

Revisiting Queer Minorities in Lexicons

Krithika Ramesh[♣]

Sumeet Kumar[♡]

Ashiqur R. KhudaBukhsh^{♣*}


[♣]Manipal University

[♡]Indian School of Business

[♣]Rochester Institute of Technology

kramesh.tlw@gmail.com, Sumeet_Kumar@isb.edu, axkvse@rit.edu

Abstract

 This paper contains words that are offensive.

Lexicons play an important role in content moderation, often being the first line of defense. However, little or no literature exists in analyzing the representation of queer-related words in them. In this paper, we consider twelve well-known English lexicons containing inappropriate words and analyze how gender and sexual minorities are represented in these lexicons. Our analyses reveal that several of these lexicons barely make any distinction between pejorative and non-pejorative queer-related words. We express concern that such unfettered usage of non-pejorative queer-related words may impact queer presence in mainstream discourse. Our analyses further reveal that the lexicons have poor overlap in queer-related words. We finally present a quantifiable measure of consistency and show that several of these lexicons are not consistent in how they include (or omit) queer-related words.

1 Introduction

On August 23, 2013, the online version of the Oxford English Dictionary updated the meaning of a word. Updates to this dictionary are not uncommon. However, the updates typically include new words in the latest edition. For instance, `Bollywood`, the notorious name for the Mumbai film industry, made its way into the dictionary in 2004. Or, for example, the ongoing pandemic forced a slew of vaccine-related words – `vaccine passport`, `vaccine hesitancy`, and `vaxxed` – into the 2021 edition. Every new edition introduces several such words reflecting the ever-changing world with intermixing cultures and acknowledging the fluid and expansive nature of English – one of the

most popular, pluricentric world languages (Leitner, 1992).

What was remarkable about the August 23, 2013, online update was that this word had its first known usage in the 14th century, and its primary meaning remained unaltered since its inclusion in the very first edition of the Oxford dictionary! `Marriage`, previously defined as the *formal union of a man and a woman, typically as recognized by law, by which they become husband and wife*, received an inclusive definition in the dictionary following the legalization of gay marriage in the UK. The new definition dispensed with the gender restriction and defined marriage as a union between two persons.

Words and their meanings exist in a continuum (Hamilton et al., 2016; Xie et al., 2019), often shifted and shaped by evolving social norms, hard-fought legal acceptances, and new world events. Lexicons proposed to aid content moderation, in turn, exhibit a rather static nature and a much narrower scope, representing a collection of words deemed as potentially hateful/harmful/abusive/toxic/offensive by a group of annotators (possibly exhibiting limited diversity and/or with under-specified expertise) at a given point of time. In this paper, we focus on twelve such lexicons aimed at aiding content moderation. A varied collection of words have been used to describe them, including being termed as abusive, offensive, profane, toxic, and hate speech lexicons. We use an umbrella term *inappropriate* to refer to any of these descriptions. In this paper, we focus on twelve inappropriate lexicons and analyze the presence (and absence) of words related to gender and sexual minorities (we call these words queer-related words) in them¹.

Our paper seeks to attract the attention of the

¹Code and additional resources are available at <https://github.com/stolenpyjak/revisiting-queer-lexicons>.

* Ashiqur R. KhudaBukhsh is the corresponding author.

broader community of psycho-linguistic experts and ethicists on the following issues.

First, our study reveals that these lexicons have limited overlap, and many of these under-specify how they were obtained. While data sets have received considerable attention for audits (Gebru et al., 2021), inappropriate lexicons have received little or no attention for quality control. Given that such lexicons often serve as the first line of defense against inappropriate content, certain omissions and inclusions can significantly influence what gets flagged as inappropriate and may impact minorities to get their voices heard. As we seek to move towards more transparent, responsible, and ethical AI systems, we need to build stronger guardrails for methods and resources that are used for content moderation/filtering.

