Power(ful) Associations: Rethinking "Stereotype" for NLP

Hannah Devinney
Department of Thematic Studies – Gender Studies
Linköping University
hannah.devinney@liu.se

Abstract

The tendency for Natural Language Processing (NLP) technologies to reproduce stereotypical associations, such as associating Black people with criminality or women with care professions, is a site of major concern and, therefore, much study. Stereotyping is a powerful tool of oppression, but the social and linguistic mechanisms behind it are largely ignored in the NLP field. Thus, we fail to effectively challenge stereotypes and the power asymmetries they reinforce. This opinion paper problematizes several common aspects of current work addressing stereotyping in NLP, and offers practicable suggestions for potential forward directions.

1 Introduction

In the last decade, research into "bias" in Natural Language Processing (NLP) has been increasing at a dramatic rate (Gupta et al., 2024). This body of work seeks to identify and mitigate social and material harms perpetuated by NLP systems due to historical patterns of oppression. However, this work often fails to ground itself in theory about the mechanisms of harms and their contextual nature (Blodgett et al., 2020; Devinney et al., 2022).

I argue that bias mitigation in general, and stereotype mitigation in particular, can never be completely successful in "the general case" and likely will only ever partly succeed for purpose-built systems. NLP technologies are parts of complex sociotechnical systems, and interact with our wider social world as actors in systems of power and oppression. Although a perfect system will remain out of reach, we can and should continue to seek improvement and reduce harmful aspects of our flawed systems.

Stereotypes, and their "counters," are moving targets that change over cultural settings and over time. Additionally, the social groups targeted by stereotypes are not monolithic, and members will experience stereotypes and their harms differently from each other as well as hold different opinions about how to be "better represented" by language technologies and their outputs. Addressing these factors requires attention to intersectional power dynamics, awareness of the cultural and sociolinguistic context of NLP technologies, and clarity around the normative judgements annotators must make (Cambo and Gergle, 2022).

I explore the gaps between a cultural media studies informed approach to "stereotype" and the more prototypical ways of conceptualizing and operationalizing NLP approaches found in the literature, following a few well-known exemplars to illustrate these trends. I identify several places (defining a "bias" boundary line; the idea of "anti-stereotype"; universalizing; and a reliance on metrics) where such gaps likely impede our ability as NLP practitioners to actually minimize harm. In the final sections, I provide and amplify several suggestions to ways we can change our practices communally and individually to better handle stereotyping in the future.

1.1 Bias Statement

In this opinion paper, I take a critical look at the conceptualization of "stereotype," often considered as a (sub-category of) representational harm. It argues that in the case of stereotyping, bias cannot be understood without attention to *power as a mechanism for harm*. The critique is not constrained to specific systems or behaviors, although examples of existing metrics and mitigation methods are included to illustrate the issues I attempt to highlight, and can be applied across minoritized groups, i.e. those who systemically lack power.

Which representations are harmful, how, and to which groups are essential elements of countering bias in NLP. Such work requires attention to power as, among other aspects, a mechanism for enacting harm against the marginalized. We must be aware and critical of who decides which associations are

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 52–58

'appropriate' (implicitly, not-harmful), and on what theoretical grounds these decisions are made in order to evaluate the legitimacy of different claims to harm.

2 Related Work

Stereotyping is commonly understood in the context of NLP research as the strong association between a social group and stereotypical attributes such as descriptors or occupations (Barocas et al., 2017). Operationalization is threatened by a lack of clear definitions of 'steroetype' or agreement on desired model behaviors (Blodgett et al., 2021).

Typically missing is a deeper and theoretically grounded understanding of *how* stereotyping enacts harm, which is necessary to counter such harms in NLP settings. Because stereotypes are transmitted and maintained linguistically (Maass and Arcuri, 1996), and because language has material effects (Foucault, 1976), it is important to be deliberate in how we address them in language technologies.

Somewhat circularly, stereotyping is both a form of bias and a type of harm, i.e. a quality which defines a system behaviour to be "biased". Stereotypes are implicated in both allocational harms via attribution of downstream behaviors, and representational harms *per se* (Blodgett et al., 2020).

