Subjective *Isms*? On the Danger of Conflating Hate and Offence in Abusive Language Detection

Amanda Cercas Curry*
MilaNLP
Bocconi University
amanda.cercas@unibocconi.it

Gavin Abercrombie*
School of Mathematical and Computer Sciences
Heriot-Watt University
g.abercrombie@hw.ac.uk

Zeerak Talat*
Mohamed Bin Zayed University of Artificial Intelligence
z@zeerak.org

Abstract

Natural language processing research has begun to embrace the notion of annotator *subjectivity*, motivated by variations in labelling. This approach understands each annotator's view as valid, which can be highly suitable for tasks that embed subjectivity, e.g., sentiment analysis. However, this construction may be inappropriate for tasks such as hate speech detection, as it affords equal validity to all positions on e.g., sexism or racism. We argue that the conflation of hate and offence can invalidate findings on hate speech, and call for future work to be situated in theory, disentangling hate from its orthogonal concept, offence.

1 Introduction

Recently, natural language processing (NLP) researchers have dedicated significant efforts towards tasks under the umbrella of online abuse detection. For example, racism (e.g. Talat, 2016; Talat and Hovy, 2016), sexism and misogyny (e.g. Jiang et al., 2022; Zeinert et al., 2021), xenophobia (e.g. Ross et al., 2016), homophobia (Dias Oliva et al., 2021), and transphobia (e.g. Chakravarthi et al., 2022) have been all been proposed as suitable for automated identification using NLP methods. Collectively these can be referred to as *isms*. We understand *isms* as prejudices, stereotyping, or discrimination on the basis on some personal characteristic. For example, sexism is defined as prejudice, stereotyping, or discrimination, typically against women, on the basis of sex or gender (Masequesmay, 2008).

This line of research has been faced with high annotator disagreement (e.g. Leonardelli et al., 2021), and as a result has conceptualised this as an indication that the concepts themselves are subjective. For example, Rottger et al. (2022) argue that labelling such phenomena is inherently subjective and can either be addressed as *descriptive*, i.e., encouraging annotator subjectivity, or *prescriptive*, i.e., discouraging it. By constructing abuse as individually subjective, social norms are disregarded in favour of an approach that is blind to existing conditions of marginalisation. This stands in contrast to early work in the field, which sought to tease apart the distinction between offensiveness and hate (Davidson et al., 2017), and sought frameworks to identify the particular vectors which indicated hate (Talat et al., 2017; Wright et al., 2017).

Discrimination is also an area subject to policy and regulatory debates. Policy often distinguishes *hate* from *offence*. For instance, in its definition of sexism, the European Institute for Gender Equality (EIGE) position sexism as the *presence* rather than the offensiveness of a gendered stereotype:

> 'Sexism is linked to beliefs around the fundamental nature of women and men and the roles they should play in society. Sexist assumptions about women and men, which manifest themselves as gender stereotypes, can rank one gender as superior to another.'

In this position paper, we consider such *isms* and how offence and hate[1] are orthogonal[2] concepts that can be mutually informative, and argue that their conflation can delegitimise research artefacts and findings. That is, we contend that the hatefulness of a statement is invariant of a reader's position on whether it should be allowed within a particular public forum. Consider for instance the use of gendered slurs: while inappropriate for a general audience (e.g., a public debate) they may be appropriate for others (e.g., academic work exploring the uses of expletives). In particular, we argue that *isms* are culturally defined, whereas offence is a subjective experience. Thus, we argue that it is the presence of a stereotype that determines if

*Equal contribution.

[1] Hate speech 'attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are' (UN), including subtle stereotyping.

[2] We use 'orthagonality' in the philosophical sense to refer to concepts that differ in scope, content, and purpose.

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 275–282

a statement is hate speech, rather than individual perceptions of its offensiveness. Understanding *isms* as culturally defined, and offence as individually subjective allows us to distinguish any offence caused to a reader from whether a message contains hate speech. We therefore call for approaches to annotating online abuse that delineate the degree of offence caused from the phenomenon itself.

2 Understanding Subjectivity

Recent efforts in NLP have constructed annotation as subjective, without attending to what other fields have understood this to mean. *Subjectivity* has been posed as the reason why 'humans (e.g. annotators) [are] sensitive to sensory demands, cognitive fatigue, and external factors that affect judgements made at a particular place and point in time' (Alm, 2011). Philosophy, however, sees *subjectivity* as concerning people's differing perspectives, formed by factors such as cultural and individual experiences (Solomon, 2005). This implies that the only valid knowledge is based on personal experiences, thereby negating the existence of objective or communal truths. In contrast, *relativism* proposes that criteria of judgement are relative to a culture or society (Baghramian, 2004). For instance, while humour may be subjective, we can understand concepts such as beauty to be culturally defined.

