Theories of "Gender" in NLP Bias Research

Hannah Devinney
Umeå University
Umeå, Sweden
hannahd@cs.umu.se

Jenny Björklund
Uppsala University
Uppsala, Sweden
jenny.bjorklund@gender.uu.se

Henrik Björklund
Umeå University
Umeå, Sweden
henrikb@cs.umu.se

ABSTRACT

The rise of concern around Natural Language Processing (NLP) technologies containing and perpetuating social biases has led to a rich and rapidly growing area of research. Gender bias is one of the central biases being analyzed, but to date there is no comprehensive analysis of how "gender" is theorized in the field. We survey nearly 200 articles concerning gender bias in NLP to discover how the field conceptualizes gender both explicitly (e.g. through definitions of terms) and implicitly (e.g. through how gender is operationalized in practice). In order to get a better idea of emerging trajectories of thought, we split these articles into two sections by time.

We find that the majority of the articles do not make their theorization of gender explicit, even if they clearly define "bias." Almost none use a model of gender that is intersectional or inclusive of nonbinary genders; and many conflate sex characteristics, social gender, and linguistic gender in ways that disregard the existence and experience of trans, nonbinary, and intersex people. There is an increase between the two time-sections in statements acknowledging that gender is a complicated reality, however, very few articles manage to put this acknowledgment into practice. In addition to analyzing these findings, we provide specific recommendations to facilitate interdisciplinary work, and to incorporate theory and methodology from Gender Studies. Our hope is that this will produce more inclusive gender bias research in NLP.

KEYWORDS

natural language processing, gender bias, gender studies

ACM Reference Format:
Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3531146.3534627

1 INTRODUCTION

Algorithmic fairness and "social bias"[1] are matters of increasing concern in the field of Natural Language Processing (NLP). The field's language models encode human prejudices and stereotypes including gender biases (see, e.g., [3, 6, 23]). Although there are meta studies on how *bias* is (or fails to be) theorized and operationalized, there is no corresponding work on how *gender* is theorized and operationalized in NLP bias research. This article seeks to fill this gap: expanding on Cao and Daumé III [7], we survey 176 articles concerned with identifying and/or mitigating "gender bias" in NLP.

We question both how papers discuss and/or define "gender," and how these definitions are implemented. In order to investigate this, we read the articles that make up our survey with the following research questions in mind: Is a theory of gender discussed, and if so, which one(s)? Do these theories draw from literature outside of NLP, such as Feminist, Gender, and Queer Studies? Where a theory of gender is not discussed, what does the underlying theory behind the method seem to be? How do gender theorization and operationalization connect to definitions and measures of "bias"?

2 BACKGROUND

2.1 Related Work

Blodgett et al. [2] survey articles analyzing and/or mitigating "bias." They find that these articles tend to lack normative reasoning; specificity regarding what "bias" is and who it harms; and grounding in theories from outside mainstream NLP. Additionally, proposed methods "are poorly matched to their motivations."

Focusing specifically on gender bias in NLP, Sun et al. [49] categorize approaches to detecting and mitigating bias. They note limitations of these approaches, including that we have little idea how they behave at scale, because many focus on small parts of larger systems and are only verified for limited applications. Cao and Daumé III [7] study cisnormativity[2] in published NLP papers, focusing in particular on coreference resolution. They find that pronouns other than *he* and *she* are rarely considered, and social or personal gender is typically not distinguished from linguistic or grammatical gender. Savoldi et al. [47] address how gender bias is conceptualized in machine translation. Their findings emphasize the need for understanding both the relationships between gender and language and the ways different factors can contribute to gender bias.

All of these articles call for more interdisciplinarity. We echo this and provide some specific recommendations to facilitate such work in Section 5.4. We also discuss incorporating theory and methodology from the fields of Feminist, Gender, and Queer Studies into NLP research.

2.2 Gender

The "Folk" Model. Discussing the operationalization of gender in Human Computer Interaction research, Keyes [34] describes the "folk understanding" of gender, where gender is derived from sex

[1] Prejudice and/or stereotyping against particular social groups, which may result in direct or indirect discrimination.

[2] The assumption that all people are cisgender, i.e. that their gender identity matches the sex they were assigned at birth.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3534627

and the two are often conflated. In this model, gender is *binary*, *immutable*, and *physiological*: there are two categories (man or woman), a person cannot alter their assigned category, and assignment is based on "externally expressed physical characteristics" [34]. This corresponds in some ways to Butler's idea of a "stable gender" in the "heterosexual matrix" - sex entails gender entails desire towards the "opposite" category [5].

This model is not accurate: "sex" is not binary (intersex people exist, see [22], among others); gender is neither immutable nor binary (trans and nonbinary people exist, and many cultures worldwide recognize more than two genders); and people do not actually assign gender to others based on physiology [34].

The "folk" model erases the existence of trans and nonbinary people. This erasure from what is "acceptable" or "normal" has material and often violent consequences for trans people [31, 48], who are not seen as having an "intelligible" gender by the people and systems they encounter. It additionally presents problems for cis people, notably cis women, reducing them to a shallow stereotype which may be at odds with their lived experiences, and defining them by their bodies (a form of objectification). This model is at odds with "fair" NLP systems, because it is guaranteed to exclude vulnerable populations from its reasoning.

Gender Performativity. In Feminist and Gender Studies, gender is understood as a social construction that varies by culture. "Sex" and "gender" are decoupled. One of the most prevalent ways of thinking about this decoupling is through *gender performativity* [5].

Gender performativity means that we construct gender via discursive practices: gender is what one *does* rather than what one *is*. Repeated acts and interactions over time create our shared understandings of gendered categories, how we ourselves fit (or do not fit) into them, and how we categorize others. This theory rejects a stable gender: both as something culturally constant (acts may be interpreted differently in different times and places) and as something that is a constant property of individuals (a person's performativity may change).

Language is a part of gender performativity, and – importantly for bias research – a key part of how we transmit and maintain stereotypes [40], (re)produce meaning [25], and navigate systems of power. On the level of interaction between individuals, language acts can be used to accomplish particular goals, known in pragmatics as speech act theory. These acts may communicate intent ("Please close the window.") or actively change the state of the world ("I now pronounce you married."). Language acts can be part of performing gender. For example, introducing yourself with the name and pronouns you would like to be called contributes to how your gender performativity is perceived [10]. Language data for NLP are typically large corpora, and in aggregate do not reflect *individual* gender performances and experiences, but rather the production of gendered categories.

Although Butler's gender performativity is not the only possible model for gender (see chapter four of Connell and Pearse [9] for examples), it is one we view as fruitful when doing gender bias research in NLP. The discussions and recommendations in this article are guided by it.

As gender is conceived differently in different contexts, we must also question whether gender can be considered as a variable independent of all others. Gender intersects with other power asymmetries, such as race, class, sexuality, and (dis)ability (see, e.g., [13, 39]), and we must account for the specific, intersectional locations of minoritized groups in NLP bias work.

3 SURVEY

3.1 Method

We collected papers concerned with "gender bias" over two phases. The first phase, collected in June and July of 2020, comprised 126 papers. We collected these from existing bibliographies of surveyed papers provided in [2, 18]. These cover papers "analyzing 'bias' in NLP systems ... [restricted] to papers about written text only" [2]. Based on titles and abstracts, we selected papers where gender was relatively significant: either the specific focus or used as a running example for more general tasks, such as bias measurement or mitigation. This yielded 115 titles, which we assume concern *bias in NLP* on some level based on their presence in Blodgett et al. [2]. An additional 11, which fit both the NLP-bias and gender-focus criteria, were added from the bibliography of Dinan et al. [18]. All of these papers are included in the first round of the survey.

In the second phase, we collected an additional 90 papers in September of 2021 following the method used by Blodgett et al. [2], by searching keywords "gender," "bias," and "NLP"/"Natural Language Processing" over several databases (the ACL anthology[3], Google Scholar[4], and arXiv[5]) and filtering for those published in 2020 and 2021. Five of these were later rejected either as duplicates or because they did not actually address gender bias in NLP. We randomly sampled 50 of the remaining papers for the second round of the survey. The full list of articles is in Appendix A.

We read the articles with the goal of identifying how they theorize and operationalize *gender* and *gender bias*. We also categorized each paper according to language investigated; what NLP technology the paper concerns; and whether gender bias is the main focus or a use case. The headings are summarized in Table 1. Papers can belong to zero categories or multiple categories under some headings, while others expect yes/no answers. For the latter, additional notes were taken (e.g. where gender bias was *not* the focus, we noted if gender is presented in combination with other biases, demonstrating a general-purpose technique, etc.).

In our analysis of the results, we focus on the categories surrounding gender (how it is theorized and operationalized; and if and how it is problematized).

3.2 Results

3.2.1 First Round (126 papers). Our findings with respect to **Technology** and **Gender Bias** are consistent with Blodgett et al.'s analysis of the NLP tasks covered and bias categories, respectively. The majority of the papers we survey (82,5%) deal only with English; seven papers (5,5%) concern machine translation; four (3,1%) are about a single non-English language; and the remainder compare multiple languages.

[3]https://aclanthology.org/
[4]https://scholar.google.com/
[5]https://arxiv.org/

Table 1: Categorization Schema for Surveyed Papers

Heading	What is of concern?
Gender	How is gender theorized?
Gender Bias	How is gender bias theorized and/or measured?
Technology	What is the technology of interest?
Gender Proxies	How is gender operationalized?
Gender Focus	Is gender bias the focus of the paper?
Language	What language(s) are investigated?
Binary Problematized	Does the paper acknowledge that gender is not a binary attribute?