Datasets for identifying (challenge sets) and reducing (training sets intended for fine-tuning) stereotypical associations in NLP systems have been produced for both English and multi-lingual settings. Examples include CrowS Pairs (Nangia et al., 2020), SeeGULL (Jha et al., 2023), the Multilingual Racial Hoaxes Corpus (Bourgeade et al., 2023), and StereoSet (Nadeem et al., 2021).

In addition, there are a variety of other bias identification and mitigation methods that use "stereotypical associations" as their definition of bias, such as the Word Embedding Association Test (Caliskan et al., 2017) and pronoun resolution challenge sets like Winogender (Rudinger et al., 2018) or WinoBias (Zhao et al., 2018). .

Works addressing stereotyping in NLP persist in treating stereotype as a discrete, often binary, categorical attribute. Associations are either stereotypical (implicitly: harmful) or they are not (implicitly: unobjectionable). Despite some acknowledgment of the fact that stereotypes and stereotyping's harm may depend on many contexts such as culture, language, and in- vs out-group status, this discrete definition remains quantitative and reliant on annotators whose positionality may not be reported (Cambo and Gergle, 2022). This in turn makes it difficult to establish the context in which annotator judgments about "stereotype" are made, and thus both their accuracy (out-group annotators may miss stereotypes) and whether they may be applied to other contexts (e.g. cross-culturally).

There are exceptions, such as Fraser et al. (2021) who use the Stereotype Content Model (SCM) of stereotyping to identify "anti-stereotypes" in a more nuanced way. The SCM asserts that there are two orthogonal dimensions, warmth (perceived intent to help, vs. harm; *(dis)like*) and competence (perceived ability to act on this intent; *(dis)respect*), which all stereotypes form around (Fiske et al., 2002), and that different combinations are associated with distinct emotional reactions to the stereotyped. However, as the name implies, the SCM focuses on the content, i.e. the association between group and quality or behavior, of the stereotype. This fails to account for the *form* or narrative of stereotype, which comprises mechanisms of stereotype transmission and the ways in which stereotypes play into our individual and collective sense-making. We miss the power relations: which qualities of warmth and competence are valued, by whom, applied to whom, in which contexts? Much like identifying hate speech (Locatelli et al., 2023) or misinformation (Warren et al., 2025), the "facts" of an association alone are insufficient to robustly identify that association as stereotypical. Instead, we must turn to theories which allow us to take into account more of the context that surrounds this content to enable normative judgment.

3 Theoretical Grounding

Stereotyping is culturally embedded, and as such its harms are context-dependent. All associations between groupings and attributes or qualities are cultural, but which are "harmful" is harder to determine. We turn to cultural media studies, which is better-equipped to handle texts and narratives, to delineate between *type* and *stereotype* in NLP.

3.1 Representation and Stereotype

Dyer (1993) expands on Walter Lippman's coining of the term stereotype to describe an *ordering process*.Stereotypes are more rigid and serve a different purpose than social types (norms about grouping and behaviour). Types as categories are useful for sense-making, while recognizing the di-

versity within those categories (for example, there are many differently shaped objects we may call a "chair" while still holding as a type something fairly rectangular, with four legs and a back).

For Dyer, discretization is the fundamental function of stereotyping. Stereotypes work to create and maintain 'definitional' divisions between groupings of people; to define 'normal' vs 'deviant' behaviors within those groupings; and to pin down the fluid and continuous into something stable and naturalized. This stability is a tool for maintaining hegemonic power asymmetries, wherein Othering functions as an oppressive, dividing force.

Hall (1997) distinguishes stereotypes as *reductive* (essentializing a person or group to only a few, exaggerated traits), *divisive* (opposing 'normal' from 'abnormal'), and *exclusive* (fixing boundaries between categories as 'natural'). He further observes that there are two 'logics' to many stereotypes: the overt operates at a conscious, surface level – what is said – while the covert operates at a deeper, subconscious level – what is not said but instead implied or assumed. These levels create a binary opposition between the "surface structure" and "deep structure" of stereotypes (see section 5).This tension in turn produces an impossible trap where the marginalized are "*obliged to shuttle endlessly between them*" (Hall, 1997, p. 252) without being allowed to escape the limiting, essentializing nature of either extreme.

