Salty.

An Investigation into Algorithmic Bias in Content Policing of Marginalized Communities on Instagram and Facebook.

Publish Date:
October 22, 2019

Table of Contents

Our Story:

Salty is a membership driven digital newsletter and platform committed to amplifying the voices and visibility of women, trans and nonbinary people. We launched on International Women's Day in 2018 - and were unceremoniously kicked off Mailchimp a few hours later, with no concrete reason, except violating "community guidelines". Since then, we've grown into a group of over 130,000 users, 45,000 newsletter subscribers and 3 million monthly impressions across our platforms.

In our first year and a half Salty has faced digital harassment, hacking, been denied access to resources and been 'accidentally' booted from platforms - including Instagram. Our lived experience tells us that there is an unconscious bias that shapes our digital environment.

Algorithms are the backbone of content moderation. Algorithmic models produce probability scores that assess whether the user-generated content abides by the platform's community guidelines. When pertaining to offensive content - social media scholar Tarleton Gillespie says "State-of-the-art detection algorithms have a difficult time discerning offensive content or behavior even when they know precisely what they are looking for…automatic detection produces too many false positives; in light of this, some platforms and third parties are pairing automatic detection with editorial oversight" (Custodians of the Internet, pg 98)

Our Story (cont):

In July 2019, Instagram algorithms rejected Salty's ads because they claimed we were promoting "escort services." The ads were simply portraits of our Salty community - women, trans and non binary people - some were disabled, some were plus sized, most were women of color.

Unable to rectify the problem via automated channels, we called this 'false flag' to the attention of our community, and as the press started to pay attention, Facebook reached out to rectify. After admitting these were falsely flagged, they reinstating the ads. Facebook publicly agreed to meet with Salty to discuss ways to make the policies more inclusive. We figured it was the beginning of a powerful conversation.

In preparation for our meeting with Facebook Policy team, we collected data from our community to better tell the story of the way these algorithms affect us, and formulate recommendations to make FB/ Instagram a safer place for women, trans and non binary people. We released a survey on our website and encouraged our readers to submit their experiences of the ways in which Instagram/Facebook rejects ads, closes down accounts, or deletes posts.

Unfortunately, over the past two months, Facebook has ceased communication with Salty, and has made no indication that they plan (or ever planned) on actually meeting with us to discuss policy development.

We believe the information included in this report is newsworthy and of public importance, and with the consent of the participants, we've decided to make it available publicly.

Survey Basics

Salty distributed a survey on to its followers on Instagram and via newsletter. As a community for and by women, nonbinary people, and queer folx, the Salty's followers is heavily composed of these demographic groups. As such, the survey surfaced issues that are affecting these communities. The survey is not meant to be representative of all Instagram users.

