adaptable
Adventurous	Affected
Aggress*	Affectionate
Ambitio*	Articulate
Analy*	Care
Assert*	Cheer*
Athlet*	Child*
Authoritative	Collaborative
Autocratic	Commit*
Autonom*	Communal
Boast*	Compassion*
Bold	Complaining
Brave	Connect*
Career aspirations	Conscientious
Challeng*	Considerate
Charismatic	Cooperat*
Coarse	Creative
Compet*	Depend*
Confident*	Down to earth
Consensus builder	Emotiona*
Courag*	Empath*
Cruel	Encouraging
Daring	Excitable
Decide	Expressive
Decision*	Feminine
Decisive	Fickle
Determin*	Flatterable
Direct	Flexible
Distinctive	Flirtatious
Domina*	Friendly
Driven	Frivolous
Dynamic	Fussy

Continued on next page

Table 6 – continued from previous page

Mainly masculine words	Mainly feminine words
Enduring	Gentle
Entrepreneurial thinking	Good listener
Entrerprising	Helpful
Focused	High-strung
Force*	Honest
Goal- oriented	Humble
Greedy	Imaginative
Gutsy	Interdependen*
Handsome	Interpersona*
Hardworking	Involved
Headstrong	Kind
Hierarch*	Kinship
Hostil*	Loving
Implusive	Loyal*
Independen*	Meek
Individual*	Mild
Influential	Modesty
Innovative	Nag
Intellect*	Nurtur*
Lead*	Open to new ideas
Logic	Passive
Loud	Patient
Masculine	Perceptive
Objective	Pleasant*
Opinion	Poised
Outspoken	Polite
Outstanding	Quiet*
Perseverance	Rattlebrained
Persist	Reasonable
Powerful	Relationship
Principle*	Reliable
Progressive	Respon*
Push	Selfless
Rational	Sensible
Realistic	Sensitiv*
Reckless	Sentimental