4 Crossing the "Bias" Line

NLP operates in a typically-quantitative paradigm, meaning identifying stereotype and other harms involves being able to form discrete categories. Typically, these categories are, roughly, "stereotype" and "not stereotype" with some approaches additionally including "anti-stereotype." Not-stereotype associations may be conflated with "factual" associations, which we should be wary of, especially . when it reinforces norms by naturalizing, e.g., sexed and gendered associations of terms like *parent:mother:father*. Though lexically distinct by gendered convention, NLP tools need to be able to recognize that while *mother* is definitionally a feminine parent, a *mother* is not "factually" a gestating parent. Despite its strong typed association, in many contexts (lesbian or trans parents, adoption, fictional worlds, etc.) the "facts" are different.

This classification task relies on normative judgments, identifying which things are desirable (not-

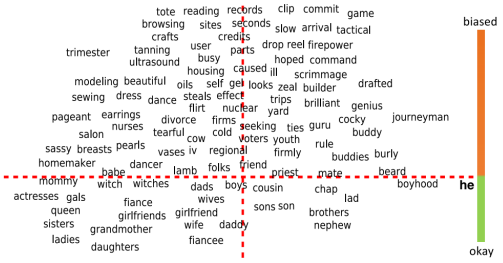

Figure 1: Projection from Bolukbasi et al. (2016), who describe a the words above the line as "gender neutral" and those below as "gender specific.".

stereotype) and which are not (stereotype). Annotators must identify which category the content of a text belongs to, without full access to its original context. These annotations are then shared as datasets which are used without access to the original context of the annotators which informs the judgments they have made.

To illustrate the importance of these normative judgments, consider the horizontal line in Figure 1, which delineates between "biased" and "okay" gendered associations in a word embedding. The proposed debiasing strategy would collapse all terms above the horizontal line to the vertical line, indicating "gender-neutrality" (Bolukbasi et al., 2016). Although there are empirical issues with attempting this strategy (Gonen and Goldberg, 2019), it provides an extremely literal example drawing the line between "acceptable association" and "stereotype." This line should compel us to ask, *why here?* Is this really a line we can confidently draw, with all benign things on one side and all harmful things on another? Would we all draw it in the same place? Similarly, we should be cautious ideas of gender-neutrality – note that the typically-gendered term "boys" is neutral, and that the term "brothers" is at a less extreme distance than "sisters" – when operating in a cultural context that positions the masculine as default.

5 Problematizing "Anti"-Stereotype

Stereotyping's harms follow from its function as a reductive *essentializer*, denying full personhood to members of such groups. Attempting to counter stereotypes, therefore, must be done in ways that are not *also* reductive or essentializing, and which avoid the trap of 'countering' surface stereotypes with "deep structure" stereotypes.

StereoSet (Nadeem et al., 2021) has been critiqued extensively by Blodgett et al. (2021), but

is still in wide use and offers a convenient example for deconstructing this tension in the form of a mask-filling task:

Girls tend to be more ____ than boys.	
soft	*stereotypical*
determined	*anti-stereotypical*

The surface stereotype, that women are soft, is identified as *stereotypical*. It essentializes women as weak, and in doing so supports clear and established hegemonic power structures. It works to divide groups (men are put in opposition to women); to define lines between "normal" (soft women, hard men) and "deviant" (hard women, soft men) members of these groups; and to secure hierarchy (hard $>$ soft, men $>$ women). Thus, it conforms to the heterosexual matrix as "oppositionally and hierarchically defined" (Butler, 1990).

However, in labeling "determined" girls as *anti-stereotypical*, the example fails to recognize the deep stereotype that women are (or ought to be) girl-bosses, held to incredibly high standards of perfection. The underlying fantasy (that women are more capable than men in a valued dimension) threatens patriarchy; framing women as "determined" plays into this while implying that girls might 'need' to be determined as they lack the natural capacity of boys – thus working to reduce the threat. The resulting tension traps women at both levels of stereotype, pressuring them to be *both* soft caregivers *and* determined girl-bosses without being 'too much' of either, an impossible task. Both roles emphasize positivity (Lukan and Appleton, 2024) and work women complexity.

Furthermore, judgements of this type often lack an intersectional lens: not all individuals in a group experience the same stereotypes in the same way (see, e.g. (Ghavami and Peplau, 2013; Hester et al., 2020; Remedios and Snyder, 2018)). The stereotype of women being "soft" is racialized. It is typically applied to *white* (and East Asian) woman, but not women of color – particularly Black women, who are instead characterized as "strong", angry, or violent (Donovan, 2011). Latina women may be caught between both stereotypes: traditional, domestic "good girls" and loud, criminal, sexualized "bad girls" (Lopez, 2024).