Table 2: Gender. How is gender theorized across papers? Note that papers may be included in multiple categories, so counts do not sum to 126 (round 1) or 50 (round 2).

Gender	Inclusion Criteria	Round 1	Round 2
binary	considers only men and women	118	41
essentialist	makes specific claim about fundamental difference	12	1
neutral	includes concept of 'neutrality' (mixed groups or unknowns)	7	6
nonbinary-inclusive	thinks & operationalizes beyond two genders	1	4
social construct	acknowledges gender varies/is a social construction	2	4
trans-inclusive	thinks & operationalizes beyond cis men and women	1	7
undefined	no framework	7	2
underspecified	does not explicitly/clearly define gender	62	24

Table 3: Gender Proxies. What data is consider representative of gender? Note that papers may use multiple strategies, so counts do not sum to 126 (round 1) or 50 (round 2).

Gender Proxy	Inclusion Criteria	Round 1	Round 2
pronouns	uses gendered pronouns e.g. *he, she, ze, they*	34	18
annotation	relies on annotated gender labels (any kind)	12	6
word lists	compiles (unordered) lists of gender-associated terms	31	15
word pairs	matches pairs of "equivalent but for gender" terms	26	9
sentence pairs	like word pairs, but only use full sentences	2	0
grammatical gender	uses grammatical (not lexical) gender e.g. morphological markers	6	4
names	uses first names	21	8
author gender	round 1: inferred by annotators; round 2: self-identification	5	1
photo (human-label)	uses profile photos (labeled by annotators)	3	1
photo (AGR)	uses profile photots (labeled by automated gender recognition)	3	0
unspecified or N/A	does not specify *or* does not use a proxy	11	2

Gender. Of the 126 papers surveyed, 77 (61,1%) are categorized as having **Gender Focus**, meaning they are specifically concerned with "gender bias" rather than using gender as an example bias. 45,2% of the articles were tagged neither *underspecified* nor *undefined*. These discuss or define "gender" to some extent within the text, although in many cases rather shallowly and without citations. 49,2% were tagged *underspecified*: these papers never outright discuss what "gender" is, and instead seem to take it for granted that the definition is obvious to the reader. A rough idea of their definition can be discerned by analyzing methods and results to see how gender is operationalized. The remaining 5,5% (7 papers) are tagged *undefined*. This indicates that not only is gender never discussed or defined, but also that the model used cannot be inferred. These papers are not tagged *binary* because by not (explicitly or implicitly) defining gender they do not meet the criterion for this tag, "considers (only) men and women."

In fact, regardless of whether or not gender is defined, the methods of most studies (118, or 93,6%) operationalize gender with a binary model that reflects the "folk understanding" of gender. This figure includes all of the 62 papers that are tagged *underspecified*. Most papers concerning English language technology use pronouns, first names, word pairs, lists of words, or some combination of these strategies to identify referent gender. Where the papers are concerned with author gender or image captioning, this information is generally labeled based on (binary) human annotation or automated gender recognition from profile photos, which means there is a very strong risk of misgendering [34].

Nearly every paper in this round fails to be inclusive of trans experiences, binary or nonbinary, in their methodology. 12 papers (9,5%) are tagged *essentialist*, indicating experimental design or analysis that, regardless of the intent of the authors, specifically *excludes* trans experiences by conflating gender presentation and

Table 4: Gender Bias. How is gender bias theorized and/or measured? Note that papers may use multiple strategies, so counts do not sum to 126 (round 1) or 50 (round 2).

Gender Bias	Inclusion Criteria	Round 1	Round 2
activities	measures relationships between verbs and gendered nouns	1	0
allocation	concerned with allocational parity (any kind)	1	0
associations	tests associations by comparing between gendered categories	26	16
clustering	word embeddings: do non-definitionally gendered words cluster?	5	0
counterfactual	a decision should be independent from gender attributes	4	2
data imbalance	the problem is in balance in number/type of datapoints	3	5
demographic parity	unequal representation in model/results	2	9
denigration	addresses misrepresentation (denigration/hate speech)	5	1
downstream task	compares performance on a downstream task	2	3
erasure	addresses lack of representation	1	1
gender unaware	gender should not be predictable	2	3
gender vector	word embedding: identify "gender" vector & locate words	17	7
multiple	explicitly considers several dimensions of bias	2	2
occupations	uses "occupation" titles as a way to measure bias	22	16
performance parity	correctness ratio, incorporates > 2 genders	11	6
perforamnce parity (M/F)	correctness ratio for women vs men	51	21
real world distributions	compare representation to 'real' statistics	1	2
sentiment	sentiment detection specifically	2	0
stereotype	addresses misrepresentation (stereotypes)	5	15
translation accuracy	asks: is this gendered translation is correct?	4	5
undefined	does not define their metrics	1	0
underspecified	does not address what 'bias' or 'fairness' is	1	0
WEAT	uses WEAT (or SEAT) as a measure, specifically	17	7
word-pair direction	word embedding: difference in "equivalent" words	2	0

bodies and/or assuming some essential difference between 'men' and 'women' exists. These binary operationalizations, and their consequences, are discussed further in Section 4.1.

Some papers acknowledge the existence of nonbinary people but note that they are excluded by the methodology used (see Section 4.3). Finally, there is one paper, [7], which is trans-inclusive in theorization and methodology. Two other papers [33, 46] meet the criteria for the **trans inclusive** and/or **nonbinary inclusive** categories but do not explicitly extend this inclusivity to their analysis (see Section 4.4).

3.2.2 Second Round (50 papers). We find some slight improvement with regards to gender inclusivity in both theory and practice over the time between the first and second rounds. Only 41 papers (82%) are tagged "binary," a noticeable improvement from the first round (93,6%); and 9 papers (18%) are actively inclusive of multiple genders at least in their theoretical positioning. Unfortunately, more than half of these inclusive papers face limitations in actually operationalizing more than two genders (for example, limited nonbinary representation in participants or data). We do not find any improvement in the number of "underspecified" papers, where gender is not explicitly theorized (48%, compared to 49,2%); although those papers that *do* include more detail on their theoretical grounding.

There is a large and encouraging increase in papers dealing with languages other than English: 42% (of which about a quarter deal with machine translation), which is more than twice as many as in the first round, although it should be noted this evidence is somewhat circumstantial. One problem that is often focused on in these papers is that methods developed to identify bias in a Western, English-language context do not perform as well in other contexts, for example due to grammatical gender or differing stereotypes.

4 DISCUSSION

Our findings are largely similar across both rounds of the survey. Thus, the remainder of our presentation draws examples from and discusses themes and patterns found throughout all 176 papers.

Generally, we find that papers concerned with "gender bias" look *only* at gender, attempting to separate it out from other variables. Although papers concerned with "bias" in a general sense are more likely to analyze multiple sources of bias, these sources are almost always analyzed separately, i.e. not in an *intersectional* manner. Papers that do attempt intersectional analysis primarily explore race and gender in a U.S. context.

Nearly every paper surveyed seems to make the assumption that gender is binary and immutable (similar to findings in [34] and [7]). Social gender and linguistic[6] gender are also consistently conflated. Many papers do not ever define gender: the title and/or abstract notes that the subject of the paper is "gender bias" and the reader must conclude that "gender" is the taken-for-granted difference between two groups, implicitly cis men and cis women. This lack of clear definition does not change across the two rounds of the survey, even as there are some changes in how gender tends to be operationalized.

[6]Both grammatical and lexical-semantic.

Although the differences between the first and second round of the survey suggest a trend towards increased inclusion of nonbinary language, as well as an increase in acknowledging the necessity of this inclusion, the field still overwhelmingly ignores non-normative genders. We discuss this, as well as some of the roadblocks faced by researchers trying to include nonbinary people and language in their work, in section 4.3.

4.1 Cisnormativity

We find that the prevalence of cisnormative assumptions constrains thinking about gender (as in [34]). The assumption of binarity becomes a standard: some authors explain that "for gender in semantics, we follow the literature and address only binary gender" [55]. This culture of internal citation, discussed in [2], reinforces the lack of engagement with the theory of gender even in research that is *about* gender. More worryingly, the binarity becomes unchallengeable, something that *must* be assumed in order to fit in with the established literature: Dayanık and Padó "assume a binary gender classification (male/female) to be compatible with existing datasets" and then immediately assert that this choice "should not be understood as a rejection of non-binary gender" [14, p. 52]. Regardless of the reasoning (the convenience of existing models, or perhaps an expectation in the field that all new datasets will resemble the old datasets), this remains a clear rejection of nonbinary genders and people.

The assumption of some inherent difference between "the" two genders, has been problematized, repeatedly, in many fields (e.g. [5, 21, 28]) but here NLP seems to fall behind. Even in papers where features other than gender (such as race or religion) are considered to have multiple categories, gender is consistently taken as "the binary case," often uncritically. Many measures to both calculate and mitigate gender bias rely on the binary, and may treat binarization as a limitation for other features, without considering it to be one for gender. The binary can be seen in common methodological choices (section 4.2), such as collecting "pairs" of words, the meanings of which differ only in gender. Masculine and feminine genders are also generally presented as "opposites" which can be "swapped" for each other; placed on each end of a linear scale of bias; or used to define vector space directions.

Although binary gender models *can* be inclusive of binary trans experiences of gender, this is not often the case in the papers we surveyed. Gender is frequently tied to bodies, with language such as "males and females" (see section 5.3). The implicit assumption that everyone is cisgender is evident in both methods and analysis. This exclusion of nonbinary genders in analyzing language technology can contribute to harms against nonbinary people, such as erasure; misgendering; derogatory associations; and allocative harms such as automatically rejecting applications to jobs or public services [15].

