Understanding Accounts That Engage in Hate and Harassment on Reddit

Deepak Kumar

Content warning: Potentially triggering language and difficult subject material ahead.


YALE

The wound that never healed

Chinese immigrants soul-searching in the post-Trump era

President of the Asian GOP in Florida, Cliff Li switched his vote to Joe Biden just days before the Nov. 3 election. Li sees a parallel between Trump and Chairman Mao Zedong, the founding father of the People's Republic of China. (Photo courtesy of Cliff Li)

Anti-Racism Protests Spark Conversations Within Chinese Immigrant Families

A New Era for Asian Americans and the Asian Diaspora around the world.

r/aznidentity

About Community

The most active Asian-American community on the web. We serve the Asian diaspora living anywhere in the West. We are Pan-Asian (East, Southeast, South) and against all forms of anti-Asian racism. We help Asians make sense out of their own life experiences, find a supportive like-minded community, and live the best possible life. We emphasize our Asian identity, not to be used as pawns by any political ideology.

Posted by u/aureolae Contributor 1 year ago 2 4

537

The ridiculousness of Eileen Huang aka 'bobacommie': It's time to take action

Activism

Eileen Huang is a student at Yale who's been making waves on social media for stupidly insulting yet sanctimonious remarks on behalf of Asian Americans. She has been discussed here.

This article does a nice job introducing people to the issue.

But it doesn't cover everything:

* She posted that “maybe it’s good to normalize racism against Asians" -- in a time of increased violence against Asians. She has since claimed it was a joke.

* She insulted other Asians like China Mac, who led a major protest against anti-Asian violence this summer, saying Asians with inner-city accents were appropriating black voices. Other than gathering clout for herself on social media, she's done nothing to help Asians.

https://twitter.com/bobacommie/status/1352289261496782853

Posted by u/kei0145 1 year ago

193

How do we cancel Eileen Huang?

Media

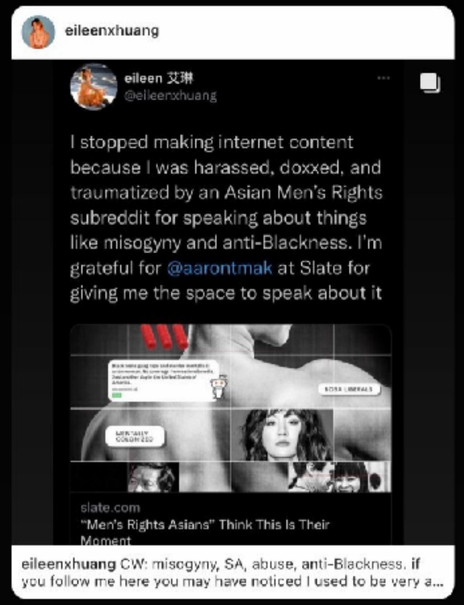
Anyone have any ideas on how to cancel this racist pos Eileen Huang? I hate all the attention that she's getting because mainstream media loves to get behind anyone who shove the "it's ok to be racist against Asians, and Asians are racist narrative down our throats", especially when it comes from some hack boba liberal.

UPDATE: I reported her twitter account for hate speech and inciting violence against a protected community, it looks like she was just suspended. Can we get her suspended on TikTok as well?

41 Comments Share Save Hide Report 96% Upvoted

8:33
◂ Chrome

hey! I haven't been actively creating content on TikTok and IG for months and months because I was psychologically abused and traumatized by a men's rights group on @reddit



it was an absolutely horrific experience that I do not wish on anyone... but that many Asian fems can unfortunately relate to. for almost a year now I've stayed silent about it, so I'm so so grateful to @slate for giving me a chance to share my experiences and speak up. go give it a read and share

Online Hate and Harassment is Ubiquitous

41% of people in US

40% of people globally

SoK: Hate, Harassment, and the Changing Landscape of Online Abuse

We still do not understand what online hate and harassment looks like at scale.

What are the behaviors of Reddit accounts that engage in hate and harassment?

Why Reddit?

- Clear connection to prior instances of harassment
- Reddit’s community structure make it unique and interesting to study
- Ease of access to historical data

Defining Online Hate and Harassment

- Online hate and harassment attacks are very broad
  - Encompasses anything from trolling to cyberstalking
- In this study, we focused on accounts that post **toxic comments**
  - "*A rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion*"

Collecting and Classifying Reddit Comments

- We collected every comment posted on Reddit from January – June 2020
  - 673M comments, 353K unique subreddits, 16.1M accounts
- Classified every comment using the Google Jigsaw Perspective API
  - Perspective API returns a score between 0 – 1 about how “toxic” an input comment is

Identifying Abusive Accounts + Comments

- Perspective API offers several different classifiers for toxicity; we chose the SEVERE_TOXICITY classifier at a reasonably high threshold which offers the best precision
- Labeled an account as an abusive account if they posted at least 1 comment above this threshold
    - Liberal threshold; but other thresholds created similar downstream results

Perspective Classifier	Threshold	Precision
IDENTITY_ATTACK	0.9	0.62
INSULT	0.9	0.53
THREAT	0.9	0.43
TOXICITY	0.9	0.51
SEVERE_TOXICITY	**0.9**	**0.75**

Dataset

156K
abusive accounts

2.6M
toxic comments posted

>100K
subreddits

What does abuse look like on Reddit?