Hate speech detection, in particular, has often been argued to be a subjective task (e.g. Almanea and Poesio, 2022; Basile, 2020). Under this framing, researchers collapse the label classes *offensive hate speech* (e.g. Leonardelli et al., 2021), thereby further conflating these concepts. For instance, Akhtar et al. (2021) posit that 'judging whether a message contains hate speech is quite subjective, given the nature of the phenomenon'. When categories of abuse are described as subjective, we understand that there is no ground truth, and wider cultural norms do not impact what constitutes hate. Within the concept of *isms*, we argue that is the wrong approach and that these are culturally defined. That is, we argue that, for a stereotype or norm, there *is* a ground truth given by the cultural and temporal context a statement is made in.

2.1 Stereotypes as Socially-defined Artefacts

Isms are a term given to various forms of marginalization and concepts such as racism, sexism, transphobia, etc. Such *isms* rely on tropes and stereotypes about a target group (Manne, 2017). They describe beliefs about the way a group is and how it ought to be (Ellemers, 2018). Although stereotypes are held by individuals, they are formed collectively (Butler, 1989). For example, stereotypes are observable: we can catalogue the content of gender stereotypes within a culture (Prentice and Carranza, 2002), suggesting these are not solely individual but instead exist in the 'collective brain'.

Haslam et al. (1997) argue that stereotypes emerge when individuals are acting in terms of a common social identity. Although the belief that stereotypes are simply an inferior representation of an unfamiliar group may be alluring, they serve to represent group-based realities: they represent (and accentuate) perceived differences between then in- and out-group (Haslam et al., 1997). Through the lens of self-categorisation theory, Haslam et al. (1997) argue that stereotypes are a social force– they reassure individuals of their belonging to a group 'by: (1) enhancing perceived in-group homogeneity; (2) providing associated expectations of mutual agreement; and (3) producing pressure to actively reach consensus through mutual influence'. Uniformity of belief is thus the very essence of a stereotype. Stereotypes cause harm by limiting people's capacity to develop personally and professionally.[3] The shared nature of stereotypes is what causes their severity, a single individual holding and acting on discriminatory beliefs is less consequential than a group holding and acting on the same beliefs. However, because stereotypes are collective, they are also fuzzy; while individuals in the in-group are at least aware of stereotypes, they do not necessarily believe in them. This is in part why the degree of offence to *isms* may vary. Group memberships and social relations play a key role in shaping cognition, leading to the application and salience of stereotypes to be context-dependent but consensual at the group level nonetheless.

2.2 Acceptability as a Social Norm

Generally speaking, some *isms* are less socially acceptable nowadays than they were a century ago due to the social justice movements of the last century. Such movements have, in some countries, resulted in an increased public awareness of the harms caused by stereotypes, making support for some of them less socially acceptable. That is, the Overton Window, a political theory that describes the spectrum of acceptable policies and discourse, has shifted to make it less socially accept-

[3] United Nations Office of the High Commissioner for Human Rights, accessed 24th April 2024

able to hold particular stereotypical beliefs. The result of such a shift is that people do not wish to label statements they agree with as an *ism* lest they be labelled as **ists* themselves. For instance, homophobia has become less tolerated in many countries, and individuals do not want their statements, or them, to be labelled as homophobic. Yet while being labelled as homophobic is perceived as undesirable, this does not mean that homophobic comments are not made, and policies not pursued. For example, in the United States of America, the American Civil Liberties Union has currently flagged more than 500 legal bills as anti-LGBTQ (American Civil Liberties Union, 2023). Thus, despite forward progress on some forms of discrimination and isms (Azcona et al., 2023; Menasce Horowitz, 2023), there are still socially acceptable *isms* that come in two general flavours: the benevolent *isms* and the scientific *isms*.