Cisnormative Methodology. Cisnormative assumptions in methodologies often determine both research questions and methods, including the proxies chosen to identify gendered categories or to gender individuals (section 4.2). Leaving out nonbinary people and genders is the most common, but not the only, way that cisnormativity is evident in most methods surveyed.

Methods that would otherwise be inclusive of binary trans people and experiences are rendered trans-exclusive by the choice of word lists or word pairs used. The word lists created by Zhao et al., which consist of "words associated with gender by definition" [54] are used in several papers to calculate gender in word embeddings. The two lists include terms like *uterus, penis, testosterone* and *ovarian cancer*. Although not based on an individual's appearance, these words incorporate the physiological assumption of the folk model.

Another problem is assuming these physiological words can be "paired" in the same way that *man:woman* or *king:queen* can be. For example, the analogy "*she* to *ovarian cancer* is as *he* to *prostate cancer*" is categorized as "gender appropriate" [3], which erases those for whom these body parts and pronouns do not 'match.' These are not equivalent diseases. Ovaries and prostates are not "equivalent" body parts (a questionable concept) differing only by "sex".

Cisnormative Analysis. Independent of method, the analysis of results can erase the lived reality of many trans people, e.g., by defining sentences such as *he gave birth* as "meaningless" [53], "nonsensical" [49], or even "biologically . . . inaccurate facts" [36]. *He gave birth* is not a statistically likely sentence in English, but it is neither meaningless nor nonsensical. It is certainly not "biologically inaccurate" – trans men, and other people who use he/him but were assigned female at birth, can and do give birth in increasing numbers [45].

The idea of semantic incorrectness *must* be understood in context, and it is necessary to draw a distinction between "flipping" gendered words that refer to *specific people*, words that refer to *unknown persons*, and words that refer to *groups*. Consider the argument for counterfactual data substitution given in (1):

(1) Flipping a gendered word when it refers to a proper noun such as *Queen Elizabeth* would result in semantically incorrect sentences. [38]

True, it would be misgendering to use *he* in reference to the real person Queen Elizabeth II of England, or to call her a *king*. However, it is not universally true to say that it is "semantically incorrect" for anyone named Elizabeth to co-refer with *king* and/or *he* (or *monarch*, *they*, or *ze*). Coreference in English depends on context, both at the level of a particular conversation and at the level of world knowledge [1]. Although names are associated with gender, they do not exclusively correspond to particular pronouns, and a system that assumes they do will ultimately commit errors including misgendering.

Gendered pronouns and nouns do not automatically correspond to bodies, nor how we dress and adorn those bodies (aspects of gender presentation), as implied in (2):

(2) . . . one would expect 'beard' to be associated with male nouns and 'bikini' to be associated with female nouns, and preserving such gender biases would be useful . . . for a recommendation system [32]

There may be a statistically strong *association* between masculine nouns and beards, and beards are often a part of masculine performativity, but that does not make it a foolproof indicator of gender for an individual. Analyzing it as such erases trans and gender

non-conforming people, and cis women with facial hair. Preserving this association in recommendation systems can be psychologically harmful for trans people who may be misgendered by ads. The particular example of *beard* with "male nouns" is also ironic, as *beard* can in specifically refer to a woman whom a gay man is dating to hide his sexuality – making it a feminine noun in these cases.

Factoring in this context is essential for analyzing NLP in a way that is accepting of gender diversity. We must know both what world (Queen Elizabeth of England, or King Elizabeth of the *Pirates of the Caribbean* franchise?) and what discourse (the beard on someone's face, or the beard they are dating?) make up the context to make a judgement about how meaningful a particular sentence or association is.

4.2 Gender Proxies

One important design choice in any system for measuring or reducing gender bias is the strategy used to find evidence of gender, i.e. what information will be used as a proxy for gender. Papers concerned with "gender bias" in NLP are, broadly speaking, either concerned with bias based on who *authored* the text, or based on who is *referred to* in the text. Within both of these categories, asking the right questions about power and bias requires knowing whether to look at the gender of *individuals* or gendered *groups*.

Individuals. For cases where a system may be biased against authors of a particular gender, it is imperative to keep in mind that each of these authors is a human with agency – including over their gender identity and expression. Larson provides four key guidelines for the responsible use of (author) gender as a variable in NLP, namely: *i*) make theory of gender explicit; *ii*) avoid using gender unless necessary; *iii*) make category assignment explicit; and *iv*) respect persons [37].

We discuss *i* in Section 5.1. Clearly in research regarding gender bias it can be necessary to use gender as a variable, covering *ii*. Therefore we primarily discuss proxy choice for individuals with respect to *iii* and *iv*. Larson makes it very clear that, where possible "participant self-identification should be the gold standard for ascribing gender categories" [37, p. 7] and discusses some of the specific challenges for obtaining this self-identification (e.g. the options do not adequately describe the participant's gender, or someone may choose not to provide their gender on platforms where it is not required). In cases where self-identification is not possible, the procedure for assigning participants to gendered categories must both be clear, reproducible, and respectful.

The most common proxies for assigning gender to an individual author are *first names*, *profile photos*, or a combination. In most cases in the papers we survey, automated methods such as facial recognition or gendering of first names based on statistics are used. These automated methods do make category assignment in theory reproducible, but fail the fourth requirement: "respect persons". The use of facial recognition for automated gender "recognition" (i.e. assignment) is trans-exclusive: "essentialising the body as the source of gender" [34, p. 11] and typically representing gender as a binary attribute. Additionally, it has been shown that commercial gender classification systems are systematically worse for people of color, in particular for dark-skinned women [4]. Many people also do not use photos of their own face as a profile photo for a variety of reasons.

Assigning gender by first name has particular pitfalls at the level of the individual. Among the surveyed papers, all strategies for using this proxy assume that gender is binary. Thus, nonbinary people will be consistently misgendered.

These strategies assume that a name has a "gender" if a majority of the population with that name belongs[7] to that gender category. There is a particular cultural bias in this assumption: while common for many Western cultures, reliably gendered names are not a universal constant. The gendered association of a particular name can change over time or by culture/language (e.g. "Jean" in French vs English). Limiting a study to names we can "confidently" gender risks skewing results towards the cultural majority for whom this data is available.

Groups. At a group level, first names have some more concrete advantages: it is possible to use them in intersectional analysis (by choosing names that are strongly associated with e.g. Black women vs white women in a certain generation in the U.S.), and in aggregate use they do not have the exact same risks of misgendering individuals. However, this method remains unreliable: not all languages and cultures have strongly gendered names, and picking only strongly gendered names therefore likely favors certain demographics.

Word pairs and lists (see Section 4.1) occasionally also use limited first names. First names are not gendered by definition and are positioned in particular generations, races, and religions. As a proxy, they must be used with many other names.

Word pair strategies are typically binary and the "semantic equivalence" of many pairs can be called into question. Consider *bachelor:spinster* – *spinster* is pejorative while *bachelor* is not; and there is no such thing as a *spinster's degree*.

As most of the papers deal only with English, grammatical gender was significantly less common and generally was not used as a proxy for social gender.[8] Where grammatical gender was a factor, articles dealt with alignments of grammatical and social gender (e.g. in generating accurate machine translation results) or developing methods to calculate gender bias in languages with grammatical gender.

In languages like English, where gender is relatively reliably marked in third person pronouns (*he, she, they, ze,* etc.), pronouns are one of the better proxies for gender, particularly when more than *he* and *she* are used. These must still come with a caveat: they are *associated* with particular genders, but not every referent of a given pronoun belongs to the same gender category.

4.3 Binary Gender as a Limitation

Around one fifth of the papers surveyed in the first round include some textual acknowledgement that gender is not binary, but do not extend their methods to reflect this definition. This rate increases to about a third of the papers in the second round. These acknowledgements are typically limited to a sentence or two, and

[7] By sex assigned at birth, or legal gender: methods for assigning "belonging" vary.
[8] For a discussion on the division(s) between grammatical and social forms of gender, see [1].

are usually included in the limitations section or formatted as footnotes. Disconnects between theory of gender (where specified) and the method chosen can give the impression that inclusivity was an afterthought rather than a limitation. For example Sun et al. note that "Non-binary genders ... should be considered in future work", but define gender bias as "the preference or prejudice toward one gender over *the other*" (emphasis added) [49].

It is relatively common to consider inclusion of genders outside the binary to be a problem for the future. This should ideally be accompanied by a specific commitment to participation in doing the necessary work, such as "we plan to extend ... genders (e.g. agender, androgyne, trans, queer, etc.)" [35].

A coherent methodology aligns theory and method, or indicates where the theory cannot fully be operationalized, for example due to limitations of the technique, as in (3), or the data, e.g. [50] where "only binary gender statistics" are available. However, unspecified "technical limitations" may give the impression that the researchers did not attempt to develop inclusive methods.

(3) Our method unfortunately could not take into account nonbinary gender identities, as it relied on she/her and he/his pronouns, and could not easily integrate the singular they/them, nor could we find sufficient examples of ze/zir or other nonbinary pronouns in our data. [8]

Encouragingly, there seems to be an increase in articles where the authors attempt to work around these limitations, one step at a time. Yeo and Chen [52] and Vig et al. [51], for example, both incorporate experiments involving singular *they* but find several confounding factors (including the tendency for *they* to be processed as plural) that prevent these results from being analyzed in the same way as their binary experiments. Ramesh et al. provide extensive bias- and gender- statements, and attempt to adapt existing metrics to include more than just masculine and feminine; but are somewhat limited by the binary nature of many of these metrics [44]. Combined with the observed increase in trans-inclusive methodologies, further discussed in the following section, it is clear that there is a desire to include minoritized groups despite how this may complicate research methods.

