

# A Survey on Gender Bias in Natural Language Processing

KAROLINA STAŃCZAK, University of Copenhagen

ISABELLE AUGENSTEIN, University of Copenhagen

Language can be used as a means of reproducing and enforcing harmful stereotypes and biases and has been analysed as such in numerous research. In this paper, we present a survey of 304 papers on gender bias in natural language processing. We analyse definitions of gender and its categories within social sciences and connect them to formal definitions of gender bias in NLP research. We survey lexica and datasets applied in research on gender bias and then compare and contrast approaches to detecting and mitigating gender bias. We find that research on gender bias suffers from four core limitations. 1) Most research treats gender as a binary variable neglecting its fluidity and continuity. 2) Most of the work has been conducted in monolingual setups for English or other high-resource languages. 3) Despite a myriad of papers on gender bias in NLP methods, we find that most of the newly developed algorithms do not test their models for bias and disregard possible ethical considerations of their work. 4) Finally, methodologies developed in this line of research are fundamentally flawed covering very limited definitions of gender bias and lacking evaluation baselines and pipelines. We see overcoming these limitations as a necessary development in future research.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine Learning**; • **Computing methodologies** → **Language resources**.

Additional Key Words and Phrases: gender bias, survey

## ACM Reference Format:

Karolina Stańczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *J. ACM* 1, 1 (December 2021), 35 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Gender bias and sexism are explicitly expressed in language and thus, have been analysed both by the linguistics and natural language processing (NLP) communities [Koolen and van Cranenburgh 2017; Sun et al. 2019]. Since the first publication on gender bias detection in 2004 in the ACL Anthology<sup>1</sup>, which indexes papers published at almost all NLP venues, there have been a total of 224 publications aiming an investigation of gender bias, showing a clear upward trend in the number of papers published every year that has started back in 2015. In particular, previous research has confirmed gender bias to be prevalent in literature [Hoyle et al. 2019], news [Wevers 2019], media [Asr et al. 2021], and communication about and directed towards people of different genders [Fast et al. 2016; Voigt et al. 2018]. Further, prior studies have shown bias in underlying NLP algorithms such as word embeddings [Bolukbasi et al. 2016] and language models [Nadeem et al. 2021], as well as in the downstream tasks they are employed for, e.g., machine translation [Savoldi et al.

<sup>1</sup><https://aclanthology.org/>

Authors' addresses: Karolina Stańczak, ks@di.ku.dk, University of Copenhagen; Isabelle Augenstein, augenstein@di.ku.dk, University of Copenhagen.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2021 Association for Computing Machinery.

0004-5411/2021/12-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

2021], coreference resolution [Rudinger et al. 2018; Webster et al. 2018; Zhao et al. 2018a], language generation [Sheng et al. 2020], and part-of-speech tagging and parsing [Garimella et al. 2019].

However, the rapid increase in research on gender bias has led to a state where the research is fractured across communities and publications often do not engage with parallel research. Thus, there is a need to summarise and critically analyse the developments hitherto, to identify the limitations of prior work and suggest recommendations for future progress. Therefore, in this paper, we present an overview of 304 papers on gender bias in natural language processing. We begin with a brief outline our methodology and explore the evolution of the field in popular NLP venues (§2). Then, we discuss different definitions of gender in society (§3). Further, we define gender bias and sexism in general and in NLP, in particular, incorporating a discussion of their ethical considerations (§4). Next, we gather common lexica and datasets curated for research on gender bias (§5). Subsequently, we discuss formal definitions of gender bias (§6). Then, we discuss methods developed for gender bias detection (§7) and mitigation (§8).

We find that existing research on gender bias has four main limitations and see addressing these limitations as necessary future focus areas of research on gender bias. Firstly, despite the wide range of research across multiple language tasks predominantly only two genders are distinguished, male and female, neglecting the fluidity and continuity of gender as a variable. Natural language has started to adopt gender-neutral linguistic forms to recognise non-binary nature of gender such as singular *they* in English and *hen* in Swedish, thus presenting a need for NLP researchers to incorporate this social development into their datasets and algorithms [Sun et al. 2021]. Otherwise, modelling gender as a binary variable can lead to a number of harms such as misgendering and erasure via invalidation or obscuring of non-binary gender identities [Behm-Morawitz and Mastro 2008; Fast et al. 2016]. Addressing this issue is critical not just to improve the quality of our systems, but more importantly to minimise these harms [Larson 2017].

Secondly, most prior research on gender bias has been monolingual, focusing predominantly on English or a small number of further high-resource languages such as Chinese [Liang et al. 2020] and Spanish [Zhao et al. 2020]. Only limited work has been conducted in a broader multilingual context with notable exceptions of analysis of gender bias in machine translation [Prates et al. 2020] and language models [Stańczak et al. 2021].

Thirdly, despite a plethora of studies showing evidence of presence of systematic gender bias in prolifically applied NLP methods [Bolukbasi et al. 2016; Nadeem et al. 2021; Nangia et al. 2020], researchers are not required to test the models they publish with respect to biases they perpetuate. In particular, still most of the recently published models do not include a study of (gender) bias and ethical considerations alongside their publication [Conneau et al. 2020; Devlin et al. 2019; Raffel et al. 2020; Zhang et al. 2020] with the noteworthy exclusion of GPT-3 [Brown et al. 2020]. In general, these methods are tested for biases only post-hoc when already being deployed in real-life applications potentially posing harm to different social groups [Mitchell et al. 2019].

Lastly, we argue that methodologies within gender bias detection often lack baselines and do not engage with parallel research. We find that similarly to research within societal biases Blodgett et al. [2020], work on gender bias in particular, is fundamentally flawed suffering incoherence in usage of evaluation metrics. Publications consider often limited definitions of bias that address only one of many ways gender bias manifests itself in language.

## 2 METHODOLOGY

The following survey is an overview of all papers identified by the authors on analysing gender bias in NLP, which spans a total of 304 papers. To collect these relevant papers, the ACL Anthology, NeurIPS, and FAccT were queried for all papers with the keywords ‘gender bias’, ‘gender’ or ‘bias’

made available prior to June 2021. Additionally, we expand the spectrum of the papers with relevant social science publications and other relevant publications cited in the collected papers.

We retained all papers about gender bias and discarded papers focusing on other definitions of the keywords (e.g., inductive bias, social bias). We review papers analysing gender bias in natural language and methods presenting an encompassing overview of gender bias in language.

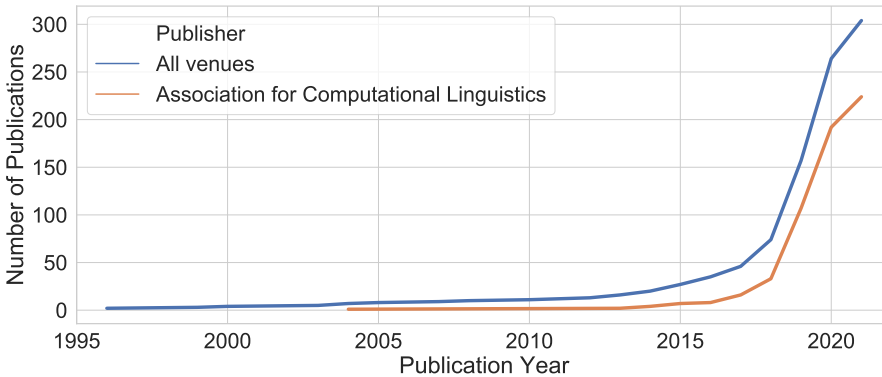


Fig. 1. Cumulative number of papers published on gender bias prior to June 2021.

We analyse the number of published papers in ACL venues mentioning the selected keywords either in the title or the abstract of the paper and present the results in Figure 1. We observe a steady increase in the number of papers since 2015 with notable peaks in 2019 (83 publications) and 2020 (a total of 107 publications). This trend suggests 2021 might end with another record in the number of papers on gender bias per year. Indeed, in 2021, we have already identified a total of 40 papers covering the topic of gender bias in NLP. This development demonstrates that the area of research has established itself within NLP research.

### 3 GENDER IN SOCIETY AND LINGUISTICS

Definitions of gender used in the linguistics literature vary substantially across subfields and are often implicit [Ackerman 2019]. Depending on the context, the concept of gender refers to a person’s self-determined identity and the way they express it, how they are perceived, and others’ social expectations of them [Ackerman 2019; Lucy and Bamman 2021]. Compared to *sex*, a term that solely refers to one’s set of physical and physiological characteristics such as chromosomes, gene expressions, and genitalia, *gender* is considered a social construct [Butler 1989; Risman 2004]. In particular, Risman [2004] argue gender is a social construct and, as such, has consequences on person’s individual development, both in interactions and institutional domains.

However, linguistic categories of gender do not map well to social categories [Cao and Daumé III 2020]. Literature on gender in linguistics often distinguishes the following types of gender that are summarised below. We note that these types are not all-encompassing and merely outline gender categories presented in the literature.

- **Grammatical gender:** refers to a classification of nouns based on a principle of a grammatical agreement into categories. Depending on the language, the number of grammatical gender classes ranges from two (e.g., *masculine* and *feminine* in French, Hindi, and Latvian) to several tens (in Bantu languages and Tuyuca) [Corbett 1991]. Many of these languages also assign grammatical gender to inanimate nouns.

- **Referential gender:** identifies referents as *female*, *male* or *neuter* [Cao and Daumé III 2020]. A very similar concept is described by conceptual gender referred to as a gender that is expressed, inferred and used by a perceiver to classify a referent [Cao and Daumé III 2020].
- **Lexical gender:** refers to an existence of lexical units carrying the property of gender, male- or female-specific words such as *father* and *waitress* [Cao and Daumé III 2020; Fuertes-Olivera 2007].
- **(Bio-)social gender:** refers to the imposition of gender roles or traits based on phenotype, social and cultural norms, gender expression, and identity (such as gender roles) [Ackerman 2019; Kramarae and Treichler 1985].

*Non-binary gender.* Since the grammatical, referential, and lexical gender are definitions widely followed in NLP research, most NLP research that includes gender as a variable in downstream tasks treats it as a categorical variable with binary values (in English) [Brooke 2019]. However, the binarisation of gender in computational studies usually does not agree with critical theorists. For instance, Butler [1989] show how gender is not simply a biological given, nor a valid dichotomy, and even though many people fit into the binary categories, there are more than two genders [Bing, Janet and Bergvall 1998]. Thus, gender can be viewed as a broad spectrum.

More recently, natural language started adopting linguistic forms to recognise the non-binary nature of gender, such as singular *they* in English, *hen* in Swedish and *hän* in Finnish. These linguistic forms are not new concepts and were used by native speakers to refer to someone whose gender is unknown. However, their popularity has increased to denote a person whose gender is non-binary. The increased popularity of gender-neutral linguistic forms in natural language presents a challenge to incorporate this social development into the datasets and algorithms [Sun et al. 2021]. However, some words that are relevant in this discussion such as *cisgender* and *binarism* are either missing or underrepresented in corpora and databases [Hicks et al. 2016].

*Determining gender.* To include gender as a variable in a NLP method, it often needs to be determined from the data first since it is often not explicitly given which is generally difficult to accomplish with high precision.

A popular method to determine gender is to infer it from a person's name, assuming that this information is given. In many languages, gender-differentiated names for men and women make gender assignment possible based on gendered name dictionaries. For instance, in Slavic languages, the ending of the last name is gender-specific (e.g., with *-i* vs. *-a*). On the other hand, gender-neutral first names are common for Chinese, Turkish, and many other languages. Additionally, names often have different gender associations depending on the country and language, such as *Andrea* being a male name in Italian and a female one in English, German and Spanish. Notably, many of the primarily Western-based name lists used for determining gender do not always generalise to names from other countries and cultures [Lucy and Bamman 2021]. Due to these aspects, all of the above methods of determining gender tend to be imprecise and neglect non-binary genders.

## 4 GENDER BIAS, SEXISM AND HARMS THEY MAKE

In the following, we state general definitions of gender bias and sexism and distinguish among their different types. Further, we outline the potential harms they might cause for individuals and society as a whole.

### 4.1 Gender Bias

Blodgett et al. [2020] warn that papers about NLP systems developed for the same task often conceptualise bias differently. Therefore, we state the most common definitions of gender bias in the following. Gender bias is defined as the systematic, unequal treatment based on one's gender

[Sun et al. 2019]. More specifically, Friedman and Nissenbaum [1996] use the term bias to refer to behaviour that systematically discriminates against specific individuals or groups in favor of others and distinguish three bias categories: pre-existing bias, technical bias, and emergent bias. **Pre-existing bias** arises when computer systems incorporate biases that appear independently and often prior to the creation of the system [Friedman and Nissenbaum 1996]. It can originate both from individuals, biased software developers or from society, private or public organizations and institutions, or especially in case of gender bias – historical and cultural context. Thus, this type of bias emerges not only through conscious decisions of individuals or institutions but can also appear unintended. On the other hand, **technical bias** emerges from models' technical design such as hardware and software limitations. While it is almost always possible to identify pre-existing bias and technical bias in a system design at the time of creation or implementation, **emergent bias** arises when the context the system was used for has changed - due to changes in society, population, or cultural values (e.g., when social media feeds are influenced by user's gender).