Continued on next page

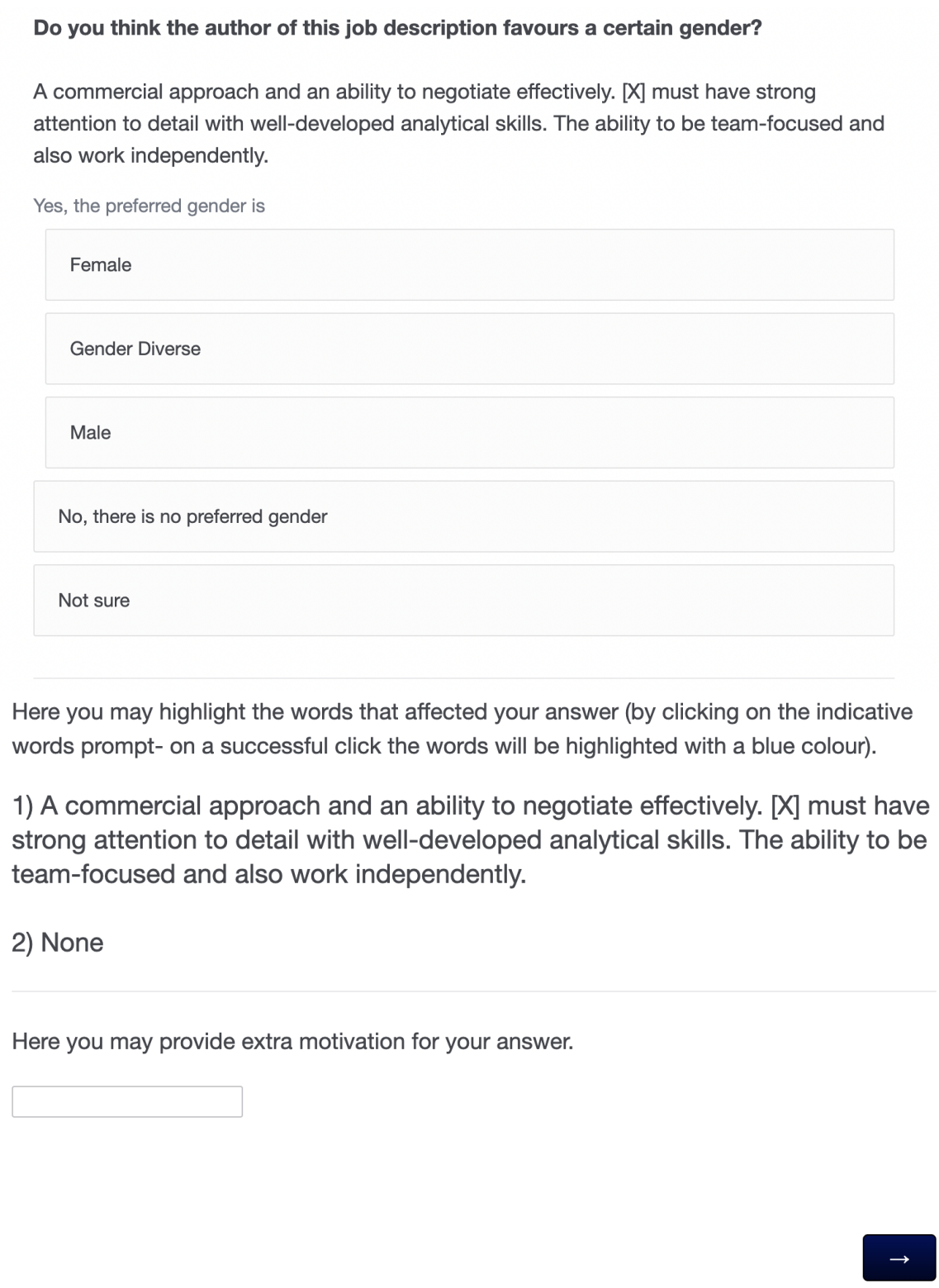
Table 6 – continued from previous page

Mainly masculine words	Mainly feminine words
Resilient	Sincere
Restrained	Sociable
Results-driven	Social
Robust	Softhearded
Self-assured	Sophisticated
Self-confiden*	Submissive
Self-relian*	Superstitious
Self-sufficien*	Support*
Severe	Sympath*
Stable	Talkative
Stern	Team player
Straightforward	Tender*
Strategic planning	Together*
Strong	Trust*
Stubborn	Understand*
Superior	Warm*
Task-oriented	Weak
Tough	Whin*
Unemotional	Yield*
Unexcitable	

Note: * denotes all possible derivatives of the word.

13 Appendix B

Figure 3: Survey Fragment



14 Appendix C

Table 7: TF-IDF signifying words

	Female As Assigned Gender
BERT	role, successful, professional, team, organisation, driven, development, management, phone, skills, attitude
Women	people, care, management, skills, recruitment, services, communication, looking, positive, experience, work, include, needs, excellent, business, customer, service, flexible, key, organisation, operation, motivated, deeply, good, professional, duties, understanding, person, relationships, advisor, career, understand, member, reporting, team
Men	team, skills, work, care, communication, role, experience, looking, able. Excellent, customer, high, needs, communications, positive, understanding, sales, successful, working, responsibilities, time, qualifications, interpersonal, player, strong, exceptional, business, development, flexible, service, day, committed, health, making, provide, good, fun, creating, verbal, casual, key, passionate, environment, responsible, doing, basic, ensure, customers, join, clinical
	Male As Assigned Gender
BERT	team, role, experience, customer, work, manager, reporting, position, business, join, working, seeking, responsible, company, skills, include, officer, leadership, service, motivated, candidate, key, sales, opportunity, support, dynamic, strong, administration, successful, management, looking, people, including, culture, excellent, individual, organisation, level, available, day, development, office, managing, safety, high, considered, responsibilities, product, assist, project
Women	work, ability, able, team, skills, business, experience, working, role, customer, quality, responsible strong, provide, duties, financial, servicing, required, including, successful, excellent, wide, range, plant, environment, highly, join, high, maintenance, key, position, applicant, professional, looking, company, opportunity, projects, enthusiastic, service, variety, fit, physically, seeking, support, time, operate, project, tasks, manner, growing
Men	work, role, ability, team, provide, experience, skills, able, strong, business, duties, working, required technical, quality, service, position, preferred, seeking, successful, environment, highly, reporting, wide, ethic, join, range, clients, opportunities, responsible, excellent, candidate, training, manager, variety, workmanship, using, communication, provided, management, job, unsupervised, including, driver, daily, person, independently, maintenance, process, project