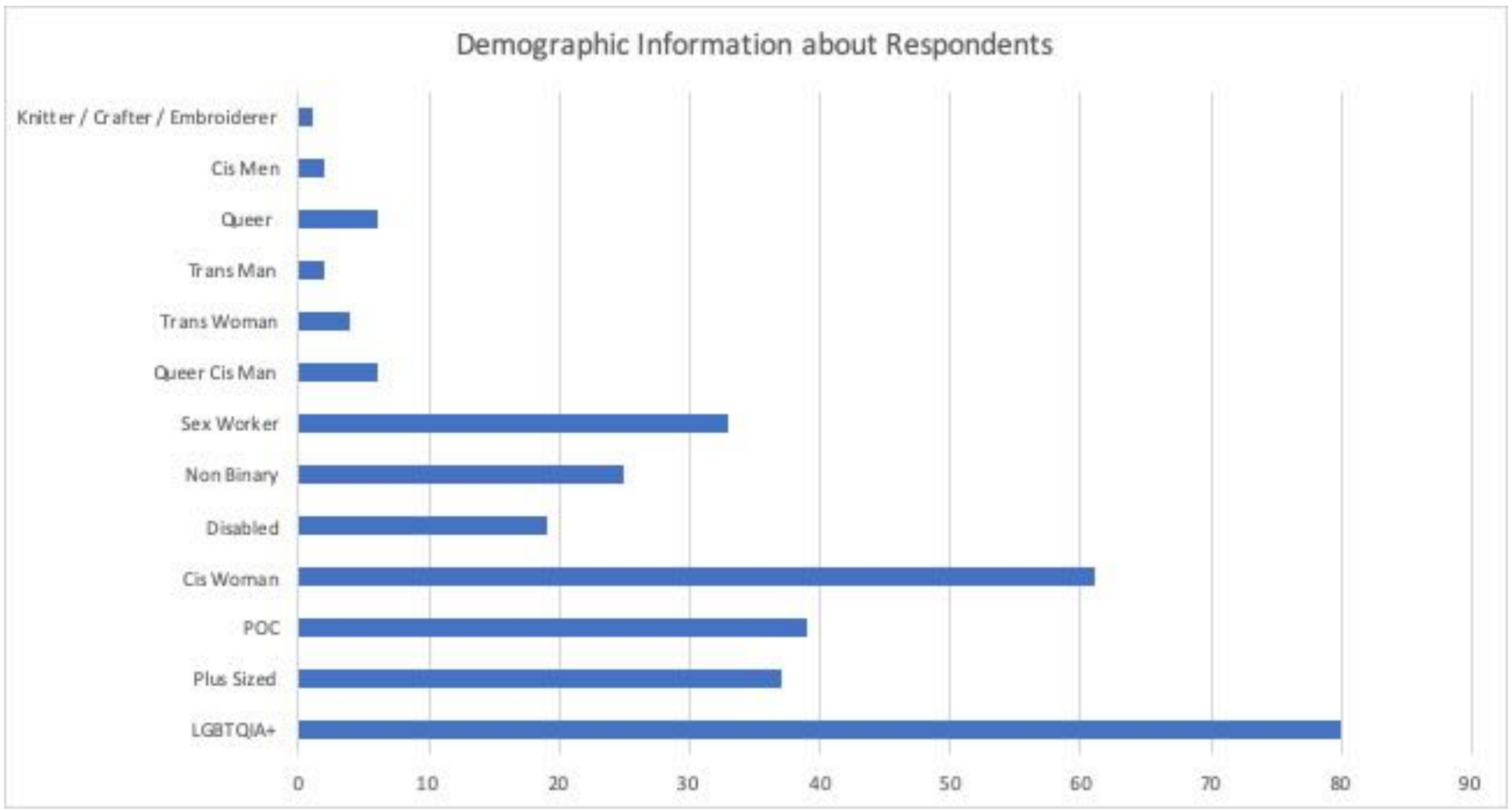
One further caveat to keep in mind with the data is that we cannot see the full universe of posts or accounts that were flagged and reviewed, so we are not making directly causal claims. As with any survey data, the response rate to certain questions was somewhat patchy, so the data visualizations must keep in mind not just the total number of respondents, but the total number of respondents per question.

These findings have been collated by the Salty Algorithmic Investigation team on behalf of The Coalition for Digital Visibility.

