Understanding Accounts That Engage in Hate and Harassment on Reddit

Deepak Kumar, Jeff Hancock, Kurt Thomas, Zakir Durumeric

Stanford University

Google

Toxic Content in Online Spaces

- Toxic content is an enormous problem affecting 48% of Internet users
  - It is the top form of online hate and harassment
  - #1 digital safety concern among Internet users
- It behooves us as a research community to understand toxic behavior online
  - Prior work focuses on experiences of targets, characterizations of events, online conversations, off-platform tactics

What are the behaviors of toxic accounts?

Can those behaviors inform nuanced defenses?

Dataset

We collected and analyzed every Reddit comment from Jan 2020 – July 2021

929K
toxic accounts

14M
toxic comments posted

>100K
subreddits

Abuser Personas

Metric	Sub-Metric	Cluster 1	Cluster 2	Cluster 3	Defense?
Cluster	Size	52K (60%)	30K (30.8%)	4.8K (5.6%)	
Toxicity	Aggregate Toxicity	5.1%	12.1%	20%	Nudges
	Fraction of subreddits with toxic comments	19.6%	36%	50%	Community-driven rules
	Norm Violated Subreddits	12.5%	24%	75%	Platform-wide bans