Abusive Accounts in Aggregate

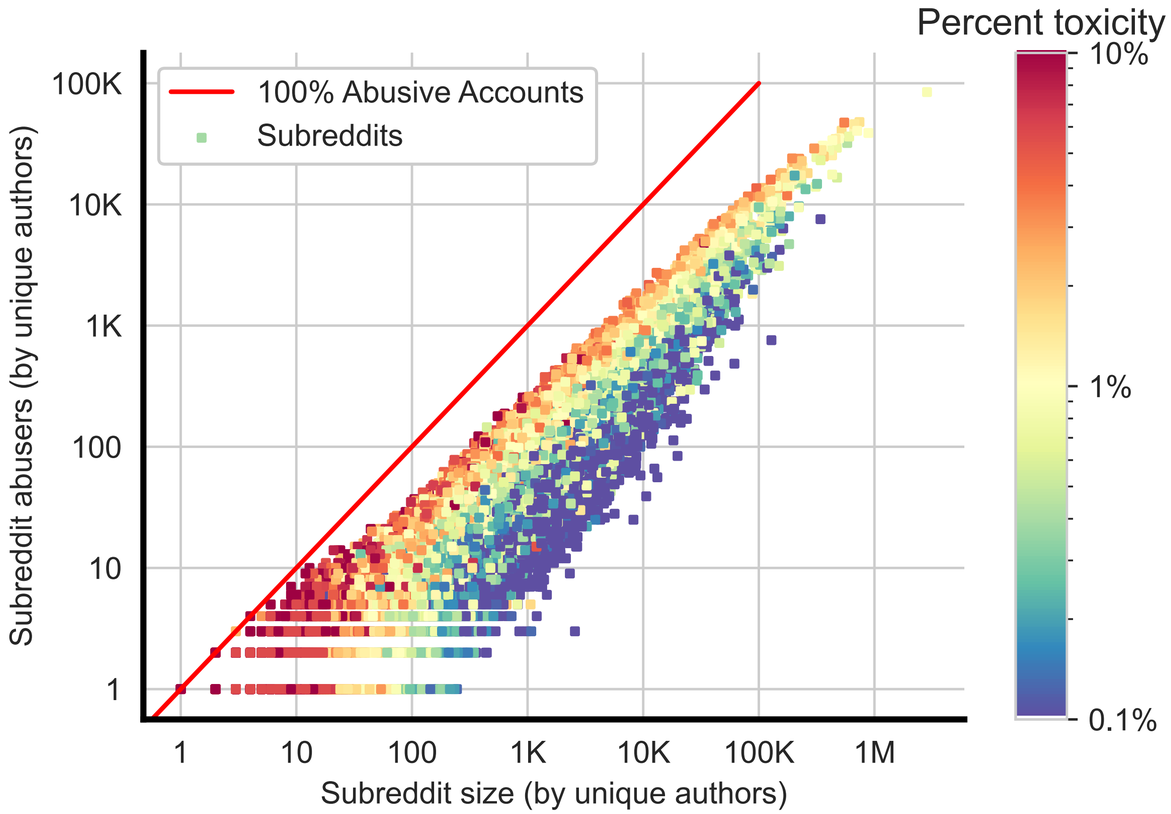
- Abusive accounts make up just ~1% of all Reddit accounts
  - However, such accounts produce significant volume: 63M (11.9%) comments are posted by abusive accounts
- 2.6M comments are toxic, accounting for 0.5% of all comments posted to the platform
  - Smaller footprint than spam, but tactics are *fundamentally different*

Abusive Accounts in Aggregate

- Abusive accounts make up just ~1% of all Reddit accounts
  - However, such accounts produce significant volume: 63M (11.9%) comments are posted by abusive accounts
- 2.6M comments are toxic, accounting for 0.5% of all comments posted to the platform
  - Smaller footprint than spam, but tactics are *fundamentally different*
- **Takeaway: Traditional defenses (e.g., banning accounts) would have outside detrimental impact on platform conversation and health**