Further literature outlines reporting and interpretation bias. **Reporting bias** refers to the phenomenon that the frequency with which situations of a certain type are described in text does not necessarily correspond to their relative likelihood in the world, or the subjective frequency captured in human beliefs [Gordon and Van Durme 2013]. On the other hand, **interpretation bias** is a phenomenon of researchers assuming that gender is a relevant variable which ultimately leads to analyses that are incapable of revealing violations of this assumption [Bamman et al. 2014; Koolen and van Cranenburgh 2017]. The results are not questioned, especially if they align with common often stereotypical knowledge [Koolen and van Cranenburgh 2017].

## 4.2 Sexism

Sexism can be defined as discrimination, stereotyping, or prejudice based on one's sex (as opposed to one's gender). According to the ambivalent sexism theory [Glick and Fiske 1996], sexism can be divided as:

- **Hostile:** follows the classic definition of prejudice - an explicitly negative sentiment that is sexist.
- **Benevolent:** subjectively positive attitude, which is sexist. Despite the seemingly positive sentiment, benevolent sexism has been shown to affect women's cognitive performance stronger than hostile sexism [Dardenne et al. 2007]. For instance, female gender associations with any word, even a subjectively positive word such as *attractive*, can cause discrimination against women if it reduces their association with other words, such as *professional*. Despite the positive sentiment of benevolent sexism, it can be backtracked to masculine dominance and stereotyping.

We note that sexism is considered a subset of hate speech [Waseem and Hovy 2016] and therefore is often analysed together with other forms of aggression [Safi Samghabadi et al. 2020].

## 4.3 Harms

Gender bias and sexism result in harms affecting individuals and society as a whole. Recently, Crawford [2017] present a framework classifying algorithmic biases by the type of harm they cause and distinguish between allocational and representational harms.

**Representational harms** refer to portrayals of certain groups that are discriminatory. In general, following Crawford [2017] representational harms can be divided into: stereotyping, under-representation, denigration, recognition, and ex-nomination. Stereotyping, in particular, perpetuates common (often negative) depictions of a certain gender. Under-representation bias is the disproportionately low representation of a specific group. Denigration refers to the use of culturally or

historically derogatory terms, while recognition bias involves a given algorithm's inaccuracy in recognition tasks. Finally, ex-nomination describes a practice where a specific category or way of being is framed as the norm by not giving it a name or not specifying it as a category in itself (e.g., 'politician' vs. 'female politician'). On the other hand, **allocational harms** refer to the unjust distribution of opportunities and resources due to algorithmic intervention. They can result in systematic differences in treatment or denial of a particular service and complete ruling out of certain groups, for instance in job applications. Allocation bias can be framed as an economic issue in which a system unfairly allocates resources to certain groups over others, while representation bias occurs when systems detract from the social identity and representation of certain groups [Crawford 2017; Sun et al. 2019].

Another harmful outcome of gender bias and sexism presents itself in **gender gaps** that arise from these asymmetrical valuations, e.g., where men are typically over-represented and have higher salaries compared to women [Mitra 2003]. The public sphere is often associated with male and agents characteristics (assertiveness, competitiveness) in domains like politics and entrepreneurship. Private or domestic domains linked to family and social relationships are traditionally related to women, although social relationships are considered more important by people independent of gender [Friedman et al. 2019].

#### 4.4 Bias in NLP

Above we have introduced gender bias and sexism as general terms. In the following, we discuss how these biases emerge in natural language and ultimately influence many downstream tasks.

Language can be used as a substantial means of expressing gender bias. Gender biases are translated from source data to existing algorithms that may reflect and amplify existing cultural prejudices and inequalities by replicating human behavior and perpetuating bias [Sweeney 2013]. This phenomenon is not unique to NLP, but the lure of making general claims with big data, coupled with NLP's semblance of objectivity, makes it a particularly pressing topic for the discipline [Koolen and van Cranenburgh 2017].

Alongside the types of biases described above, there are forms of bias that apply specifically in NLP research. In particular, Hitti et al. [2019] define gender bias in a text as the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender. Further, Hitti et al. [2019] note that gender bias can manifest itself structurally, contextually or in both of these forms. **Structural bias** arises when the construction of sentences shows patterns that are closely tied to the presence of gender bias. It encompasses gender generalisation (i.e., when a gender-neutral term is assumed to refer to a specific gender-based on some (stereotypical) assumptions) and explicit labeling of sex. On the other hand, **contextual bias** manifests itself in a tone, the words used, or the context of a sentence. Unlike structural bias, this type of bias cannot be observed through grammatical structure but requires contextual background information and human perception. Contextual bias can be divided into societal stereotypes (which showcase traditional gender roles that reflect social norms) and behavioral stereotypes (attributes and traits used to describe a specific person or gender). Therefore, gender bias can be detected using both linguistic and extra-linguistic cues, and can manifest itself with different intensities, which can be subtle or explicit, posing a challenge in this line of research.

Gender bias is known to perpetuate to models and downstream tasks posing harm for the end-users [Bolukbasi et al. 2016]. These harms can emerge as representational and allocational harms and gender gaps. **Allocation harm** is reflected when models often perform better on data associated with the majority gender. In the context of NLP, this is often the case for machine translation [Sap et al. 2017] and coreference resolution [Webster et al. 2018] (see §7.3). **Representation harm** is reflected when associations between gender with certain concepts are captured in word embeddings

and model parameters [Sun et al. 2019], for instance, as shown in [Bolukbasi et al. 2016; Zhao et al. 2018b]. On the other hand, **gender gap** is a phenomenon influencing gender bias in the text. Since women are underrepresented in most areas of society, it is not surprising that available texts mainly discuss and quote men [Asr et al. 2021], which leads, for example, to biased corpora researchers train their models on.

## 5 RESOURCES

Comprehensive data resources are crucial in probing for gender bias in language. However, many of the datasets in NLP are inadequate for measuring gender bias since they are often severely gender imbalanced with a substantial under-representation of female and non-binary instances. Further, analysing gender bias often requires a dataset of a specific structure or including certain information to enable proper isolation of the effect of gender [Sun et al. 2019]. Thus, evaluation on widely-used datasets (e.g., SNLI [Rudinger et al. 2017]) might not reveal gender bias due to inherent biases encoded in the data, presenting a need in research for targeted datasets for gender bias detection.

We note that the choice of a dataset is dependent on the considered definition of bias (discussed in §4) that needs to be targeted specifically, the NLP task at hand, domain, etc. Here, we describe the most popular publicly available lexica (§5.1) and datasets (§5.2) that have been used to analyse gender bias in NLP with respect to the above-mentioned aspects.

### 5.1 Gender lexica

Lexicon matching is an interpretable and technically simple approach, and thus, it has been frequently adopted by NLP practitioners. In particular, in gender bias detection, lexica representing genderness, sentiment, and the affect dimensions of valence, arousal, and dominance have been widely employed since these measures are often used as proxies for bias. In Table 1, we present the most popular lexica used for gender bias detection, and in the following, we describe measures they quantify.

Lexicon	No. of words	Measure
Gender Ladeness Lexicon [Ramakrishna et al. 2015]	10 000	Genderness
Gender Predictive Lexicon [Sap et al. 2014]	7 136	Genderness
Gender Ladeness Lexicon [Clark and Paivio 2004]	925	Genderness
Williams and Best [Williams and Best 1990]	300	Genderness
NRC VAD Lexicon [Mohammad 2018]	20 000	Valence, Arousal and Dominance
Valence, Arousal, and Dominance [Warriner et al. 2013]	13 915	Valence and Dominance
NRC Emotion Lexicon [Mohammad and Turney 2013]	10 170	Emotion and Sentiment
Connotation Frames [Sap et al. 2017]	2 155	Power and Agency

Table 1. List of popular lexica used in gender bias research.

**5.1.1 Sentiment.** Differences in sentiment towards people of different genders have been analysed in the context of gender bias in numerous papers [Cho et al. 2019; Hoyle et al. 2019; Stańczak et al. 2021; Touileb et al. 2020], which have exploited sentiment lexica for this purpose. Since creating a comprehensive overview of sentiment lexica is outside the scope of this paper, we refer the reader to Taboada et al. [2011] for such an overview. However, we note that sentiment is indicative solely of hostile biases rather than more nuanced ones.

**5.1.2 Gender Ladeness.** Gender ladeness is a measure to quantitatively represent a normative rating of the perceived feminine or masculine association of a word [Paivio et al. 1968]. In particular, this metric indicates the gender specificity of individual words, with extreme values assigned to

highly stereotypical concepts. For instance, in Ramakrishna et al. [2015]’s lexicon, which is based on movie scripts, the word *bride* would be assigned the gender ladenness value of 0.84 on a scale from -1 (most masculine) to 1 (most feminine). Similarly, Williams and Best [1990] use a list of pre-selected adjectives, Sap et al. [2014] use words collected on social media, and Clark and Paivio [2004] select a list of nouns to create a genderness lexicon.

**5.1.3 Valence, Arousal, and Dominance.** Based on social psychology, NLP research has identified three primary affect dimensions: power/dominance (strength/weakness), valence (goodness/badness), and agency/arousal (activeness/passiveness of an identity) [Field and Tsvetkov 2019]. Since a common stereotype associates female gender with weakness, passiveness, and submissiveness, lexica reporting measures for these dimensions are a valuable resource in gender bias analysis, and going beyond sentiment, they can be applied in unveiling benevolent biases.

**5.1.4 Limitations.** By their nature, lexicon approaches are limited to known words [Field et al. 2019], and they assume that the context of the words remains constant [Lucy et al. 2020]. However, collecting exhaustive lexica can be very resource-consuming since they rely on human-generated annotations [Lucy et al. 2020]. Moreover, we note that all the lexica listed in Table 1 are created solely for English. There has been very little research enabling multi-lingual gender bias analysis employing lexica, with the notable exception of Stańczak et al. [2021].

## 5.2 Datasets

Dataset	Size	Data	Gender	Task	Bias
EEC [Kiritchenko and Mohammad 2018]	8 640 sent.	sent. templates	b	SA	stereotyping
WinoBias [Zhao et al. 2018a]	3 160 sent.	sent. templates	nb	cor. res.	occ. bias
WinoGender [Rudinger et al. 2018]	720 sent.	sent. templates	b	cor. res.	occ. bias
WinoMT [Stanovsky et al. 2019]	3 888 sent.	sent. templates	b	MT	occ. bias
Occupations Test [Escudé Font and Costa-jussà 2019]	2 000 sent.	sent. templates	b	MT	occ. bias
GAP [Webster et al. 2018]	8 908 ex.	Wikipedia	b	cor. res.	stereotyping
KNOWREF	8 724 sent.	Wikipedia & other	b	cor. res.	stereotyping
BiosBias [De-Arteaga et al. 2019]	397 340 bios	CommonCrawl	b	classification	occ. bias
GeBioCorpus	2 000 sent.	Wikipedia	b	MT	occ. bias
StereoSet [Nadeem et al. 2021]	2 022 sent.	human-generated	b	probing LMs	stereotyping
CrowS-Pairs	1508 ex.	human-generated	b	probing LMs	stereotyping

Table 2. List of common probing datasets for gender bias in language.

In order to measure gender bias in NLP methods and downstream applications, a number of datasets have been developed. We list the well-established datasets in Table 2 together with the tasks they can probe and biases they provide a testbed for. Below we discussed three groups of datasets: those based on simple template structures, those based on natural language data, and datasets that have been developed to detect gender bias in language models.

**5.2.1 Template-Based Datasets.** A number of studies accounting for gender bias in natural language processing have been conducted on benchmark datasets consisting of template sentences of simple structures such as “*He/She is a/an [occupation/adjective].*” where *[person/adjective]* is populated with occupations or positive/negative descriptors [Bhaskaran and Bhallamudi 2019; Cho et al. 2019; Prates et al. 2020; Saunders and Byrne 2020]. Similarly, the EEC dataset Kiritchenko and Mohammad [2018] includes sentence templates such as *[Person] feels [emotional state word].* and *The [person] has two children.* The EEC dataset has been widely used in other projects [Bhardwaj et al. 2021] and has been extended with German sentences by Bartl et al. [2020]. Another multilingual dataset has been proposed by Nozza et al. [2021] that create a template-based dataset in 6 languages (English, Italian, French, Portuguese, Romanian, and Spanish) similarly consisting of a subject and a predicate.



Another strain of work has utilised the structure of Winograd Schemas [Levesque et al. 2012]: WinoBias [Zhao et al. 2018a], WinoGender [Rudinger et al. 2018], and WinoMT [Stanovsky et al. 2019]. Since Winograd Schema Challenge is a coreference resolution task with human-generated sentence templates which requires reasoning with commonsense knowledge, it has been employed to analyse if reasoning of coreference system is dependent on a gender of a pronoun in a sentence and to measure stereotypical and non-stereotypical gender associations for different occupations.