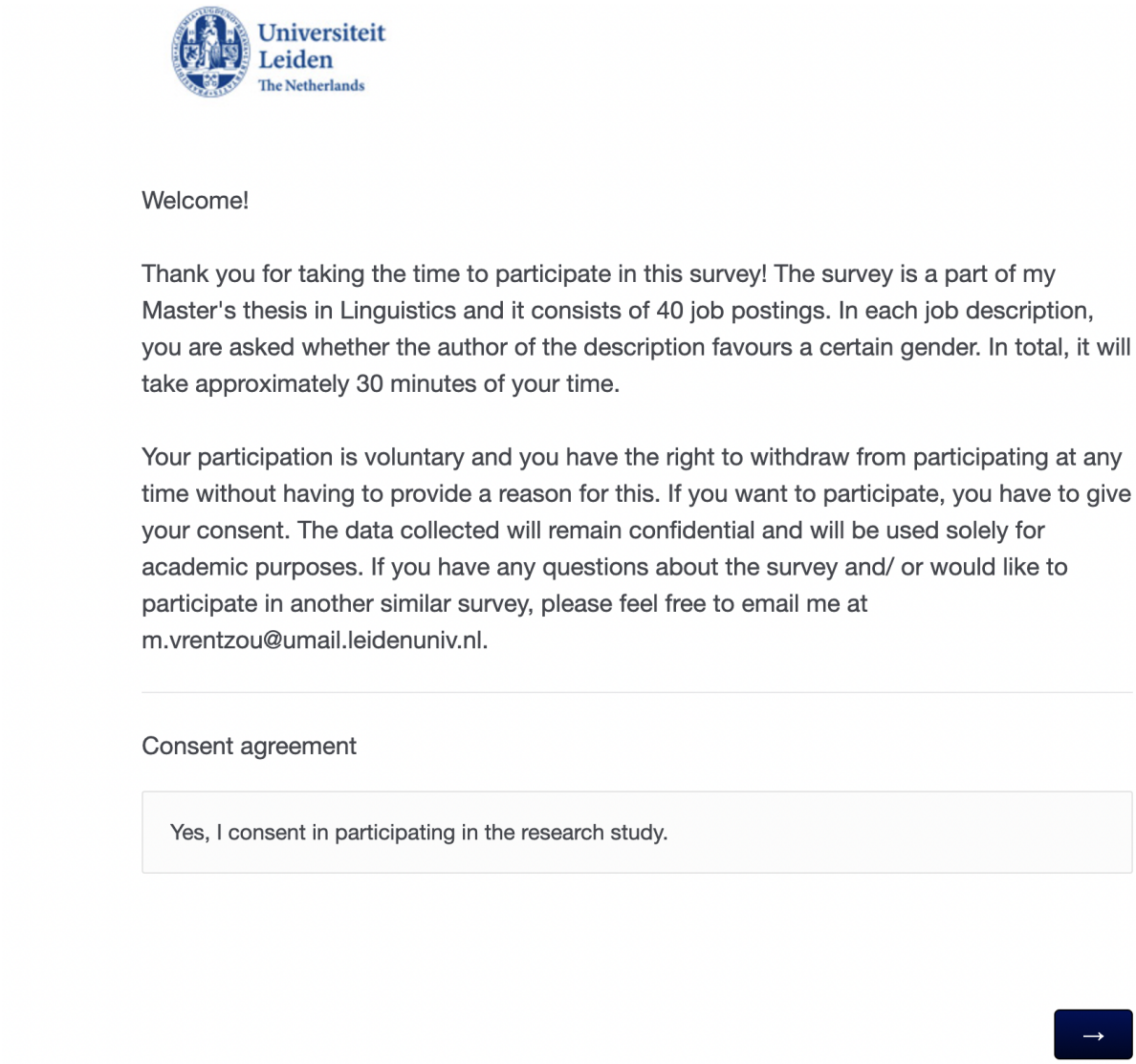
Continued on next page

Table 7 – continued from previous page

	No Preferred Gender
BERT	team, work, skills, ability, experience, role, able, business, working, successful, strong, communication, customer, service, excellent, environment, support, clients, provide, new looking, high, responsible, management, company, level, required, opportunity, experienced position, services, sales, join, candidate, quality, highly, good, key, range, food, previous interpersonal, professional, care, including, technical, manage, learn, providing, people
Women	team, work, skills, role, experience, ability, business, able, working, successful, customer, strong, communication, service, excellent, support, clients, environment, position, new, looking, high, provide, responsible, company, management, reporting, sales, level, candidate, experienced, opportunity, join, required, highly, key, manage, ensure, motivated, manager, managing, providing, leadership, seeking, food, customers, technical, services, previous, development
Men	team, work, experience, skills, role, ability, business, customer, working, able, successful, support, service, strong, excellent, environment, clients, communication, management, new, looking, position, company, responsible, level, high, opportunity, experienced, reporting, join, sales, professional, key, manage, leadership, people, motivated, candidate, organisation, required, highly, previous, manager, lead. food, services, including, seeking, providing, attitude

15 Appendix D

Figure 4: Informed Consent Form



16 Appendix E

This section provides more information on the dataset and its structure. The dataset consists of 5 smaller datasets, each one named "SurveyX_dataresults.tsv" where X denotes the incrementing number of the dataset. Each dataset contains the answers of the participants from each survey. Personal information that may directly or indirectly identify individuals was omitted. Figure 5 provides a partial snapshot of one of the datasets' sheets. The full table is omitted due to its large number of attributes. The code and its accompanying datasets can be found at https://github.com/MarinaVrentzou/thesis

Figure 5: Dataset Snapshot

Column9	Column10	Column11	Column12
Do you think the author of tl	Here you may highlight the words th	Here you may provide extra r	Do you think the author of this j
Not sure	45: communication,46: skills.	-	No, there is no preferred gende
No, there is no preferred ge	48: None	None	No, there is no preferred gende
No, there is no preferred ge	48: None	use only person and people	Gender Diverse
No, there is no preferred ge	48: None	Not preferred gender	No, there is no preferred gende
No, there is no preferred ge	48: None	-	No, there is no preferred gende
Female	8: life,9: at,10: home,11: solutions,1	__	Gender Diverse
No, there is no preferred ge	48: None	none	Not sure
No, there is no preferred ge	3: Salespeople	Salespeople is a word that re	Gender Diverse
No, there is no preferred ge	48: None	None	No, there is no preferred gende
No, there is no preferred ge	48: None	None	No, there is no preferred gende
Not sure	18: best,19: outcome,22: customers.	This ad contains a lot of gene	Not sure
Female	6: understanding,8: life,10: home,11	No	No, there is no preferred gende