

# CS 6474/CS 4803 Social Computing: Online Content Moderation

*Munmun De Choudhury*

[munmund@gatech.edu](mailto:munmund@gatech.edu)

Week 8 | March 2, 2021



# Christchurch shooting: Gunman Tarrant wanted to kill 'as many as possible'

🕒 24 August 2020



Christchurch mosque shootings



- 1 44 ↑ Any women have experience with having a baby while in grad school? (self.GradSchool)  
submitted 12 hours ago by HigHog  
29 comments share
- 2 10 ↑ Hitler doesn't get a postdoc (youtube.com)  
submitted 5 hours ago by Epistaxis (PhD, genetics)  
5 comments share
- 3 6 ↑ I think it's time to quit (self.GradSchool)  
submitted 5 hours ago \* by Amelorn  
11 comments share
- 4 ↑ Advice: Masters or go straight for PhD (self.GradSchool)  
submitted 12 minutes ago by Cagedcrab  
comment share
- 5 ↑ For Profit or Not University (self.GradSchool)  
submitted 19 minutes ago by lilfermat  
comment share
- 6 ↑ First time writing an state of purpose !! pls some help (self.GradSchool)  
submitted 25 minutes ago by nautiluz92  
comment share
- ↑ Does it make sense to take a mortgage (if I have the downpayment) than rent out a studio for my 5 year PhD program? (self.GradSchool)



From the SW Mods | If you see abuse, trolling, or guideline violations, click here to message us!

- 156 ↑ Update on PM trolls, self-appointed enforcers, and SW Moderation Practices. If you got a PM that was abusive, read, thx! (self.SuicideWatch)  
submitted 11 months ago \* by SQLwitch - - stickied post  
39 comments share
- 45 ↑ Something New - an automod message to helpers (don't panic!) (self.SuicideWatch)  
submitted 3 months ago \* by skyqween - [M] - stickied post  
57 comments share
- 1 ↑ No one will want to read the note, so I'm leaving it here (self.SuicideWatch)  
submitted 2 hours ago \* by spacedoughnut  
5 comments share
- 2 ↑ I need some help here (self.SuicideWatch)  
submitted an hour ago \* by imightneedhelphere  
2 comments share
- 2 ↑ Tired of the pain (self.SuicideWatch)

ADRIAN CHEN BUSINESS 10.23.14 6:30 AM

# THE LABORERS WHO KEEP DICK PICS AND BEHEADINGS OUT OF YOUR FACEBOOK FEED

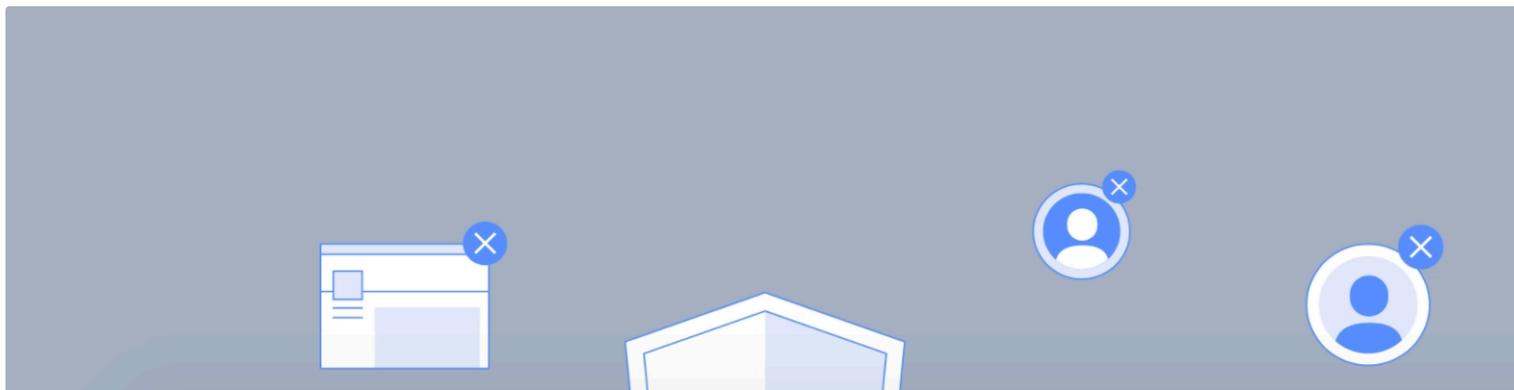
**WIRED**

 [Back to Newsroom](#)

