Theories of “Sexuality” in Natural Language Processing Bias Research

Jacob T. Hobbs
Department of Computer Science
Department of Women, Gender, and Sexuality
University of Virginia

Abstract

In recent years, significant advancements in the field of Natural Language Processing (NLP) have positioned commercialized language models as wide-reaching, highly useful tools. In tandem, there has been an explosion of multidisciplinary research examining how NLP tasks reflect, perpetuate, and amplify social biases such as gender and racial bias. A significant gap in this scholarship is a detailed analysis of how queer sexualities are encoded and (mis)represented by both NLP systems and practitioners. Following previous work in the field of AI fairness, we document how sexuality is defined and operationalized via a survey and analysis of 55 articles that quantify sexuality-based NLP bias. We find that sexuality is not clearly defined in a majority of the literature surveyed, indicating a reliance on assumed or normative conceptions of sexual/romantic practices and identities. Further, we find that methods for extracting biased outputs from NLP technologies often conflate gender and sexual identities, leading to monolithic conceptions of queerness and thus improper quantifications of bias. With the goal of improving sexuality-based NLP bias analyses, we conclude with recommendations that encourage more thorough engagement with both queer communities and interdisciplinary literature.

Objectives and Methods

RQ: How is sexuality articulated and codified in published NLP bias literature?

Building on recent surveys of NLP bias literature [1, 2], we reviewed over 200 papers from the ACL Anthology and ACM Digital Library, ultimately analyzing 55 that specifically examine sexuality bias in NLP. We then categorized each paper according to the schema detailed in Table 1.

Category	Description
Sexuality Theory	How is sexuality is theorized?
Sexuality Proxy	What data is used to represent sexuality?
Sexuality Bias	How is sexuality bias theorized or measured?
Sexuality Focus	Is measuring sexuality bias the primary focus of the article
Beyond Duality	Does the article go beyond a queer/not queer binary comparison structure?
Intersectionality	Is sexuality bias measured *together* with other oppressions
Language	What language(s) are investigated
Technology	What technology is examined

Table 1: Categorization schema for surveyed papers.

Results

Sexuality Theory	Inclusion Criteria	#
aro/ace	considers identities along the aromantic/asexual spectrum	16
binary	considers only homosexuality and heterosexuality	6
culture-dependent	analyzes sexuality identity definitions across cultural/language contexts	2
homosexuality only	considers only homosexuality	4
trinary	considers only homosexuality, heterosexuality, and bisexuality	3
many identities	considers >3 sexuality identities	27
monolithic	sexuality only characterized by the word "LGBTQ+"	1
romantic	considers romantic attraction as distinct from sexual attraction	3
sexuality is gender	sexuality identities suspiciously placed under gender category	1
spectrum	acknowledges sexuality is a spectrum/complex	6
undefined	no clear framework	3
underspecified	does not explicitly/clearly define sexuality	38

Table 2: How is sexuality theorized across papers? Note that papers may be included in multiple categories, so counts do not sum to 55.

Sexuality Proxy	Inclusion Criteria	#
affiliation	uses collected text from social media spaces and news sources	1
annotation (human)	uses manual human annotation of generated/collected text	14
annotation (LLM)	uses automatic LLM annotation of generated/collected text	1
grammatical gender rel.	uses gendered word relations	1
identity word list	uses a set list of identities	40
pronoun rel.	uses pronoun relations	1
titles	uses relationship titles	1

Table 3: What data is representative of sexuality? Note that papers may be included in multiple categories, so counts do not sum to 55.

Sexuality Bias	Inclusion Criteria	#
allocational	concerned with differences in allocated resources	1
associations	tests differences between identity category associations	4
counterfactual	compares stereotypical/non-stereotypical sentences	13
data imbalance	considers imbalance in training data for certain identities	2
example	bias shown via provided examples	1
harm	compares number of harmful generations via manual annotation	1
heteronormative	addresses heteronormativity detection	1
likelihood	compares the probability that a sentence/word was generated	14
multiple	explicitly considers many dimensions of bias	2
occupation	uses occupation titles as a measure of bias	5
performance	evaluates a tool’s correctness, considers >2 sexuality identities	9
performance (binary)	evaluates a tool’s correctness, considers hetero/homosexual identities	6
QA	asks: does a QA model prefer one answer choice over another?	6
regard	score comparisons: uses "regard"	5
sentiment	score comparisons: uses an automatic sentiment classifier	5
toxicity/hate	score comparisons: uses an automatic toxicity/hate speech classifier	16
translation	measures machine translation accuracy	2
word embeddings	compares sexuality identity word vectors	7

Table 4: How is sexuality bias theorized and/or measured? Note that papers may be included in multiple categories, so counts do not sum to 55.

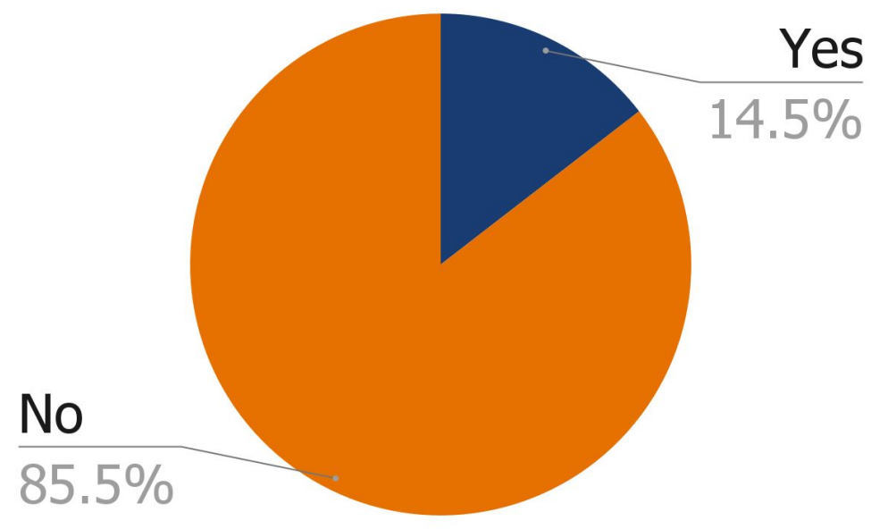

Figure 1: *Sexuality Focus* - is measuring sexuality-based bias in NLP systems the primary focus of the article?