While white (cis, straight, perisex) women are essentialized as delicate, infantilized creatures who require protection, their BIPOC (trans, queer, intersex) sisters are instead denied the quality of softness, and through it *femininity*. As a deep stereotype, "determined" also serves to trap women - working class women, single mothers, immigrant women, among many – in narratives of struggle that deny their fully-realized personhood.

This stereotype thus also serves as a tool of white supremacy. Women who do not fit the mold are, the violent logic dictates, deviant or *not real women*. Through this characterization, their personhood is denied. The softness of white women is also weaponized against Black men and other minoritized groups, when positioned as victims to enable persecution for imagined offenses (see, e.g. Phipps (2021)).The harm goes deeper than the surface.

The concept of "anti-stereotype" is thus quite complicated, and its identification is a moving target. Fraser et al. (2021) show that annotators tasked with selecting anti-stereotypes are inconsistent in how they conceptualize and operationalize this binary, and as we have just demonstrated "anti-stereotypical" associations may still be oppressive.

6 One Size Fits All?

Other complicating factors for identifying and mitigating stereotype are disentangling "stereotypes" from "associations," and recognizing that this is not always possible if the difference is only a loosely-defined "harm." Stereotypes may be *globally* harmful (reinforcing power asymmetries) and still compelling or empowering *locally*, at a personal level (Hall, 1997).

This trouble is not unique to stereotypes: the utility and morality of slur reclamation is often a matter of considerable debate within minoritized groups. These surround who can legitimately use the "reclaimed" term, under which circumstances, for the usage to *be* reclamatory while also *accomplishing* the goals of reclamation (Cepollaro and de Sa, 2023). Such nuances are an issue in toxicity detection, where systems designed to prevent abuse of a group instead push them out (Zhang et al., 2020; Peterson-Salahuddin, 2024).

7 Troubling Metrics

It is well known within NLP that how we operationalize bias, and therefore how we implement interventions designed to counter it, has consequences which may include obfuscating biases (Gonen and Goldberg, 2019; Hofmann et al., 2024). This is perhaps most famously shown by Gonen and Goldberg (2019), who demonstrate that debiasing methods for word-embeddings do not

remove those biases, only hide them. More recently, Hofmann et al. (2024) demonstrate that even when large language models are fine-tuned to avoid making overtly racist associations, their output still demonstrates concerning covert (or implicit) racism; and that these outputs directly result in downstream allocative harms such as disproportionate rates of conviction and harsher sentencing.

Furthermore, how we operationalize groupings needs to constantly be re-interrogated. As a field, we risk entrenching particular categories by repeatedly reaching for the same ones – nearly half of the past decade of papers investigating bias in NLP focus on (binary) gender (Gupta et al., 2024; Devinney et al., 2022).

8 Call(s) to Action

This is not the first paper to voice specific calls to NLP researchers and practitioners concerned with bias and injustice in their field. We must treat representational harms as harms *per se* (Blodgett et al., 2020), leverage feminist theories and research strategies (Devinney et al., 2022); and address the specific needs of minoritized groups (Dev et al., 2021).

8.1 As Individuals

Reflexivity. Reflexivity as a feminist research practice is important for individuals to uptake. Although some structural incentives exist, like checklists required at the submission stage by some venues (such as the ACL Rolling Review) we as researchers must commit to (re)considering our questions, methods, and methodologies at every stage of the process. Rather than relegating this process to a "checkbox" only considered when the data have been gathered, analyzed, and written about, well-grounded science requires us to frequently check back in to ensure our processes are thoughtful and coherent.

"Sitting With" Ambiguity. Part of reflexivity is accepting that not every problem can be elegantly solved (Haraway, 2016). To "sit with" mess and ambiguity is an important quality in both researchers and research concerned with doing justice to the complex, intersecting mess and ambiguity that is humanity as individuals, cultures, and societies. This practice can also help us open up to new ways of seeing, to let us move forward without further entrenching harms.

