# Non-Binary Gender Expression in Online Interactions

Rebecca Dorn, Negar Mokhberian, Julie Jiang, Jeremy Abramson, Fred Morstatter, and Kristina Lerman

University of Southern California, Information Science Institute, Marina del Rey, CA
rdorn@usc.edu, nmokhber@usc.edu, yioujian@usc.edu, abramson@isi.edu, fredmors@isi.edu, lerman@isi.edu

**Abstract.** *Trigger Warning: Profane Language, Slurs*
The presence of openly non-binary gender individuals on social networks is growing. However, the relationship between gender, activity, and language in online interactions has not been extensively explored. Lack of understanding surrounding this interaction can result in the disparate treatment of non-binary gender individuals on online platforms. We investigate patterns of gender-based behavior identity on Twitter, focusing on gender expression as represented by users' expression of pronouns from eight different pronoun groups. We find that non-binary gender groups tend to receive substantially less attention in the form of likes and followers compared to binary groups. Additionally, non-binary users send and receive tweets with higher toxicity scores than other groups. This study identifies differences in the language and online activity of users with non-binary gender identity, and highlights a need for further evaluation of potential disparate treatment by algorithms used by online platforms.

**Keywords:** Big Data applications · Social computing · Text analysis

## 1 Introduction

An individual's identity, defined along multiple dimensions such as age, gender, and race, influences self expression and social connection [7]. As social interactions continue to migrate online, social media platforms play an increasingly critical role in identity formation [15]. While gender has been traditionally conceptualized in Western society as binary—specifically 'male' vs 'female'—two recent developments have transformed how we think about gender. First, there is growing recognition of gender as a cultural construct, distinct from the biologically-based sex [1]; second, there is growing awareness that gender forms a spectrum, rather than a binary identity [17].

We study individual identity on X (formerly Twitter), and how it mediates online expression and interactions. We focus on gender, one of the core dimensions of individual identity. As a proxy of gender expression, we study pronouns users choose to display in their online profile or biography. These pronouns range

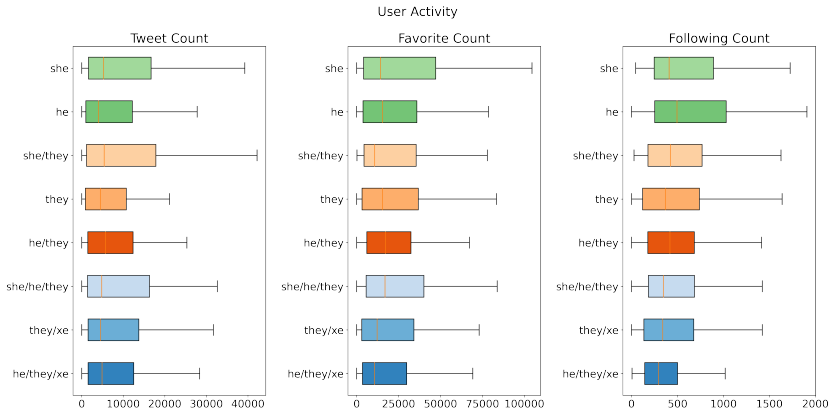


**Fig. 1.** User activity by pronoun group. Number of followers and favorites is slightly lower for non-binary pronoun groups. Pronoun groups are ordered from top to bottom by representation in the `Seed Set` of tweets.

from the traditional binary gender categories, such as 'she/her' and 'he/him', to non-binary and gender nonconforming [12] categories 'they/them', 'she/ze', 'she/they/xe', etc.

Creating safe online spaces for gender minorities is crucial, as individuals often turn to online communities for support [16] in the face of discrimination and social isolation they experience in real life [25]. In this work, we collect and study `NB-TwitCorpus3M`: a dataset of 3 million tweets annotated with author's self-provided pronouns, majority with non-binary pronouns. We analyze toxicity in tweets and replies to understand the role offensive content plays in online dynamics in gender minority communities.

To explore differences in language and activity of groups we investigate the following research questions:

- **RQ1**: How do users in different pronoun groups vary in their level of online activity and the online attention they receive from others? Are there systematic differences across the spectrum of gender identity?
- **RQ2**: Which user pronoun groups convey and experience more toxicity on Twitter?