We see our work as a voice in the scientific conversation focusing on the treatment of the queer community in language technologies (Dev et al., 2021; Nozza et al., 2022; Dodge et al., 2021). Among these recent prominent studies, Dev et al. (2021) discuss the potential erasure of non-binary identities due to stereotypical harms propagated by language models; Nozza et al. (2022) reveal that large language models exhibit discriminative behavior by producing harmful text completions for subjects from the queer community; and Dodge et al. (2021) demonstrate how blacklist-based filters have been shown to remove content related to the queer community, particularly when it contains terms related to sexual orientation. Our work focusing on queer-related terms in inappropriate lexicons complements these aforementioned important studies.

Second, our study raises a question that we believe is timely and important. We observe that several non-pejorative words representing gender and sexual minorities (e.g., `gay`, `queer`, `lesbian`, `trans`) are present in these inappropriate lexicons. However, these lexicons often do not make any clear distinction between the targets for harm and targeted harms. We worry that unfettered use of `gay`, `lesbian` or `trans` along with their pejorative versions (e.g., `faggot`²) within the same lexicon may hinder the inclusion of sexual minorities into mainstream discourse. Thus we seek guidance

²In this paper, we have not censored any of these historically charged words. There is a broad range of opinions and practices on censoring (or not censoring) historically charged words (Cannon, 2005; Stephens-Davidowitz and Pabon, 2017; Sap et al., 2020; Schick et al., 2021).

from true experts on this issue that may significantly influence how a safe web may look like for sexual minorities in the future.

Third, continuing the same thread of discussion surrounding the inclusion or omission of non-pejorative versions representing gender and sexual minorities, we present a first step towards quantifying inconsistencies in lexicons with respect to queer-related words. Our study reveals that these lexicons exhibit inconsistencies that can potentially influence content moderation outcomes if these lexicons are used as an aid.

2 Design Considerations

2.1 Classification of Lexicons into Abusive, Offensive, and Hate Speech

As mentioned in Davidson et al. (2017), the difference between hate speech, offensive language, and abusive language is that hate speech tends to be directed toward specific communities so as to disparage or disadvantage them. Davidson et al. (2017) also state that their definition of hate speech may not include all instances of offensive language, as it is possible that these derogatory terms that target certain communities may be used in a manner that is not necessarily motivated by the intention to deride the said community. This includes words that have been reclaimed by the very same groups they were meant to stigmatize. This distinction is important as the resulting lexicon used in offensive/abusive language detection may vary from those used in hate speech detection, as the latter may contain more relevant pejoratives targeted at specific demographics. Caselli et al. (2020) explore the distinction between abusive language and offensive language. According to Caselli et al. (2020), abusive language focuses more on the intention of the message conveyed, and offensive language emphasizes more on the target’s sentiment and the profanity in the message. However, profane language is shown to fall under both these categories. Additionally, we find that the source for some of our lexicons uses the terms *profane*, *abusive* and *offensive* interchangeably. The term *toxicity* is also used for one of these lexicons, which Mohan et al. (2017) use to refer to various forms of harassment, such as hate speech, cyber threats, cyberbullying, etc. As our lexicons are obtained from multiple sources with various such classifications and definitions of their own, we thereby deem it necessary to classify all these words as *inappropriate words*

that cover a broad taxonomy of potentially harmful language.

2.2 Development of Queer Lexicon

In order to carry out our analysis across these English lexicons, we survey several web sources to identify terms that are commonly used among the queer community. We compile terms based on both gender and sexuality (including any pejorative terms encountered) from multiple online resources ³.

The non-pejorative version of the lexicon was obtained by eliminating terms that are considered pejorative from multiple sources, including ⁴. Overall, our list of queer-specific words, \mathcal{L}^Q , consists of 115 terms. Of this, we identify 28 as pejorative (denoted as \mathcal{L}_p^Q) and 87 as non-pejorative terms (denoted as \mathcal{L}_{np}^Q). These 115 terms have consensus labels from two annotators, one cis-female and one cis-male, of whom one identifies as a queer.