4.4 Trans-Inclusive Methodology

In their work on gender bias in coreference resolution, Cao and Daumé III present an encouraging example of what gender-inclusive NLP can look like [7]. A detailed theoretical grounding on the nature of both "gender" (in society and in language) and gendered harms (including trans-specific harms stemming from exclusion) motivates and supports the paper. They focus on the data side of coreference resolution tasks, and a key aspect of their method is how they source data: both obtaining permission from authors and working with stakeholder communities to develop a dataset "by and about trans people".

Two papers have at least one annotator identifying themself as nonbinary or "other" [33, 46]. These papers meet the criteria for the *trans* and/or *nonbinary inclusive* tags, but do not necessarily analyze gender or their results in ways that account for this gender diversity. Sap et al.'s annotation method leave spaces for trans and gender nonconforming issues, via free-text responses [46], but this relies on annotators being familiar with trans people, issues, and stereotypes.

Papers in the second round (published after the spring of 2020) increasingly feature trans-inclusive methodologies. Hansson et al. [26] present a Wino-gender style for Swedish which incorporates the neutral third-person singular neo-pronoun *hen*. Dinan et al. [19] offer a framework for categorizing gender in dialogs along multiple dimensions, classifying genders as {masculine, feminine, neutral, unknown}. Munro and Morrison [43] address the performance gap between *his* and both *hers* and *theirs* in part of speech taggers. In addition to papers included by random sample, we are aware of several other papers with trans- and nonbinary-inclusive methodologies.

5 RECOMMENDATIONS

Although it will never be possible to perfectly model the world in all its complexity, simplification of identity aspects such as gender willfully ignores the existence and experience of marginalized groups and can contribute to material harms done to already vulnerable populations. As researchers and practitioners concerned with the risks evident in biased NLP tools, we must be attentive to these harms and incorporate that knowledge into our work.

To address this, we make a series of recommendations for addressing theories of gender, choosing bias measures, writing about gender in a respectful manner, and incorporating feminist research methodologies. In addition to these recommendations, we provide the start of an Open Bibliography, *Gender Theory for Computer Scientists*[9]. This is a community resource for collecting readings and other reference material we find useful for addressing questions of gender applied to NLP.

5.1 Make Theorization of Gender Explicit

First, we extend Larson's first guideline for using gender as a variable ("make theory of gender explicit") to all work on gender bias in NLP. The way that the operationalization of gender is grounded in this theory should also be specifically discussed. This could be considered to already be a part of Blodgett et al.'s call for explicit definitions of bias, which includes both how particular systems are harmful and *to whom*. However, it is not enough to, for example, claim a particular behavior is harmful to women without some understanding of who this encompasses, because "women" may not be a coherent category (see, among others, [5, 30]). Grounding the conceptualization of gender in theory can also help specify the scope of the work and allows for better comparisons between studies.

At the Second Workshop on Gender Bias in Natural Language Processing, a *bias statement* was required for all submissions [12]. Authors were asked to clarify what harms, against whom, their research was concerned with. These statements were evaluated by reviewers from the Humanities and Social Sciences. In addition to providing an opportunity for reflexive analysis on the parts of the authors, these statements make normative assumptions explicit to reviewers, helping them evaluate the research in context. For example, the bias statement in Falenska and Çetinoğlu [20] clearly defines the way that the 'bias' in question harms people, and how this connects to their lived experiences, and makes reference to

[9] https://github.com/hdevinney/open-gender-bib

relevant literature outside of NLP. They explain exactly why they have limited their work to binary gender, but continue to revisit the question of nonbinary inclusion throughout the paper.

Making normative assumptions explicit is helpful both for reviewers and readers to evaluate and make sense of research, and for researchers themselves to formulate the most appropriate research questions and methods for their work. The latter works best when the requirement becomes part of the culture of research, rather than a checkbox when submitting a paper. When the model of gender used is explicitly defined from the start, researchers can "check back" as they develop their methods to ensure alignment (see Section 5.4). For example, from the research side, when exploring gendered associations in news corpora for Devinney et al. [16], we realized that despite our attempts to analyze gendered categories, our dataset contained little to no nonbinary representation. This resulted in the "gendered, but not men or women" category being primarily about mixed-gender groups. Because we checked back and realized early in the process that our data and methods did not align, we were able to gather more text in a different corpus which contained sufficient nonbinary representation to allow for our desired analysis.

5.2 Understand Your Goal

The concept of "bias" is a complicated issue rooted in society and culture, and in order to make progress in correcting for unfairness in a technical domain it is essential that we understand what we are trying to accomplish. Just as bias can be defined in many conflicting ways [2], so can its absence. Technology attempting to reach this state is often referred to as "ethical" or "fair," but definitions vary. From a data science perspective, D'Ignazio and Klein [17] divide concepts around removing "bias" into two types. Those that *secure power* provide some short term solutions, but ultimately do not go far enough. Those that *challenge power*, however, "acknowledge structural power differentials and work towards dismantling them" [17, p. 60].

There are a number of measures of algorithmic fairness, including *fairness through unawareness*, *fairness through awareness*, *demographic parity*, and *counterfactual fairness*. They can broadly be classified as measures that work on a group level and measures that work on an individual level. (For a summary, see, e.g., [41].) There are also domain-specific measures of, e.g., how biased a particular word embedding is. In order to determine what measure is best suited for a particular project, we require clear goals. What harms are we trying to prevent and against whom? If we are trying to ensure fair representation, group level measures seem appropriate; whereas if we want to avoid unfair hiring practices in a system, an individual level measure might be better suited.

Bias in language is a complicated topic, and we must not overstate the efficacy of mitigation measures. Claims to "significantly reduce" or "eliminate" gender bias must acknowledge that this can only be known for the measures of bias tested. Claiming otherwise overreaches the scope of our research and misrepresents the field of algorithmic bias research as something that *has* a purely technical solution [24]. The fact remains that we (as researchers, and as a global society) do not know what a truly "unbiased" system is. Incremental work towards an unknown goal is difficult, so it can be useful to ask ourselves open questions, such as: In an unbiased world, what would an unbiased system look like? In our biased world, what does an unbiased system look like? Is the goal of NLP to understand and use language as humans *currently* do? Are there things we would like NLP to be better than us at?

5.3 Use Consistent, Respectful, and Accurate Language

One challenge we faced in conducting this survey was that, in addition to individual papers failing to define their theorization of gender, there is no standard vocabulary within the field of "gender bias in NLP". This makes comparing papers "within" this field difficult, but also presents difficulties for connecting to "outside" research. The vocabulary chosen to describe different gender categories, aspects of grammatical gender, and individual people is not consistent with related literature from the fields of linguistics and gender studies. It is necessary to decouple the various sociolinguistic aspects of gender, i.e. "the different ways in which gender can be realized linguistically" [7], in order to develop coherent methodologies and accurately analyze results.

In particular, we note the problematic use of the terms *male* and *female* in two circumstances: 1) as nouns describing people and 2) as adjectives describing grammatical or lexical gender. Although it is not grammatically incorrect in English to use these terms as nouns, the implications and context of this usage are important because our work purports to try to counter gender biases in language. Such language is regularly used to dehumanize vulnerable groups, including women, trans people of all genders, and people of color.

As adjectives describing grammatical or lexical gender, this is simply not the preferred terminology in linguistics. For languages with grammatical gender (i.e. noun classes defined by syntactic agreement) it may be traditional to refer to *feminine*, *masculine*, *neuter* (etc.) words; in other languages these classes are numbered [11].

We thus consider *male* and *female* to be inappropriate terms within NLP and suggest instead that we follow linguistics praxis when discussing grammatical and lexical gender, and gender studies when discussing people. This encompasses adjectives such as *masculine, feminine, genderless, nonbinary, unknown, neutral* and nouns such as *men, women, nonbinary people*. Where individual terms (e.g. pronouns) are investigated, the term can simply be specified without categorizing.

5.4 Use Feminist Research Methodologies

In addition to incorporating theories of gender from other fields, we recommend borrowing and adapting methodologies from Feminist research, including reflexive research and situated knowledges. When the model of gender used is explicitly defined from the start, researchers can "check back" as they develop their methods to ensure alignment. We also recommend collaboration between researchers and other stakeholders to open up new avenues of inquiry.

Situate Knowledge. NLP research has a tendency to use what Haraway describes as the "god trick" [27]. By aggregating enough language data, the logic goes, we can actually achieve the "view from nowhere" and thereby produce objective truth. But knowledge does not come from nowhere, it is produced by people who are *situated*

in particular locations, with different backgrounds, experiences, viewpoints, etc. These positions inform how we think and what we notice, and are always partial and subjective. By interrogating the world from our various positions, we produce partial and situated knowledges, which can be compared and combined to triangulate something that approaches "objective" truth. We can also reflexively analyze our research, to better understand how our research methods and goals align.

Positionality statements are one way to situate ourselves and contextualize our research, problems, data, and solutions. They are tools which allow us the opportunity to reflect on what we see (and what we might miss); assumptions we may be making; and potential gaps in our knowledge. Research lacking in reflexivity is "likely to reproduce the exact forms of social oppression" we are trying to prevent [24, p. 2].