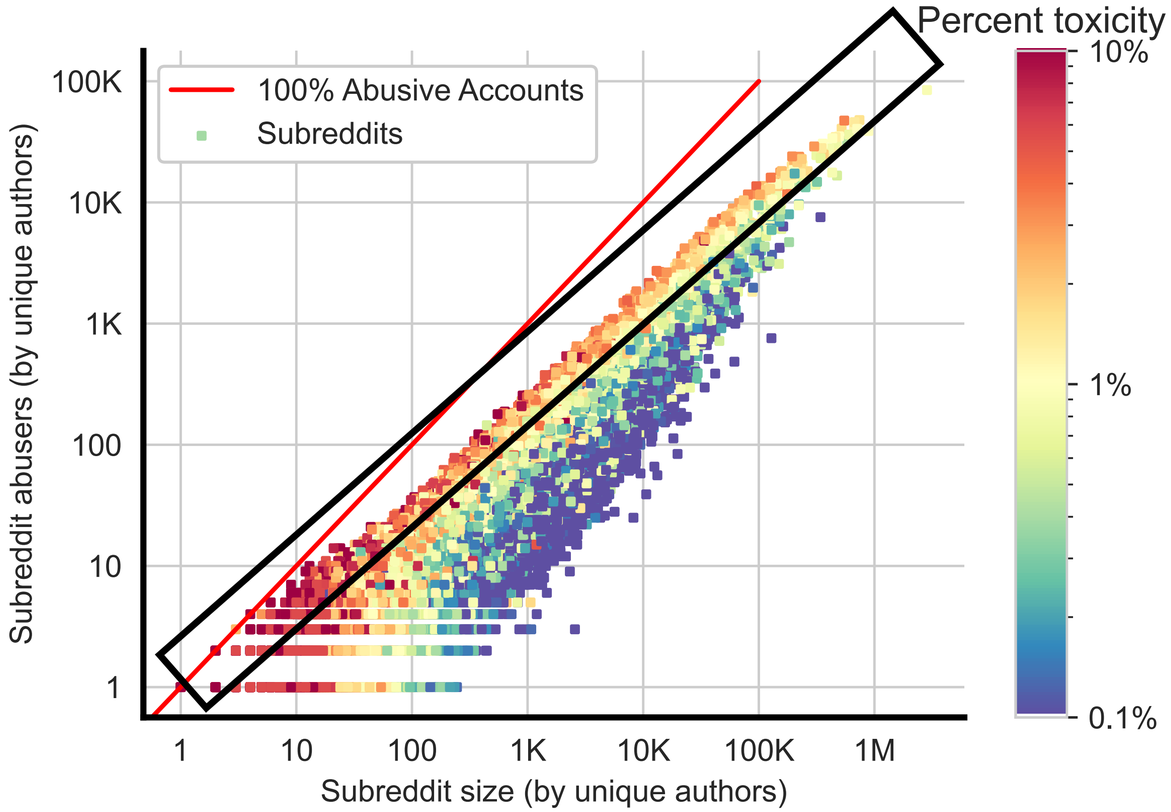
Footprint of Abusive Accounts

- Abusive accounts post in 100K subreddits; 43K (42.6%) contain at least one toxic comment



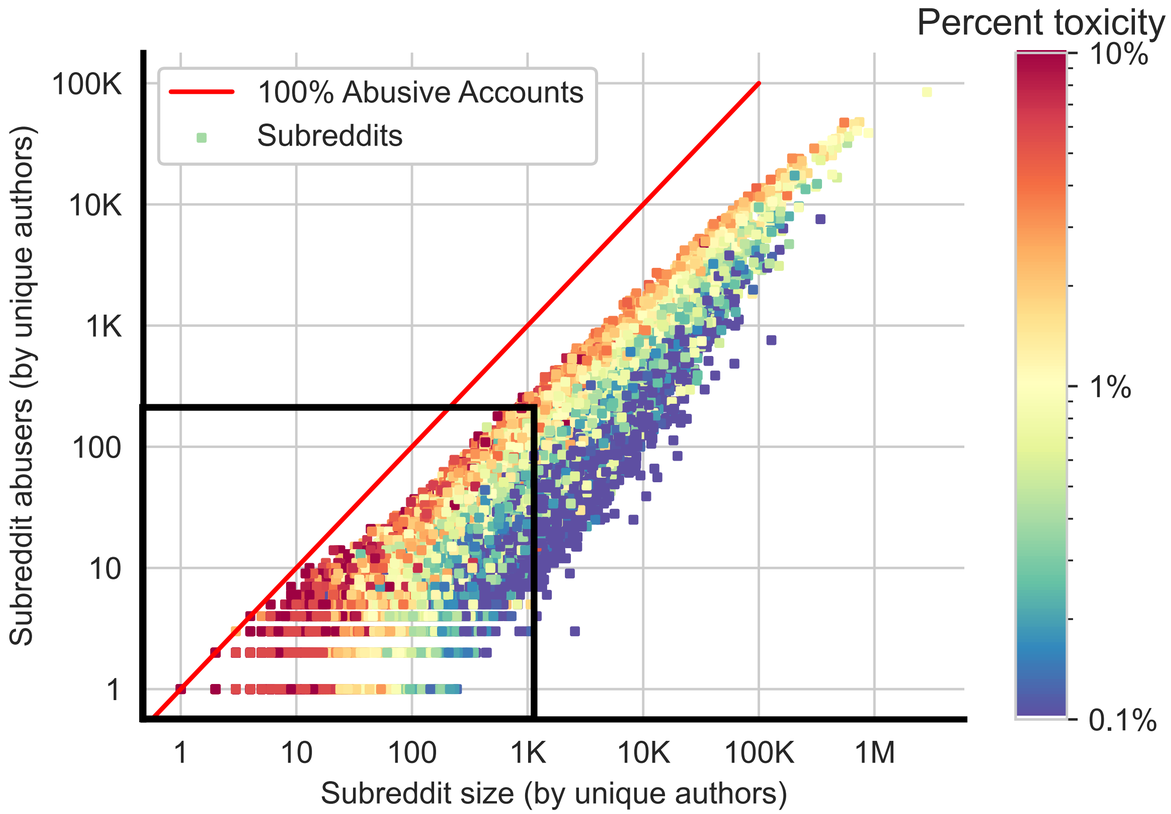
Footprint of Abusive Accounts

- Abusive accounts post in 100K subreddits; 43K (42.6%) contain at least one toxic comment