WinoBias [Zhao et al. 2018a] contains two types of sentences that require the linking of gendered pronouns to either male or female stereotypical occupations. None of the examples can be disambiguated by the gender of the pronoun, but this cue can potentially distract the model. The WinoBias sentences have been constructed so that, in the absence of stereotypes, there is no objective way to choose between different gender pronouns. In parallel, Rudinger et al. [2018] develop a WinoGender dataset [Levesque et al. 2012]. As in the WinoBias dataset, each sentence contains three variables: *occupation*, *person* and *pronoun*. For each occupation, WinoGender includes two similar sentence templates: one in which *pronoun* is coreferent with *occupation*, and one coreferent with *person*. Notably, WinoGender sentences unlike WinoBias also include gender-neutral pronouns. Finally, sentences in WinoGender are not resolvable from syntax alone, unlike in the WinoBias dataset, which might enable better isolation of the effect of gender bias. Both of these datasets have been employed in a number of analysis on gender bias in coreference resolution [de Vassimon Manela et al. 2021; Jin et al. 2021; Tan and Celis 2019; Vig et al. 2020].

Building on WinoGender and WinoBias, Stanovsky et al. [2019] curate WinoMT, a probing dataset for machine translation, with sentences with stereotypical and non-stereotypical gender-role assignments. WinoMT has become widely applied as a challenge dataset for gender bias detection in MT systems [Basta et al. 2020; Renduchintala et al. 2021; Saunders and Byrne 2020; Stefanovičs et al. 2020] with Saunders et al. [2020] developing a version of the WinoMT dataset with binary templates filled with singular *they* pronoun. Similarly, the Occupations Test dataset [Escudé Font and Costa-jussà 2019] contains template sentences to test MT systems on. Ultimately, both Occupations Test and WinoMT test if the grammatical gender of the translation is aligned with the gender of the pronoun in the original sentence which limits the aspects of gender bias they can probe for.

**5.2.2 Natural Language Based Datasets.** Probing datasets utilise also available natural language resources and extend them with annotations to tune it for the gender bias detection task. Importantly, these datasets can be applied to analyse gender bias in natural language and in algorithms, and are not limited by artificial structures of the template-based approaches to collecting data.

A number of popular datasets rely on data collected from Wikipedia. For instance, GAP [Webster et al. 2018] is a human-labeled corpus derived from Wikipedia including sentences relevant for coreference resolution task. Unlike WinoGender and WinoBias, GAP focuses on relations where the antecedent is a named entity instead of pronouns [Webster et al. 2018] and thus, can be used to unravel biases towards entities. Similarly, to analyse gender bias in coreference resolution, Emami et al. [2019] develop the KNOWREF dataset, which is scraped from Wikipedia together with OpenSubtitles, and Reddit comments. Then, after initial filtering they infer the genders of antecedents based on their first names and ask human annotators to predict which antecedent was the correct coreferent of the pronoun. Due a relatively large size of these datasets, both GAP and KNOWREF can be used as an alternative to sentence template based datasets.

Another line of work is analysing gender bias in biographies. [De-Arteaga et al. 2019] develop the BiosBias dataset, which consists of biographies with labelled occupations and gender identified within Common Crawl. The dataset has been created for the task of correctly classifying the subject's occupation from their biography assuming that there are differences between mens'

and womens' online biographies other than gender indicators De-Arteaga et al. [2019]. Further, GeBioCorpus [Costa-jussà et al. 2020] present a dataset with biography and gender information from Wikipedia which has been widely used to analyse gender bias in MT (for English, Spanish, and Catalan) [Basta et al. 2020; Escudé Font and Costa-jussà 2019; Vanmassenhove et al. 2018].

Datasets employ also other online data sources. For instance, RtGender [Voigt et al. 2018] is a dataset of online communication to enable research in communication directed to people of a specific gender. Studies on detecting misogynist or toxic language on social media released Twitter-based datasets [Anzovino et al. 2018; Hewitt et al. 2016]. Bentivogli et al. [2020] develop MuST-SHE, a multilingual benchmark based on TED data for gender bias detection in machine and speech translation. Recently, Marjanovic et al. [2021] create a dataset with Reddit comments to study gender biases that appear in online political discussion.

*5.2.3 Probing Language Models.* A significant, though relatively recent and thus undiscovered, research direction has concentrated on analysing gender bias in language models. To this end, specific datasets have been curated. In particular, Nadeem et al. [2021] present StereoSet, which is a dataset to measure stereotypical biases in gender, among other domains. It consists of triplets of sentences with each instance corresponding to a stereotypical, anti-stereotypical or a meaningless association. This dataset enables ranking language models based on probabilities they assign to each of these triplets. In parallel, Nangia et al. [2020] introduce CrowS-Pairs, a crowdsourced, template-based challenge set for measuring social biases, including gender bias, that are present in current language models. In CrowS-Pairs, each example consists of a pair of sentences, a stereotypical and anti-stereotypical. Both of these datasets are a significant starting point for creating a benchmark for evaluating gender bias in language models. Notably, Stańczak et al. [2021] propose a method for generating multilingual datasets for analysing gender bias towards named entities in LMs.

### 5.3 Summary

Above we have discussed popular datasets employed for analysing gender bias. We note that datasets based on simple template structures allow for a controlled experiment environment. However, we warn that the limitations they impose might include artificial biases, and the results of models tested on them may not map to a more natural environment. Since the above datasets provide means of conducting diagnostic tests for gender bias, they have a high positive and low negative predictive value for the presence of gender bias [Rudinger et al. 2018]. Therefore, using these datasets, it is only possible to demonstrate the presence of gender bias in a system but not to prove its absence. Although datasets based on natural language obviate the downsides of the benchmark datasets with simple patterns, they often concentrate on data from one domain, e.g., social media, Wikipedia, or news. Therefore, the results might not generalise well to other domains and should be treated with caution. We note that natural language data might encode gender bias itself so that it is impossible to isolate bias from the data and the tested model. For instance, Chaloner and Maldonado [2019] find evidence of bias in word embeddings trained on the GAP dataset when testing on a standard bias benchmark. They assume that this is due to gender bias on Wikipedia, GAP's underlying data.

However, irrespectively if based on natural language or sentence templates, most of these lexica and datasets are only available for English. Only datasets to analyse gender bias in machine translation, due to the nature of the task, are available in other languages. However, they often consider high-resource languages such as Spanish or German. Similarly, most of these datasets restrict themselves to the binary view on gender presenting a major gap in the research. Thus, we encourage data collection for gender inclusive task-specific datasets. Further, many of the popular publications have focused solely on occupational biases without accounting for a nuanced nature of gender bias. Finally, despite a number of datasets curated specifically to assess for gender

bias, only a few can be considered as benchmarks for a targeted downstream task and they come predominantly from the machine translation and coreference resolution domain. Therefore, we strongly encourage further research along the lines of establishing evaluation benchmarks for the underlying models such as Nadeem et al. [2021]; Nangia et al. [2020].

## 6 DEFINING BIAS

In the following, we list the common formal definitions of bias that are utilised to quantify the social concepts presented in Section 4 and divide them into definitions used for detecting gender bias in language (§6.1), either natural or generated, and in NLP methods (§6.2).

### 6.1 Measuring Gender Bias in Language

Gender bias manifests itself in texts in many ways and can be identified using both linguistic and extra-linguistic cues [Marjanovic et al. 2021]. Already structure of the data, e.g., the distribution of genders mentioned in the text, can be a bias indicator and the differences in these distributions can be used as a measure for bias. However, in the following, we focus on more complex textual biases, *i.e.*, lexical biases, and discuss measures for quantifying differences in portrayals of genders, and their stereotypical depictions.

*6.1.1 Differences in Gender Descriptions.* Differences in depictions of men and women have been prolifically quantified using point-wise mutual information (PMI) [Hoyle et al. 2019; Rudinger et al. 2017; Stańczak et al. 2021]. In particular, PMI investigates the co-occurrence of words with a particular gender. In PMI descriptors (such as adjectives or verbs) linked to a gendered entity are counted and the probability of their co-occurrence to a gender across entity is calculated. More formally, PMI is defined as:

$$PMI(\text{gender}, \mathbf{word}) = \ln \left( \frac{P(\text{gender}, \mathbf{word})}{P(\text{gender})P(\mathbf{word})} \right) \quad (1)$$

In general, words with high PMI values for one gender are suggested to have a high gender bias. However, Rudinger et al. [2017] note that bias at the level of word co-occurrences is likely to lead to overgeneralisation when applied to a heterogenous dataset. Notably, PMI can also be used to measure differences in word choice for genders beyond the binary [Stańczak et al. 2021].

Further, Hoyle et al. [2019] extend the PMI approach and propose an unsupervised model that jointly represents descriptors with their sentiment to investigate gender bias in words used to describe men and women together with word's sentiment.

*6.1.2 Stereotypical and Occupational Bias.* Occupational gender segregation and stereotyping is a major problem in the labor market often caused by gender roles and stereotypes present in society and as such has been in focus in a numerous research [Lu et al. 2020]. To this end, Qian [2019] calculate an overall stereotype score of a text as the sum of stereotype scores of all the by definition gender-neutral words with gendered words in the text, divided by the total count of words calculated. Then, Qian [2019] define the gender stereotype score of a word:

$$\text{bias}(\mathbf{word}) = \left| \log \frac{c(\mathbf{word}, m)}{c(\mathbf{word}, f)} \right|$$

where  $f$  is a set of female words (e.g., she, girl, and woman), and  $c(\mathbf{word}, g)$  is the number of times a gender-neutral  $\mathbf{word}$  co-occurs with gendered words. A word is used in a neutral way, if the stereotype score is 0, which means it occurs equally frequently with male words and females word in the text. Qian [2019] assess occupation stereotypes score in a text as the average stereotype score

of a list of gender-neutral occupations in the text. These definitions of stereotypical and occupational bias have been employed in subsequent research [Bordia and Bowman 2019; Qian et al. 2019].

## 6.2 Measuring Gender Bias in Methods

With the prevalence of NLP systems and their increasing application areas, researchers have developed measures to probe for gender biases encoded in these methods. In the following, we discuss different definitions used for bias detection in NLP methods.

**6.2.1 Bias influencing Performance.** For downstream tasks where there exists a gold gender, researchers have utilised performance-based measures to quantify bias. In particular, these measures are relevant for applications such as machine translation and coreference resolution where the objective involves the correct handling of gendered (pro-)nouns.

Then, the amount of bias encoded in NLP systems can be quantified using: accuracy (percentage of observations with the correctly gendered entity) [Saunders and Byrne 2020]; difference in accuracy between the set of sentences with anti-stereotypical and stereotypical sentences;  $F_1$  score and difference in  $F_1$  score between the stereotypical and anti-stereotypical gender role assignments [de Vassimon Manela et al. 2021; Webster et al. 2018; Zhao et al. 2018a]; log-loss of the probability estimates [Webster et al. 2019]; false positive rates [Jin et al. 2021; Kennedy et al. 2020]; ratio of observations with masculine and feminine predictions; gender differences in distributions of and within occupations Kirk et al. [2021].

Depending on the downstream task, task-specific performance measures are used to evaluate gender bias. For instance, to assess gender bias in dependency parsing, the labeled attachment score that measures the percentage of tokens that have a correct assignment and the correct dependency relation has been applied [Garimella et al. 2019]. Next, BLEU is used in machine translation to assess the quality of the translated text [Saunders and Byrne 2020]. If the MT system is gender biased, the system produces an incorrect gender prediction even when no ambiguity exists [Costa-jussà and de Jorge 2020]. Thus, the lower the bias, the better the translation quality in terms of BLEU score and accuracy [Basta et al. 2020; Escudé Font and Costa-jussà 2019; Stanovsky et al. 2019]. However, Bentivogli et al. [2020] point out that previously obtained BLEU gains [Moryossef et al. 2019; Vanmassenhove et al. 2018] cannot be ascribed with certainty to a better control of gender features and following previous research [Elaraby et al. 2018; Vanmassenhove et al. 2018] underlie the importance of applying gender-swapping in BLEU-based evaluations focused on gender translation.

**6.2.2 Stereotypical Bias.** Another stream of research attempts to quantify gender bias in terms of stereotypical associations that a method conveys. For instance, Zhao et al. [2018a] consider a system gender biased if it links pronouns to occupations more accurately for the stereotypical pronoun, rather than the anti-stereotypical one. Next, in order to assess stereotypical associations encoded in NLP methods, Kurita et al. [2019] suggest to measure how much more a model prefers the male association with a certain attribute, e.g., a programmer, compared to the female gender. To this end, Kurita et al. [2019] propose to create template sentences, similar to the ones discussed in §5.2.1, and calculate a log probability bias score for BERT predictions when filling in a template with the gendered words and the target word. This measure has been widely applied in numerous research [Bartl et al. 2020; Vig et al. 2020]. Building up on this approach, Munro and Morrison [2020] calculate the ratio of the actual probabilities instead of log probabilities, claiming that ratios allow for more transparent comparisons.

For datasets where each instance contains at least two versions of the same template sentence, e.g., male and female, the paired t-test has been used to measure if the mean predicted class probabilities are different across genders [Bhaskaran and Bhallamudi 2019; Kiritchenko and Mohammad 2018]. Similarly, Nangia et al. [2020] propose a metric that calculates the percentage of examples for which

the language model is in favor of the more stereotyping sentence. To measure this, Nangia et al. [2020] first break each sentence in an example into two parts: the modified tokens that appear only in one of the sentences and the unmodified part that is shared. Then, using pseudo-log-likelihood masked language model scoring [Salazar et al. 2020], they estimate the probability of the unmodified tokens conditioned on the ones.