Like explicit descriptions of theoretical groundings, when shared they can help the audience interpret how and why the author(s) developed their methods and reached their conclusions. For example, with respect to this paper: all three authors of this paper are white academics based in Sweden, with EU citizenship. This positionality influences how we think about, e.g., race or immigration. We can read and learn from others but we cannot directly know or access their experiences. Our academic backgrounds (in computer science, linguistics, and gender studies) and personal experiences both inform what we are likely to notice when analyzing "gender" in NLP. As a queer, gender non-conforming person, the first author is more likely to notice cisnormativity due to the extent that they confront and navigate it in their everyday life.

Although this reflexive examination of our positions and work should be a component of our research processes, it is not necessary to include it in reporting on that research. Particularly for minoritized researchers in a field where positionality statements are not an unremarked standard, it can be dangerous to make ones status as part of a minoritized group public. Therefore, it is *not* our recommendation that such statements be disclosed publicly; rather, we encourage researchers to consider their own positionalities at every step of the research process, and reflexively interrogate their research in light of these positionalities.

Work with Others. One way to access other positions and knowledges is collaboration: with researchers across disciplinary lines and with stakeholders outside of academia and industry. Stakeholders offer lived expertise and experience and minoritized groups (such as queer and trans people) in particular must be included in NLP research that concerns us. The phrase "nothing about us without us", originating in disability rights activism, applies. If we are the ones at risk, we must have a hand in shaping the solutions: beyond surface-level inclusion which fails to challenge the social order [29]. One place this applies directly is in collecting more data to compensate for sparsity, which must be done in collaboration with communities and with respect for autonomy. Being made visible and countable within the data, may both benefit minoritized groups and put them at risk.

Doing inter-disciplinary work requires time and deliberation. It specifically strives to "develop a joint understanding of the problem" and melds approaches from the researchers involved [42, p. 4]. Thus, much of this time is best spent explaining and listening, which allows all participants to make connections and figure out how best to blend them and reconcile any fundamental differences of opinion or understanding. We have found in our collaborative work that (global health situations allowing) sitting down together for longer periods of time leads to more progress, as this gives us opportunities to ask questions and make sure we understand the answers.

6 CONCLUSION

We surveyed nearly 200 papers concerned with gender bias in NLP. Throughout both time-spans surveyed, the vast majority of papers do not discuss how they theorize gender and/or operationalize gender following the cisnormative "folk model". However, differences between the two rounds suggest that there is more and more research inclusive of nonbinary genders, as well as an increase in papers noting that the binary model of gender is a methodological limitation. In analyzing some of the problems with these papers, we detail how they exclude trans people and experiences and the impact this has on NLP bias research as a whole.

Finally, we provide some recommendations for doing better gender bias research in NLP, such as explicitly defining gender (and using respectful language to do so); selecting methods that work well with this definition; and borrowing from feminist research methodologies. To help with the last aspect, we contribute the start of a community-sourced list of recommended texts and other resources concerning gender, language, and feminist theory.

Funding/Support. The authors declare no additional sources of funding.

REFERENCES

[1] Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics* 4, 1 (oct 2019), 27 pages. https://doi.org/10.5334/gjgl.721

[2] 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://aclanthology.org/2020.acl-main.485/

[3] 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 *Advances in Neural Information Processing Systems*. NeurIPS, Barcelona, Spain, 4356–4364. arXiv:1607.06520 http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-d

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, New York, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[5] Judith Butler. 1999. *Gender Trouble*. Routledge, New York.

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)* 356, 6334 (apr 2017), 183–186. https://doi.org/10.1126/science.aal4230

[7] 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

[8] Serina Chang and Kathleen McKeown. 2019. Automatically Inferring Gender Associations from Language. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Hong Kong, 5746–5752. https://doi.org/10.18653/v1/D19-1579

[9] Raewyn W. Connell and Rebecca Pearse. 2015. *Gender: In World Perspective*. Polity Press, Cambridge, UK.

[10] Kirby Conrod. 2020. How to do things with gender. https://www.youtube.com/watch?v=jVr8NJwcMH4 Talk given for English Language and Linguistics at the University of Kent.

[11] Greville G. Corbett. 2013. Number of Genders. In *The World Atlas of Language Structures Online*, Matthew S. Dryer and Martin Haspelmath (Eds.). Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/chapter/30

[12] Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster (Eds.). 2020. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online). https://www.aclweb.org/anthology/2020.gebnlp-1.0

[13] Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241–1299.

[14] Erenay Dayanık and Sebastian Padó. 2021. Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Online, 50–61. https://aclanthology.org/2021.wassa-1.6

[15] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings ofthe 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1968–1994. https://aclanthology.org/2021.emnlp-main.150/

[16] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 79–92. https://www.aclweb.org/anthology/2020.gebnlp-1.8

[17] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism*. The MIT Press, Cambridge, Massachusetts.

[18] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8173–8188. https://doi.org/10.18653/v1/2020.emnlp-main.656

[19] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23

[20] Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing Gender Bias in Wikipedia: Inequalities in Article Titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Online, 75–85. https://doi.org/10.18653/v1/2021.gebnlp-1.9

[21] Anne Fausto-Sterling. 1992. *Myths of Gender* (2 ed.). Basic Books, New York, New York.

[22] Anne Fausto-Sterling. 2000. *Sexing the Body* (1 ed.). Basic Books, New York, New York.

[23] 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 of the United States of America* 115, 16 (apr 2018), E3635–E3644. https://doi.org/10.1073/pnas.1720347115 arXiv:1711.08412

[24] Ben Green. 2019. "Good" isn't good enough. In *NeurIPS Joint Workshop on AI for Social Good*. NeurIPS, Vancouver, Canada, 7 pages. https://aiforsocialgood.github.io/neurips2019/accepted/track3/pdfs/67_aisg_neurips2019.pdf

[25] Stuart Hall. 2013. The Work of Representation. In *Representation*, Stuart Hall, Jessica Evans, and Sean Nixon (Eds.). Sage, 1–59.

[26] Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender Dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 452–459. https://aclanthology.org/2021.nodalida-main.52

[27] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. https://doi.org/10.2307/3178066

[28] Myra J. Hird. 2000. Gender's nature: Intersexuality, transsexualism, and the 'sex'/'gender' binary. *Feminist Theory* 1, 3 (2000), 347–34.

[29] Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12 (2021), 3539–3556. https://doi.org/10.1177/1461444820958725

[30] bell hooks. 2000. *Feminism is for Everybody*. Pluto Press, London, UK.

[31] Human Rights Council. 2011. *Discriminatory laws and practices and acts of violence against individuals based on their sexual orientation and gender identity: REport of the United Nations High Commissioner for Human Rights*. Technical Report. United Nations. Report No. A/HRC/19/41.

[32] 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 (ACL), Florence, Italy, 1641–1650. https://doi.org/10.18653/v1/P19-1160 arXiv:1906.00742

[33] Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (Male, Bachelor) and (Female, Ph.D) have different connotations: Parallelly Annotated Stylistic Language Dataset with Multiple Personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Langauge Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, 1696–1706. https://www.aclweb.org/anthology/D19-1179/

[34] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 22. https://doi.org/10.1145/3274357

[35] Svetlana Kiritchenko and Saif M. 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, USA, 43–53. https://www.aclweb.org/anthology/S18-2005/

[36] Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. 2020. The Gap on GAP: Tackling the Problem of Differing Data Distributions in Bias-Measuring Datasets. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI, Online, 9727–9736. arXiv:2011.01837 https://arxiv.org/abs/2011.01837v3

[37] 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, Stroudsburg, PA, USA, 1–11. https://doi.org/10.18653/v1/W17-1601

[38] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. In *Logic, Language, and Security*. Springer International Publishing, Cham, 189–202. https://doi.org/10.1007/978-3-030-62077-6_14

[39] Helma Lutz, Maria Teresa Herrera Vivar, and Linda Supik (Eds.). 2011. *Framing Intersectionality*. Ashgate Publishing Limited, Farnham, England.

[40] Anne Maass and Luciano Arcuri. 1996. Language and Stereotyping. In *Stereotypes and Stereotyping*, C. Niel Macra, Charles Strangor, and Miles Hewstone (Eds.). Guilford Press, New York, NY, Chapter 6, 193–225.

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. *A Survey on Bias and Fairness in Machine Learning*. Technical Report. University of Southern California, Information Sciences Institute. arXiv:1908.09635 http://arxiv.org/abs/1908.09635

[42] Malin Mobjörk, Camilla Berglund, Mikael Granberg, Magnus Johansson, Margareta Dahlström, Jon Moen, Lars Nyberg, and Mariele Evers. 2019. *Facilitating Doctoral Education in Cross-disciplinary Milieus: Experiences from PhD-candidates*. Technical Report. Karlstads University.