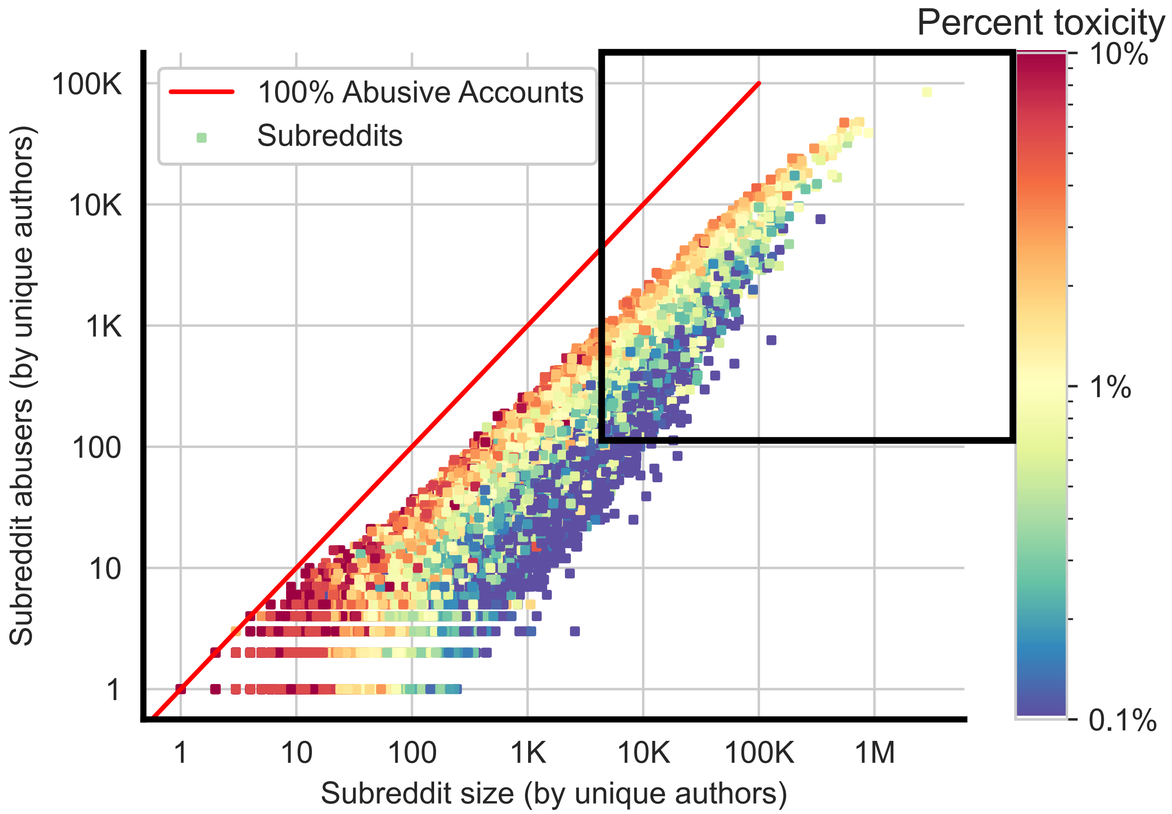
Footprint of Abusive Accounts

- Abusive accounts post in 100K subreddits; 43K (42.6%) contain at least one toxic comment
  - Subreddits in the red band are hotspots of toxic activity; 1% of subreddits consist entirely of abusive accounts



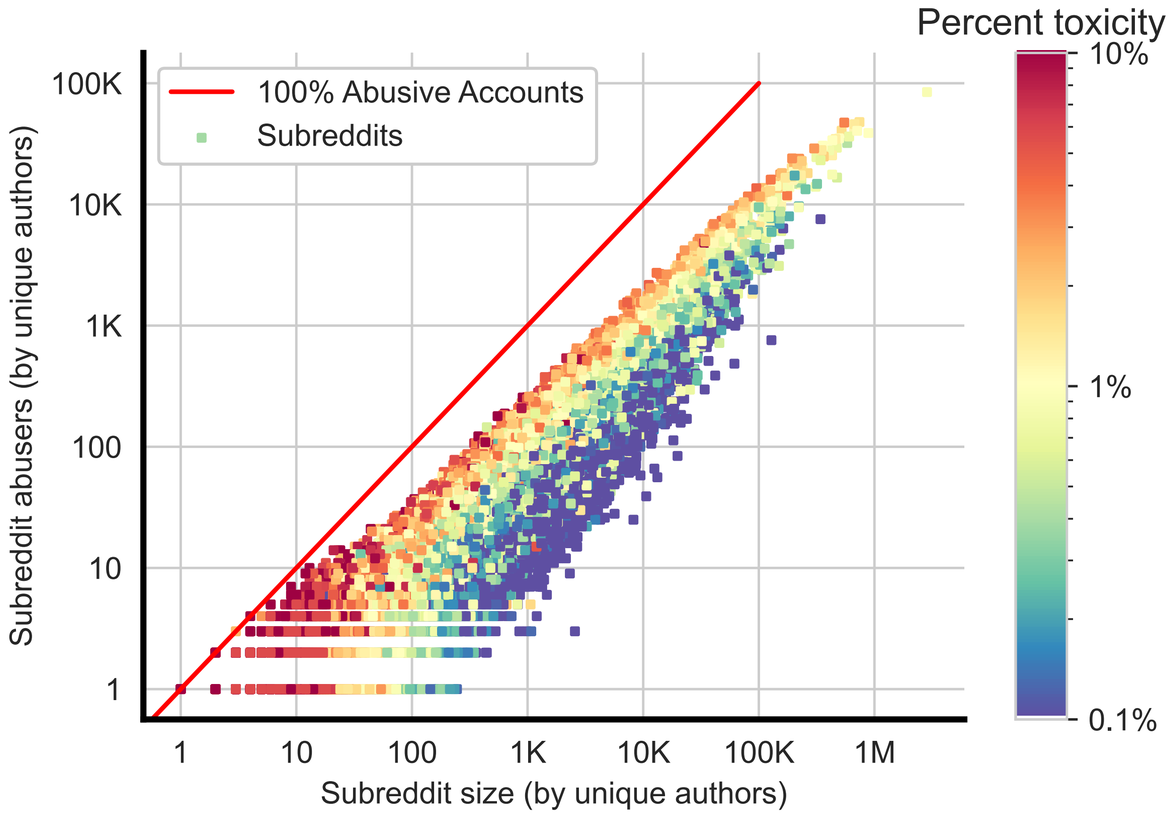
Footprint of Abusive Accounts

- Abusive accounts post in 100K subreddits; 43K (42.6%) contain at least one toxic comment
  - Subreddits in the red band are hotspots of toxic activity; 1% of subreddits consist entirely of abusive accounts
- Some large subreddits consist of upwards of 20% toxic comments!