Due to their simplicity and interpretability the above measures have been widely adopted to measure gender bias. However, these methods cover only stereotypical bias neglecting many other ways in which gender bias can be expressed.

**6.2.3 Causal Bias.** Causal testing presents another way of measuring gender bias in NLP systems. Then, gender bias is defined as the disparity in the output when model is feeded with different genders [Qian et al. 2019]. Lu et al. [2020] define bias as the expected difference in scores assigned to expected absolute bias across different genders. Later, Qian et al. [2019] limit the above bias evaluation to a set of gender-neutral occupations and measure how the probabilities of occupation words depend on the gendered word and in reverse, how the probabilities of gendered words depend on the occupation words. Similarly, Emami et al. [2019] propose consistency as a bias metric, where they duplicate the dataset by switching the candidate antecedents each time they appear in a sentence. If a coreference model relies on knowledge and contextual understanding, its prediction should differ between the two versions. Emami et al. [2019] define the consistency score as the percentage of predictions that change from the original instances to the switched instances.

Causal testing in gender bias detection has been used to define bias in terms of stereotypical bias, rather than approaching other possible harms, which sets a possible ground for future work.

**6.2.4 Male Default.** Gender bias can be defined as the deviation of the distribution of gender pronouns in an output of an NLP system from a gender distribution of demographics of an occupation [Prates et al. 2020]. These differences occur more often in a presence of the male default phenomenon (§4). Especially in machine translation systems, male defaults lead to overestimating the distribution of male instances over female ones.

To account for male default in MT, Cho et al. [2019] propose a translation gender bias index (TGBI) and apply it to Korean-English translations. Let  $p_i^f$  be the portion of a sentence translated to a female pronoun,  $p_i^m$  as male and  $p_i^n$  as gender-neutral pronouns in any set of sentences  $S_i \in S$ .

$$TGBI = \frac{1}{n} \sum_{i=1}^n \sqrt{p_i^f p_i^m + p_i^n}$$

where  $p_i^f + p_i^m + p_i^n = 1$  and  $p_i^f, p_i^m, p_i^n \in [0, 1]$  for each  $i$ . TGBI is equal to 1 in optimum when all the predictions incorporate gender-neutral terms. Cho et al. [2019] expect TGBI to be a representative measure for inter-system comparison, especially if the gap between the systems is noticeable. Recently, Ramesh et al. [2021] extend TGBI to Hindi. In general, this is a suitable method for applications where male default is the predominant risk.

**6.2.5 Bias in Word Embeddings.** In recent years, a myriad of publications have approached quantifying bias in word embeddings. In the following, we present the according to our judgement most influential research in this field.

*Projection-Based Measures.* In the initial work on gender bias in word embeddings, Bolukbasi et al. [2016] distinguish between two types of bias, direct and indirect. Following Bolukbasi et al. [2016] direct bias of a word embedding  $\vec{w}$  can be quantified as:

$$DirectBias_c = \frac{1}{|N|} = \sum_{w \in N} |\cos(\vec{w}, g)|^c$$

where  $N$  is a set of gender neutral words,  $g$  is the gender direction and  $c$  is a parameter determining how strict bias is defined. The direct bias manifests itself in relative similarities between gendered and gender-neutral words. However, since gender bias could also affect the relative geometry between gender neutral words themselves, Bolukbasi et al. [2016] introduce notion of indirect gender bias which manifests as associations between gender neutral words that are arising from gender. In particular, if word such as *businessman* and *genius* are closer to *football*, a word with an embedding closer in the gender subspace to a man, it can indicate indirect gender bias. However, Gonen and Goldberg [2019] argue that the indirect bias has been disregarded to some extent and complain that mitigation methods are not provided.

Another researched distance-based metric to measure gender bias in word embeddings uses the relative norm distance between two groups [Garg et al. 2018]:

$$d = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2$$

where  $M$  is the set of neutral word vectors and  $v_i$  is the average vector for group  $i$ . The more positive (negative) that the relative norm distance is, the more associated the neutral words are towards group two (one). Thus, the above metric captures the relative distance (*i.e.*, relative strength of association) between the group words and the neutral word list of interest. Similarly, Friedman et al. [2019] compute bias as the average axis projection of a neutral word set onto the male-female axis and evaluate it for any region's word embedding computing its correlation to gender gaps.

Since the above definitions are straightforward and geometrically grounded, they have been often employed to quantify gender bias in word embeddings. However, bias is much more profound and systematic than the projection of words [Gonen and Goldberg 2019].

*Word Embedding Association Test (WEAT)*. The WEAT has been developed as a benchmark for testing gender bias in word embeddings via semantic similarities. In particular, the WEAT compares set of target concepts (e.g., male and female words) denoted as  $X$  and  $Y$  (each of equal size  $N$ ), with a set of attributes to measure bias over social attributes and roles (e.g., career/family words) denoted as  $A$  and  $B$ . The resulting test statistics is defined as a permutation test over  $X$  and  $Y$ :

$$S(X, Y, A, B) = [\text{mean}_{x \in X} \text{sim}(x, A, B) - \text{mean}_{y \in Y} \text{sim}(y, A, B)]$$

where *sim* is the cosine similarity. The resulting effect size is then the measure of association:

$$d = \frac{S(X, Y, A, B)}{\text{std}_{t \in X \cup Y} S(t, A, B)}$$

The null hypothesis suggests there is no difference between  $X$  and  $Y$  in terms of their relative similarity to  $A$  and  $B$ . In Caliskan et al. [2017], the null hypothesis is tested through a permutation test, *i.e.*, the probability that there is no difference between  $X$  and  $Y$  (in relation to  $A$  and  $B$ ) and therefore, that the word category is not biased. However, we note that results obtained with WEAT should be treated with a grain of salt since Ethayarajh et al. [2019] prove that WEAT systematically overestimates bias.

*Sentence Embedding Association Test (SEAT)*. Based on the WEAT, May et al. [2019] develop an analogous method, SEAT, that compares sets of sentences, rather than words. In particular, May et al. [2019] apply WEAT to the sentence representation. Thus, WEAT can be seen as a special case of SEAT in which the sentence is a single word. To extend a word-level test to sentence contexts, May et al. [2019] slot each word into each of several semantically bleached sentence templates.

*Bias Amplification.* Previous research has shown that NLP models are able not only to perpetuate biases extant in language, but also to amplify them [Zhao et al. 2017]. In particular, Zhao et al. [2017] interpret gender bias as correlations that are potentially amplified by the model and define gender bias towards a *man* for each word as:

$$b(\text{word}, \text{man}) = \frac{c(\text{word}, \text{man})}{c(\text{word}, \text{man}) + c(\text{word}, \text{woman})} \quad (2)$$

where  $c(\text{word}, \text{man})$  is the number of occurrences of a word and male gender in a corpus. If  $b(\text{word}, \text{man}) > 1/|G|$  ( $G = \{\text{man}, \text{woman}\}$  under gender binarity assumption), then a word is positively correlated with gender and may exhibit bias. To evaluate the degree of bias amplification, Zhao et al. [2017] propose to compare bias scores on the training set,  $b^*(\text{word}, \text{man})$ , with bias scores on an unlabeled evaluation set. We note that this method is applicable solely to individual words and would require an extension to be used as a general evaluation metric.

*6.2.6 Qualitative Assessment.* Alongside the above discussed quantitative gender bias measures, some research includes qualitative measures to analyse the extent of gender bias. For instance, Moryossef et al. [2019] conduct a syntactic analysis of generated translations examining inflection statistics for sentence templates from the dataset. Escudé Font and Costa-jussà [2019] introduce clustering as a measure of gender bias. Then, the higher the clustering accuracy for stereotypically-gendered words, the more bias the word embeddings trained on the dataset have. We find this line of work particularly interesting as it encourages better model understanding and interpretability.

### 6.3 Summary

Gender bias can be expressed in language in many nuanced ways which poses stating a comprehensive definition as one of the main challenges in this research field. In this section, we have examined different gender bias definitions. We find that they vary dramatically across and within algorithms and tasks, which supports findings made by Blodgett et al. [2020] that analyse bias definitions in general. Bias is often described only implicitly without any formal definition. Even when a paper states a formal definition, it essentially covers only one type of bias which oversimplifies the task and thus, makes it impossible to detect all harmful signals in language. In particular, we discuss a number of methods to quantify bias in word embeddings which are utilised in many downstream tasks. However, most of them consider only one way of defining bias and do not engage enough parallel research to combine these methods. We here support [Silva et al. 2021]’s claim that solely using one bias metric or test is not enough – diversifying metrics to ensure robustness of the evaluations is thus important. Additionally, we strongly encourage developing standard evaluation measures and tests to enhance comparability.

Another limitation we see is that defining bias in terms of decreasing performance, however straight-forward, carries a risk of capturing bias only as long as it influences the performance. This way bias detection is only a means of enhancing model’s performance instead of being a goal on its own which can raise ethical considerations. Moreover, some of the performance measures have been previously criticised as evaluation benchmarks for tasks they address. For instance, it is widely acknowledged in machine translation that BLEU score is a coarse and indirect indicator of a machine translation system’s performance [Callison-Burch et al. 2006].

Finally, similarly to our observations regarding datasets, most of the measures developed for quantifying gender bias are created and calculated only for binary genders. Even if a specific metric allows for analysing non-binary genders, it usually remains unmentioned.

## 7 DETECTING GENDER BIAS

Armed with datasets (§5) suitable for gender bias analysis and formal gender bias definitions (§6), we focus herein on research on detecting and analysing the nature of gender bias in natural language, NLP algorithms, and downstream tasks. We discuss its challenges, and influential lines of work.

### 7.1 Detecting Gender Bias in Natural Language

Natural language is known to exhibit societal biases. Gender bias, in particular, has been studied in a broad spectrum of texts such as portrayals of characters in movies, books, news and media.

Choueiti et al. [2014]; Ramakrishna et al. [2015, 2017] examined gender differences in portrayal of characters in movies and consistently show that female characters appear to be more positive in language use with fewer references to death and fewer swear words compared to male characters. However, Sap et al. [2017] find that, high-agency women frames are rare in modern films. Rashkin et al. [2018] use commonsense inference tasks on movie scripts' corpus to unveil presence of gender bias finding that women's looks and sexuality are highlighted, while men's actions are motivated by violence, with strong negative reactions. Moreover, Bamman and Smith [2014] employ a probabilistic latent-variable model to extract event classes from biographies and find that characterisation bias on Wikipedia with biographies of women containing significantly more emphasis on events of marriage and divorce than biographies of men. Field and Tsvetkov [2019] show that although powerful women are frequently portrayed in the media, they are typically described as less powerful than their actual role in society. However, Asr et al. [2021] report that there, in fact, is a gender gap in coverage of women in Canadian news outlets. Further, Hoyle et al. [2019] use an unsupervised model to find that differences between descriptions of males and females in literature align with common gender stereotypes: Positive adjectives used to describe women are more often related to their bodies than adjectives used to describe men.

However, Garg et al. [2018] show that gender bias has decreased in the last 100 years and that the women's movement in the 1960s and 1970s had a significant effect on women's portrayals in literature and culture. To this end, Garg et al. [2018] use word embeddings as a tool to observe the development of adjectives associated with men and women. This is possible since word embeddings learn harmful associations and stereotypes from the underlying data and thus, may serve as a means to extract implicit gender associations from a corpus to detect gender associations present in society [Bolukbasi et al. 2016]. Similarly, Wevers [2019] show that word embeddings can be used to investigate shifts in language related to gender, while Friedman et al. [2019] prove that word embeddings are able to characterise and predict statistical gender gaps in education politics, economics and health across cultures.

A number of research has investigated differences in language directed towards men and women. For instance, Tsou et al. [2014] find that comments on TED talks are more likely to be about the presenter than the content if the presenter is a woman. Fu et al. [2016] analyse questions directed at male and female tennis players, finding that questions to men are rather about the game while questions directed at women are often about their appearance and relationships. Further, Voigt et al. [2018] corroborate the former findings, such as remarks on appearance being more often targeted towards women, responses to women being more emotive (non-neutral sentiment) and of higher sentiment in general which can be ascribed to benevolent sexism.

While the above research unveils some of the ways gender bias is manifested in natural language, it gives only a limited view since most of this research has concentrated on binary gender identities and was mostly conducted in English. We note that there exist real-life applications with societal implications to algorithms detecting gender bias in natural language such as warning systems classifying texts as biased to notify readers.