The Benevolent **Ism* Some stereotypes may be seen as 'positive' and therefore not recognised by some as hateful. The existence of 'benevolent' stereotypes (Jha and Mamidi, 2017), such as 'neosexism' (Tougas et al., 1995)—those without clear negative connotations—means that annotators may be unlikely to recognise them as harmful. For example, the seemingly positive stereotype in Western nations that Asians are successful, high-achievers leads to their vilification (for being *too* high-achieving) and the perception that they lack interpersonal skills (Wong and Halgin, 2006). These stereotypes may also cause indirect harm to the individuals who may feel they are not living up to what is expected from them (Haslam et al., 1997). We might be tempted to only oppose or target stereotypes that imply or directly state that a certain group is inferior, however this approach would leave many of the issues of stereotyping unaddressed. For example, not addressing claims such as 'women need to be protected' or that 'women's bodies are more aesthetically pleasing' suggests that the perception of women as inferior, or inherently sexualised, should remain acceptable.

The Scientific **Ism* This *ism* uses evolutionary biology as evidence for stereotypes. In this case, different groups are proposed as differing on the basis of *natural* differences, such as physiology. One such example is the idea that women are naturally more nurturing than men due to imaginations of gender roles of the past. However, investigations of hunter-gatherer societies indicate that this idea may not be an accurate reflection of past societies and social evolution (Hewlett and Macfarlan, 2010). The idea of evolutionary psychology as evidence stems from Social Darwinism (Miller, 2011), which argues that one cannot accuse nature of being *-ist*, and therefore any generalisation based on biology cannot be labelled as such. Such pseudo-scientific *isms* are commonly used as a rationalisation for the 'objective' differences between dominant and marginalised groups (e.g. Browne (2006)).

2.3 Separating *Isms* and Offensiveness

So far, we have established that *isms* are rooted in socio-cultural contexts, and, while not necessarily factual or objective, exist as normative and therefore stable concepts, given their socio-cultural and temporal situations. As norms, *isms* can cause harms to members of targeted groups, present barriers to harmonious community relations, or pose threats to law and order (Barendt, 2019).

Offensiveness can be understood as moral outrage or disgust (Sneddon, 2020). As *isms* can be harmful, it is tempting to suggest that they should always be constructed as offensive. However, this would not afford the high levels of disagreement often observed in their annotation. Such disagreement can be accounted for by considering the degree of offence taken as subjective. That is, the degree of offence is knowable only by each annotator. According to Sneddon (2020), we tend to give claims of offensiveness more credence than they deserve. That is, offence itself does not pose a moral harm. People get more offended about topics that particularly matter to them, and these are impacted by one's identity: A citizen of the USA is more likely to be offended by the burning of their national flag than a European. That is to say, when we are offended, we take the object of offence as a personal affront. This has material consequences when it comes to modelling *isms* as offensive.

3 Annotator Competency

Dataset labelling in NLP is typically performed by annotators recruited either as crowd-sourced workers (e.g. Abercrombie et al., 2023a; Basile et al., 2019; Fersini et al., 2018), academics or students available to the researchers (e.g. Cercas Curry et al., 2021; Fanton et al., 2021; Jiang et al., 2022), or people deemed to hold expertise in the phenomena (e.g. Talat, 2016; Vidgen et al., 2021; Zeinert et al., 2021). However, Standpoint Theory (Harding, 1991) argues that annotators, can largely only be competent within their own lived experiences, regardless of training. Without lived experience, annotators may not be able to gain a full understanding of the *ism* under consideration. For instance, Larimore et al. (2021) found that white annotators were far less competent in identifying anti-Black racism than Black annotators. Guidelines and labelling taxonomies, no matter how thoroughly and carefully constructed are not capable of adjusting for a lifetime of lived experience. It is not, therefore, inherent subjectivity within the task, but rather differences in annotator ability due to their personal standpoint that impact on annotators' ability to recognise whether hate speech or abuse is present. Sometimes even if an individual does recognise the target phenomenon, they may choose to ignore it for political reasons (Marable, 1995).

4 Towards a New Formulation of *Isms* as Cultural Formation of Societal Norms

Given our understanding of *isms* as culturally relative constructions and *offence* as an individually subjective concept, we propose that *isms* can best be understood as cultural formations of societal norms. That is, *isms* encode norms, which are inherently fuzzy at the border (Hall, 1997). When creating data for *isms*, researchers often work at the fuzzy borders of acceptability. In operating at these borders, and developing computational methods to draw them, research delineates what is acceptable from that which is not. While such borders are inherently messy, through an understanding of determining acceptability as cultural norms, we can refocus our attention towards the question of how such norms and borders should be drawn.