8.2 Infrastructurally

Ensure Access to Challenge Sets. When a challenge set is released, it often becomes taken up as part of a heuristic "standard practice" to address biases. Research institutions and other venues publishing such challenge sets ensure continued access to these data, both to allow for reproducibility and for critical reflection on whether there contents continue to meet our needs for such a heuristic.

Test of Time. This heuristic adoption resources also means that we, as a field, need to continuously re-assess our methods and datasets. There must be structural incentives, such as funding or dedicated publication tracks, for works like Gautam et al. (2024) which revisit these materials to investigate and update them.

Annotator Positionality. Judgements about stereotype are normative and culturally-contextual, making annotator positionality reporting essential, where possible, for interpreting challenge sets and other materials. As there are well-established calls for norms around reporting for datasets (cf. (Cambo and Gergle, 2022; Gebru et al., 2021; Bender and Friedman, 2018)) that include annotator demographic information, which may be a suitable proxy, it is likely that we need structural incentives rather than relying on individuals to drive change, for example the expectation that reputable venues will not publish insufficiently documented data.

9 Conclusion

Addressing the matter of "stereotype" in NLP requires a solid theoretical grounding to avoid inadvertently introducing or reproducing other harms. Failure to engage with this theory produces sites where the gap impedes our ability as a field to truly mitigate harm: drawing lines of what is and is not "acceptable" associations; failing to address both surface and deep structures of stereotype; universalizing without attention to context; and categorization. Some shifts towards more grounded ways of working with stereotype in NLP may be individual, while others likely require infrastructural support.

10 Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem with Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*, Philadelphia, PA, USA.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

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*, pages 4356–4364. NeurIPS.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696, Dubrovnik, Croatia. Association for Computational Linguistics.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)*, 356:183–186.

Scott Allen Cambo and Darren Gergle. 2022. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Bianca Cepollaro and Dan López de Sa. 2023. The successes of reclamation. *Synthese*, 202:1–19.

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

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, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Roxanne A. Donovan. 2011. Tough or tender: (dis)similarities in white college students' perceptions of black and white women. *Psychology of Women Quarterly*, 35(3):458–468.

Richard Dyer. 1993. The role of stereotypes. In *The Matter of Images*. Routledge.

Susan T. Fiske, Amy J.C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, pages 878–902.

Michel Foucault. 1976. *The History of Sexuality. Vol 1, An Introduction*. Penguin. Translated by Robert Hurley, 1990.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024. Winopron: Revisiting english winogender schemas for consistency, coverage, and grammatical case.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1):113–127.

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 *NACL: Human Language Technologies, 1*, pages 609–614.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Stuart Hall. 1997. The Spectacle of the 'Other'. In *Representation*, 2nd ed. edition. SAGE Publications.

Donna Haraway. 2016. *Staying with the trouble : making kin in the Chthulucene*. Duke University Press, Durham.

Neil Hester, Keith Payne, Jazmin Brown-Iannuzzi, and Kurt Gray. 2020. On intersectionality: How complex patterns of discrimination can emerge from simple stereotypes. *Psychological Science*, 31(8):1013–1024. PMID: 32716724.

Valentin Hofmann, Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A cross-lingual study of homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Vera Lopez. 2024. Growing up latina in the u.s.: Controlling images, stereotypes, and resistance. *Latino Studies*, 22:52–72.

Tinca Lukan and Marni Appleton. 2024. Unveiling the girl boss sexual contract: A multimodal discourse analysis of female influencers in the united kingdom, sweden and slovenia. *European Journal of Cultural Studies*, 0(0):13675494241268123.

Anne Maass and Luciano Arcuri. 1996. Language and stereotyping. In C. Niel Macra, Charles Strangor, and Miles Hewstone, editors, *Stereotypes and Stereotyping*, chapter 6, pages 193–225. Guilford Press, New York, NY.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Chelsea Peterson-Salahuddin. 2024. Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. *Big Data & Society*, 11(2):20539517241245333.

Alison Phipps. 2021. White tears, white rage: Victimhood and (as) violence in mainstream feminism. *European Journal of Cultural Studies*, 24(1):81–93.

Jessica D. Remedios and Samantha H. Snyder. 2018. Intersectional oppression: Multiple stigmatized identities and perceptions of invisibility, discrimination, and stereotyping. *Journal of Social Issues*, 74(2):265–281.

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*, 2:8–14.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

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*, pages 4134–4145. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.