We measure *activity* with number of tweets a user interacts with, and *attention* through external engagement with a user's messages. We find that non-binary groups have lower rates of activity online, and messages from non-binary groups get less attention through likes. We find that non-binary users tend to receive more toxic replies to their tweets. Surprisingly, we find that non-binary groups also post more toxic messages. These findings signal an important avenue for future work, as the consequences of gender variant communities having higher toxicity scores can lead to increased social exclusion.

## 2 Related Works

**Gender as Identity.** Gender is one of the earliest forms of social stratification and is a core dimension of identity. Inspired by second-wave feminism [23] people have started to draw a distinction between sex (biologically-produced) and gender (culturally-produced) identity. This has helped resolve the tension between the traditional conceptualization of gender in Western society and science as binary (i.e., 'male' and 'female'), and historical and societal evidence of the presence of non-binary individuals [19]. In English, third-person singular pronouns are used to express some form of gender identity [21]. The traditional two-sex naming system uses pronouns 'he' and 'she' to convey gender-binary identity. Individuals falling outside of the two-sex system (such as folks identifying with neither, both or a fluctuating set of binary gender identities) have begun to adopt pronouns such as 'they', 'xe' and 'ze' to convey their non-socially normative gender identity [26].

**LGBTQ+ Online Interactions.** Transgender adolescents participate in virtual communities for emotional and information support [24] and to explore their gender identity and the process of coming out [11]. In fact, queer adolescents demonstrate a tendency to participate in online LGBTQ+ communities over in-person communities [16]. With online communities playing such a large role in so many LGTBQ+ peoples lives, promoting safe queer online spaces is crucial for promoting long and happy LGBTQ+ lives. However, since 2018, there has been a stark rise in the visibility of LGBTQ+ hate speech[1] that has been reflected in social media. For example, despite its purported hate speech detection methods, X is wrought with transphobia [14], LGBTQ+ friendly on the platform 'Gab' have been overtaken with queer-phobic fetishization in the name of free speech [4] and Facebook comments against the LGBTQ+ community have a pattern of denying any gender conception outside of a two-sex gender system [2]. The need for understanding and mitigating discrimination towards queer people online is imminent.

**Gender Behavior on X.** X biographies measure expressions of personal identity and cultural trends. Longitudinal Online Profile Sampling (LOPS) measures identity formation through the evolution of a user's X (then Twitter) biography, finding that the tokens with the highest prevalence within biographies over 5 years were *he*, *him*, *she* and *her* [20]. The LOPS studies relied upon the notion of *personally expressed identity* where individuals declare *their own* attributes. Previous work assessing X activity across a spectrum of gender relied upon Census data to infer gender, a practice which excludes gender variance [3]. In this work, we allow users to conceptualize their own gender through their X biographies.

[1] https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023

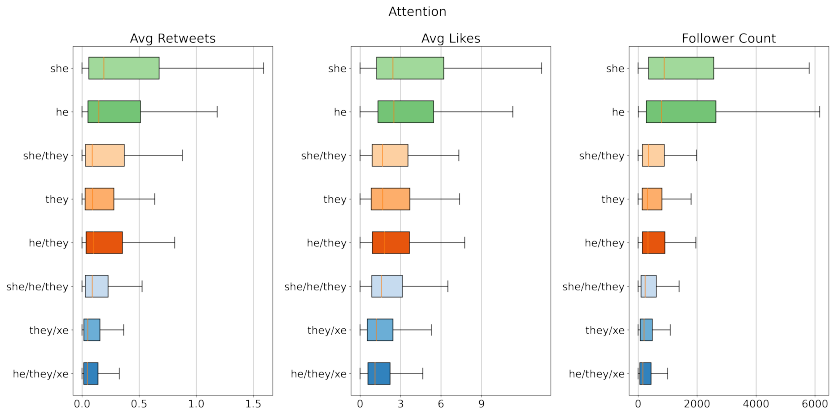


**Fig. 2.** Attention. Users with more representation receive more attention in retweet and like averages, as well as followers.

# 3 Methodology

## 3.1 Gender Spectrum

In this work, we will use the term **non-binary** to describe anyone using pronouns outside of 'she/her/hers' or 'he/him/his'. While this conceptualization of non-binary does not include many gender-queer populations (such as transgender people with a binary gender and non-binary individuals without pronouns in their biography) we find that this conceptualization is interoperable for downstream applications while maximizing user ability to self-describe their gender.