We acknowledge that our list is not comprehensive and may (inadvertently) fail to include terms pertaining to several sexualities and genders across the spectrum. We further note that some of the terms in this non-pejorative version of the lexicon (such as *gay*) can be considered derogatory based on context. Similarly, as mentioned in Section 2.1, some of the terms not present in the non-pejorative version of this lexicon have been reclaimed by some parts of the queer community and, therefore, may not be considered derogatory in a given context. Ideally, we feel that studies that aim to construct and utilize lexicons should provide information regarding the same (see, e.g., Pamungkas et al. (2022)), as opposed to imposing a blanket statement (via their lexicon) that dictates that terms like *gay* are considered offensive language or hate speech.

Overall, we use 12 well-known lexicons listed in Table 1. In addition, we also present the overlap of individual lexicons with \mathcal{L}^Q , \mathcal{L}_{np}^Q and \mathcal{L}_p^Q along with any publicly available annotation details.

³<https://www.smcgov.org/lgbtq/lgbtq-glossary>
<https://www.itspronouncedmetrosexual.com/2013/01/a-comprehensive-list-of-lgbtq-term-definitions/>
<https://www.healthline.com/health/different-types-of-sexuality#takeaway>

⁴<https://www.advocate.com/arts-entertainment/2017/8/02/21-words-queer-community-has-reclaimed-and-some-we-havent>

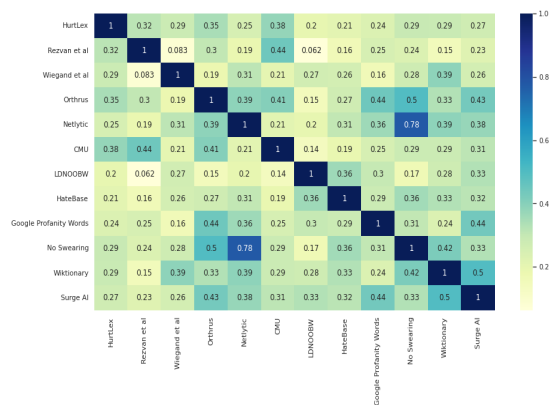


Figure 1: Jaccard similarity of all queer-related words in the inappropriate lexicons. Jaccard similarity is a statistic to gauge similarity between two sets, \mathcal{A}, \mathcal{B} , expressed as $\frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}$.

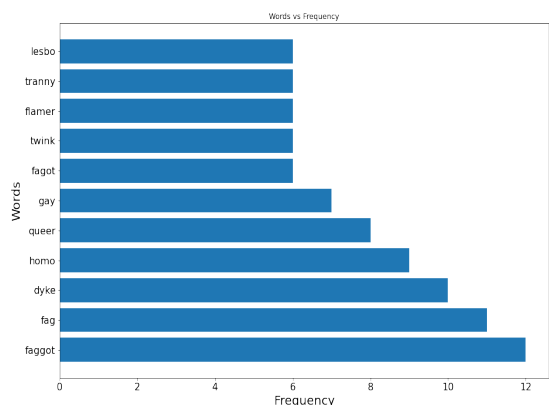


Figure 2: Some of the most frequently occurring queer-related words in the English lexicons.

3 Analysis

We now present an analysis of these lexicons considering the following aspects.

Coverage: We first note that the overlap between \mathcal{L}^Q and the twelve inappropriate lexicons is minimal, with the CMU Lexicon achieving the highest overlap (23.48%), indicating that a vast majority of the queer lexicon is not incorporated into any of the well-known lexicons. When we combine all lexicons, the resulting lexicon has a slightly higher overlap of 40.87%. As shown in Figure 1, within the lexicons, limited overlap of these queer-related terms exists. These findings point to the following observations. First, lexicons can benefit from further inclusive efforts in identifying pejorative (if the sole intended purpose is to detect harm) and non-pejorative (if the purpose also involves detecting targets of harm) queer-related terms. Second, given that there is poor overlap within lexicons with