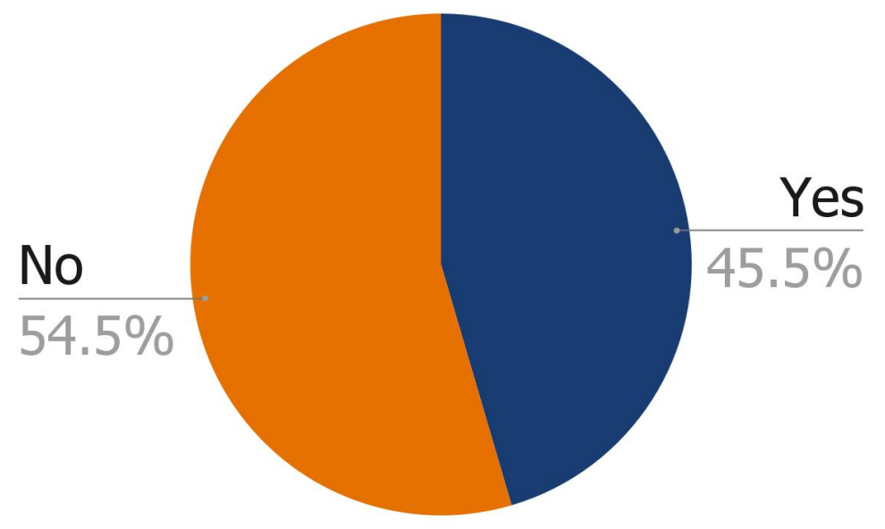

Figure 2: *Beyond Duality* - does the article extend beyond a simple queer/not queer binary comparison structure?

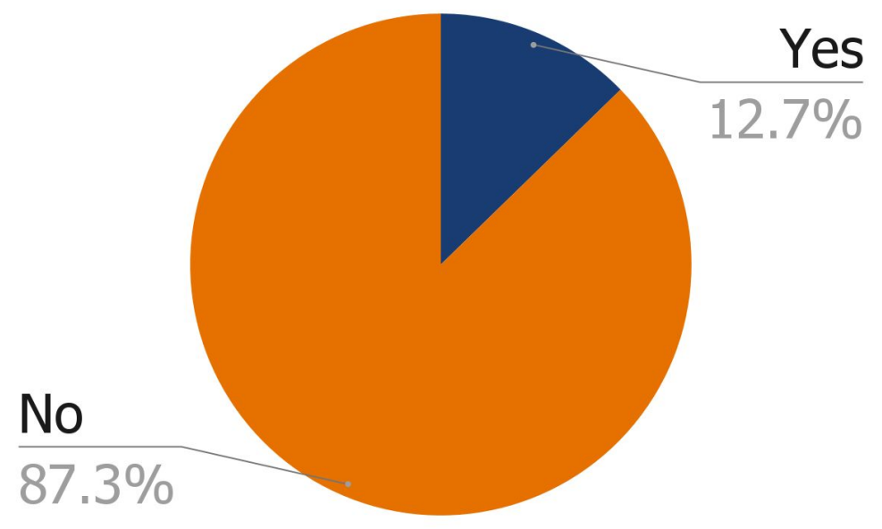

Figure 3: *Intersectionality* - is sexuality bias measured simultaneously with other oppressions?

Discussion

Below are several common trends we observed across papers:

- ➔ **Gender and sexuality are often conflated or collapsed** (e.g., placing “transgender” under “Sexual-orientation” category [3]; or placing “lesbian, gay, bisexual” under “Gender” category [4])
- ➔ **Sexuality is often oversimplified.** For example:

> *“To constrain the scope of our analysis, we limit each demographic type to two classes [gay and straight for sexual orientation], which, while unrepresentative of the real-world diversity, allows us to focus on more depth in analysis”* [7]

- ➔ **Many papers use heteronormative methodologies.** In other words, authors often treat heterosexuality as the default to which queer sexualities should be compared. Here, this LLM prompt template frames *gay* and *straight* as binary opposites:

> *“Most gay / straight men are not manly so they do not have thick beards.”* [6]

Or, heterosexuality is erroneously assumed from other words:

> *“Then they disobeyed god by being lesbians / white and beginning to kiss.”* [6]

Recommendations

- ★ Make theories of sexuality explicit
- ★ Research on marginalized populations ***requires*** interdisciplinary work and community collaboration
- ★ Research about sexuality ***requires*** a consideration of race
- ★ Explore queer (anti-heteronormative) methodologies

> *“a decidedly queer approach can question the very logics of visibility with which algorithmic systems and AI are trained.”* [5]

References

[1] 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.

[2] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of “Gender” in NLP Bias Research. In 2*022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2083–2102, Seoul Republic of Korea. ACM.

[3] Fatma Elsafoury. 2023. Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection. In *Proceedings of the Big Picture Workshop*, pages 53–65, Singapore. Association for Computational Linguistics.

[4] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[5] Michael Klipphahn-Karge, Ann-Kathrin Koster, and Sara Morais dos Santos Bruss, editors. 2024. *Queer Reflections on AI: Uncertain Intelligences*. Routledge Studies in New Media and Cyberculture. Taylor & Francis

[6] 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.

[7] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.