Non-Binary Gender Expression in Online Interactions

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

Summary

★ To best avoid disparate treatment, we need to understand *how* non-binary users act on social media.

★ We explore individual identity of non-binary Twitter users through likes, follows, and toxicity scores

★ We find that **non-binary individuals receive less attention online**, and **non-binary individuals score higher for inferred toxicity by SoTA.**

★ Our work highlights the need for further research into how social media platforms can prevent new harms towards genderqueer communities.

Motivation

There is growing recognition of gender (1) as a cultural construct, distinct from the biologically-based sex and (2) as a spectrum, rather than a binary identity. This new visibility enables a more accurate study in how gender expression interacts with online behavior. This also enables us to identify patterns in non-binary (NB) social media behavior to prevent anti-NB bias in large language models.

Data Collection

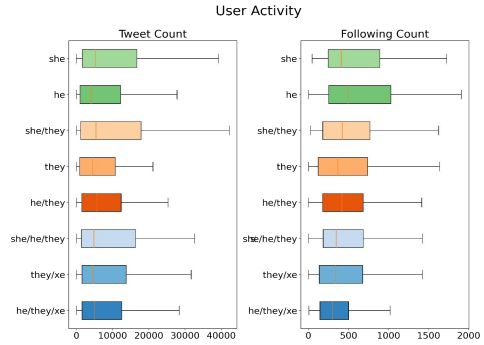
A **pronoun group** is a group of users sharing the same pronoun series in their biographies (e.g. users with she/they and she/her/they/them share a pronoun group). We take a sample of ~2M Twitter users with pronouns in their biography, randomly sample >600 users from each pronoun group and extract 1000 posts (tweets) from each user.

Group	Original Users	Sample Users	Sample 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**

Pronoun group composition within collected sample. Users show number of unique users in sample, tweets shows total tweets.

Activity

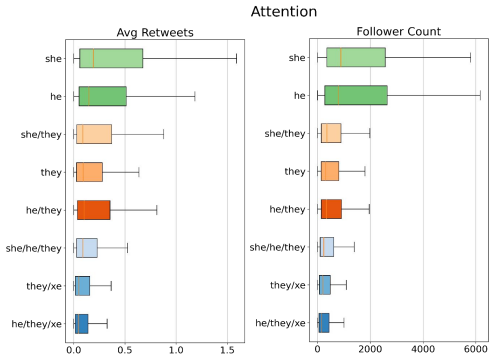
We measure pronoun group engagement through status updates, likes and accounts followed. Number of likes and following is relatively similar between pronoun groups.



User activity by pronoun group. Outliers excluded. Number of followers is relatively similar between groups.

Attention

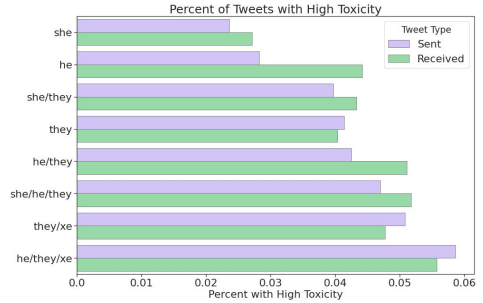
Attention (external engagement) is measured through retweets received, likes received, number of followers and percent verified. Pronoun groups with less representation in our initial sample receive less engagement.



Attention by pronoun group. Users with more representation receive more attention in retweets and followers.

Toxicity

We collect <10 replies to each tweet in our dataset, then look at toxicity in posts and replies for each pronoun group. Toxicity is measured by the Detoxify model. We find that toxicity decreases with pronoun group representation in our initial dataset.



Percent of tweets posted and percent of replies received labeled as toxic (toxicity > 0.9).

User	Tweet	Toxicity
binary	RT @user: i'll die for my niggas, i ride my niggas	.996
binary	Solid, got my first death threat today! Filled with such language as "faggot", "I will shoot you in themotherfucking mouth", "you dumb ass..."	.998
non-binary	I will start t one day One Fuckin Day	.981
non-binary	these fucking ladybug cockroach monsters will be the death of me	.998

Examples of tweets flagged as highly toxic. Usernames are swapped with "user" to preserve privacy.

Conclusions

We find that NB pronoun groups with less overall representation on Twitter receive less attention via retweets, likes and follows than groups with higher overall representation. These low-response users engage on Twitter's platform by sending and favoriting tweets at about an equal rate as other users.

We find that, in comparison with binary pronoun users, NB groups have higher levels of toxicity detected in their tweets. We hypothesize that the toxicity classifier may contain *dialect bias* against NB users.