Name	Year	Size	Annotation Method	Overlap with \mathcal{L}^Q	Overlap with \mathcal{L}_{np}^Q	Overlap with \mathcal{L}_p^Q	Classification
HurtLex	2019	5,963	Experts	11.3%	6.9%	25%	Offensive, aggressive, and hateful words
Rezvan et al. (2018)	2018	700	Crowdsourced sources, compiled by a Native English speaker	10.43%	8.05%	17.86%	Offensive/Profane words
Wiegand et al. (2018)	2018	7,049	Experts	12.17%	4.6%	35.71%	Abusive words
Palomino et al. (2021)	2021	1,924	Compiled from multiple lexicon sources	15.65%	9.2%	35.71%	Toxic/Profane words
Kwon and Gruzd (2017)	2017	426	Crowdsourced with custom expert additions	6.09%	1.15%	21.43%	Offensive words
CMU Lexicon	Not specified	1,383	Not specified	23.48%	16.09%	46.43%	Offensive/Profane words
LDNOOBW	2019	403	Not specified	4.35%	0%	17.86%	Offensive/Profane words
HateBase	2019	1,522	Crowdsourced	8.7%	2.3%	28.57%	Hate speech lexicon
Google Profanity Words	2022	451	Not specified	6.96%	0%	28.57%	Offensive/Profane words
NoSwearing	2022	361	Partially crowdsourced list	7.83%	3.45%	21.43%	Offensive/Profane words
Wiktionary	2022	4,738	Crowdsourced	15.65%	3.45%	53.57%	Offensive/Profane words
Surge AI	Not specified	1,598	Not specified	13.04%	1.15%	50%	Offensive/Profane words

Table 1: Details about English lexicons and their overlap with \mathcal{L}^Q and \mathcal{L}_p^Q .

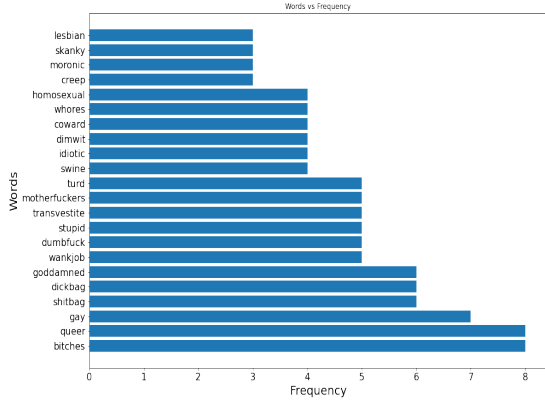


Figure 3: Most frequently occurring queer-related words juxtaposed with similarly frequently occurring slurs from the lexicons.

respect to queer-related terms, consulting multiple lexicons can improve coverage.

Annotation: We note that four lexicons have not specified how they are annotated. Of the remaining, only three are vetted by experts. Existing lexicons with an unspecified annotation that can potentially decide the content outcome for minorities is a major concern, and we identify this as an area where future lexicons can substantially improve.

Presence of pejorative and non-pejorative terms: We note that ten lexicons have more pejorative queer-related words than non-pejorative queer-related words (in terms of absolute value). We argue that putting the pejorative and non-pejorative terms together in the same lexicon potentially con-

Name	Consistency %
(Bassignana et al., 2018)	55.56
(Rezvan et al., 2018)	66.67
(Wiegand et al., 2018)	100.0
(Palomino et al., 2021)	66.67
(Kwon and Gruzd, 2017)	88.89
CMU Lexicon	88.89
LDNOOBW	88.89
HateBase	77.78
Google Profanity Words	88.89
NoSwearing	77.78
Wiktionary	77.78
Surge AI	100.0

Table 2: Consistency % of the English Lexicons

flates between targets of harm and words to inflict harm. As shown in Figure 2, among the most-frequent queer-related words in the lexicon, `gay` and `queer` are present. To emphasize our point further, Figure 3 juxtaposes a few words from \mathcal{L}_{np}^Q along with other similarly frequent words across the lexicons. We note that words like `motherfuckers` or `whores` have appeared less frequently than `queer` or `gay`! We believe that unless these lexicons present concrete examples distinguishing between pejorative and non-pejorative usage of `gay` as presented in Pamungkas et al. (2022), unfettered use of non-pejorative queer-related terms can seriously limit queer presence in mainstream discourse.