Footprint of Abusive Accounts

- Abusive accounts post in 100K subreddits; 43K (42.6%) contain at least one toxic comment
  - Subreddits in the red band are hotspots of toxic activity; 1% of subreddits consist entirely of abusive accounts
- Some large subreddits consist of upwards of 20% toxic comments!
- 13.1M (78.2%) accounts participate in a thread where toxic comments appear, suggesting such comments are a pervasive part of Reddit



Toxic Comment Breakdown

Category	% Comments
Insult	63.4%
Identity Attack	14.2%
Call to Leave Conversation	12%
Threat	5.5%
Sexual Aggression	2.8%
Identity Misrepresentation	1.6%

Toxic Comment Breakdown

Category	% Comments
Insult	**63.4%**
Identity Attack	14.2%
Call to Leave Conversation	12%
Threat	5.5%
Sexual Aggression	2.8%
Identity Misrepresentation	1.6%

“I don’t know what’s more salty. Your mouth or your asshole. Not like there is a big difference between them in your case, diarrhea and entitlement come out of both ends, and they’re both just as pathetic.”

Toxic Comment Breakdown

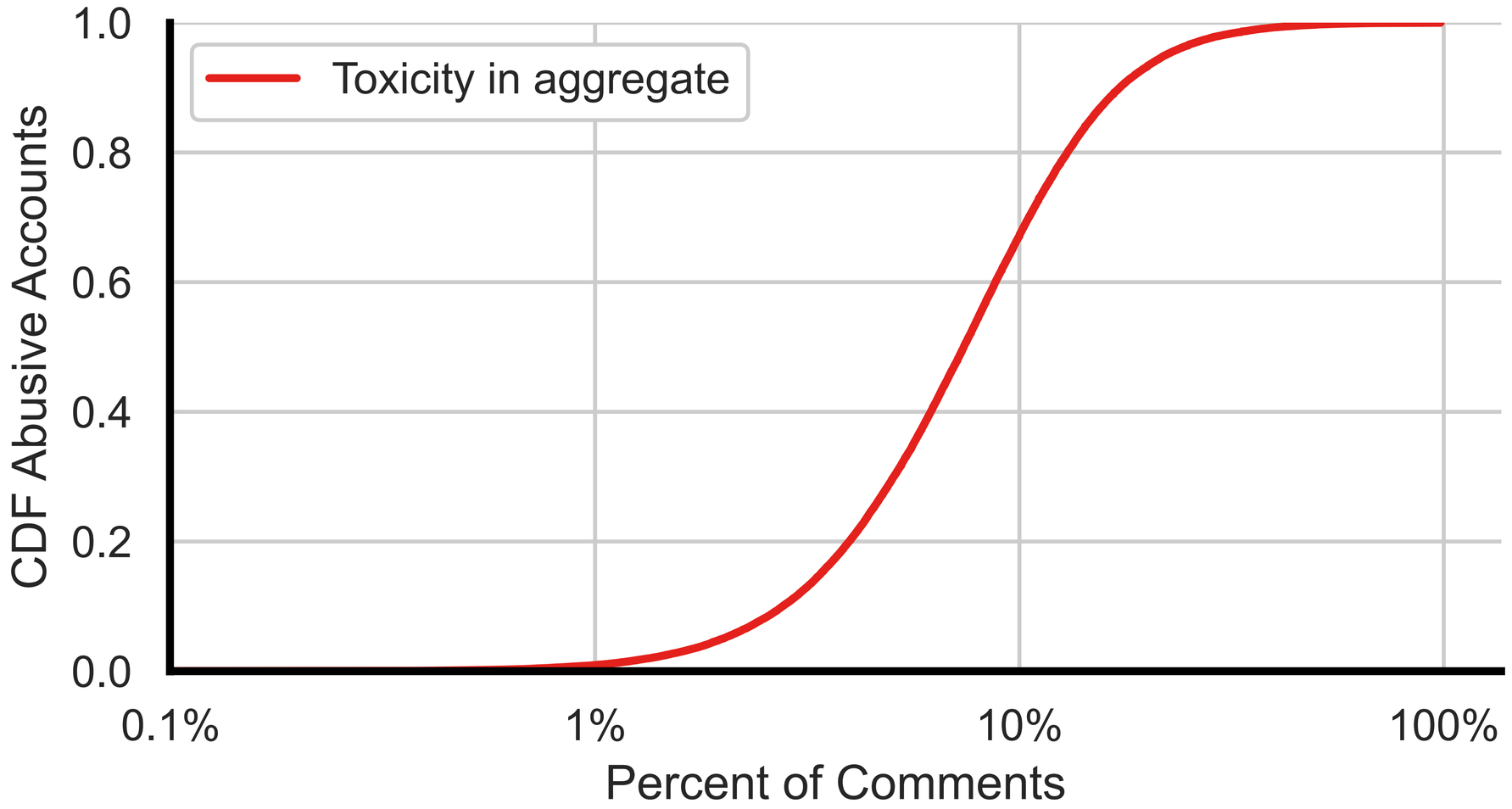
Category	% Comments
Insult	63.4%
Identity Attack	14.2%
Call to Leave Conversation	**12%**
Threat	5.5%
Sexual Aggression	2.8%
Identity Misrepresentation	1.6%

"...Get the fuck off this subreddit. You clearly don't really care about how severe NVLD can be. There's literally no fucking help out there for us."