[43] 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*. Association for Computational Linguistics (ACL), Online, 2011–2017. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.157

[44] 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 (ACL), Online, 16–23. https://doi.org/10.18653/V1/2021.GEBNLP-1.3

[45] Damien W Riggs, Carla A Pfeffer, Ruth Pearce, Sally Hines, and Francis Ray White. 2020. Men, trans/masculine, and non-binary people negotiating conception: Normative resistance and inventive pragmatism. *International Journal of Transgender Health* 22, 1-2 (sep 2020), 6–17. https://doi.org/10.1080/15532739.2020.1808554

[46] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486 arXiv:1911.03891

[47] 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 (aug 2021), 845–874. https://doi.org/10.1162/TACL_A_00401 arXiv:2104.06001

[48] Chase Strangio. 2018. Deadly Violence Against Transgender People Is on the Rise. The Government Isn't Helping. https://www.aclu.org/blog/lgbt-rights/criminal-justice-reform-lgbt-people/deadly-violence-against-transgender-people-rise

[49] 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 *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), Florence, Italy, 1630–1640. arXiv:1906.08976 http://arxiv.org/abs/1906.08976

[50] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the Biases in My Word Embedding?. In *Artificial Intelligence, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 305–311. https://doi.org/10.1145/3306618.3314270

[51] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language

Models Using Causal Mediation Analysis. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. NeurIPS, Vancouver, Canada, 14 pages. https://www.cs.technion.ac.il/~belinkov/assets/pdf/neurips2020.pdf

[52] Catherine Yeo and Alyssa Chen. 2020. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop at ACL*. Association for Computational Linguistics, Seattle, USA, 107–109. https://doi.org/10.18653/v1/2020.winlp-1.27 arXiv:2008.01548

[53] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4134–4145. https://doi.org/10.18653/v1/2020.acl-main.380

[54] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. 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

[55] 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 *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, Hong Kong, 5276–5284. arXiv:1909.02224 http://arxiv.org/abs/1909.02224

A FULL BIBLIOGRAPHY OF SURVEYED PAPERS

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics (ACL), Online, 24–33. https://doi.org/10.18653/V1/2021.GEBNLP-1.4

[2] Artem Abzaliev. 2019. On GAP coreference resolution shared task: insights from the 3rd place solution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 107–112. https://doi.org/10.18653/v1/W19-3816

[3] Felipe Alfaro, Lois José, A R Fonollosa, and Marta R Costa-Jussà. 2019. BERT Masked Language Modeling for Co-reference Resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 76–81. https://doi.org/10.18653/v1/W19-3811

[4] Ananya, Nitya Parthasarthi, and Sameer Singh. 2019. GenderQuant: Quantifying Mention-Level Genderedness. 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, 2959–2969. https://doi.org/10.18653/V1/N19-1303

[5] Sandeep Attree. 2019. Gendered Ambiguous Pronouns Shared Task: Boosting Model Confidenceby Evidence Pooling. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 134–146. https://doi.org/10.18653/v1/W19-3820

[6] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying Gender Bias in Different Corpora | Enhanced Reader.pdf. In *Companion Proceedings of the Web Conference 2020 (WWW'20 Companion)*. ACM, New York, New York, USA, 752–759. https://dl-acm-org.proxy.ub.umu.se/doi/abs/10.1145/3366424.3383559

[7] David Bamman, Sejal Popat, and Sheng Shen. 2019. An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Minneapolis, Minnesota, 2138–2144. https://doi.org/10.18653/v1/N19-1220

[8] 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

[9] 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

[10] Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 33–39. https://doi.org/10.18653/v1/W19-3805

[11] Christine Basta, Marta R Costa-Jussà, and • Noe Casas. 2020. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications* 33 (2020), 3371–3384. https://doi.org/10.1007/s00521-020-05211-z

[12] Christine Raouf Saad Basta, Marta Ruiz Costa-Jussà, and José Adrián Rodríguez 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

[13] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating Gender Bias in BERT. *Cognitive Computation* 13 (2021), 1008–1018. https://doi.org/10.1007/s12559-021-09881-2

[14] Shruti Bhargava and David Forsyth. 2019. *Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models*. Technical Report. University of Illinois at Urbana-Champaign. arXiv:1912.00578 http://arxiv.org/abs/1912.00578

[15] 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

[16] 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 *Advances in Neural Information Processing Systems*. NeurIPS, Barcelona, Spain, 4356–4364. arXiv:1607.06520 http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-d

[17] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and Reducing Stereotypes in Word Embeddings. In *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*. ICML, New York, New York, USA, 41–45. arXiv:1606.06121 http://arxiv.org/abs/1606.06121

[18] 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 arXiv:1904.03035

[19] Clothilde Breger, Ghasem Elyasi, Guillermo Infante, and Mariano Zarza. 2020. Reducing Gender Bias in Natural Language Processing methods. In *Meta Research Research Methods Course: Master in Intelligent and Interactive Systems*, Davinia Hernández-Leo and Judit Martínez-Moreno (Eds.). Universitat Pompeu Fabra, Barcelona, Spain, 17–24. http://repositori.upf.edu/

[20] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *36th International Conference on Machine Learning*, Vol. 97. Proceedings of Machine Learning Research, Irvine, CA, USA, 803–811. arXiv:1810.03611 http://proceedings.mlr.press/v97/brunet19a.html

[21] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the European Conference on Computer Vision*. ECCV, Munich, Germany, 771–787. arXiv:1803.09797 http://arxiv.org/abs/1803.09797

[22] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)* 356, 6334 (apr 2017), 183–186. https://doi.org/10.1126/science.aal4230

[23] 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

[24] Rakesh Chada. 2019. Gendered Pronoun Resolution using BERT and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 126–133. https://doi.org/10.18653/v1/W19-3819

[25] 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

[26] Serina Chang and Kathleen McKeown. 2019. Automatically Inferring Gender Associations from Language. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Hong Kong, 5746–5752. https://doi.org/10.18653/v1/D19-1579

[27] Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender Bias and Under-Representation in Natural Language Processing Across Human Languages; Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, Virtual Event, 11 pages. https://doi.org/10.1145/3461702.3462530

[28] Won Ik Cho, Ji Won Kim, 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/W19-3824

[29] Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. Association for the Advancement of Artificial Intelligence (AAAI), New York, NY, USA, 386–393. https://doi.org/10.1609/aaai.v34i01.5374

[30] Marta R Costa-Jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence* 1, 11 (2019), 495–496. https://doi.org/10.1038/s42256-019-0105-5

[31] 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, 26–34. https://aclanthology.org/2020.gebnlp-1.3

[32] Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020. *Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters*. Technical Report. TALP Research Center, Universitat Politècnica de Catalunya, Barcelona. arXiv:2012.13176 https://arxiv.org/abs/2012.13176v1

[33] Amanda Cercas Curry and Verena Rieser. 2018. #MeToo: 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

[34] Karan Dabas, Nishtha Madan, Vijay Arya, Sameep Mehta, Gautam Singh, and Tanmoy Chakraborty. 2019. Fair Transfer of Multiple Style Attributes in Text. In *2019 Grace Hopper Celebration India, GHCI*. IEEE, Bangalore, India, 1–5. https://doi.org/10.1109/GHCI47972.2019.9071799 arXiv:2001.06693

[35] Jamell Dacon and Haochen Liu. 2021. Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles. In *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*, Vol. 2. Association for Computing Machinery, Ljubljana, Slovenia, 385–392. https://doi.org/10.1145/3442442.3452325

[36] Erenay Dayanık and Sebastian Padó. 2021. Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Online, 50–61. https://aclanthology.org/2021.wassa-1.6

[37] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Atlanta, Georgia, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[38] Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI press, New York, NY, USA, 7659–7666. https://doi.org/10.1609/aaai.v34i05.6267

[39] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. In *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*. PLMR, Naha, Okinawa, Japan, 879–887. arXiv:1901.07656 http://proceedings.mlr.press/v89/dev19a.html

[40] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8173–8188. https://doi.org/10.18653/v1/2020.emnlp-main.656

[41] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23

[42] Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender Bias in Text: Origin, Taxonomy, and Implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics (ACL), Online, 34–44. https://doi.org/10.18653/V1/2021.GEBNLP-1.5

[43] Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring Human Gender Stereotypes with Word Association Test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Langauge Processing*. Association for Computational Linguistics, Hong Kong, 6133–6143. https://doi.org/10.18653/v1/D19-1635

[44] Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, Jackie Chi, and 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

[45] Kawin Ethayarajh. 2020. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2914–2919. https://doi.org/10.18653/v1/2020.acl-main.262 arXiv:2004.12332

[46] Kawin Ethayarajh, David Duvenaud, 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 arXiv:1908.06361

[47] Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020. Measuring Social Bias in Knowledge Graph Embeddings. In *Workshop on Bias in Automatic Knowledge Graph Construction (KG-BIAS)*. Automated KNowledge Base Construction, Online. arXiv:1912.02761 http://arxiv.org/abs/1912.02761

[48] Eve Fleisig. 2021. *Adversarial Learning for Bias Mitigation in Machine Translation*. Bachelor of Science in Engineering. Princeton University.