## 7.2 Detecting Gender Bias in Methods

Biased datasets used in the training process are the primary source of gender bias in NLP methods [Zhao et al. 2017]. Tan and Celis [2019]; Zhao et al. [2019] examine datasets that were used as training corpora for the popular NLP methods and find that the occurrence of male pronouns is consistently higher across all datasets and evidence of stereotypical associations. These gender imbalances lead to gender bias in the NLP systems, such as coreference resolution Zhao et al. [2018a]. It has been shown that the level of bias encoded in a model differs depending on the training data. For instance, Chaloner and Maldonado [2019] study differences in bias in a number of word embeddings trained on corpora from four domains showing the lowest bias in word embeddings trained on a biomedical corpus and the highest bias when trained on news data (higher than social media and Wikipedia-based corpus). Surprisingly, Lauscher and Glavaš [2019]’s findings confirm that gender bias seems to be less pronounced in embeddings trained on social media texts.

A common phenomenon leading to gender bias is a generic masculine pronoun which arises when the masculine form is taken as the generic form to designate all persons of any gender. This is especially the case in the gendered languages [Carl et al. 2004]. Generic masculine poses a challenge in text interpretation since it is unclear if a given person denotation refers to a particular person or a generic form to describe all people in a specific group. For instance, in a sentence “*A researcher must always test his model for biases.*”, it is ambiguous if a particular researcher is considered or researchers in general. In particular, Hitti et al. [2019] analyse data from Project Gutenberg and IMDB to identify such gender generalisations and detect that even 5% of each corpus is affected.

Due to simple interpretation and ability to capture gender stereotypes occupation words have become a common domain for gender bias detection [Garg et al. 2018]. Bolukbasi et al. [2016] project the occupation words onto the *she-he* axis and find that the projections are strongly correlated with the stereotypicality estimates of these words, suggesting that the geometric bias of word embeddings is aligned with crowd judgment of gender stereotypes. Sahlgren and Olsson [2019] show that male names are on average more similar to stereotypically male occupations with an according observation applying to female names. Rudinger et al. [2018] demonstrate how occupation-specific bias is correlated with employment statistics and often so magnified.

Although the majority of the research has focused on analysing gender bias in methods developed on English corpora, there have been some advances in extending this line of work to other languages. Developing language-specific methods to assess language model’s limitations is crucial to prevent bias propagation to downstream tasks in the analysed language [Bartl et al. 2020; Sun et al. 2019]. Findings made for English do not automatically extend to other languages, especially if those exhibit morphological gender agreement [Nozza et al. 2021]. In particular, gender bias in word embeddings of languages with grammatical gender can be expressed in different ways, such as in a discrepancy in semantics between the masculine and feminine forms of the same noun in word embeddings. For example, it has been shown that when aligning Spanish to English word embeddings, the word “*abogado*” (male lawyer) is closer to “*lawyer*” than “*abogada*” (female lawyer) [Zhou et al. 2019]. Interestingly, Lauscher and Glavaš [2019] find that the level of bias in cross-lingual embedding spaces can roughly be predicted from the bias of the corresponding monolingual embedding spaces.

Model architecture is analysed as one of the influencing factors for bias in algorithms. For instance, Lauscher and Glavaš [2019] hypothesise that the bias effects reflected in the distributional space depend on the preprocessing steps of the embedding model. Additionally, discovering bias in transformer models has proven to be more nuanced than bias-discovery in word embedding models [Kurita et al. 2019; May et al. 2019]. Nadeem et al. [2021] hypothesise that an ideal language model should not only be able to perform the task of language modeling, but also cannot exhibit stereotypical bias – it should avoid ranking stereotypical contexts higher than anti-stereotypical

contexts. Recent research has aimed to rank language models in terms of bias they perpetuate [Nangia et al. 2020; Silva et al. 2021]. However, these studies present partially contradictory results presenting a need for more exhaustive testing. The influence of the model's size on the encoded (gender) bias has been examined. For instance, Silva et al. [2021] find that distilled models almost always exhibit statistically significant bias and that the bias effect sizes are often much stronger than in the original models. Vig et al. [2020] show that gender bias increases with the size of a model. Recently, Bender et al. [2021] confirm this claim warning from potential risks associated with large language models. However, in a study of gender bias in cross-lingual language models Stańczak et al. [2021] do not find significant results to support this claim.

It is difficult to understand the nature of biases encoded in large language models due to their complexity. However, applying interpretability methods can shed light on the models and biases preserved. For instance, Vig [2019] use visualisations to reveal attention patterns generated by GPT-2 in the task of conditional language generation and show that the model's coreference resolution might be biased. Vig et al. [2020] probe neural models to analyse the role of individual neurons and attention heads in mediating gender bias and find out that the source of gender bias is concentrated in a small part of the model. Moreover, Bhardwaj et al. [2021] identify gender informative features (and discard them from the model as a mitigation technique).

Until now research has aimed to detect gender bias in a strictly binary setting. We want to highlight the importance of a gender-inclusive research and discuss below publications that have step up to this task. Hicks et al. [2015] collect a data set and develop visualisation tools that show relative frequency and co-occurrence networks for American English trans words on Twitter. Manzini et al. [2019] extend the method presented in Bolukbasi et al. [2016] and use their approach to find non-binary gender bias by aggregating n-tuples instead of gender pairs. Saunders et al. [2020] explore applying tagging to indicate gender-neutral referents in coreference sentences with a gender-neutral pronoun. Recently, Vig et al. [2020] test the probability of a model to generate the pronoun *they* for a number of templates. The probability of the pronoun *they* is relatively low, however constant across probed professions.

### 7.3 Detecting Gender Bias in Downstream Tasks

Bias in the above methods influences many downstream tasks for which these methods are used, which presents a risk of propagating and amplifying gender bias [Zhao et al. 2017, 2018a]. Thus, in the following, we analyse literature on gender bias in downstream applications.

*Machine Translation.* Popular online machine learning services, such as Google Translate or Microsoft Translator, were shown to exhibit biases and to default to the masculine pronoun [Escudé Font and Costa-jussà 2019]. NLP models may learn associations of gender-specified pronouns (for a gendered language) and gender-neutral ones for lexicon pairs that frequently collocate in the corpora [Cho et al. 2019]. This kind of phenomenon threatens the fairness of a translation system since it lacks generality and inserts social bias to the inference. Moreover, the output is not fully correct (considering gender-neutrality) and poses ethical considerations.

When translating from a language without grammatical gender to a gendered one, the required clue about the noun's gender is missing which poses a challenge for MT systems. Saunders et al. [2020] find that existing approaches tend to overgeneralise and incorrectly use the same inflection for every entity in the sentence. However, gender is incorrectly predicted not only in the absence of the gender information. MT methods produce stereotyped translations even when gender information is present in the sentence. Schiebinger [2014] argue that scientific research fails to take this issue into account. Recently, Prates et al. [2020] show that Google Translate still exhibits a strong tendency towards male defaults, in particular for fields typically associated with unbalanced gender

distribution or stereotypes such as STEM (Science, Technology, Engineering, and Mathematics) jobs. Prates et al. [2020] hypothesise that gender neutrality in language and communication leads to improved gender equality. Thus, translations should aim gender-neutrality, instead of defaulting to male or female variants.

*Coreference Resolution.* Various aspects of gender are embedded in coreference inferences, both because gender can show up explicitly (e.g., pronouns in English, morphology in Arabic) and because societal expectations and stereotypes around gender roles may be explicitly or implicitly assumed by speakers or listeners [Cao and Daumé III 2020]. Although existing corpora have promoted research into coreference resolution, they suffer from gender bias [Zhao et al. 2018a].

Webster et al. [2018] find that existing resolvers do not perform well and are biased to favour better resolution of masculine pronouns. Rudinger et al. [2018] show how overall, male pronouns are more likely to be resolved as occupation than female or neutral pronouns across all systems. Moreover, Zhao et al. [2018a] demonstrate that neural coreference systems all link gendered pronouns to stereotypical entities with higher accuracy than anti-stereotypical entities. Zhao et al. [2018a] warn that bias encoded in word embeddings leads to sexism in coreference resolution. Further, Bao and Qiao [2019] show significant gender bias when using popular NLP methods for coreference resolution on both sentence and word level, indicating that women are associated with family while men are associated with career.

*Language Generation.* Henderson et al. [2018] suggest that, due to their subjective nature and goal of mimicking human behaviour, data-driven dialogue models are prone to implicitly encode underlying biases in human dialogue, similar to related studies on biased lexical semantics derived from large corpora [Bolukbasi et al. 2016; Caliskan et al. 2017]. Cercas Curry and Rieser [2018] estimate that as many as 4% of conversations with chatbased systems are sexually charged. Further, Bartl et al. [2020] find that the monolingual BERT reflects the real-world bias of the male- and female-typical profession groups through stereotypical associations. Stories generated by GPT-3 differ based on a perceived gender of the character in a prompt with female characters being more often associated with family, emotions and appearance, even in spite of a presence of power verbs in a prompt [Lucy and Bamman 2021].

*Sentiment Analysis.* Kiritchenko and Mohammad [2018] test 219 automatic sentiment analysis systems that participated in SemEval-2018 Task 1 *Affect in Tweets* [Mohammad et al. 2018]. In particular, Kiritchenko and Mohammad [2018] examine a hypothesis that a system should equally rate the intensity of the emotion expressed by two sentences that differ only in the gender of a person mentioned and find that the majority of the systems studied show statistically significant bias. In particular, they consistently provide slightly higher sentiment intensity predictions for sentences associated with one gender (gender with more positive sentiment varies based on a task and system used). When predicting anger, joy, or valence, the number of systems with consistently higher sentiment for sentences with female noun phrases is higher than for male noun phrases. Bhaskaran and Bhallamudi [2019] show that analysed sentiment analysis methods exhibit differences in mean predicted class probability between genders, though the directions vary again.

## 7.4 Summary

As seen above, NLP methods tend to be consistently biased and associate harmful stereotypes with genders. Despite this fact, most of the papers that have focused on detecting gender bias in natural language, methods, or downstream tasks, have seen bias detection as a goal in itself or a means of analysing the nature of bias in domains of their interest. Some of this research has been followed up with bias mitigation methods (discussed in §8). However, often enough, findings of this line

of research are treated solely as a fact statement and not an action trigger. In particular, despite a number of evidence showing that NLP methods encode gender bias, developers are not required to provide any formal testing prior to releasing new models. Widely acknowledged models that have led in recent years to significant gains on many NLP tasks have not included any study of bias alongside the publication [Conneau et al. 2020; Devlin et al. 2019; Peters et al. 2018; Radford et al. 2019]. Since these models were probed for gender bias only after their release, they might have already caused harm in real life applications. We strongly encourage including bias detection into the model development pipeline and see it as a necessary future development.

So far, research has predominantly aimed to detect bias towards male and female gender, ignoring non-binary gender identities. However, it is crucial to design studies on gender bias detection that are gender-inclusive at all stages, from defining gender and bias, dataset choice to selecting bias detection method.

As discussed in §3, gender manifests itself in different ways across languages. Hence, it can be expected that it's also the case for gender bias. For instance, languages such as German, Hebrew and Russian use gender inflections that reflect grammatical genders of the nouns. Further, gender bias is grounded in societal and cultural views on gender and thus, its expressions potentially vary across languages. Expanding research to languages beyond English and including data from outside of the Anglosphere would lead to gaining a broader view on gender bias in societies. In particular, analysing cross-lingual data might enable a comparative studies of gender bias.

## 8 MITIGATING GENDER BIAS

While it is impossible to altogether remove gender bias from language or from NLP algorithms, research on gender bias mitigation is a significant step towards developing fair systems. In specific applications, one might argue that gender biases in algorithms could capture valuable statistics such as a higher probability of a nurse being a female. Nevertheless, given the potential risk of employing machine learning algorithms that amplify gender stereotypes, Bolukbasi et al. [2016] recommend erring on the side of neutrality and using debiased methods. However, following D'Ignazio [2021], mitigating gender bias in AI systems is a short-term solution that needs to be combined with higher-level long-term projects in challenging the current social status quo.

The main challenge in debiasing task is to strike a trade-off between maintaining model performance on downstream tasks while reducing the encoded gender bias [de Vassimon Manela et al. 2021; Zhao et al. 2018a]. Further, Bartl et al. [2020]; Sun et al. [2019] emphasise the need for more typological variety in NLP research as well as for language-specific solutions. Many of the mitigation methods rely on pre-defined words lists that are not scalable in a multilingual setup and are tedious to create. However, recent work on dictionary definitions for debiasing might obviate the need for predefined word lists [Kaneko and Bollegala 2021b]. While prior work has mainly focused on mitigating gender bias in English, more recently, researchers have started to apply methods initially developed for English to other languages as well. Naturally, a significant chunk of work for multilingual settings has been researched in the context of neural machine translation [Prates et al. 2020; Vanmassenhove et al. 2018]. This stream of research has confirmed that language-specific solutions are required, since gender is expressed in different ways across languages. For instance, transferring a method successful in gender bias mitigation for English to German may be ineffective which emphasises the need for more typological variety in research as well as language-specific solutions [Bartl et al. 2020]. Therefore, it is crucial to develop (language-specific) debiasing methods, especially for relatively new methods, to assess these limitations. Next, Kiritchenko and Mohammad [2018] observed that different debiasing approaches have varying effects on the analysed word embedding architectures. Many of the current debiasing methods are evaluated only on selected downstream tasks without testing them in a broader scope. Hence, additional and potentially costly

tests are required before applying these techniques to other, previously un-tested tasks since their effectiveness there is unclear [Jin et al. 2021]. Therefore, we encourage research on debiasing methods in the early modelling stages.