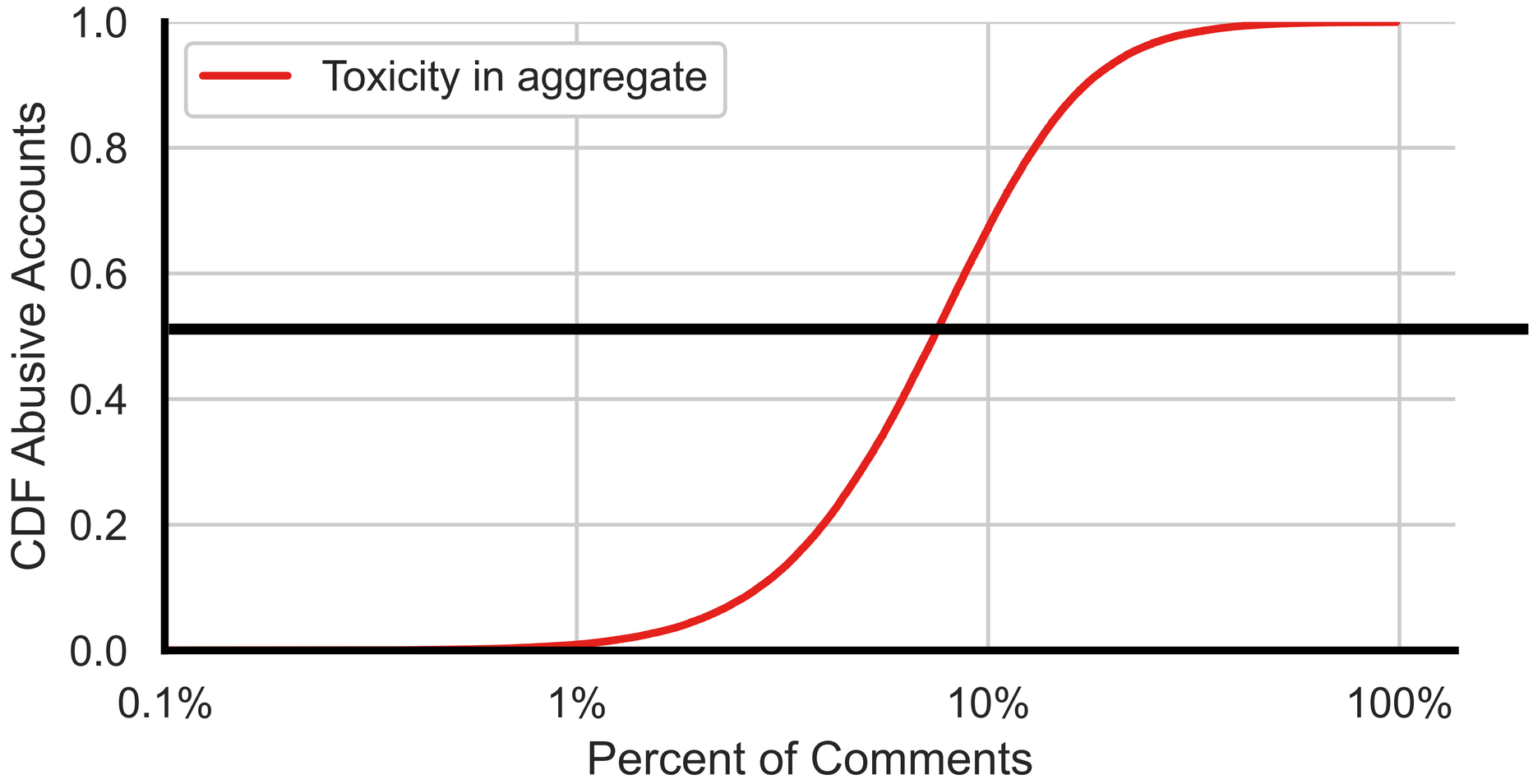
Toxic comments are a highly visible and pervasive part of Reddit.

What are the toxicity behaviors of abusive accounts?