Facebook

# Banning More Dangerous Organizations from Facebook in Myanmar

February 5, 2019

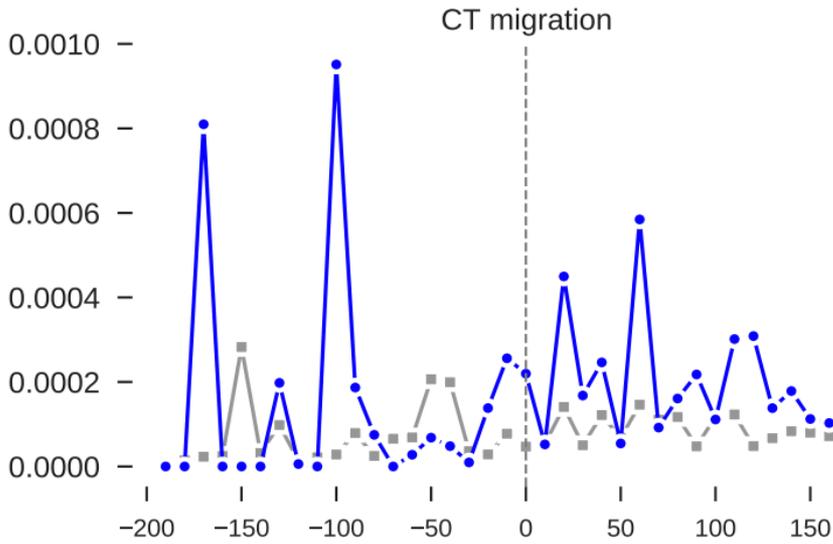


# Deviance

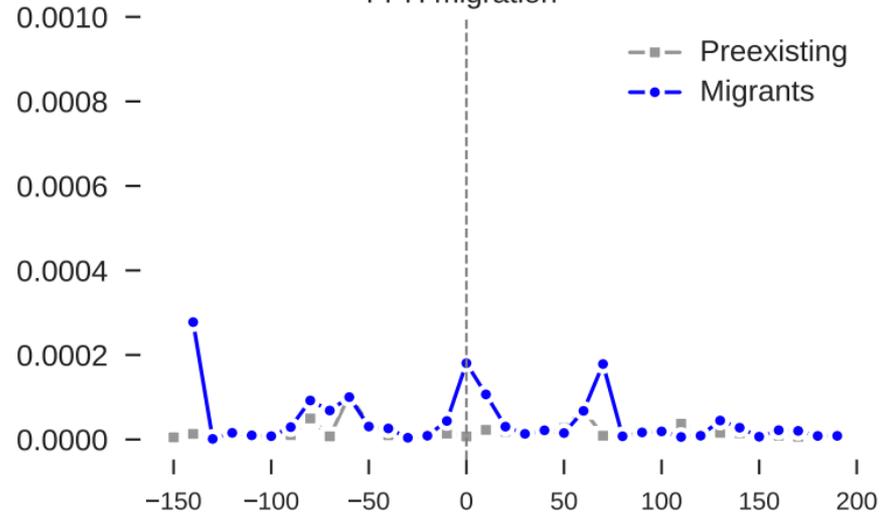
- Behaviors that violate the norms of a group
  - Akers, 1977; Suler and Phillips 1998.
- Sociological concept
  - Classically comes from Durkeim's *Anomie* book
- Online content moderation and the connection to deviance

# You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

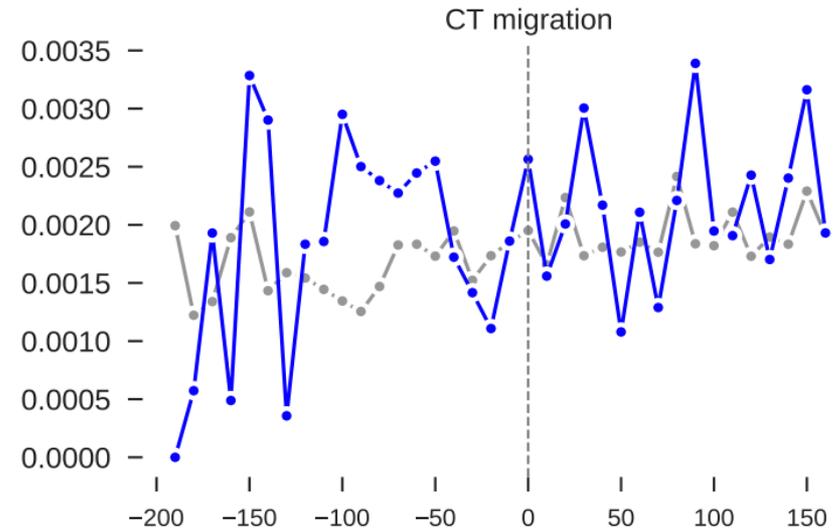
Mean Hate Speech (manually filtered words)