For instance, Douglas (1978) argues that determining what is 'dirt' is a cultural process which strengthens communities and builds community cohesion. That is, while encountering an offensive instance, i.e., an instance of sexism, can be destabilising to a community, the process with which the community makes a determination, and the determination itself, allows for the community to reify itself. This is particularly important as we can come to understand that *isms* are culturally defined objects, and identifying the borders of acceptability necessitates an ongoing negotiation with the communities in question (Thylstrup and Talat, 2020). Within this formulation of *isms*, we can come to understand *isms* as distinct from *offence*. Thus, this formulation of *isms* provides space for both a cultural understanding of *isms* whilst making space for *offence* as an individual and subjective notion.

5 Recommendations

We have argued that conflation of *isms* and offence stems from annotation **task construction**. We recommend that schema be designed to carefully delineate these concepts, by e.g., creating distinct categories, and labelling them separately. Researchers should be clear about the phenomenon they are investigating. If the task is *offensiveness*, a subjective framing may very well be appropriate. In the case of *isms*, given the confusion surrounding them, the question posed to annotators may be better phrased as whether the instance makes reference to stereotypes about a particular group.

As guidelines cannot meaningfully offset gaps between annotators and any missing lived experience required to identify *isms*, we recommend that **annotator recruitment** target people with relevant profiles to label the data in question. We recommend subject-area experts, such as feminist scholars or those working in the target area such as relevant NGO and activist stakeholders, be involved at every stage of the data annotation process and their expertise to be carefully incorporated into the schema (Abercrombie et al., 2023b). In the case where experts are out of reach, annotators should be recruited to label data for which they have lived experience. Where this is not possible, schema should allow annotators the option of indicating where they do not have the necessary lived experience to label specific items.

6 Conclusion: Implications for NLP

If, as we propose, identifiying *isms* is not subjective, we must conclude that annotator differences are irrelevant at the individual level for such tasks. Rather, they are symptoms of disagreement on the degree to which *isms* offend individual annotators.

At the group level, we must take care not to treat conflicting responses equally. If a minority with the necessary lived experience (e.g. to recognise misogyny) disagree with the majority who don't, that matters. For example, Gordon et al. (2022) attempt to pick out the 'correct' minority perspectives from the wider pool of annotators for each instance, and Fleisig et al. (2023) specifically assume that the majority of annotators are likely 'wrong', i.e., they will not recognise the target phenomenon. However, belonging to the targeted group is not necessarily sufficient.

Construction of the desired classification schema based on societal norms comes with its challenges. While prescriptivist annotation based on agreed societal norms may be desired, it can be difficult or even impossible to implement comprehensively in practice. One reason for this is that it is probably not possible to recruit annotators with the correct standpoint or competencies to recognise every instance—or indeed to know what those characteristics might be. Another is the nature of building classification schema. While a clearly defined, unambiguous, comprehensive and static *Aristotelian* classification scheme may be desired rather than *prototypical* classification, it can be hard or even impossible to implement, and people generally resort to the latter (Bowker and Star, 2000, p. 61-62).

Despite this, we believe that it is vital that *isms* like misogyny and other hate and abuse not be constructed as individually subjective, but rather as culturally formed societal norms. While there may be much to gain from examining the responses of individual annotators to these tasks, NLP researchers should be careful not to conflate individual differences with inherent subjectivity of tasks.

Limitations

We have presented a position on the modelling of hate speech in NLP backed by existing literature in philosophy, gender studies, and critical race theory. While we have made actionable recommendations for NLP researchers working on hate speech and related phenomena, schema definition and annotator recruitment to exactly capture a phenomenon are known to be challenging. We encourage researchers to follow best practices and involve interdisciplinary researchers and other stakeholders given the nature of the particular task.

Ethical Considerations

This paper presents a re-framing of tasks related to hate speech and abusive language detection. In this new frame, we delineate between that which causes offence at an individual level and that which is hate, defined at a societal level with regard to concepts such as sexism, racism, and so forth, collectively referred to as *isms*. From this understanding of *isms*, it becomes clear that current practices reinforce social norms of desirability and respectability. The implications of disentangling offence from *isms*, is then to disentangle individual desirability from our understanding and modelling of *isms*. Consequently, our framing makes space for marginalised communities to name the discrimination that they are subject to, without also making determinations on whether discriminative messages should be moderated for *all* potential viewers. This affords space for marginalised communities, in particular, to call out the discrimination that they are subject to, regardless of whether others recognise that discrimination. Furthermore, by disentangling offence from *isms*, public policy analysis and decisions on what should be regulated and what should be subject to individual preference can disregard whether content causes offence, and instead pay attention to whether the content constitutes a discriminatory statement on its own merits. Data and models that arise from disentangling offence from *isms* thus afford individuality in terms of what causes offence to an individual, and therefore what they would wish to (not) be exposed to, without making inference as to whether that content constitutes an *ism*. Further, our framing of *isms* removes sovereignty to individually define and operationalise *isms*. Instead, we follow Butler (1989) in their understanding that *isms* arise from the socio-cultural citations of past events, i.e., from the norms that are established and reused in a given society over time. Thus, establishing what constitutes an *ism* is a task that must be conducted by examining the social and political conditions in a given society and is liable to change with society.