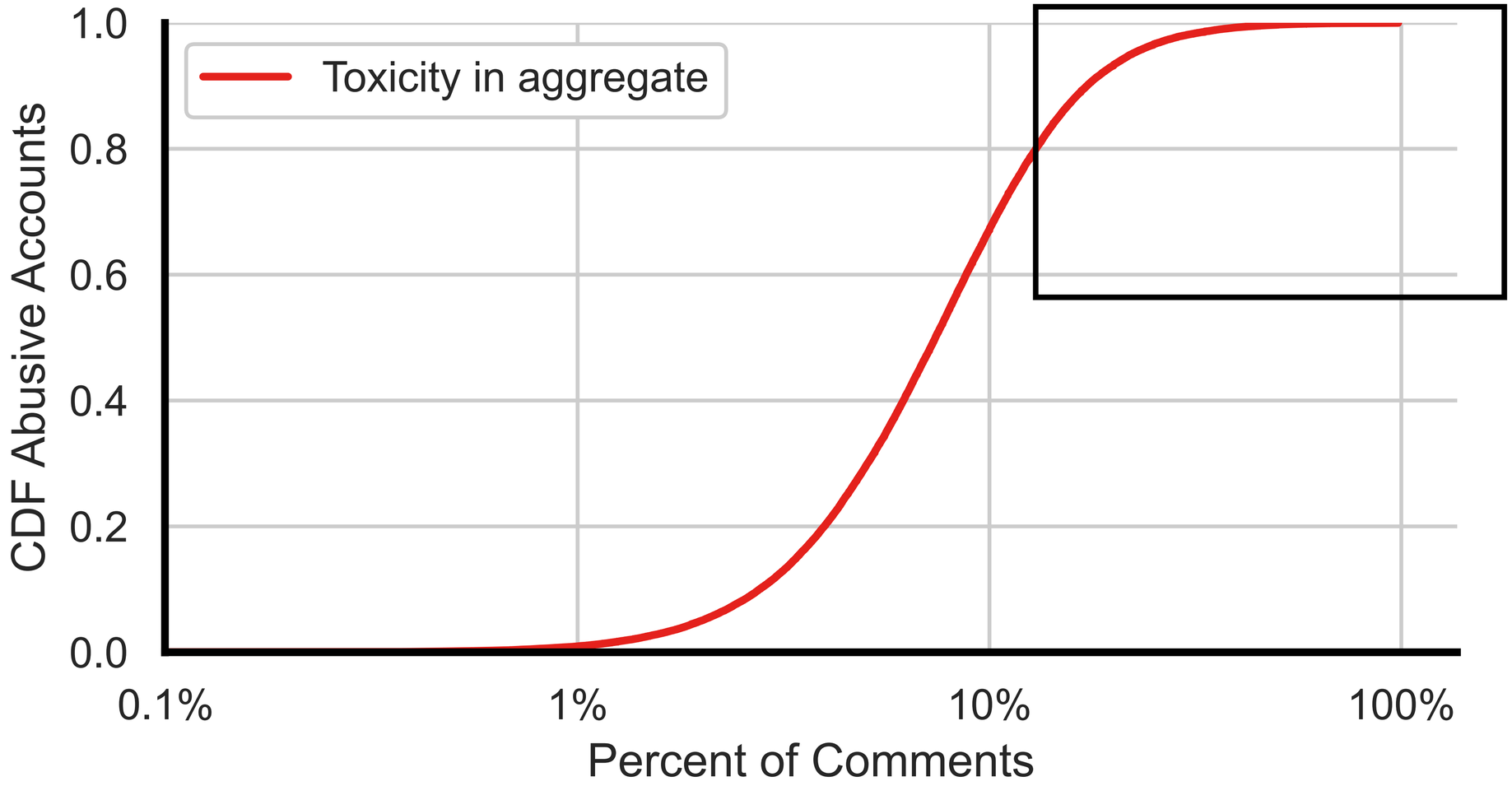
Abusive Account Toxicity



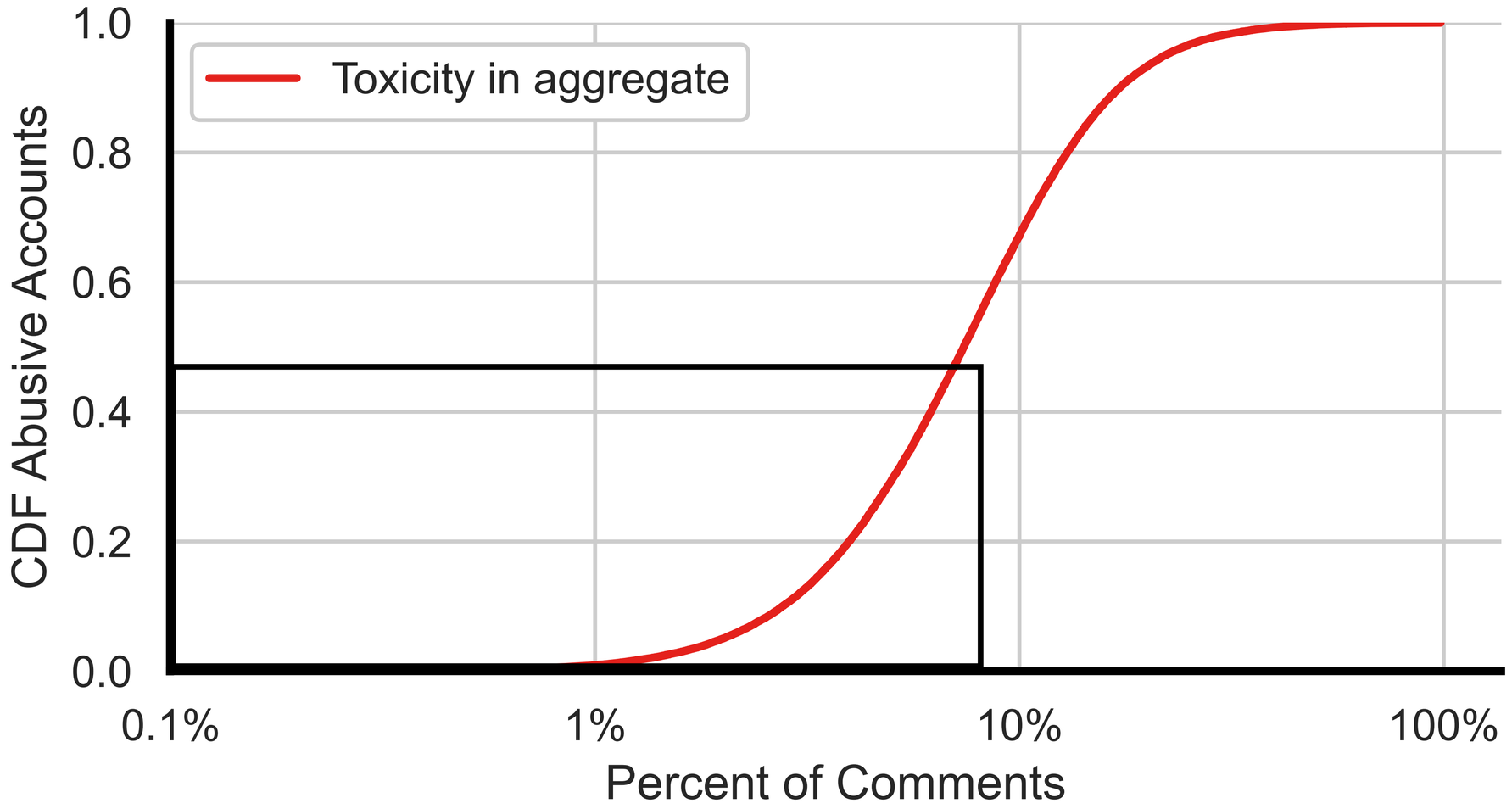