Data Manipulation			
Data Augmentation	Gender Tagging	Balanced Fine-Tuning	Adding Context
Madaan et al. [2018]; Park et al. [2018]	Moryossef et al. [2019]; Vanmassenhove et al. [2018]	Park et al. [2018]; Saunders and Byrne [2020]	Basta et al. [2020]
Hall Maudslay et al. [2019]; Zhao et al. [2018a]	Habash et al. [2019]; Stefanovičs et al. [2020]	Costa-jussà and de Jorge [2020]	
Emami et al. [2019]; Zmigrod et al. [2019]	Saunders et al. [2020]		
Bartl et al. [2020]; Zhao et al. [2019]			
de Vassimon Manela et al. [2021]; Sen et al. [2021]			
Methodological Adjustment			
Projection-Based Debiasing	Adversarial Learning	Constraining Output	Other
Bolukbasi et al. [2016]; Schmidt [2015]	Li et al. [2018]; Zhang et al. [2018]	Ma et al. [2020]; Zhao et al. [2017]	Qian et al. [2019]; Zhao et al. [2018b]
Bordia and Bowman [2019]; Park et al. [2018]			Jin et al. [2021]; Kaneko and Bollegala [2019]
Ethayarajah et al. [2019]; Sahlgren and Olsson [2019]			
Karve et al. [2019]; Sedoc and Ungar [2019]			
Liang et al. [2020]; Probst et al. [2019]			
Dev et al. [2020]; Kaneko and Bollegala [2021a]			

Table 3. Classification of gender bias mitigation methods with respective publications.

Different approaches have been developed to mitigate gender bias in NLP. In this paper, we classify each of these methods following the two main categories, similarly to Sun et al. [2019] – debiasing using data manipulation 8.1 and by adjusting algorithms 8.2 – while extending the scope of our analysis with recent publications and incorporating word embeddings mitigation methods into the methodological adjustment category. We summarise the identified lines of gender bias mitigation methods in Table 3 together with the respective publications.

## 8.1 Debiasing Using Data Manipulation

Debiasing using data manipulation commonly refers to counterfactual data augmentation, gender tagging, adding context, and balanced fine-tuning. Below we describe these approaches in detail.

**8.1.1 Data Augmentation.** Many concerns have been posed regarding modern NLP systems having been trained on potentially biased datasets, as as these biases can be perpetuated to downstream tasks and eventually society in the form of allocational harms [Hovy and Prabhumoye 2021]. Therefore, Costa-jussà and de Jorge [2020] claim that developing methods trained on balanced data is a first step to eliminating representational harms.

In order to attenuate the impact of gender bias from the dataset used, Zhao et al. [2018a] propose a rule-based approach to generate an auxiliary dataset where all-male entities are replaced by female entities (and vice-versa) and suggest to train methods on the union of the original and augmented dataset. Thus, both male and female genders are equally represented in the dataset. For instance, a sentence *My son plays with a car.* would be transformed into *My daughter plays with a car.* Therefore, to apply this method, a list of gendered pairs (such as *son–daughter*) is required. Similarly, Emami et al. [2019] propose to extend a training set for coreference resolution by switching every entity pair. A method for debiasing gender-inflected languages is demonstrated in Zmigrod et al. [2019], where sentences are reinflected from masculine to feminine (and vice-versa) in a counterfactual data augmentation (CDA) scheme. Since this method analyses each word separately, it is not applicable to more complex sentences involving coreference resolution. However, it introduces a feasible debiasing approach for languages beyond English. Hall Maudslay et al. [2019] develop a name-based version of CDA, in which the gender of words denoting persons in a training corpus are swapped probabilistically in order to counterbalance bias.

Due to its simple implementation, counterfactual data augmentation has been widely applied to mitigate gender bias. Since the model observes the same scenario in the doubled (for binary gender) sentences, it can learn to abstract away from the entities to the context [Emami et al. 2019]. This method has shown encouraging results in mitigating bias in contextualised word representations

such as ELMo and monolingual BERT [Bartl et al. 2020; de Vassimon Manela et al. 2021; Sen et al. 2021; Zhao et al. 2019], and for hate speech detection [Park et al. 2018]. Nonetheless, collecting annotated lists for gender-specific pairs can be expensive, and the method essentially doubles the size of the training data. To this end, de Vassimon Manela et al. [2021] compare fine-tuning contextualised word representation on augmented and un-augmented datasets and show that fine-tuning solely on an augmented corpus successfully decreases gender bias.

Another method of gender bias mitigation via data augmentation is presented in Stanovsky et al. [2019] who suggest a simple approach of “fighting bias with bias” and add stereotypical adjectives to describe entities of the respective gender, e.g., “*The pretty doctor asked the nurse to help her in the procedure.*”. However impractical this method is, relying on accurate coreference resolution, it has shown to reduce bias in the tested languages.

**8.1.2 Gender Tagging.** Another stream of work has concentrated on incorporating external or internal gender information during training. This method has been widely employed in debiasing neural machine translation models to mitigate the issue of male default. Moryossef et al. [2019] append a short phrase at inference time which acts as an indicator for the speaker’s gender, e.g., “*She said:*”, while similarly, Vanmassenhove et al. [2018] use sentence-level annotations. In order to extend the mitigation method to be applicable to sentences with more than one gendered entity, Stafanovičs et al. [2020] utilise token-level annotations for the subject’s grammatical gender. Habash et al. [2019] propose a post-processing method that is an intersection of gender tagging and CDA and test it on Arabic. In gender-aware debiasing, a gender-blind system is being turned into a gender-aware one by identifying gender-specific phrases in the system’s output and subsequently offering alternative reinflections. In the domain of machine translation, Saunders et al. [2020] propose an approach based on fine-tuning a model on a small, artificial dataset of sentences with gender inflection tags which are then replaced by placeholders. However, the results of this scheme are ambiguous, and this method is not well suited for translating sentences with multiple entities.

Methods relying on gender tagging are a flexible tool for controlling for bias. However, we note that these methods do not inherently remove gender bias from the system [Cho et al. 2019]. Additionally, gender tagging requires meta-information on the gender of the speaker, which is often either expensive or unavailable.

**8.1.3 Adding Context.** Alongside including the speaker’s information as in the above examples, Basta et al. [2020] concatenate the previous sentence from a corpus to increase the context. Using the additional information only in the decoder part of the Transformer architecture ultimately reduces training parameters, simplifies the model, and requires no further information for training or inference. Basta et al. [2020] show that this method improves the performance of machine translation with coreference resolution tasks. However, Savoldi et al. [2021] note that this improvement might not be due to the added gender context, but for instance, a regularisation effect.

**8.1.4 Balanced Fine-Tuning.** Balanced fine-tuning incorporates transfer learning from a less biased dataset. In the first step, a model is trained on a large, unbiased dataset for the same or similar downstream task and is then fine-tuned on a target dataset which is more biased [Park et al. 2018]. Such a training regime obviates potential over-fitting to a biased dataset. This method suffers from a severe limitation, namely assuming an existence of an unbiased dataset in its initial step, which is usually infeasible to obtain and thus, not applicable in real-life applications. On the other hand, Saunders and Byrne [2020] consider gender bias in machine translation as a domain adaptation task and use a handcrafted gender-balanced dataset together with a lattice re-scoring module to mitigate the consequences of initial training on unbalanced data. Saunders and Byrne [2020] consider three aspects of the adaptation problem: creating less biased adaptation data, parameter adaptation using

this data, and inference with the debiased models produced by adaptation. However, the need for a gender-balanced dataset for a specific domain might be a drawback of this approach. Costa-jussà and de Jorge [2020] use an inverse approach and train their model on a larger corpus and fine-tune it with a gender-balanced corpus showing that their approach successfully mitigates gender bias and increases performance quality even if the balanced dataset is coming from a different domain. However, Savoldi et al. [2021] note that this approach does not account for the qualitative differences in how men and women are portrayed [Savoldi et al. 2021].

## 8.2 Debiasing by Adjusting Algorithms

Instead of manipulating the underlying data, a number of gender debiasing methods have been implemented to approach the issue via algorithm adjustment. Such techniques can be categorised into the following groups: projection-based debiasing, constraining models' predictions, applying adversarial learning approaches, and other.

*8.2.1 Projection-Based Debiasing.* To the best of our knowledge, Schmidt [2015] propose the first word embedding debiasing algorithm and remove multiple gender dimensions from word vectors. In parallel, instead of completely removing gender information, Bolukbasi et al. [2016] suggest shifting word embeddings to be equally male and female in terms of their vector direction. For instance, a debiased embeddings for *grandmother* and *grandfather* will be equally close to *babysit* without neglecting the analogy to gender. More formally, Bolukbasi et al. [2016] propose two debiasing methods, hard- and soft-debiasing. The first step for both of them consists of identifying a list of gender-neutral words and a direction of the embedding that captures the bias. **Hard-debiasing** (or 'Neutralise and Equalise method') ensures that gender-neutral words are zero in the gender subspace and equalises sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set (a set of words which differ only in the gender component). For instance, if (grandmother, grandfather) and (guy, gal) were two equality sets, then after equalisation, 'babysit' would be equidistant to grandmother and grandfather and also to gal and guy, but closer to the grandparents and further from the gal and guy. This approach is suitable for applications where one does not want any such pair to display any bias with respect to neutral words. The disadvantage of equalising sets of words outside the subspace is that it removes particular distinctions that are valuable in specific applications. For instance, one may wish a language model to assign a higher probability to the phrase to 'grandfather a regulation' since it is an idiom, unlike 'grandmother a regulation'. The **soft-debiasing** algorithm reduces differences between these sets while maintaining as much similarity to the original embedding as possible, with a parameter that controls for this trade-off. In particular, soft-debiasing applies a linear transformation that seeks to preserve pairwise inner products between all the word vectors while minimising the projection of the gender-neutral words onto the gender subspace.

Both hard- and soft-debiasing approaches have been applied in research to word embeddings and language models. Bordia and Bowman [2019] validate the soft-debiasing approach to mitigate bias in LSTM based word-level language models. Park et al. [2018] compare hard-debiasing method to other methods in the context of abusive language detection. Sahlgren and Olsson [2019] apply hard-debiasing to Swedish word embeddings and show that this method does not have the desired effect when tested on selected downstream tasks. Sahlgren and Olsson [2019] argue that these unsatisfactory results are due to including person names in their training process. Interestingly, Ethayarajh et al. [2019] show that debiasing word embeddings post hoc using subspace projection is, under certain conditions, equivalent to training on an unbiased corpus. Similarly to Bolukbasi et al. [2016], Karve et al. [2019]; Sedoc and Ungar [2019] aim to identify the bias subspace in word

embeddings using a set of target words and a **debiasing conceptor**, a mathematical representation of subspaces that can be operated on and composed by logic-based manipulations.

However, these methods strongly rely on the pre-defined lists of gender-neutral words Sedoc and Ungar [2019]. Moreover, Zhao et al. [2018b] prove that an error in identifying gender-neutral words affects the performance of the downstream model. Bordia and Bowman [2019]; Zhao et al. [2018b] notice a trade-off between perplexity and gender bias as in an unbiased model, male and female words are predicted with an equal probability. This can be undesirable in domains such as social science and medicine. While Gonen and Goldberg [2019] claim that debiasing is primarily superficial since a lot of the supposedly removed bias can still be recovered due to the geometry of the word representation of the gender neutralised words, Prost et al. [2019] show that soft-debiasing can even increase the bias of a downstream classifier by removing noise from gender-neutral words and ultimately providing a less noisy communication channel for gender clues.

Recently, Liang et al. [2020] use DensRay [Dufter and Schütze 2019], an interpretable method for identifying the embedding subspace using projections and then evaluate gender bias in masked language models by comparing the difference in the log-likelihood between male and female pronouns in a template “[MASK] is a/an [occupation].”. However, this method heavily relies on a list of occupations and a simple template. Further, Dev et al. [2020] employ an orthogonal projection to gender direction [Dev and Phillips 2019] to debias contextualised embeddings and test it on a NLI task with gender-biased hypothesis pairs. However, this method can only be applied to the model’s non-contextualised layers. Kaneko and Bollegala [2021a] obviate this limitation in a fine-tuning setting. Their method applies an orthogonal projection only in the hidden layers and proves to outperform Dev et al. [2020]. Additionally, this method is independent of model architectures or their pre-training method. However, this approach requires a list of attribute words (e.g., she, man, her) and target words (e.g., occupations) to extract relevant sentences from the corpus, making their method highly reliant on this list.

**8.2.2 Constraining Output.** A simple approach to debiasing algorithms is to constrain model output post-hoc. To this end, Zhao et al. [2017] propose a debiasing technique that constrains model predictions to follow a distribution from a training corpus, e.g., the ratio of male and female pronouns. Thus, this method is highly dependent on the gender balance and bias in the underlying data.

In the field of language generation, Ma et al. [2020] introduce *controllable debiasing* as an unsupervised text revision task that aims to correct the implicit bias against or towards a specific character portrayed in a language model generated text. For this purpose, they create an encoder-decoder model that rewrites a text to portray females as more agent (in terms of Sap et al. [2017]’s connotation frames). However, their approach relies strongly on an external corpus of paraphrases.

