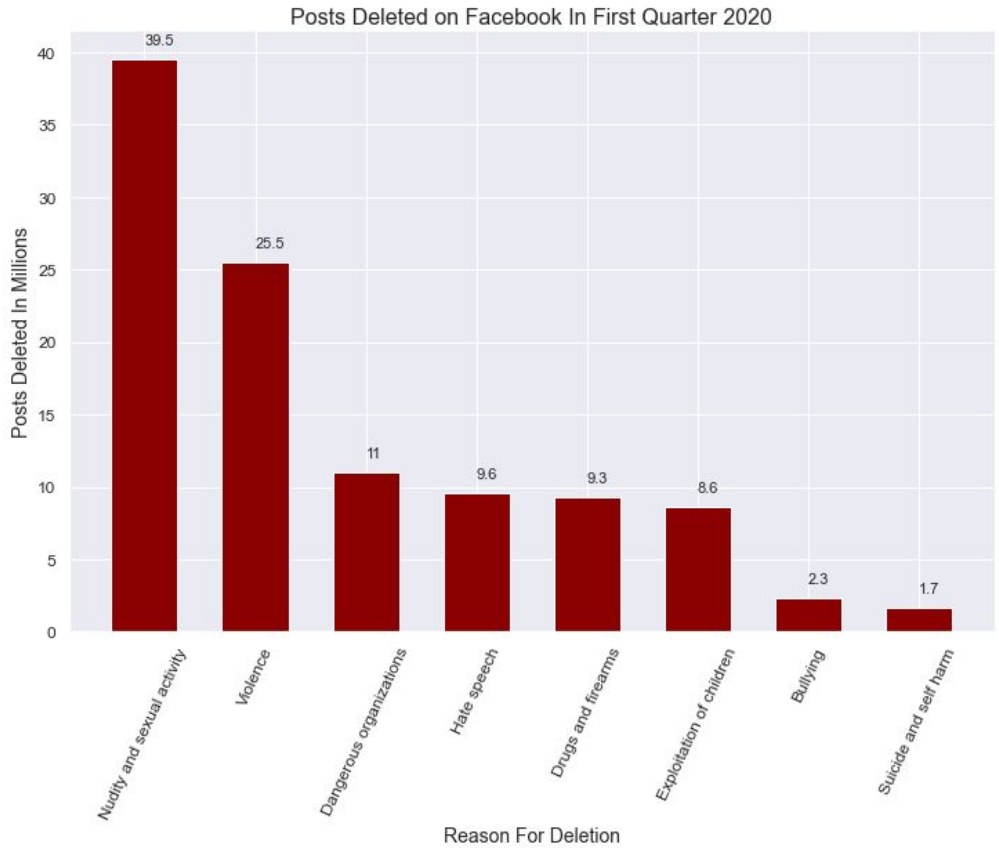
Inclusive Content Filtering

Samantha Baltodano

Agenda

1. **Content filter for explicit images**
2. **Public outrage and potential growth areas**
3. **Image classifications and workarounds**
4. **An outsized impact, a happier community**

There Will Always Be A Need For Content Filtering



“Women are shamed for just being women, for being mothers... That’s not OK.”

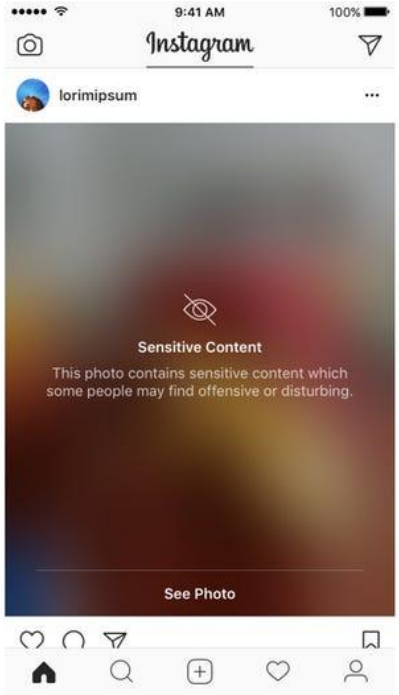
- Heather Bays, Post

Black Instagram users are 50% more likely than white users to have their accounts automatically disabled.

- Business Insider

Biased Content Filters Prevent Inclusivity And Freedom Of Expression

- Racial bias in image and account banning
- #freethenipple
- Freedom of expression



Final Model Performance

	Neutral	Sexy	Explicit	Breast Feeding	Violent
Recall	87%	54%	83%	17%	0%
Accuracy					80.5%

Content Filtering For A Better Tomorrow

1. Explicit content filtering
2. Filtering equipped to prevent racial bias
3. #Freethenipple (just a little)
4. Continued success with violent content

Recommendations

- Initiate change
  - Less rigid restrictions on breastfeeding photos
  - Child filters for sexy photos
  - Free the nipple responsibly

Resources

1. Koetsier, John. “Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day.” *Forbes*, Forbes Magazine, 30 June 2021, www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=34f48f2954d0.

Thank you!

Check out the full project on Github.

Samantha Baltodano

LinkedIn

Github