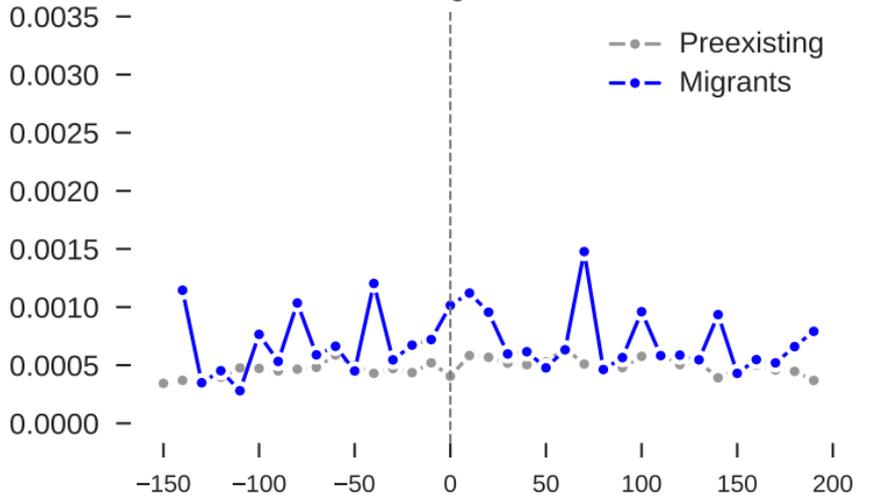
FPH migration



Mean Hate Speech (automatically generated)



FPH migration



MEDIA

JANUARY 17, 2021

## Deplatforming Trump Is Already Having a Huge Impact

*A new report finds election misinformation online has fallen 73 percent since the president's ban from Twitter.*



**MADISON PAULY**

Reporter

[Bio](#) | [Follow](#)