[49] Omar U. Florez. 2019. *On the Unintended Social Bias of Training Language Generation Models with Data from Local Media*. Technical Report. arXiv:1911.00461 http://arxiv.org/abs/1911.00461

[50] 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

[51] 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 of the United States of America* 115, 16 (apr 2018), E3635–E3644. https://doi.org/10.1073/pnas.1720347115 arXiv:1711.08412

[52] 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

[53] Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2943–2953. https://doi.org/10.18653/v1/2020.acl-main.265 arXiv:1911.03642

[54] Oguzhan Gencoglu. 2021. Cyberbullying Detection with Fairness Constraints. *IEEE Internet Computing* 25, 1 (may 2021), 20–29. https://doi.org/10.1109/MIC.2020.3032461 arXiv:2005.06625

[55] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz, Mu~ Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics (ACL), 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150 arXiv:2012.15859

[56] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061 arXiv:1903.03862

[57] Hila Gonen and Kellie Webster. 2020. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics (ACL), 1991–1995. https://doi.org/10.18653/v1/2020.findings-emnlp.180 arXiv:2004.14065

[58] 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

[59] Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender Dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 452–459. https://aclanthology.org/2021.nodalida-main.52

[60] Reyhaneh Hashempour, Aline Villavicencio, Barbara Plank, and Renato Amorim. 2019. A Deep Learning Approach to Language-independent Gender Prediction on Twitter. In *Proceedings of the 2019 Workshop on Widening NLP*. https://doi.org/10.1016/j.diin.2011.04.002

[61] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. *Can You Translate that into Man? Commercial Machine Translation Systems Include Stylistic Biases*. Technical Report. https://www.cia.gov/library/

[62] Alexander Hoyle, 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, 1706–1716. https://doi.org/10.18653/v1/P19-1167 arXiv:1906.04760

[63] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 65–83. https://doi.org/10.18653/v1/2020.findings-emnlp.7 arXiv:1911.03064

[64] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *Proceedings of the 12th Conference on Langauge Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), Marseille, France, 1440–1448. arXiv:2002.10361 http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.180.pdf

[65] Matei Ionita, Yury Kashnitsky, Ken Krige Kitsong, Vladimir Larin, Pjsc Sberbank, Denis Logvinenko, and Atanas Atanasov. 2019. Resolving Gendered Ambiguous Pronouns with BERT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 113–119. https://doi.org/10.18653/v1/W19-3817

[66] Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating Gender Augmented Data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics (ACL), Online, 93–102. https://doi.org/10.18653/V1/2021.GEBNLP-1.11

[67] Hailey James and David Alvarez-Melis. 2019. Probabilistic Bias Mitigation in Word Embeddings. In *Human-Centric Machine Learning Workshop, NeurIPS 2019*. arXiv:1910.14497 http://arxiv.org/abs/1910.14497

[68] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating Gender Bias Amplification in Distribution by Posterior Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2936–2942. https://doi.org/10.18653/v1/2020.acl-main.264 arXiv:2005.06251

[69] May Jiang and Christiane Fellbaum. 2020. Interdependencies of Gender and Race in Contextualized Word Embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online), 17–25. https://aclanthology.org/2020.gebnlp-1.2

[70] Meichun Jiao and Ziyang Luo. 2021. Gender Bias Hidden Behind Chinese Word Embeddings: The Case of Chinese Adjectives. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics (ACL), Barcelona, Spain (Online), 8–15. https://doi.org/10.18653/V1/2021.GEBNLP-1.2

[71] Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment. In *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*. Association for Computational Linguistics, Hong Kong, 1–11. https://doi.org/10.18653/v1/K19-1001 arXiv:1909.08975

[72] 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 (ACL), Florence, Italy, 1641–1650. https://doi.org/10.18653/v1/P19-1160 arXiv:1906.00742

[73] Masahiro Kaneko and Danushka Bollegala. 2021. 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://aclanthology.org/2021.eacl-main.16

[74] Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (Male, Bachelor) and (Female, Ph.D) have different connotations: Parallelly Annotated Stylistic Language Dataset with Multiple Personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Langauge Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, 1696–1706. https://www.aclweb.org/anthology/D19-1179/

[75] Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor Debiasing of Word Representations Evaluated on WEAT. (jun 2019), 40–48. arXiv:1906.05993 http://arxiv.org/abs/1906.05993

[76] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. (may 2020). arXiv:2005.05921 http://arxiv.org/abs/2005.05921

[77] Svetlana Kiritchenko and Saif M. 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, USA, 43–53. https://www.aclweb.org/anthology/S18-2005/

[78] Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. 2020. The Gap on GAP: Tackling the Problem of Differing Data Distributions in Bias-Measuring Datasets. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI, Online, 9727–9736. arXiv:2011.01837

https://arxiv.org/abs/2011.01837v3

[79] Vaibhav Kumar, Tenzin Singhay Bhotia, and Vaibhav Kumar. 2020. Fair Embedding Engine: A Library for Analyzing and Mitigating Gender Bias in Word Embeddings. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS),*. Association for Computational Linguistics (ACL), 26–31. https://doi.org/10.18653/v1/2020.nlposs-1.5 arXiv:2010.13168

[80] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. (jun 2019), 166–172. arXiv:1906.07337 http://arxiv.org/abs/1906.07337

[81] Anne Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. (apr 2019), 85–91. arXiv:1904.11783 http://arxiv.org/abs/1904.11783

[82] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vuli´cvuli´c. 2020. *A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces.* Technical Report. www.aaai.org

[83] Nayeon Lee, Yejin Bang, Jamin Shin, and Pascale Fung. 2019. Understanding the Shades of Sexism in Popular TV Series. In *Proceedings of the 2019 Workshop on Widening NLP*. 177–180. https://tvtropes.org/

[84] Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring Social Bias in Chatbots using Stereotype Knowledge. In *Proceedings of the Workshop on Widening NLP*. Florence, Italy, 177–180.

[85] Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Towards Debiasing Sentence Representations. In *Proceedings of the NeurIPS Workshop on Human-Centric Machine Learning*. Vancouver, Canada.

[86] Bo Liu. 2019. *Anonymized BERT: An Augmentation Approach to the Gendered Pronoun Resolution Challenge.* Technical Report. 120–125 pages. https://github.com/boliu61/gendered-pronoun-resolution

[87] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does Gender Matter? Towards Fairness in Dialogue Systems. (oct 2019). arXiv:1910.10486 http://arxiv.org/abs/1910.10486

[88] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. In *Logic, Language, and Security*. Springer International Publishing, Cham, 189–202. https://doi.org/10.1007/978-3-030-62077-6_14

[89] 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 (ACL), Virtual, 48–55. https://doi.org/10.18653/V1/2021.NUSE-1.5

[90] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (apr 2019), 615–621. https://doi.org/10.18653/v1/n19-1062 arXiv:1904.04047

[91] Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender Bias in Natural Language Processing Across Human Languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*. Association for Computational Linguistics (ACL), Online, 45–54. https://doi.org/10.18653/V1/2021.TRUSTNLP-1.6

[92] 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 *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, 5267–5275. arXiv:1909.00871 http://arxiv.org/abs/1909.00871

[93] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. (mar 2019). arXiv:1903.10561 http://arxiv.org/abs/1903.10561

[94] Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. (may 2020). arXiv:2005.08864 http://arxiv.org/abs/2005.08864

[95] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. (oct 2019). arXiv:1910.10872 http://arxiv.org/abs/1910.10872

[96] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), Marseille, France, 6462–6468. https://aclanthology.org/2020.lrec-1.796

[97] Joshua R. Minot, Nicholas Cheney, Marc Maier, Danne C. Elbers, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. *Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance.* Technical Report. arXiv:2103.05841 https://arxiv.org/abs/2103.05841v1

[98] Inom Mirzaev, Anthony Schulte, Michael Conover, and Sam Shah. 2019. *Considerations for the Interpretation of Bias Measures of Word Embeddings.* Technical Report. arXiv:1906.08379 http://arxiv.org/abs/1906.08379

[99] Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating Bias In Dutch Word Embeddings. , 56–71 pages. https://aclanthology.org/2020.gebnlp-1.6

[100] 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*. Association for Computational Linguistics (ACL), Online, 2011–2017. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.157

[101] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. (apr 2020). arXiv:2004.09456 http://arxiv.org/abs/2004.09456

[102] Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. *Computational Linguistics* (mar 2020), 1–11. https://doi.org/10.1162/coli_a_00379 arXiv:1905.09866

[103] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *2019 IEEE/WIC/ACM International Conference on WEb Intelligence (WI)*. 149–155.

[104] Enoch Opanin Gyamfi, Yunbo Rao, Miao Gou, and Yanhua Shao. 2020. deb2viz: Debiasing gender in word embedding data using subspace visualization. In *Eleventh International Conference on Graphics and Image Processing (ICGIP 2019)*, Zhigeng Pan and Xun Wang (Eds.), Vol. 11373. SPIE, Hangzhou, China. https://doi.org/10.1117/12.2557465

[105] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos, Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. (2020), 12. https://doi.org/10.1145/3351095.3372843

[106] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* (aug 2018), 2799–2804. arXiv:1808.07231 http://arxiv.org/abs/1808.07231

[107] Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2020. Joint Multiclass Debiasing of Word Embeddings. (mar 2020). arXiv:2003.11520 http://arxiv.org/abs/2003.11520

[108] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2019. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications* (mar 2019), 1–19. https://doi.org/10.1007/s00521-019-04144-6

[109] Rasmus Précenth. 2019. *Word Embeddings and Gender Stereotypes in Swedish and English.* Master's Thesis. Uppsala University.