Consistency: If a lexicon contains both `dyke` and `faggot` in it yet omits `tranny`, content moder-

ation outcomes (that considers this lexicon) could affect the transgender minority. Similarly, notwithstanding our earlier point that speculates if non-pejorative queer specific words should be at all present in an inappropriate lexicon, presence of `gay` in the lexicon but absence of `lesbian` could potentially trigger differential content moderation treatment for the two communities. In what follows, we develop simple constraints and quantify how consistent published lexicons are. We acknowledge that our choice of lexicon subsets and defined constraints are somewhat over-simplified and a far more nuanced treatment is possible, our primary goal in this experiment is to attract the research community’s attention about addressing these potential inconsistencies that can pave the way towards better practices in future lexicons.

Let \mathcal{L}_{np} and \mathcal{L}_p denote two disjoint lexicon subsets where \mathcal{L}_{np} contains non-pejorative queer-related words and \mathcal{L}_p contains pejorative queer-related words; i.e., $\mathcal{L}_{np} \cap \mathcal{L}_p = \emptyset$. Further, let a bijective mapping f from \mathcal{L}_{np} to \mathcal{L}_p exist, i.e., for each element in \mathcal{L}_{np} , a corresponding unique element in \mathcal{L}_p exists and vice versa. Let the function, f , returns the corresponding pejorative word.

We define $\mathcal{L}_{np} = \{\text{gay}, \text{lesbian}, \text{trans}\}$ and $\mathcal{L}_p = \{\text{faggot}, \text{dyke}, \text{tranny}\}$. Next, we define the following constraints with respect to a lexicon \mathcal{L} :

1. $\forall w_1, w_2 \in \mathcal{L}_{np}$, if $w_1 \in \mathcal{L}$ then $w_2 \in \mathcal{L}$
2. $\forall w_1, w_2 \in \mathcal{L}_p$, if $w_1 \in \mathcal{L}$ then $w_2 \in \mathcal{L}$
3. $\forall w \in \mathcal{L}_{np}$, if $w \in \mathcal{L}$ then $f(w) \in \mathcal{L}$. If $f(w) \notin \mathcal{L}$, we impose a penalty of equal weight. That is, if `gay` exists in the lexicon, but its pejorative counterpart `faggot` does not, we penalize the consistency score by the same weight awarded to a lexicon with both the pejorative and non-pejorative versions.

The consistency of these lexicons based on these constraints are depicted in Table 2, with lexicons that contain neither words from \mathcal{L}_p or \mathcal{L}_{np} being declared completely consistent as well. The lexicons from Wiegand et al. (2018) and the Surge AI profanity lexicon⁵ do not fall under this category, and are the most consistent. It is worth noting that neither of these lexicons contains words from the non-pejorative set \mathcal{L}_{np} .

⁵<https://www.surgehq.ai/datasets/profanity-dataset>

4 Conclusions and Discussions

In this paper, we analyze the presence of queer-related words in several well-known inappropriate English language lexicons. Our analysis identifies possible avenues to provide stronger guardrails against potential harm through (1) expanding lexicons with additional terms; (2) setting more transparent annotation guidelines; (3) distinguishing between pejorative and non-pejorative queer related terms; and (4) improving lexicon consistency concerning queer-related terms.

We believe our most important contribution is raising the question of whether non-pejorative queer-related terms should appear in inappropriate lexicons to begin with. With the current disturbing situation in US politics, where six states are considering passing what the proponents of minority rights dub as the *Don’t say gay bill*⁶, we strongly feel that including non-pejorative queer-related words merits serious discussion. We believe our paper will motivate a scientific dialogue by setting better guidelines to encourage queer presence in mainstream discourse.

Our work raises several important points to ponder.