Acknowledgements

Amanda Cercas Curry was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). She is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA). Gavin Abercrombie was supported by the EPSRC projects 'Equally Safe Online' (EP/W025493/1) and 'Gender Bias in Conversational AI' (EP/T023767/1).

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023a. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023b. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

American Civil Liberties Union. 2023. Mapping Attacks on LGBTQ Rights in U.S. State Legislatures.

Ginette Azcona, Antra Bhatt, Guillem Fortuny Fillo, Yongyi Min, Heather Page, and Sokunpanha You. 2023. *Progress on the Sustainable Development Goals: The gender snapshot 2023*. United Nations Entity for Gender Equality and the Empowerment of Women (UN Women) Department of Economic and Social Affairs (DESA).

Maria Baghramian. 2004. *Relativism*. Routledge.

Eric Barendt. 2019. What is the harm of hate speech? *Ethical Theory and Moral Practice*, 22:539–553.

Valerio Basile. 2020. It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. In *Proceedings of The 19th International Conference of the Italian Association for Artificial Intelligence*, volume 2776, pages 31–40. CEUR-WS.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

Kingsley R Browne. 2006. Sex, power, and dominance: The evolutionary psychology of sexual harassment. *Managerial and Decision Economics*, 27(2-3):145–158.

Judith Butler. 1989. *Excitable Speech: A Politics of the Performative*, 1 edition. Routledge.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25(2):700–732.

Mary Douglas. 1978. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*, repr edition. Routledge, London.

EIGE. Sexism at work handbook. https://eige.europa.eu/publications-resources/toolkits-guides/sexism-at-work-handbook/part-1-understand/what-sexism?language_content_entity=en. Accessed June 20 2023.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69:275–298.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, Online. Association for Computational Linguistics.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. In *Ibereval@ sepln*, volume 2150, pages 214–228.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Stuart Hall. 1997. Race, the Floating Signifier.

Sandra Harding. 1991. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

S Alexander Haslam, John C Turner, Penelope J Oakes, Craig McGarty, and Katherine J Reynolds. 1997. The group as a basis for emergent stereotype consensus. *European review of social psychology*, 8(1):203–239.

Barry S. Hewlett and Shane J. Macfarlan. 2010. Fathers' roles in hunter-gatherer and other small-scale cultures. In Michael E. Lamb, editor, *The Role of the Father in Child Development*, pages 413–434. John Wiley & Sons, Inc., Hoboken, NJ.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

M. Marable. 1995. *Beyond Black and White: Transforming African-American Politics*. Verso.

Gina Masequesmay. 2008. Sexism. In Jodi O'Brien, editor, *Encyclopedia of gender and society*. Sage.

Juliana Menasce Horowitz. 2023. *Martin Luther King Jr.'s Legacy 60 years after the March on Washington: Views of the country's progress on racial equality*. Pew Research Centre.

Geoffrey Miller. 2011. *The mating mind: How sexual choice shaped the evolution of human nature*. Anchor.

Deborah A Prentice and Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, 26(4):269–281.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Andrew Sneddon. 2020. *Offense and offensiveness: a philosophical account*. Routledge.

R. C. Solomon. 2005. Subjectivity. In Ted Honderich, editor, *The Oxford Companion to Philosophy*, page 900. Oxford University Press.

Zeerak Talat. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour.

Francine Tougas, Rupert Brown, Ann M Beaton, and Stéphane Joly. 1995. Neosexism: Plus ça change, plus c'est pareil. *Personality and social psychology bulletin*, 21(8):842–849.

UN. What is hate speech? Accessed February 14 2024.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Frieda Wong and Richard Halgin. 2006. The "Model Minority": Bane or Blessing for Asian Americans? *Journal of Multicultural Counseling and Development*, 34(1):38–49.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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 3181–3197, Online. Association for Computational Linguistics.