Part I, The Data

Who is Affected by Censorship

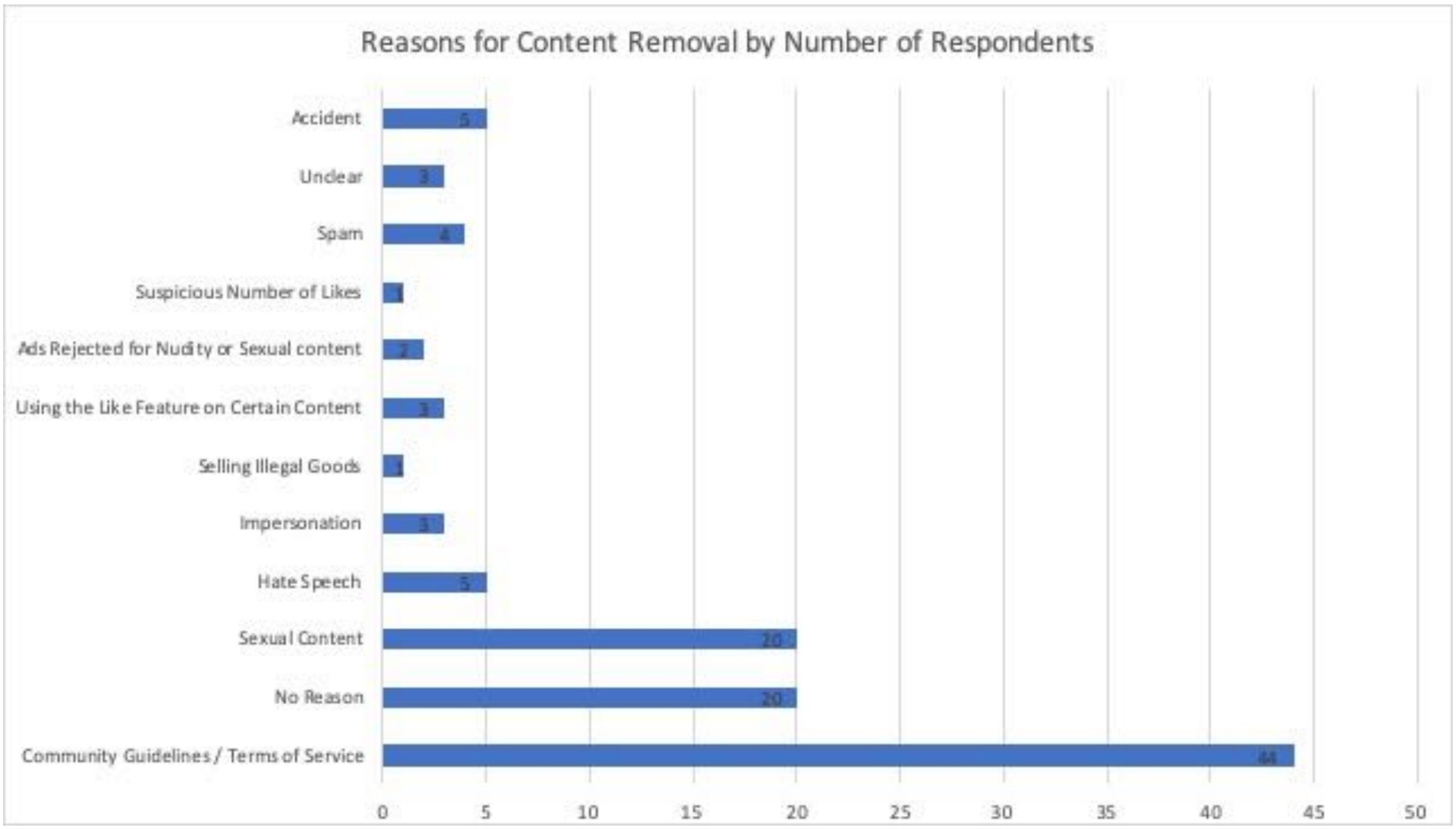
The demographics of our survey respondents (118 people total) reflects our readership. Many of the respondents identified as LGBTQIA+, people of color, plus sized, and sex workers or educators. All of these Instagram users experienced friction with the platform in some form, such content taken down, disabled profiles or pages, and/or rejected advertisements.



Part I, The Data

Reasons for Disabling Account or Rejecting Ads

The vast majority of respondents said they were given no reason for for actions against their account, and were simply told they violated "community guidelines," an extremely vague response.



Part II, Community Findings:

a) BIPOC users, body positive advocates, women and queer folx feel Instagram is targeting them for their perceived identity terms.

- “I haven't had my profile deleted or disabled, but they delete my posts all the time and threaten to disable me. Sometimes I get a notification telling me that the post didn't follow community guidelines, sometimes there is a blurred thumbnail of the post, but sometimes it doesn't even show me what the post is. Sometimes it gives me no notification or feedback at all and posts just disappear. And the thing that annoys me the most is that it's always pictures of my partner and I that get reported and deleted and he is a POC. Even if I censor any nudity, they still get removed. Not that I even agree with censoring it, as my posts are about disability, intimacy, care, and sex.”

- “My posts have been banned from showing up in hashtags because I used a hashtag associated with pornography. So my account has been labeled as pornographic and no longer shows up in searches. I use tags such as body positive, body image, HAES, black girl magic.”

- “They said I violated the nudity guidelines when I didn’t show any private parts at all. I’m a brown fat man so people don’t like to see my skin.”

Part II, Community Findings:

b) Policies Harm those they are Supposed to Protect.
Users who come under targeted harassment due to their identity are the ones getting reported and banned.