## 3.2 Data

Rather than collecting a key-word based sample, representation of a tweet sample can be increased by sampling users from a seed data set [9]. Users are first collected from a collection of over 2 billion tweets related to the Covid-19 pandemic collected between January 21, 2020, and November 5, 2021 [5]. From here on out we will call this dataset `Seed Data`. As most platform-engaged Twitter users tweeted about Covid-19 at some point, this generates a sample of active Twitter users. This dataset includes tweets from 2,066,165 users with specified pronouns in their Twitter profiles or biographies. The presence of pronouns is determined by whether a user has specified any combination of {he, him, his, she, her, hers, they, them, theirs, their, xe, xem, ze, zem} separated by forward slashes or commas, with any or no white space in their profile descriptions [13]. Profile descriptions of the users are recorded at the time of the first tweet collected for the sample.

We group the pronouns into five different series: she/her/hers, he/him/his, they/them/theirs, xe/xem and ze/zem. We encode the combinations of these pronouns series via a 5-digit dummy variable that is malleable to a range of gender representations and computationally efficient (e.g. 'she/they' represented as

10100). We encode the pronouns of all ∼2 million users into this 5-digit schema. We identify all pronoun groups with at least 1,000 members, and randomly sample up to 600 valid users from each group. This yields eight pronoun groups with at least 375 users. Table 1 reports groups in decreasing order of their size within `Seed Data`. For each user in our sample, we collect at most 1,000 of their most recent tweets posted before September 30, 2022. Tweets are retrieved using the API's `user_timeline` call. Table 1 reports the total tweets in our sample authored by each pronoun group. The resulting collection of tweets is the `NB-TwitCorpus3M` dataset.

**Table 1.** Pronoun Group composition. 'Original Users' denotes the number of users within each pronoun group of `Seed Data`, 'Sample Users' shows user number in our new dataset, and 'Tweets' shows the number of tweets collected for each pronoun group.

| Group | Original Users | Sample Users | Tweets |
|---|---|---|---|
| She | 1,194,565 | 508 | 464,262 |
| He | 461,264 | 559 | 503,780 |
| She/They | 158,025 | 508 | 463,599 |
| They | 132,374 | 560 | 506,064 |
| He/They | 77,951 | 514 | 469,328 |
| She/He/They | 20,882 | 557 | 611,227 |
| They/Xe | 1,312 | 468 | 462,775 |
| He/They/Xe | 1,015 | 377 | 387,722 |
| **Total** | **2,047,388** | **4,051** | **3,868,757** |

### 3.3 Toxicity Inference

To measure toxicity we use the *Detoxify* model [10] , a RoBERTa model trained on open source data emphasizing toxicity towards specific identities [6]. This model has an AUC score of 92.11 on the Kaggle dataset. The model outputs a continuous value between zero and one that captures the toxicity of language in the tweet. Values close to one are associated with high toxicity.

## 4 Results

### 4.1 RQ1: Activity and Attention

LGBTQ+ data is notoriously sparse and often low-quality [22]. Leveraging online engagement provides a novel opportunity to improve the representation and quality of queer-related data. We focus on measuring user activity, which encompasses a user's outward engagement on Twitter, including the number of original tweets, likes, and accounts followed.

Our analysis reveals lower levels of activity among non-binary users compared to binary users. Figure 1 illustrates the distribution of activity for each pronoun group, with outliers excluded to highlight group differences. Visually, non-binary

pronoun groups exhibit slightly lower levels of outward engagement on Twitter compared to binary pronoun groups.

We investigate the interaction between pronoun group representation in the `Seed Set` and group-level online activity, as representation in the user pool may portray overall representation on the X platform. We find a negative correlation between the number of tweets sent out by users and the size of the group in the `Seed Set` (Spearman's $\rho$ = -0.272, $p < 0.01$). Similarly, the favorite count (the number of tweets liked by the user) and the following count (the number of other users someone follows) show negative correlations with group size (Spearman's $\rho$ = -0.216, $p < 0.01$ and Spearman's $\rho$ = -0.352, $p < 0.01$, respectively). Further, we observe that the pronoun group with the highest median following count is *he* (494), while the lowest is *she/he/they* (346). Overall, our findings suggest that user activity, as measured by likes, retweets, and following count, tends to be lower for minority non-binary groups compared to groups with higher representation. This raises concerns regarding the potential bias in randomly sampled Twitter data, which may disproportionately represent binary users over non-binary users, thus posing challenges for achieving adequate representation in downstream analyses.