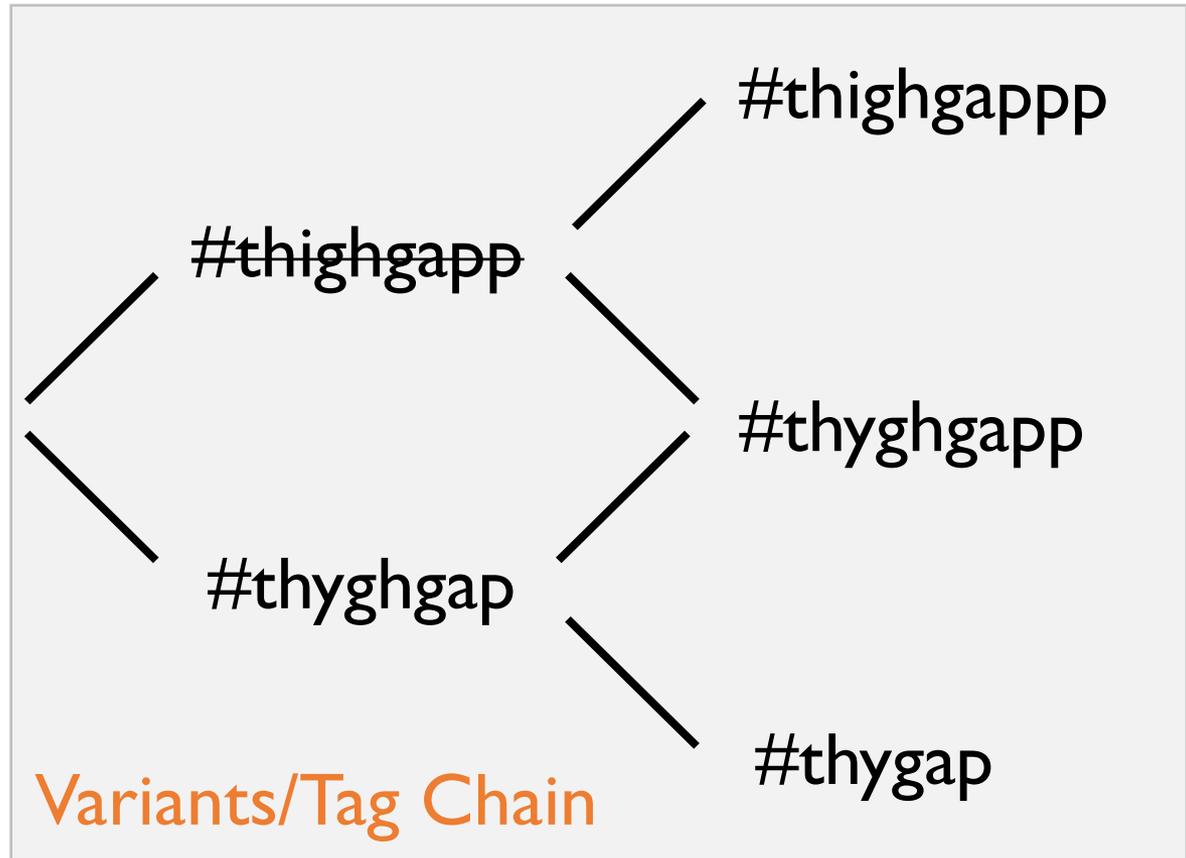


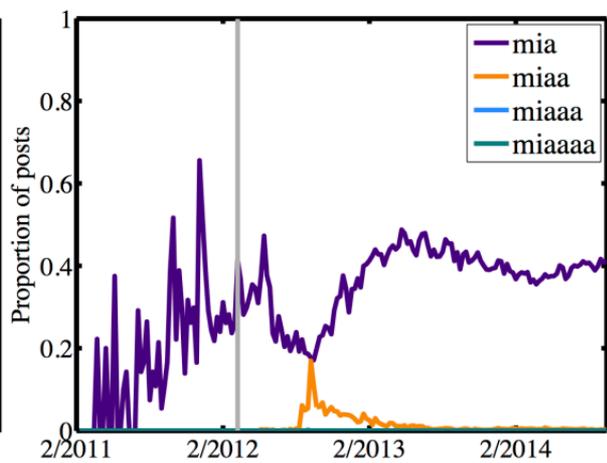
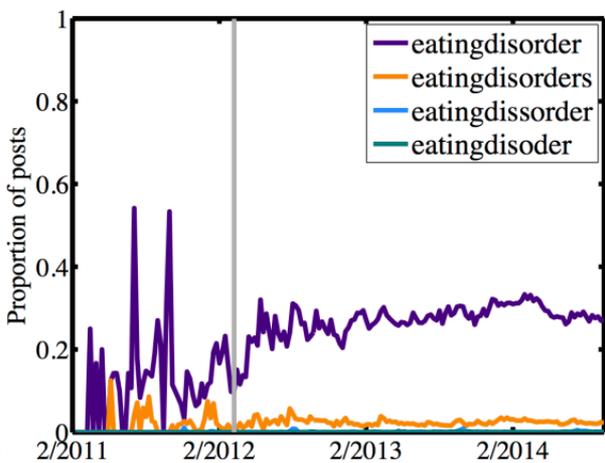
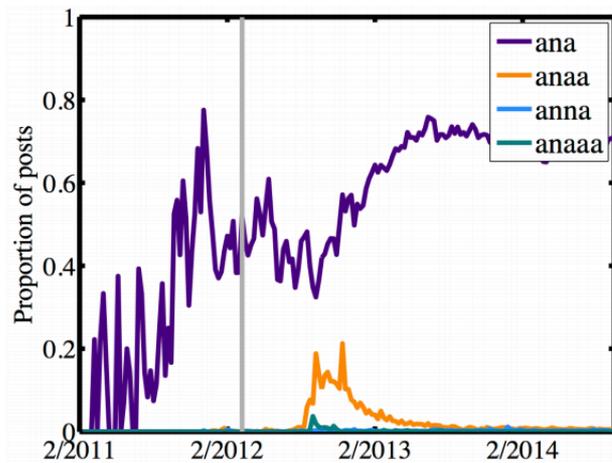
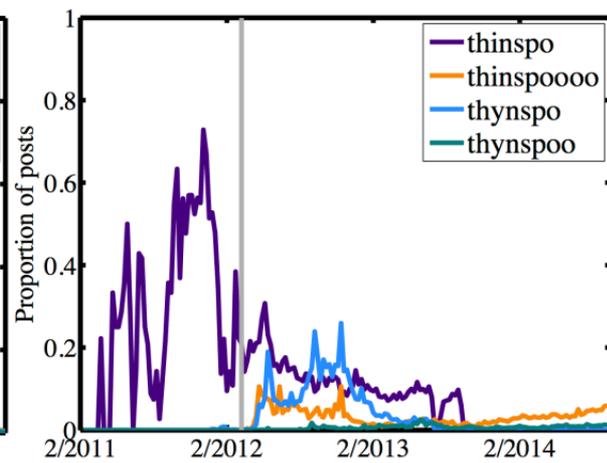
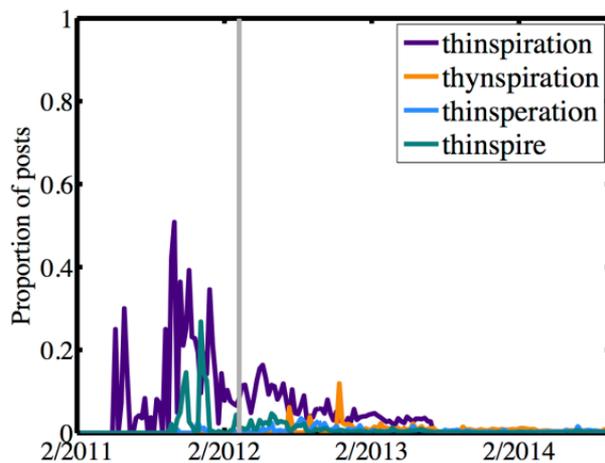
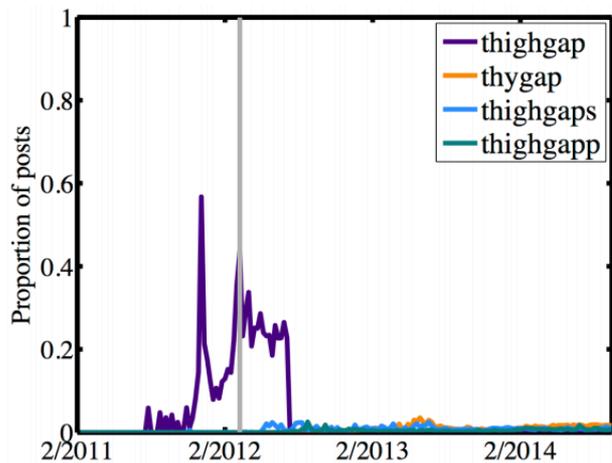
# The Fallacy of Deplatforming