- “They stated ‘I violated community guidelines.’ When I inquired how, I was met with silence. I believe I was unfairly targeted by fatphobic bigots who reported my profile and IG did nothing to stop it, instead choosing to delete my account, the person who was harassed. My new account is now being unfairly targeted, including shadowbanning a photo of me in a one-piece bathing suit.”

- “Shadow-banned and taken off for a few days several times with the threat of being deleted next time. Because of female nipples (which is transphobic), showing art content, critical political posts, and feminist and queer art.”

- “Captions, comments, and follows have been disabled with messages like ‘This action has been blocked. We restrain certain actions and content to protect our community.’ Anti-racist posts have been deleted as well for ‘violating community standards.’”

- “Our feminist page was automatically disabled when it was targeted by a group of MRA’s.”

Part II, Community Findings:

c) Distrust of Platform: Preventing users from easily appealing action and not clarifying why the decision was reached leads to greater distrust of the platform and fosters the idea of arbitrary shadow-banning.

- "They said, ‘No user found but email taken.’ After I sent several emails, they barely answered and then stopped answering me. They never gave me a reason why my account was deleted. I believe it was because of my photos in lingerie. I’m a lingerie designer trying to start my own business, so most of the time I'm my own model. I worked so much on my art to be deleted without explanation.”

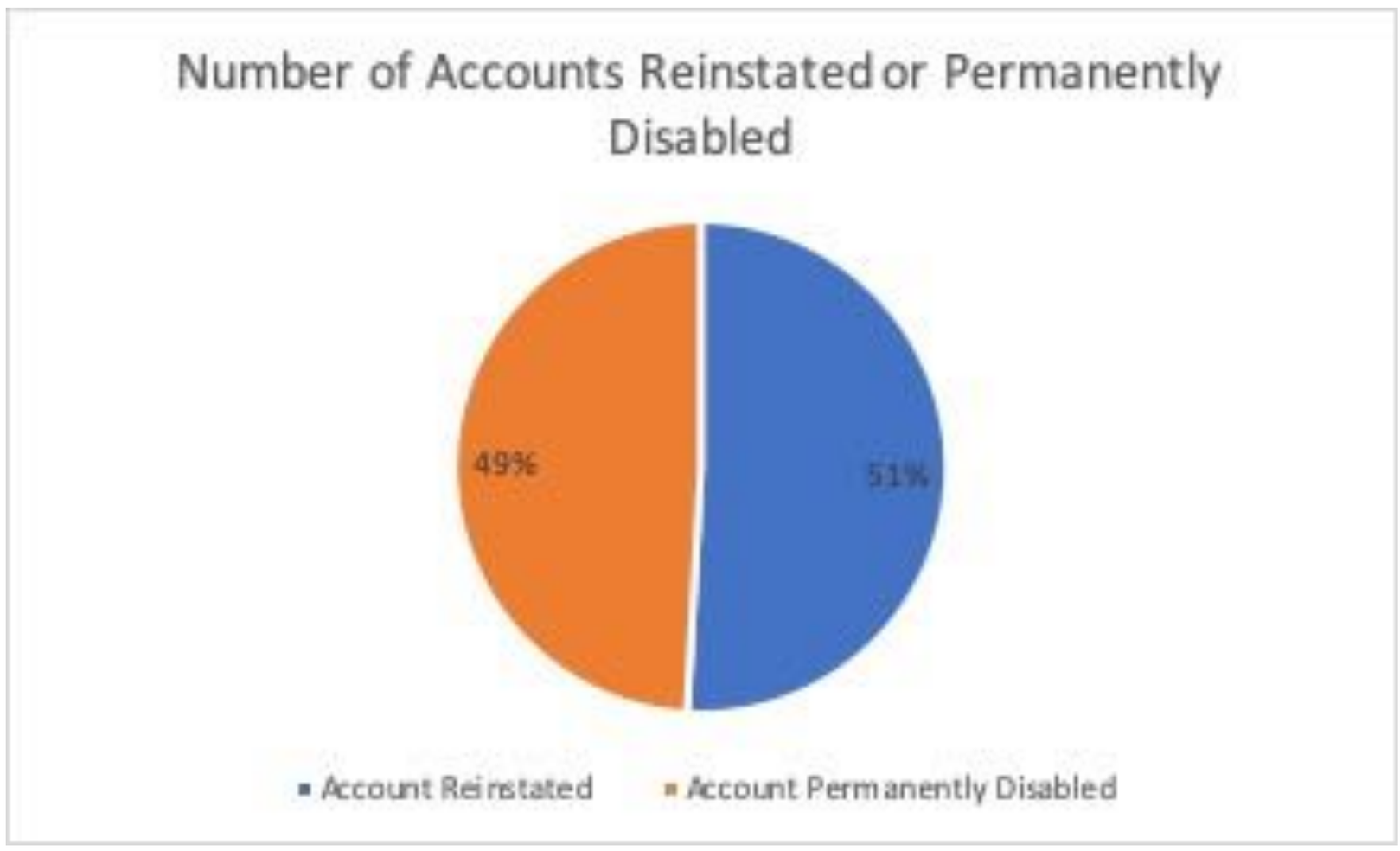
- “When reinstated, they said, ‘We deleted you by accident.’ I was only reinstated after sending in multiple verified accounts with nudity on them. Ads denied because I'm ‘pornographic.’ Photos in swimwear taken down as pornographic. I've been shadow-banned for two years and all my previous posts on my self-created hashtag are gone. I'm also unsearchable.”