The allocation of attention on social media platforms carry significant political, economic, and social ramifications [18]. In this study, we operationalize *attention* as the level of inward engagement and prominence that users attain on Twitter. We quantify attention using metrics such as the number of followers, average retweets, average likes, and the percentage of verified users. Our analysis reveals a notable trend: as the representation of pronoun groups in the `Seed Set` decreases, the corresponding amount of attention also diminishes.

Figure 2 illustrates the distributions of attention across different gender pronoun groups. Visual inspection suggests that smaller pronoun groups tend to receive comparatively lower levels of attention. We observe a substantial negative correlation between the lack of representation and the median user retweets (Spearman's $\rho$ = -0.215, $p < 0.01$), with the pronoun group *she* exhibiting the highest median retweets (0.192) and *he/they/xe* displaying the lowest (0.045). Similar trends are evident in the correlation between median likes and representation (Spearman's $\rho$ = -0.229, $p < 0.01$), with the pronoun group *he* receiving the highest median likes (2.510) and *he/they/xe* receiving the lowest (1.111). The number of followers demonstrates an even more pronounced version of this trend (Spearman's $\rho$ = -0.363, $p < 0.01$). Our findings strongly indicate that pronoun groups with less representation tend to receive diminished attention online.

We analyze the pronoun group composition of verified users as displayed in Figure 3. Verification status was obtained before X released an option for users to purchase verification status. Over 75% of verified users in our sample are in the pronoun groups *he* or *she*, the two most represented groups. The only groups with no verified users, *he/they/xe* and *she/he/they*, are non-binary. This strong disparity in verification indicates a gap in social validity and visibility of the users in our sample.

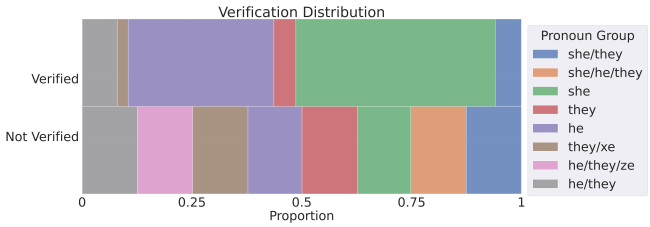


**Fig. 3.** Pronoun group decomposition or verified users in our sample. Groups *he/they/xe* and *she/he/they* have no verified users.

### 4.2 RQ2: Toxicity in Tweets and Replies

Next, we investigate the toxicity of tweets in both tweets posted (sent tweets) and a sample of replies received (received tweets). To generate received tweets, we randomly sample 100 users from each pronoun group in our dataset and collect replies to their original tweets using the tweet's `conversation_id`. This process results in 95,381 replies from 29,537 unique `conversation_id`s. We look at incidence of highly toxic tweets, considering a tweet as highly toxic if its toxicity score exceeds the threshold of 0.9.

Figure 4 (a) displays the proportion of tweets deemed highly toxic among both sent and received tweets. Our analysis reveals that all groups send out tweets where a small fraction (less than 1%) are highly toxic. However, the share of tweets estimated to be highly toxic increases monotonically with lack of representation of the pronoun group in `Seed Set`. Notably, five out of eight pronoun groups receive more tweets estimated to be highly toxic than what they post, with only *they/xe*, *he/they/xe* and *they/them* posting more toxic tweets than they receive. We observe that these three groups are all non-binary.

In Figure 4 (b) we present distributions of toxicity scores of tweets after removing outliers. Notably, no pronoun group exhibits a higher median toxicity score for sent tweets than for received tweets. The highest median toxicity scores for sent tweets are *he/they/xe* (.0032) and *they/xe* (.0028). These two particular groups also exhibit the highest median toxicity scores for received tweets as well (.0048, .0047, respectively). Remarkably, these are the two groups with the lowest initial representation. Simultaneously, the groups *he* and *she* boast the lowest median toxicity for both sent (.0009, .0010, respectively) and received tweets (.0016, .0023, respectively). We observe that these two groups reflect binary conceptualizations of gender.

The toxicity scores for binary and non-binary users as two distinct groups exhibit significant differences (T-statistic = -125.72, $p < .01$). Additionally, there is a strong correlation between pronoun group representation and the toxicity scores assigned to their tweets (Spearman's $\rho = .16$, $p < 0.01$). These observations surprisingly suggest that the non-binary pronoun groups tend to post more highly toxic tweets than binary groups. Given that content moderation decisions often rely on automated toxicity classification, this finding implies that tweets by gender minorities are more likely to be flagged or removed by these algorithms.