[110] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. *Debiasing Embeddings for Reduced Gender Bias in Text Classification.* Technical Report. 69–75 pages. https://github.com/tolga-b/debiaswe

[111] Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 450–456. https://doi.org/10.1145/3377713.3377792

[112] 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. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop* (may 2019), 223–228. arXiv:1905.12801 http://arxiv.org/abs/1905.12801

[113] 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 (ACL), Online, 16–23. https://doi.org/10.18653/V1/2021.GEBNLP-1.3

[114] Prashanth Rao and Maite Taboada. 2021. Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. *Frontiers in Artificial Intelligence* 4 (jun 2021). https://doi.org/10.3389/FRAI.2021.664737/FULL

[115] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. (apr 2020). arXiv:2004.07667 http://arxiv.org/abs/2004.07667

[116] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. (apr 2019). arXiv:1904.05233 http://arxiv.org/abs/1904.05233

[117] Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring Social Biases in Grounded Vision and Language Embeddings. (feb 2020). arXiv:2002.08911 http://arxiv.org/abs/2002.08911

[118] David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLOS ONE* 15, 4 (apr 2020), e0231189. https://doi.org/10.1371/journal.pone.0231189

[119] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. *Social Bias in Elicited Natural Language Inferences.* Technical Report. 74–79 pages. https://github.com/cjmay/snli-ethics

[120] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 2 (apr 2018), 8–14. arXiv:1804.09301 http://arxiv.org/abs/1804.09301

[121] 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, Sweden, Turku, Finland, 35–43. https://aclanthology.org/W19-6104

[122] Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. 2018. Is there Gender bias and stereotype in Portuguese Word Embeddings?. In *International Conference on the Computational Processing of Portuguese*. arXiv:1810.04528 http://arxiv.org/abs/1810.04528

[123] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486 arXiv:1911.03891

[124] Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. (apr 2020). arXiv:2004.04498 http://arxiv.org/abs/2004.04498

[125] 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 (aug 2021), 845–874. https://doi.org/10.1162/TACL_A_00401 arXiv:2104.06001

[126] Ramit Sawhney, Arshiya Aggarwal, and Rajiv Shah. 2021. An Empirical Investigation of Bias in the Multimodal Analysis of Financial Earnings Calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Association for Computational Linguistics (ACL), Online, 3751–3757. https://doi.org/10.18653/V1/2021.NAACL-MAIN.294

[127] Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. *Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings.* Technical Report. TU Delft. 94–103 pages. https://doi.org/10.18653/v1/2020.nlpcss-1.11

[128] João Sedoc and Lyle Ungar. 2019. *The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations.* Technical Report. 55–61 pages. https://github.com/jsedoc/

[129] Procheta Sen and Debasis Ganguly. 2020. *Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning.* Technical Report. https://ojs.aaai.org//index.php/AAAI/article/view/5654

[130] Sima Sharifirad, Alon Jacovi, and Stan Matwin. 2019. Learning and Understanding Different Categories of Sexism Using Convolutional Neural Network's Filters. In *Proceedings of the 2019 Workshop on Widening NLP.* 21–23. https://doi.org/10.1177/1741659016652445

[131] Sima Sharifirad and Stan Matwin. 2019. Using Attention-based Bidirectional LSTM to Identify Different Categories of Offensive Language Directed Toward Female Celebrities. In *Proceedings of the 2019 Workshop on Widening NLP.* Association for Computational Linguistics, Florence, Italy, 46–48. https://aclanthology.org/W19-3616

[132] Shanya Sharma, Manan Dey, and Koustuv Sinha. 2020. Evaluating Gender Bias in Natural Language Inference. In *NeurIPS 2020 Workshop on Dataset Curation and Security.* arXiv:2105.05541 https://arxiv.org/abs/2105.05541v1

[133] Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. 2018. *Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis.* Technical Report. http://textblob.readthedocs.io/en/dev/

[134] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (sep 2019), 3407–3412. arXiv:1909.01326 http://arxiv.org/abs/1909.01326

[135] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. (may 2020). arXiv:2005.00268 http://arxiv.org/abs/2005.00268

[136] Eli Sherman, Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access.* Association for Computational Linguistics (ACL), Online, 217–223. https://doi.org/10.18653/V1/2021.CLPSYCH-1.23

[137] Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing Gender Bias in Word Embedding with Latent Disentanglement and Counterfactual Generation. (apr 2020). arXiv:2004.03133 http://arxiv.org/abs/2004.03133

[138] 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 (ACL), Online, 2383–2389. https://doi.org/10.18653/V1/2021.NAACL-MAIN.189

[139] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (jun 2019), 1679–1684. arXiv:1906.00591 http://arxiv.org/abs/1906.00591

[140] 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 *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.* Association for Computational Linguistics (ACL), Florence, Italy, 1630–1640. arXiv:1906.08976 http://arxiv.org/abs/1906.08976

[141] Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased Embeddings from Wild Data: Measuring, Understanding and Removing. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11191 LNCS (jun 2018), 328–339. arXiv:1806.06301 http://arxiv.org/abs/1806.06301

[142] Chris Sweeney and Maryam Najafian. 2020. Reducing Sentiment Polarity for Demographic Attributes in Word Embeddings using Adversarial Learning. (2020). https://doi.org/10.1145/3351095.3372837

[143] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the Biases in My Word Embedding?. In *Artificial Intelligence, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 305–311. https://doi.org/10.1145/3306618.3314270

[144] Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can Existing Methods Debias Languages Other than English? First Attempt to Analyze and Mitigate Japanese Word Embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing.* Association for Computational Linguistics, Barcelona, Spain (Online), 44–55. https://aclanthology.org/2020.gebnlp-1.5

[145] Yi Chern Tan and L Elisa Celis. 2019. *Assessing Social and Intersectional Biases in Contextualized Word Representations.* Technical Report. 13230–13241 pages. https://github.com/openai/gpt-2-output-dataset

[146] Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review* 42, 1 (2018), 45–57. https://doi.org/10.1108/OIR-05-2017-0139

[147] Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2021. Using Gender- and Polarity-Informed Models to Investigate Bias. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing.* Association for Computational Linguistics (ACL), Online, 66–74. https://doi.org/10.18653/V1/2021.GEBNLP-1.8

[148] Ameya Vaidya, Feng Mai, and Yue Ning. 2019. Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection. (sep 2019). arXiv:1909.09758 http://arxiv.org/abs/1909.09758

[149] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting Gender Right in Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* (sep 2019), 3003–3008. https://doi.org/10.18653/v1/D18-1334 arXiv:1909.05088

[150] Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. , 2203–2213 pages. https://aclanthology.org/2021.eacl-main.188

[151] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. (apr 2020). arXiv:2004.12265 http://arxiv.org/abs/2004.12265

[152] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. NeurIPS, Vancouver, Canada, 14 pages. https://www.cs.technion.ac.il/$\sim$belinkov/assets/pdf/neurips2020.pdf

[153] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. (may 2020), 5443–5453. arXiv:2005.00965 http://arxiv.org/abs/2005.00965

[154] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2021. [RE] Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. (apr 2021), 5443–5453. https://doi.org/10.18653/v1/2020.acl-main.484 arXiv:2104.06973

[155] Zili Wang. 2019. *MSnet: A BERT-based Network for Gendered Pronoun Resolution*. Technical Report. 89–95 pages. https://github.com/google-research-

[156] Kellie Webster, Marta R Costa-Jussà, Christian Hardmeier, and Will Radford. 2019. *Gendered Ambiguous Pronouns (GAP) Shared Task at the Gender Bias in NLP Workshop 2019*. Technical Report. 1–7 pages. https://www.kaggle.com/c/

[157] 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 (dec 2018), 605–617. https://doi.org/10.1162/tacl_a_00240 arXiv:1810.05201

[158] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, Duen Horng, and Chau. 2020. RECAST: Interactive Auditing of Automatic Toxicity Detection Models. (jan 2020). arXiv:2001.01819 http://arxiv.org/abs/2001.01819

[159] Yinchuan Xu and Junlin Yang. 2019. Look Again at the Syntax: Relational Graph Convolutional Network for Gendered Ambiguous Pronoun Resolution. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 96–101. https://www.aclweb.org/anthology/W19-3814/

[160] Kai-Chou Yang, Timothy Niven, Tzu-Hsuan Chou, and Hung-Yu Kao. 2019. Fill the GAP: Exploiting BERT for Pronoun Resolution. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 102–106. https://www.aclweb.org/anthology/W19-3815/

[161] Zekun Yang and Juan Feng. 2020. A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI press, New York, NY, USA, 9434–9441. https://doi.org/10.1609/AAAI.V34I05.6486 arXiv:1911.10787

[162] Catherine Yeo and Alyssa Chen. 2020. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop at ACL*. Association for Computational Linguistics, Seattle, USA, 107–109. https://doi.org/10.18653/v1/2020.winlp-1.27 arXiv:2008.01548

[163] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, Inc, 335–340. https://doi.org/10.1145/3278721.3278779 arXiv:1801.07593

[164] Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double Perturbation: On the Robustness of Robustness and Counterfactual Bias Evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics (ACL), Online, 3899–3916. https://doi.org/10.18653/V1/2021.NAACL-MAIN.305

[165] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4134–4145. https://doi.org/10.18653/v1/2020.acl-main.380

[166] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying Biases in Clinical Contextual Word Embeddings. In *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*. Association for Computing Machinery, Inc, 110–120. https://doi.org/10.1145/3368555.3384448

[167] Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local Group Bias Detection by Clustering. In *Proceedings ofthe 2020 Conference on Empirical Methods in Natural Language Processing,*. Association for Computational Linguistics (ACL), 1968–1977. https://doi.org/10.18653/v1/2020.emnlp-main.155 arXiv:2010.02867

[168] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. (may 2020), 2896–2907. arXiv:2005.00699 http://arxiv.org/abs/2005.00699

[169] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1. Association for Computational Linguistics (ACL), Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/V1/N19-1064

[170] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (apr 2019), 629–634. arXiv:1904.03310 http://arxiv.org/abs/1904.03310

[171] 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. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (jul 2017), 2979–2989. arXiv:1707.09457 http://arxiv.org/abs/1707.09457

[172] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 2 (apr 2018), 15–20. arXiv:1804.06876 http://arxiv.org/abs/1804.06876

[173] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. 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

[174] 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 *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, Hong Kong, 5276–5284. arXiv:1909.02224 http://arxiv.org/abs/1909.02224

[175] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (jun 2019), 1651–1661. arXiv:1906.04571 http://arxiv.org/abs/1906.04571