- “Instagram hasn’t given us feedback, we’ve been shadowbanned, and it’s really hurting us because we’re a new collective unable to reach our own audience.”

Part III, General Observations in Our Survey:

- High Instance of False Flagging

A significant number of accounts in the survey seem to be disabled in error. For example, close to half of accounts that were disabled were later reinstated. This shows there seems to a problem with false flagging.



Part IV, Advertising Bias Against Women Led Business:

a) Respondents in our survey that were unable to advertise were more likely to identify as cis women than any other identity type.

- “They can’t promote the nudity in my ad, but they are paintings. I’ve even tried to promote plain art-show flyers with no nudity and just text and they rejected my ads.”

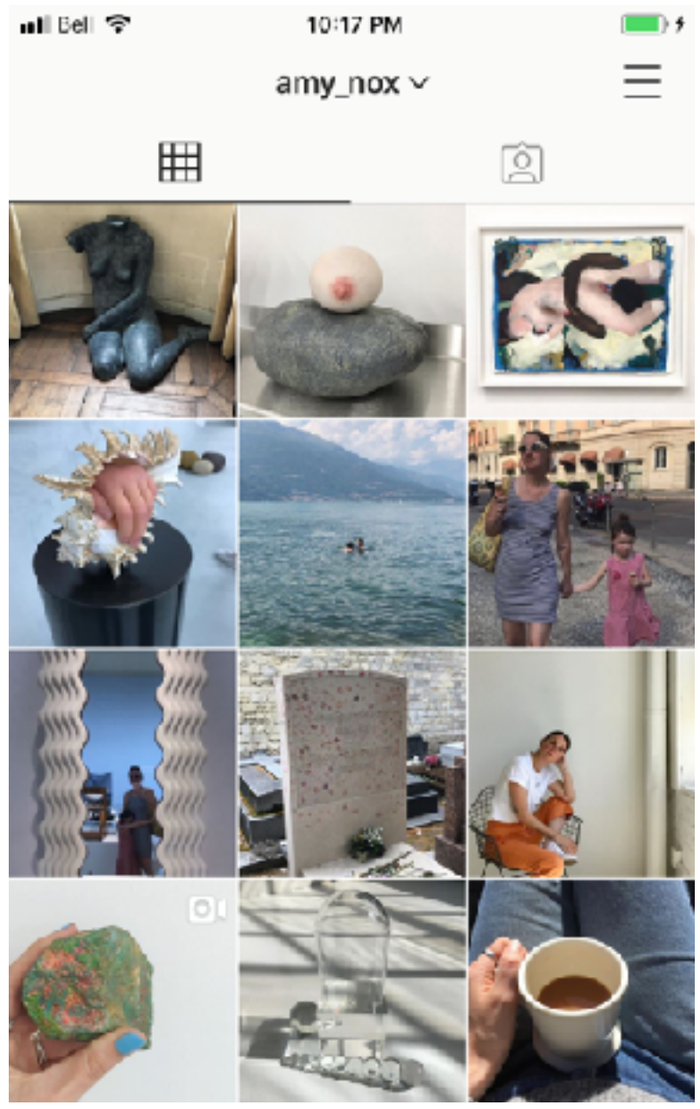
- “My account has been blocked from appearing on hashtags I use to promote my work for the past six months because it’s body-positive and women in plus-sized bodies, even though I don’t break any of IG’s rules. I don’t even try to run ads because I know they won’t be approved.”