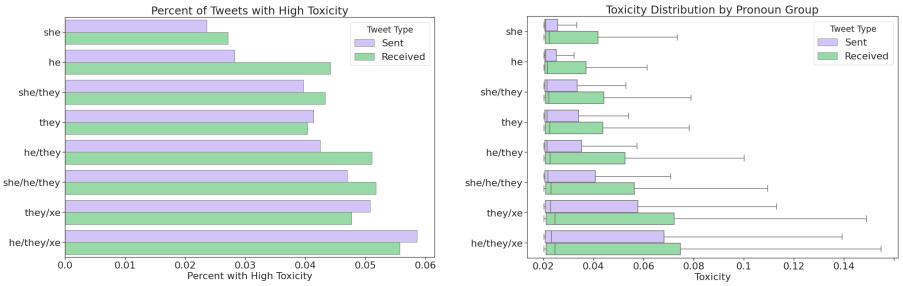


**Fig. 4.** (a) Percent of Tweets Posted and Percent of Replies Received Labeled as Toxic (toxicity $> 0.9$). (b) Toxicity distribution for tweets posted and received. Tweets from non-binary receive higher toxicity scores from Detoxify than those from binary users

## 5 Discussion

We compiled and analyzed the `NB-TwitCorpus3M` dataset, comprising of approximately 3 million tweets from users with pronouns in their biographies, primarily those using non-binary pronouns. Our exploratory analysis investigates online behavior for eight pronoun groups including both binary and non-binary pronoun series. Specifically, we examine outgoing activity, incoming attention, and toxicity levels in both posted tweets and received replies.

Our analysis reveals that non-binary pronoun groups exhibit lower activity levels on Twitter compared to binary pronoun groups. Given the sparse data focusing on gender-queer populations [22], this finding suggests a potential underrepresentation of users with non-binary pronouns in social media data. Addressing this disparity is crucial to ensure adequate representation of gender minority groups in online spaces.

Further, we observe that non-binary pronoun groups receive less attention through retweets, likes and followers when compared to binary pronoun groups. Attention plays a pivotal role in information spread in the digital age. The lower attention towards non-binary groups may indicate reduced social influence within Twitter's online ecosystem, potentially hindering the political power and visibility of non-binary communities. This underscores the importance of amplifying the voices of gender-queer users to bolster activist causes within the non-binary community.

Surprisingly, we find that non-binary users exhibit higher levels of toxicity detected in their posted tweets compared to users with binary pronouns.We posit that this discrepancy may stem from dialect bias within the toxicity classifier, where dialect commonly used in queer communities is erroneously flagged as expressing toxicity. This would align with prior evidence suggesting that social media content from gender-variant groups, such as drag queens, is disproportionately classified as hate speech [8], highlighting the need for further investigation into the impact of dialect bias on toxicity detection mechanisms.

**Ethical Limitations** Inherently, this data set is sensitive due to its collection of individuals with historically marginalized gender identities. To safeguard privacy, we

focus analyses on aggregated data and remove identifiable information from provided examples. Our analysis is constrained by the specificity of the population under study, which may inadvertently exclude certain binary and non-binary Twitter users who do not include pronouns in their biographies. Additionally, our study does not differentiate between pronoun order (e.g., *she/they* versus *they/she*).

This study could be enhanced by incorporating a group of users without pronouns in their biographies, or using multiple seed topics to ensure better representation of active X users. We acknowledge the limitations stemming from the relatively low number of replies collected and recognize the importance of further research into the dynamics of reply senders. Exploring how results evolve with an increased number of replies would contribute to a more nuanced understanding of user interactions. This study was reviewed by authors' IRB and designated exempt. Authors declare no competing interests.

**Acknowledgments.** This work was funded in part by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002 and Contract No. HR00112290021.

