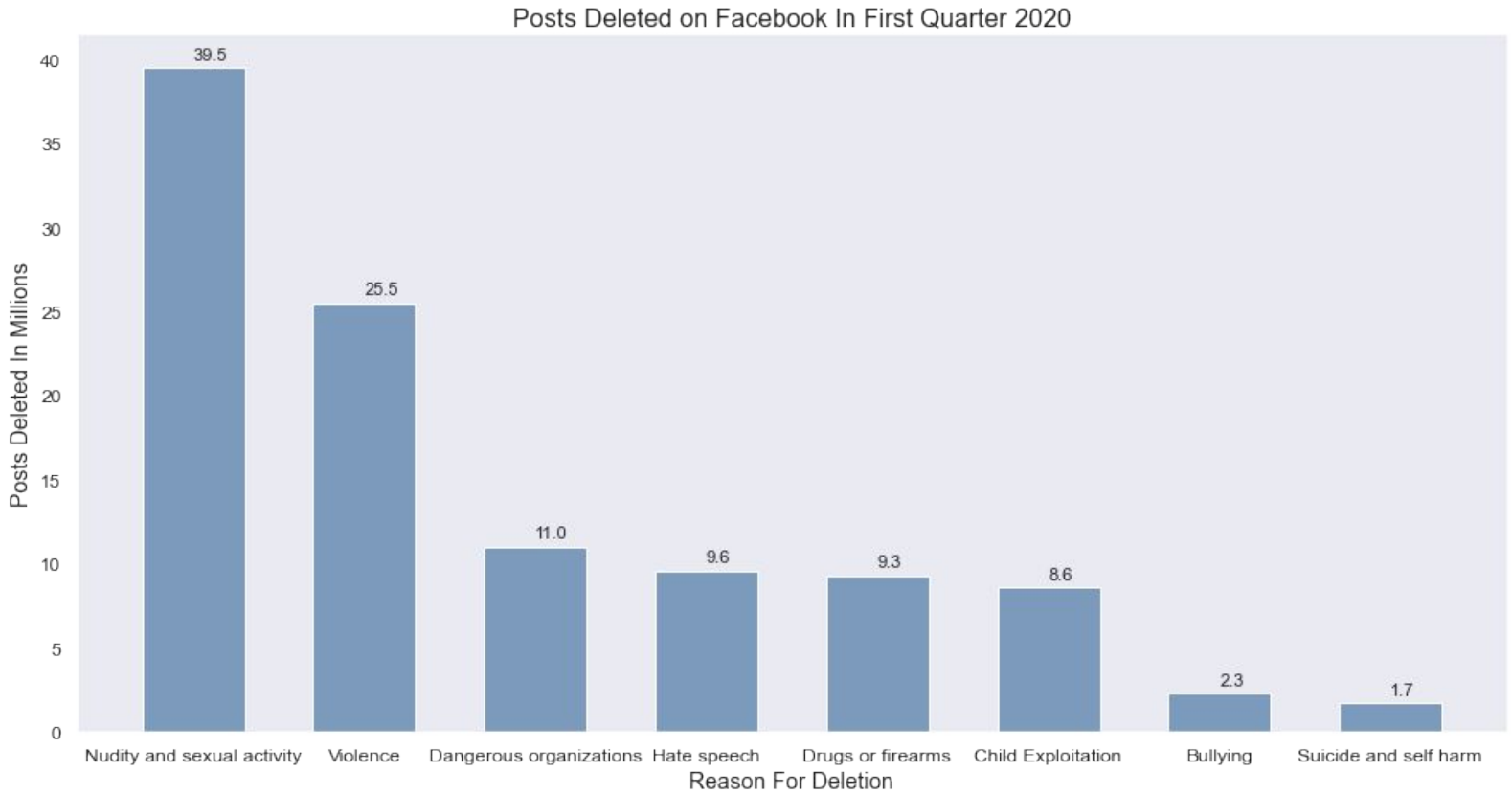
Inclusive Content Filtering

Samantha Baltodano

Agenda

1. **Content filter for explicit images**
2. **Public outrage and potential growth areas**
3. **Image classifiers**
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.”

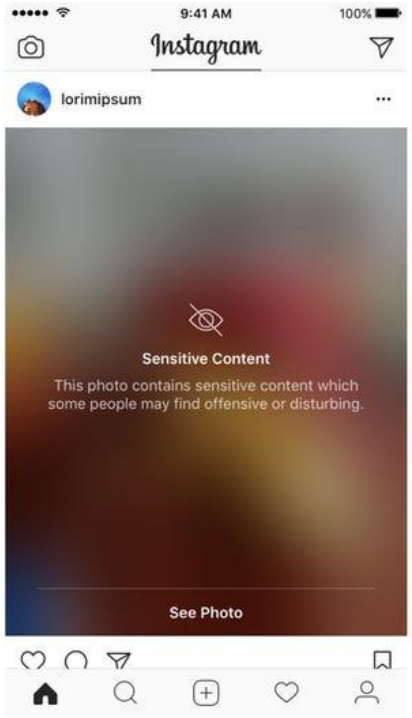
- Heather Bays, Post

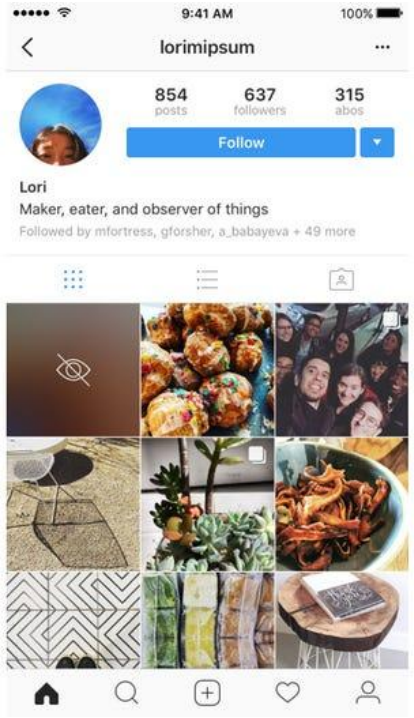
Black Instagram users are 50% more likely than white users to have their accounts automatically disabled.

- Business Insider

Biased Content Filters Prevent Inclusivity And Inhibit Freedom Of Expression

1. Explicit content filtering
2. Filtering equipt to prevent racial bias
3. #Freethenipple (just a little)





The Final Binary Model Can Predict 'SFW' Images With 91% Recall

KPI for Week Starting

Models	Image Classes	Recall	Precision	F1-Score	Accuracy
Base Models					
Binary	SFW, NSFW	**84.5%**	**84.5%**	**84.5%**	**84.5%**
Multi-Class	Neutral, Sexy, Explicit	**72.0%**	**80.0%**	**74.7%**	**81.5%**
Inclusive Models					
Binary	SFW, NSFW	**82.5%**	**84.0%**	**83.0%**	**80.2%**
Multi-Class	Breastfeeding, Neutral, Sexy, Explicit	**60.3%**	**77.0%**	**63.3%**	**80.5%**

Binary Inclusive Model
'SFW' Recall: **91%**

Recommendations & Next Steps

- Initiate change:
  - Less rigid restrictions on breastfeeding photos
  - Grayscale images
- Next:
  - Filter for violent or bullying photos
  - Collect more data for breastfeeding category

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