Grounding Other Research Efforts: Apart from aiding content moderation, inappropriate lexicons can lend grounding to other research efforts. For example, a recent paper (Ramesh et al., 2022) has consulted the CMU Lexicon and another lexicon listing *taboo-words* for kids (Jay, 1992) to construct a set of inappropriate words for kids. Ramesh et al. (2022) take a rather passive stance in their treatment of queer-related words. Ramesh et al. (2022) state that the authors extensively debated whether non-pejorative queer-related words such as `gay` or `queer` should be in the lexicon, but since these words were already present in both lexicons, they retain them, seeking more inputs from developmental psychologists. Unless the research community takes a more definitive stance on when and how non-pejorative queer-related words should be included in these inappropriate language lexicons, we may see more research efforts sidestepping this important issue.

Cultural Effect: Our study is limited to English lexicons. We notice the non-uniform presence of queer-related words across lexicons even within

⁶<https://www.npr.org/2022/04/10/1091543359/15-states-dont-say-gay-anti-transgender-bills>

that. Different countries and cultures have varying degrees of legal, social, and cultural acceptance of the queer community. We believe our study will open the gates for a multi-lingual, multi-cultural analysis of queer presence in inappropriate lexicons.

In-The-Wild Impact Assessment: We hypothesize that lexicon variations can influence content outcome when deployed in the wild to decide the moderation fate of web users. While some anecdotal evidence already exists⁷, an extensive in-the-wild impact assessment of how different lexicons can affect content moderation outcomes can further strengthen our findings.

A List To Criticize Other Lists: Regardless of how well-meaning our intentions are, the 115 queer-related terms chosen by our annotators affect our analyses. Nonetheless, we point out that several of our findings are unaffected (or minimally affected) by \mathcal{L}^Q . For example, the annotation details (or lack thereof) of the inappropriate lexicons have nothing to do with \mathcal{L}^Q . Second, our consistency analysis focuses on a handful of pejorative and non-pejorative queer-related words that are well-recognized by the community. Finally, using well-recognized non-pejorative words such as `gay` and `queer` to substantiate our argument, we show that certain non-pejorative queer-related words are more frequently listed than unambiguously inappropriate non-queer-related words.

5 Acknowledgements

We thank the anonymous reviewers for their thoughtful suggestions. We thank Joseph W. Hostetler for his valuable input.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurltlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

Kevin D Cannon. 2005. “ain’t no faggot gonna rob me!”: Anti-gay attitudes of criminal justice undergraduate majors. *Journal of Criminal Justice Education*, 16(2):226–243.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language.

⁷<https://www.wired.com/story/ai-list-dirty-naughty-obscene-bad-words/>

Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

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.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Timothy Jay. 1992. *Cursing in America*, volume 10. Philadelphia: John Benjamins.

K. Hazel Kwon and Anatoliy Gruzd. 2017. *Interpersonal swearing dictionary*.

Gerhard Leitner. 1992. English as a pluricentric language. *Pluricentric languages: Differing norms in different nations*, 62:178–237.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. pages 51–56.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Marco Palomino, Dawid Grad, and James Bedwell. 2021. GoldenWind at SemEval-2021 task 5: Orthrus - an ensemble approach to identify toxicity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 860–864, Online. Association for Computational Linguistics.

- Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2022. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, pages 1–34.
- Krithika Ramesh, Ashiqur R. KhudaBukhsh, and Sumeet Kumar. 2022. “Beach” to “Bitch”: Inadvertent Unsafe Transcription of Kids’ Content on YouTube. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, page to appear. AAAI Press.
- Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. 2018. Publishing a quality context-aware annotated corpus and lexicon for harassment research.
- 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, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.
- Seth Stephens-Davidowitz and Andrés Pabon. 2017. *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. HarperCollins New York.
- Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. [Inducing a lexicon of abusive words – a feature-based approach](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.
- Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. [Text-based inference of moral sentiment change](#). 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 4654–4663, Hong Kong, China. Association for Computational Linguistics.