## References

1. Allen, K., Cuthbert, K., Hall, J.J., Hines, S., Elley, S.: Trailblazing the gender revolution? young people's understandings of gender diversity through generation and social change. Journal of Youth Studies **25**(5), 650–666 (2022)
2. Aperocho, M.D., Aliñabon, G., Camia, K., Tenorio, A.J.: Exploring the hate language: An analysis of discourses against the lgbtqia+ community. Psychology and Education: A Multidisciplinary Journal **8**(6), 729–737 (2023)
3. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics **18**(2), 135–160 (Apr 2014). https://doi.org/10.1111/josl.12080
4. Brody, E., Greenhalgh, S.P., Sajjad, M.: Free speech or free to hate?: Anti-lgbtq+ discourses in lgbtq+-affirming spaces on gab social. Journal of Homosexuality pp. 1–26 (2023)
5. Chen, E., Lerman, K., Ferrara, E.: Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance **6**(2), e19273 (may 2020). https://doi.org/10.2196/19273, https://doi.org/10.21962F19273
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
7. Deaux, K.: Reconstructing social identity. Personality and social psychology bulletin **19**(1), 4–12 (1993)
8. Dias, O.T., Antonialli, D.M., Gomes, A.: Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. Sexuality & Culture **25**(2), 700–732 (04 2021), http://libproxy.usc.edu/login?url=https://www.proquest.com/scholarly-journals/fighting-hate-speech-silencing-drag-queens/docview/2495185828/se-2
9. Garcia, D., Rimé, B.: Collective emotions and social resilience in the digital traces after a terrorist attack. Psychological Science **30**(4), 617–628 (2019). https://doi.org/10.1177/0956797619831964, pMID: 30865565

10. Hanu, L.: Unitary team. detoxify. github (2020), https://github.com/unitaryai/detoxify
11. Herrmann, L., Bindt, C., Hohmann, S., Becker-Hebly, I.: Social media use and experiences among transgender and gender diverse adolescents. International Journal of Transgender Health pp. 1–14 (2023)
12. Hicks, A., Rutherford, M., Fellbaum, C., Bian, J.: An analysis of WordNet's coverage of gender identity using Twitter and the national transgender discrimination survey. In: Proceedings of the 8th Global WordNet Conference (GWC). pp. 123–130. Global Wordnet Association, Bucharest, Romania (27–30 Jan 2016), https://aclanthology.org/2016.gwc-1.19
13. Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H.C.H., Ferrara, E.: What are your pronouns? examining gender pronoun usage on twitter. arXiv preprint arXiv:2207.10894 (2022)
14. Locatelli, D., Damo, G., Nozza, D.: A cross-lingual study of homotransphobia on twitter. In: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). pp. 16–24 (2023)
15. Manago, A.M.: Identity development in the digital age: The case of social networking sites. (2015)
16. McInroy, L.B., McCloskey, R.J., Craig, S.L., Eaton, A.D.: Lgbtq+ youths' community engagement and resource seeking online versus offline. Journal of Technology in Human Services **37**(4), 315–333 (2019)
17. Monro, S.: Non-binary and genderqueer: An overview of the field. International Journal of Transgenderism **20**(2-3), 126–131 (2019)
18. Pedersen, M.A., Albris, K., Seaver, N.: The political economy of attention. Annual Review of Anthropology **50**, 309–325 (2021)
19. Richards, C., Bouman, W.P., Barker, M.J.: Genderqueer and non-binary genders. Palgrave (2017). https://doi.org/10.1057/978-1-137-51053-2
20. Rogers, N., Jones, J.J.: Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time? Journal of Social Computing **2**(1), 1–13 (2021). https://doi.org/10.23919/JSC.2021.0002
21. Rose, E., Winig, M., Nash, J., Roepke, K., Conrod, K.: Variation in acceptability of neologistic english pronouns. Proceedings of the Linguistic Society of America **8**(1), 5526–5526 (2023)
22. Ruberg, B., Ruelos, S.: Data for queer lives: How lgbtq gender and sexuality identities challenge norms of demographics. Big Data & Society **7**(1), 2053951720933286 (2020)
23. Sanz, V.: No way out of the binary: A critical history of the scientific production of sex. Signs: Journal of Women in Culture and Society **43**, 1–27 (09 2017). https://doi.org/10.1086/692517
24. Selkie, E., Adkins, V., Masters, E., Bajpai, A., Shumer, D.: Transgender adolescents' uses of social media for social support. Journal of Adolescent Health **66**(3), 275–280 (2020)
25. Tabaac, A., Perrin, P.B., Benotsch, E.G.: Discrimination, mental health, and body image among transgender and gender-non-binary individuals: constructing a multiple mediational path model. Journal of gay & lesbian social services **30**(1), 1–16 (2018)
26. Wayne, L.D.: Neutral pronouns: A modest proposal whose time has come. Canadian Woman Studies/les cahiers de la femme (2005)