Stringent moderation such as deplatforming works. But does it always?

#thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities

#thighgap  
Root





Other examples where stringent content moderation and deplatforming didn't help

# Challenges of reliance on AI moderation tools

## Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit

SHAGUN JHAVER, Georgia Institute of Technology

AMY BRUCKMAN, Georgia Institute of Technology

ERIC GILBERT, University of Michigan

When posts are removed on a social media platform, users may or may not receive an explanation. What kinds of explanations are provided? Do those explanations matter? Using a sample of 32 million Reddit posts, we characterize the removal explanations that are provided to Redditors, and link them to measures of subsequent user behaviors—including future post submissions and future post removals. Adopting a topic modeling approach, we show that removal explanations often provide information that educate users about the social norms of the community, thereby (theoretically) preparing them to become a productive member. We build regression models that show evidence of removal explanations playing a role in future user activity. Most importantly, we show that offering explanations for content moderation reduces the odds of future post removals. Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots for reducing future post removals. We propose design solutions that can promote the efficient use of explanation mechanisms, reflecting on how automated moderation tools can contribute to this space. Overall, our findings suggest that removal explanations may be under-utilized in moderation practices, and it is potentially worthwhile for community managers to invest time and resources into providing them.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation; content regulation; platform governance; post

# Platform Governance outside of the US

# Decentralized platform governance

## Article Menu

[Close](#) ^[Download PDF](#)[Open EPUB](#)

Accessing resources off campus can be a challenge. Lean Library can solve it

[Full Article](#)

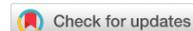
### Content List

[Introduction](#)[Turning to AI for moderation at scale](#)

## Algorithmic content moderation: Technical and political challenges in the automation of platform governance

Robert Gorwa , Reuben Binns, Christian Katzenbach

First Published February 28, 2020 | Research Article



<https://doi.org/10.1177/2053951719897945>

[Article information](#) ▾

120



### Abstract

As government pressure on major technology companies builds, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation. Automated hash-matching and predictive machine learning tools – what we define here as *algorithmic moderation systems* – are increasingly being deployed to conduct content moderation at scale by major platforms for user-generated content such as Facebook, YouTube and Twitter. This article provides an accessible technical primer on how algorithmic moderation works; examines some of the existing automated tools used by major platforms to handle copyright infringement, terrorism and toxic speech; and identifies key political and ethical issues for these systems as the reliance on them grows. Recent events suggest that algorithmic moderation has become necessary to manage growing

Class Exercise – alternative strategies for content moderation that does not involve outright banning or deplatforming