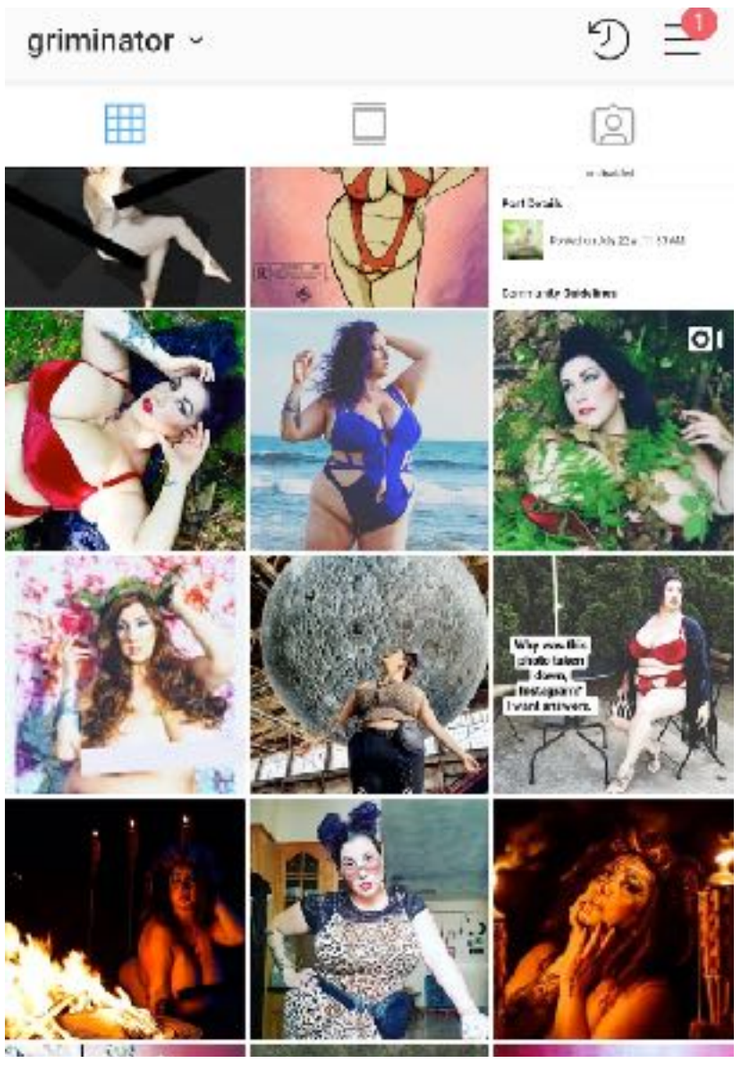
- “The fact that our adverts are not approved means that the growth of our women and nonbinary focused business is seriously hindered.”

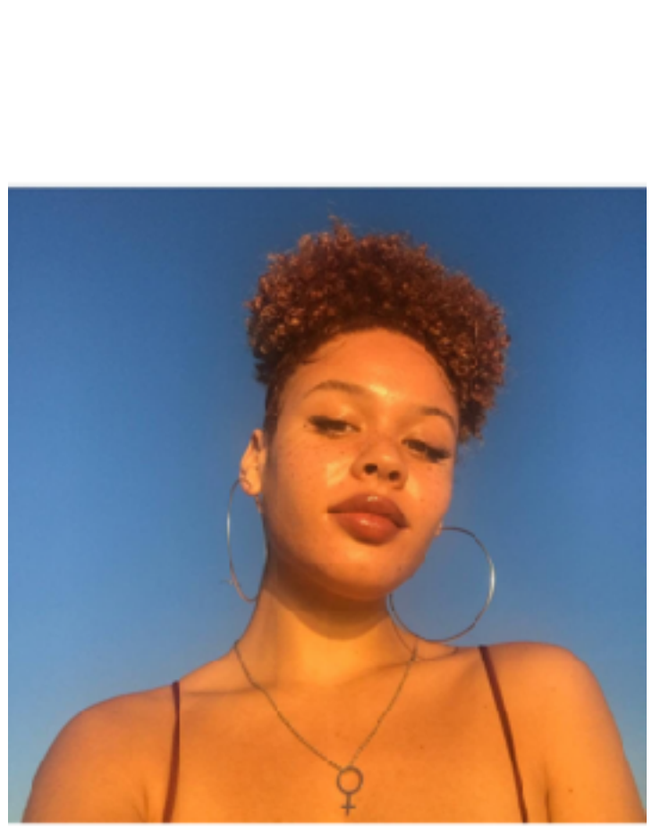
- “We sell breast pumps and are unable to advertise on Facebook.”