**8.2.3 Adversarial Learning.** Another strain of work has employed adversarial learning as a debiasing method. Li et al. [2018] propose a method for removing model biases by explicitly protecting demographic information (such as gender) during model training. However, Elazar and Goldberg [2018] claim that word representations preserve traces of the protected attributes and recommend external verification of the method. Similarly, Zhang et al. [2018] apply adversarial learning by including gender as a protected variable and having the generator learn with respect to it. In general, the objective of such a model is to maximise the predictor’s ability to predict a variable of interest while fooling the adversary to predict the protected attribute. However, in general, adversarial learning is often an unstable method and can only be used when gender is a protected attribute rather than a variable of interest.

**8.2.4 Other.** Several other methods have been tested to mitigate gender bias in NLP methods.



Alongside projection-based methods for debiasing word embeddings, another approach to debiasing word embeddings has aimed to learn their gender-neutralised variant. In particular, Zhao et al. [2018b] propose to train word embeddings such that protected attributes are neutralised in some of the dimensions, resulting in gender-neutral word representations. Restricting the information of protected attributes in certain dimensions enables its removal from an embedding. Additionally, other than the method presented in Bolukbasi et al. [2016] gender-neutral words are learned jointly in the training process instead of being manually created. However, Sun et al. [2019] note that it is unclear if gender-neutralised word embeddings are applicable to languages with grammatical genders.

Adjusting the loss function has proven to be another viable method for gender bias mitigation. In particular, Qian et al. [2019] introduces a new term to the loss function, which attempts to equalise the probabilities of male and female words (based on a pre-defined list) in the output and evaluate it on a text generation task. We see two main limitations of this approach. First, it relies on a straightforward definition of bias (*i.e.*, an equal number of gender mentions). Second, as with many other methods, it requires a list of gender pairs, a limitation we discuss above.

Gender-preserving debiasing has been introduced to mitigate gender bias, accounting that not all gender associations are stereotypical. Kaneko and Bollegala [2019] split a given vocabulary into four mutually exclusive sets of word categories: words that are female-biased but non-discriminative, male-biased but non-discriminative, gender-neutral words, and words perpetuating stereotypes. Kaneko and Bollegala [2019] learn word embeddings that preserve the information for the gendered but non-stereotypical words protects the neutrality of the gender-neutral words while removing the gender-related biases from stereotypical words. The embedding is learnt using an encoder in a denoising autoencoder, while the decoder is trained to reconstruct the original word embeddings from the debiased embeddings. However, creating a word list with the above-mentioned categories of words is time-consuming, and word categorisation might not be straightforward.

Jin et al. [2021] investigate incorporating bias mitigation into the model's objective. First, an upstream model is fine-tuned with a bias mitigation objective which consists of a text encoder and a classifier head. Subsequently, the encoder from the upstream model, jointly with the new classification layer, are again fine-tuned on a downstream task. Jin et al. [2021] note that upstream bias mitigation, while less effective, is more efficient than direct bias mitigation methods without fine-tuning. However, it requires a tailored evaluation for the downstream task.

## 9 DISCUSSION

After presenting probing datasets, formal definitions, detection, and mitigation methods, we next present the main findings we make throughout this survey. We find that existing research on gender bias has four main limitations and discuss them in the following.

*Gender in NLP.* It is not uncommon for studies about gender to be reported without any explanation of how gender labels are ascribed, and the ones that do, explain the imputation of gender categories in a debatable way [Larson 2017]. Using gender as a variable in NLP is an ethical issue, thus unreflectively assigning gender category labels may violate ethical frameworks that demand transparency and accountability from researchers [Larson 2017]. Therefore, it is crucial to ask how researchers can use NLP tools to investigate the relationship between gender and text meaningfully, yet without harmful stereotypes Koolen and van Cranenburgh [2017]. To obviate this risk, Larson [2017] suggest formulating research questions with explicit definitions of *gender*, avoiding using gender as a variable unless it is necessary. Not being explicit about the ascription of the category of gender as a variable to participants brings into question the internal and external validity of

research findings because it makes it difficult to near-impossible for other scholars to reproduce, test, or extend study findings [Larson 2017].

We find that researchers often decide to define gender in their study as binary. However, making this assumption is an oversimplification of gender complexity and can perpetuate harms to non-binary people [Behm-Morawitz and Mastro 2008; Fast et al. 2016]. We encourage researchers to define gender in a transparent and inclusive manner, to expand corpora with inclusive pronouns, and evaluate models on non-binary pronouns as well to mitigate these harms. So far models' performance on downstream tasks has been consistently lower for non-binary pronouns compared to the binary pronouns [Cao and Daumé III 2020; Sun et al. 2021].

*Monolingual focus.* Gender bias is grounded in societal and cultural views on gender, and thus, its expressions vary across languages. Expanding research to languages beyond English and including data from outside of the Anglosphere would lead to gaining a broader view on gender bias in societies which we strongly encourage. However, most prior research on gender bias has been monolingual, focusing predominantly on English or a small number of other high-resource languages such as Chinese [Liang et al. 2020] and Spanish [Zhao et al. 2020] with the notable exception of a broader multilingual analysis of gender bias in machine translation [Prates et al. 2020] and language models [Stańczak et al. 2021].

*Need for formal testing.* Most of the papers that have focused on detecting gender bias in natural language, methods, or downstream tasks, have seen bias detection as a goal in itself or a means of analysing the nature of bias in domains of their interest. Widely acknowledged models that have led in recent years to significant gains on many NLP tasks have not included any study of bias alongside the publication [Conneau et al. 2020; Devlin et al. 2019; Peters et al. 2018; Radford et al. 2019]. In general, these methods are tested for biases only post-hoc when already being deployed in real-life applications, potentially posing harm to different social groups [Mitchell et al. 2019]. Since these models were probed for gender bias only after their release, they might have already caused societal harms. We find that bias detection should be included in the model development pipeline at early stages and see enforcing this change as a primary challenge. The way to ensure that researchers abide by ethical principles is to hold them accountable when research projects are planned, *i.e.*, requiring project proposals and publications to include ethical considerations and, later, during the peer review process.

*Limited definitions.* However, to introduce formal testing comprehensive and multi-faceted bias measures are required. We find that similarly to research within societal biases Blodgett et al. [2020], work on gender bias in particular, suffers from incoherence in usage of evaluation metrics. Most of the publications on gender bias consider only one way of defining bias and do not engage enough parallel research to combine these methods. Gender bias can be expressed in language in many nuanced ways which poses stating a comprehensive definition as one of the main challenges in this research field. Finally, we strongly encourage developing standard evaluation benchmarks and tests to enhance comparability.

## 10 CONCLUSION

In this paper, we present a comprehensive survey of 304 papers on gender bias in natural language and NLP methods published since gender bias has been studied in NLP. We find four major limitations in the existing research and see overcoming these limitations as crucial for further development of this field.

First, most research lacks transparent and inclusive gender and gender bias definitions. Gender is mainly treated as a binary variable which disagrees with social science position. Next, the majority

of the work disregards low-resource languages, concentrating solely on English and other high-resource languages such as Spanish and Chinese, which imposes a strongly restricted view on the nature of gender bias in NLP. Moreover, despite a myriad of papers on gender bias in NLP methods, most of the newly developed algorithms do not test their models for bias and disregard possible ethical considerations of their work. This leads to deployment of models that lead to potential societal harms. Finally, we find that the methodology used in this research field is fundamentally flawed, covering only limited aspects of gender bias and lacking baselines for evaluation and testing pipelines.

## REFERENCES

- Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics* 4 (10 2019). <https://doi.org/10.5334/gjgl.721>
- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. *Automatic Identification and Classification of Misogynistic Language on Twitter*. 57–64. [https://doi.org/10.1007/978-3-319-91947-8\\_6](https://doi.org/10.1007/978-3-319-91947-8_6)
- Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE* 16, 1 (01 2021), 1–28. <https://doi.org/10.1371/journal.pone.0245533>
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 2 (2014), 135–160. <https://doi.org/10.1111/josl.12080> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12080>
- David Bamman and Noah A. Smith. 2014. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics* 2 (2014), 363–376. [https://doi.org/10.1162/tacl\\_a\\_00189](https://doi.org/10.1162/tacl_a_00189)
- Xingce Bao and Qianqian Qiao. 2019. Transfer Learning from Pre-trained BERT for Pronoun Resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 82–88. <https://doi.org/10.18653/v1/W19-3812>
- Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 1–16. <https://aclanthology.org/2020.gebnlp-1.1>
- Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. Association for Computational Linguistics, Seattle, USA, 99–102. <https://doi.org/10.18653/v1/2020.winlp-1.25>
- Elizabeth Behm-Morawitz and Dana Mastro. 2008. Mean Girls? The Influence of Gender Portrayals in Teen Movies on Emerging Adults’ Gender-Based Attitudes and Beliefs. *Journalism & Mass Communication Quarterly - JOURNALISM MASS COMMUN* 85 (03 2008), 131–146. <https://doi.org/10.1177/107769900808500109>
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT ’21)*. Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6923–6933. <https://doi.org/10.18653/v1/2020.acl-main.619>
- Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating Gender Bias in BERT. *Cognitive Computation* 13 (2021), 1008–1018.
- Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 62–68. <https://doi.org/10.18653/v1/W19-3809>
- M Bing, Janet and Victoria L Bergvall. 1998. The question of questions: Beyond binary thinking. In *Language and Gender: A Reader*, Jennifer Coates (Ed.). Blackwell, Oxford, 496–510.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural*

- Information Processing Systems* (Barcelona, Spain) (*NIPS'16*). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.
- Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 7–15. <https://doi.org/10.18653/v1/N19-3002>
- Sian Brooke. 2019. “Condescending, Rude, Assholes”: Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 172–180. <https://doi.org/10.18653/v1/W19-3519>
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- Judith Butler. 1989. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (Apr 2017), 183–186. <https://doi.org/10.1126/science.aal4230>
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Trento, Italy. <https://aclanthology.org/E06-1032>
- Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4568–4595. <https://doi.org/10.18653/v1/2020.acl-main.418>
- Michael Carl, Sandrine Garnier, Johann Haller, Anne Altmayer, and Bärbel Miemietz. 2004. Controlling Gender Equality with Shallow NLP Techniques. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. COLING, Geneva, Switzerland, 820–826. <https://aclanthology.org/C04-1118>
- Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, New Orleans, Louisiana, USA, 7–14. <https://doi.org/10.18653/v1/W18-0802>
- Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 25–32. <https://doi.org/10.18653/v1/W19-3804>
- Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 173–181. <https://doi.org/10.18653/v1/2019-3824>
- Marc Choueiti, Dr. Katherine Pieper, and Yu-Ting Liu. 2014. Gender Bias Without Borders An Investigation of Female Characters in Popular Films Across 11 Countries. *Gender Bias Without Borders*. <https://seejane.org/symposiums-on-gender-in-media/gender-bias-without-borders/>
- James M. Clark and Allan Paivio. 2004. Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers* 36, 3 (2004), 371–383. <https://doi.org/10.3758/bf03195584>
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Greville G. Corbett. 1991. *Gender*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139166119>
- Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 26–34. <https://aclanthology.org/2020.gebnlp-1.3>
- Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4081–4088. <https://aclanthology.org/2020.lrec-1.502>
- Kate Crawford. 2017. The Trouble with Bias. [https://www.youtube.com/watch?v=fMym\\_BKWQzk&ab\\_channel=TheArtificialIntelligenceChannel](https://www.youtube.com/watch?v=fMym_BKWQzk&ab_channel=TheArtificialIntelligenceChannel) Conference on Neural Information Processing Systems (NIPS) – Keynote.
- B. Dardenne, M. Dumont, and T. Bollier. 2007. Insidious dangers of benevolent sexism: consequences for women’s performance. *Journal of personality and social psychology* 93 5 (2007), 764–79.

- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnamurthy Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan 2019). <https://doi.org/10.1145/3287560.3287572>
- Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2232–2242. <https://aclanthology.org/2021.eacl-main.190>
- Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7659–7666. <https://doi.org/10.1609/aaai.v34i05.6267>
- Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word vectors. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 879–887. <https://proceedings.mlr.press/v89/dev19a.html>
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Catherine D’Ignazio. 2021. *Data Feminism: Teaching and Learning for Justice*. Association for Computing Machinery, New York, NY, USA, 3. <https://doi.org/10.1145/3430665.3456388>
- Philipp Dufter and Hinrich Schütze. 2019. Analytical Methods for Interpretable Ultradense Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1185–1191. <https://doi.org/10.18653/v1/D19-1111>
- Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender Aware Spoken Language Translation Applied to English-Arabic. *CoRR abs/1802.09287* (2018). arXiv:1802.09287 <http://arxiv.org/abs/1802.09287>
- Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 11–21. <https://doi.org/10.18653/v1/D18-1002>
- Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3952–3961. <https://doi.org/10.18653/v1/P19-1386>
- Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 147–154. <https://doi.org/10.18653/v1/W19-3821>
- Kawin Ethayarajh, David Duvinaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1696–1705. <https://doi.org/10.18653/v1/P19-1166>
- Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. arXiv:1603.08832 [cs.CL]
- Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. arXiv:1904.04164 [cs.SL]
- Anjalie Field and Yulia Tsvetkov. 2019. Entity-Centric Contextual Affective Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2550–2560. <https://doi.org/10.18653/v1/P19-1243>
- Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. <https://doi.org/10.1145/230538.230561>
- Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating Word Embedding Gender Biases to Gender Gaps: A Cross-Cultural Analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 18–24. <https://doi.org/10.18653/v1/W19-3803>
- Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. *CoRR abs/1607.03895* (2016). arXiv:1607.03895 <http://arxiv.org/abs/1607.03895>
- Pedro A. Fuertes-Olivera. 2007. A corpus-based view of lexical gender in written Business English. *English for Specific Purposes* 26 (2007), 219–234.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (Apr 2018), E3635–E3644. <https://doi.org/10.1073/pnas.1720347115>

- Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women’s Syntactic Resilience and Men’s Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3493–3498. <https://doi.org/10.18653/v1/P19-1339>
- Peter Glick and Susan Fiske. 1996. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology* 70 (03 1996), 491–512. <https://doi.org/10.1037/0022-3514.70.3.491>
- Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. <https://doi.org/10.18653/v1/N19-1061>
- Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (San Francisco, California, USA) (AKBC ’13)*. Association for Computing Machinery, New York, NY, USA, 25–30. <https://doi.org/10.1145/2509558.2509563>
- Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 155–165. <https://doi.org/10.18653/v1/W19-3822>
- Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It’s All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5267–5275. <https://doi.org/10.18653/v1/D19-1530>
- Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (New Orleans, LA, USA) (AIES ’18)*. Association for Computing Machinery, New York, NY, USA, 123–129. <https://doi.org/10.1145/3278721.3278777>
- Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The Problem of Identifying Misogynist Language on Twitter (and Other Online Social Spaces). In *Proceedings of the 8th ACM Conference on Web Science (Hannover, Germany) (WebSci ’16)*. Association for Computing Machinery, New York, NY, USA, 333–335. <https://doi.org/10.1145/2908131.2908183>
- Amanda Hicks, William Hogan, Michael Rutherford, Bradley Malin, Mengjun Xie, Christiane Fellbaum, Zhijun Yin, Daniel Fabbri, Josh Hanna, and Jiang Bian. 2015. Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms. *AMIA Annual Symposium Proceedings* 2015 (11 2015), 611–620. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765681/pdf/2217289.pdf>
- Amanda Hicks, Michael Rutherford, Christiane Fellbaum, and Jiang Bian. 2016. An Analysis of WordNet’s Coverage of Gender Identity Using Twitter and The National Transgender Discrimination Survey. In *Proceedings of the 8th Global WordNet Conference (GWC)*. Global Wordnet Association, Bucharest, Romania, 123–130. <https://www.aclweb.org/anthology/2016.gwc-1.19>
- Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 8–17. <https://doi.org/10.18653/v1/W19-3802>
- Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (2021), e12432. <https://doi.org/10.1111/lnc3.12432> <https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432>
- Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1706–1716. <https://doi.org/10.18653/v1/P19-1167>
- Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3770–3783. <https://doi.org/10.18653/v1/2021.naacl-main.296>
- Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1641–1650. <https://doi.org/10.18653/v1/P19-1160>
- Masahiro Kaneko and Danushka Bollegala. 2021a. Debiasing Pre-trained Contextualised Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1256–1266. <https://doi.org/10.18653/v1/2021.eacl-main.107>

- Masahiro Kaneko and Danushka Bollegala. 2021b. Dictionary-based Debiasing of Pre-trained Word Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 212–223. <https://doi.org/10.18653/v1/2021.eacl-main.16>
- Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor Debiasing of Word Representations Evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 40–48. <https://doi.org/10.18653/v1/W19-3806>
- Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5435–5442. <https://doi.org/10.18653/v1/2020.acl-main.483>
- Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 43–53. <https://doi.org/10.18653/v1/S18-2005>
- Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv:2102.04130 [cs.CL]
- Corina Koolen and Andreas van Cranenburgh. 2017. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 12–22. <https://doi.org/10.18653/v1/W17-1602>
- Cheris Kramarae and Paula A. Treichler. 1985. *A Feminist Dictionary*.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 166–172. <https://doi.org/10.18653/v1/W19-3823>
- Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 1–11. <https://doi.org/10.18653/v1/W17-1601>
- Anne Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM 2019)*. Association for Computational Linguistics, Minneapolis, Minnesota, 85–91. <https://doi.org/10.18653/v1/S19-1010>
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (KR'12)*. AAAI Press, 552–561.
- Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards Robust and Privacy-preserving Text Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 25–30. <https://doi.org/10.18653/v1/P18-2005>
- Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5082–5093. <https://doi.org/10.18653/v1/2020.coling-main.446>
- Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*. Springer International Publishing, Cham, 189–202. [https://doi.org/10.1007/978-3-030-62077-6\\_14](https://doi.org/10.1007/978-3-030-62077-6_14)
- Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*. Association for Computational Linguistics, Virtual, 48–55. <https://doi.org/10.18653/v1/2021.nuse-1.5>
- Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. *AERA Open* 6, 3 (2020), 2332858420940312. <https://doi.org/10.1177/2332858420940312> arXiv:https://doi.org/10.1177/2332858420940312
- Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7426–7441. <https://doi.org/10.18653/v1/2020.emnlp-main.602>
- N. Madaan, S. Mehta, Tanea S. Agrawaal, Vrinda Malhotra, A. Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *FAT*.
- Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 615–621. <https://doi.org/10.18653/v1/N19-1062>

- Sara Marjanovic, Karolina Stańczak, and Isabelle Augenstein. 2021. Quantifying Gender Biases Towards Politicians on Reddit. arXiv:2112.12014 [cs.CL]
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 622–628. <https://doi.org/10.18653/v1/N19-1063>
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT\* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Aparna Mitra. 2003. Establishment size, employment, and the gender wage gap. *Journal of Socio-Economics* 32 (07 2003), 317–330. [https://doi.org/10.1016/S1053-5357\(03\)00042-8](https://doi.org/10.1016/S1053-5357(03)00042-8)
- Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 174–184. <https://doi.org/10.18653/v1/P18-1017>
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, 1–17. <https://doi.org/10.18653/v1/S18-1001>
- Saif Mohammad and Peter Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29 (08 2013). <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Amit Moryossef, Roei Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 49–54. <https://doi.org/10.18653/v1/W19-3807>
- Robert Munro and Alex (Carmen) Morrison. 2020. Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2011–2017. <https://doi.org/10.18653/v1/2020.emnlp-main.157>
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5356–5371. <https://doi.org/10.18653/v1/2021.acl-long.416>
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1953–1967. <https://doi.org/10.18653/v1/2020.emnlp-main.154>
- Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2398–2406. <https://doi.org/10.18653/v1/2021.naacl-main.191>
- A. Paivio, J. Yuille, and S. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology* 76 1 (1968), Suppl:1–25.
- Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2799–2804. <https://doi.org/10.18653/v1/D18-1302>
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. <https://doi.org/10.18653/v1/N18-1202>
- Marcelo Prates, Pedro Avelar, and Luis Lamb. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications* 32 (05 2020). <https://doi.org/10.1007/s00521-019-04144-6>
- Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 69–75. <https://doi.org/10.18653/v1/W19-3810>
- Yusu Qian. 2019. Gender Stereotypes Differ between Male and Female Writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 48–53. <https://doi.org/10.18653/v1/P19-2007>
- Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. In *Proceedings of the 57th Annual Meeting of the Association for Computational*



- Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 223–228. <https://doi.org/10.18653/v1/P19-2031>
- A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 140:1–140:67. <http://jmlr.org/papers/v21/20-074.html>
- Anil Ramakrishna, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. 2015. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1996–2001. <https://doi.org/10.18653/v1/D15-1234>
- Anil Ramakrishna, Victor R. Martinez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1669–1678. <https://doi.org/10.18653/v1/P17-1153>
- Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating Gender Bias in Hindi-English Machine Translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Online, 16–23. <https://doi.org/10.18653/v1/2021.gebnlp-1.3>
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 463–473. <https://doi.org/10.18653/v1/P18-1043>
- Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during Speed-Quality optimization in Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 99–109. <https://doi.org/10.18653/v1/2021.acl-short.15>
- Barbara Risman. 2004. Gender As a Social Structure. *Gender & Society - GENDER SOC* 18 (08 2004), 429–450. <https://doi.org/10.1177/0891243204265349>
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 74–79. <https://doi.org/10.18653/v1/W17-1609>
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 8–14. <https://doi.org/10.18653/v1/N18-2002>
- Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Tamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France, 126–131. <https://www.aclweb.org/anthology/2020.trac-1.20>
- Magnus Sahlgren and Fredrik Olsson. 2019. Gender Bias in Pretrained Swedish Embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press, Turku, Finland, 35–43. <https://www.aclweb.org/anthology/W19-6104>
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2699–2712. <https://doi.org/10.18653/v1/2020.acl-main.240>
- Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1146–1151. <https://doi.org/10.3115/v1/D14-1121>
- Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2329–2334. <https://doi.org/10.18653/v1/D17-1247>
- Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7724–7736. <https://doi.org/10.18653/v1/2020.acl-main.690>
- Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural Machine Translation Doesn’t Translate Gender Coreference Right Unless You Make It. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for

- Computational Linguistics, Barcelona, Spain (Online), 35–43. <https://aclanthology.org/2020.gebnlp-1.4>
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics* 9 (08 2021), 845–874. [https://doi.org/10.1162/tacl\\_a\\_00401](https://doi.org/10.1162/tacl_a_00401) arXiv:[https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\\_a\\_00401/1957705/tacl\\_a\\_00401.pdf](https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00401/1957705/tacl_a_00401.pdf)
- Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature* 507 (03 2014), 9. <https://doi.org/10.1038/507009a>
- Ben Schmidt. 2015. Rejecting the gender binary: a vector-space operation. *Ben’s Bookworm Blog* (2015).
- João Sedoc and Lyle Ungar. 2019. The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 55–61. <https://doi.org/10.18653/v1/W19-3808>
- Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 325–344. <https://aclanthology.org/2021.emnlp-main.28>
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3239–3254. <https://doi.org/10.18653/v1/2020.findings-emnlp.291>
- Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2383–2389. <https://doi.org/10.18653/v1/2021.naacl-main.189>
- Artūrs Stefanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online, 629–638. <https://aclanthology.org/2020.wmt-1.73>
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1679–1684. <https://doi.org/10.18653/v1/P19-1164>
- Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. arXiv:2104.07505 [cs.CL]
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. <https://doi.org/10.18653/v1/P19-1159>
- Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *ArXiv abs/2102.06788* (2021).
- Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising. *Queue* 11, 3 (mar 2013), 10–29. <https://doi.org/10.1145/2460276.2460278>
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* 37, 2 (June 2011), 267–307. [https://doi.org/10.1162/COLI\\_a\\_00049](https://doi.org/10.1162/COLI_a_00049)
- Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.), 13209–13220. <https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html>
- Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2020. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 125–138. <https://aclanthology.org/2020.gebnlp-1.11>
- Andrew Tsou, Mike A Thelwall, P. Mongeon, and Cassidy R. Sugimoto. 2014. A Community of Curious Souls: An Analysis of Commenting Behavior on TED Talks Videos. *PLoS ONE* 9 (2014).
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3003–3008. <https://doi.org/10.18653/v1/D18-1334>
- Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 37–42. <https://doi.org/10.18653/v1/P19-3007>
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias.

arXiv:2004.12265 [cs.CL]

- Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. <https://aclanthology.org/L18-1445>
- Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45 (02 2013). <https://doi.org/10.3758/s13428-012-0314-x>
- Zeeraq Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. <https://doi.org/10.18653/v1/N16-2013>
- Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered Ambiguous Pronoun (GAP) Shared Task at the Gender Bias in NLP Workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 1–7. <https://doi.org/10.18653/v1/W19-3801>
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics* 6 (2018), 605–617. [https://doi.org/10.1162/tacl\\_a\\_00240](https://doi.org/10.1162/tacl_a_00240)
- Melvin Wevers. 2019. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Florence, Italy, 92–97. <https://doi.org/10.18653/v1/W19-4712>
- John E. Williams and Deborah L. Best. 1990. Measuring sex stereotypes: a multination study. Sage, Newbury Park, Calif.
- B. Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018).
- Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, YuSheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. CPM: A Large-scale Generative Chinese Pre-trained Language Model. *CoRR* abs/2012.00413 (2020). arXiv:2012.00413 <https://arxiv.org/abs/2012.00413>
- Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2896–2907. <https://doi.org/10.18653/v1/2020.acl-main.260>
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. <https://doi.org/10.18653/v1/N19-1064>
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. <https://doi.org/10.18653/v1/D17-1323>
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 15–20. <https://doi.org/10.18653/v1/N18-2003>
- Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4847–4853. <https://doi.org/10.18653/v1/D18-1521>
- Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5276–5284. <https://doi.org/10.18653/v1/D19-1531>
- Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1651–1661. <https://doi.org/10.18653/v1/P19-1161>