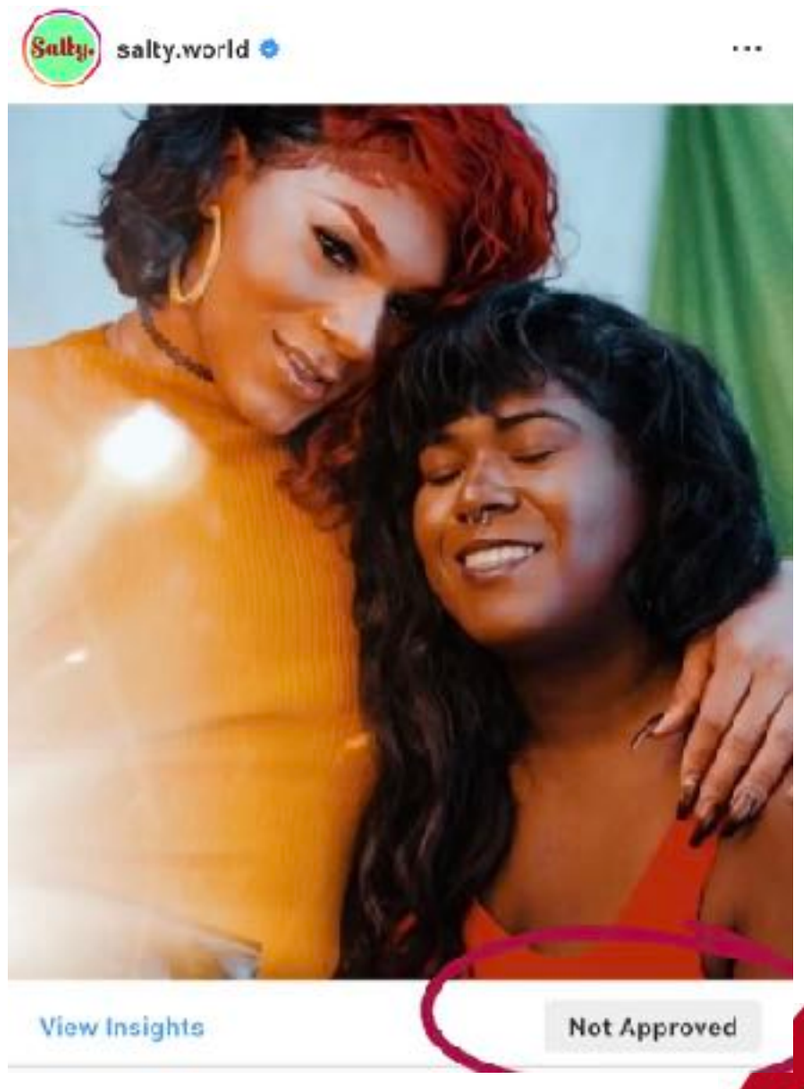
- “Unbound is consistently banned from advertising on Facebook and Instagram and it’s debilitating to our business. It’s also infuriating, because we see endless ads on the same platforms for erectile dysfunction medication, penis pumps, and “manscapping” razors. Why are penises normal but the female and non-binary body considered a threat?”

Submitted Censored Content









Part V, Further Inquiry

Salty will be continuing this investigation - to take part in our next survey, please Click Here.

Part VI, Support

Salty relies on contributions and volunteers to survive. If you believe the work we are doing with this kind of research has value- please click here to become a Member.

Or click here to make a one off contribution.

For follow up questions regarding this report - email saltyalgorithmicbiasteam@gmail.com