Abusive Account Toxicity



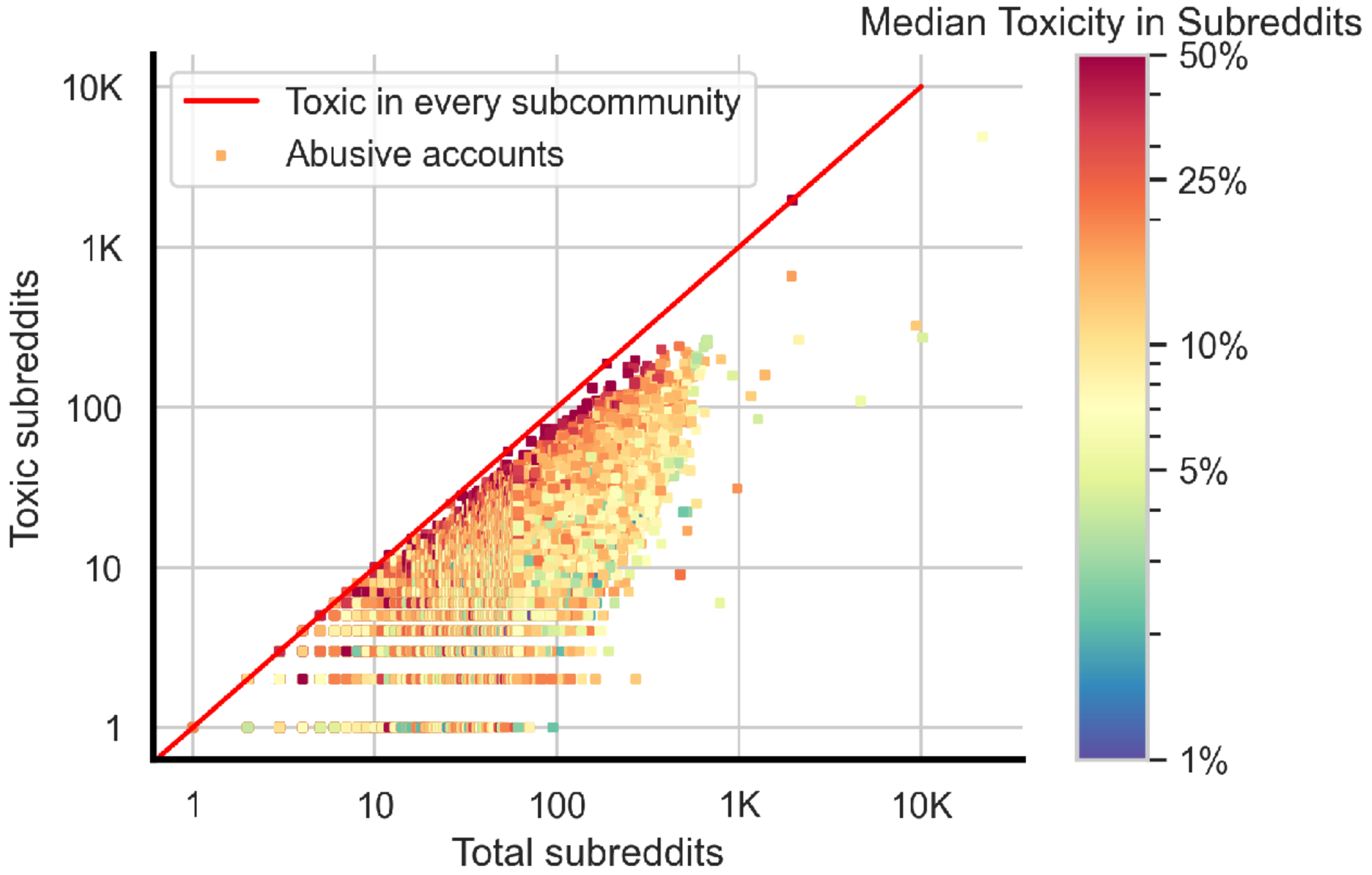
Abusive Account Toxicity



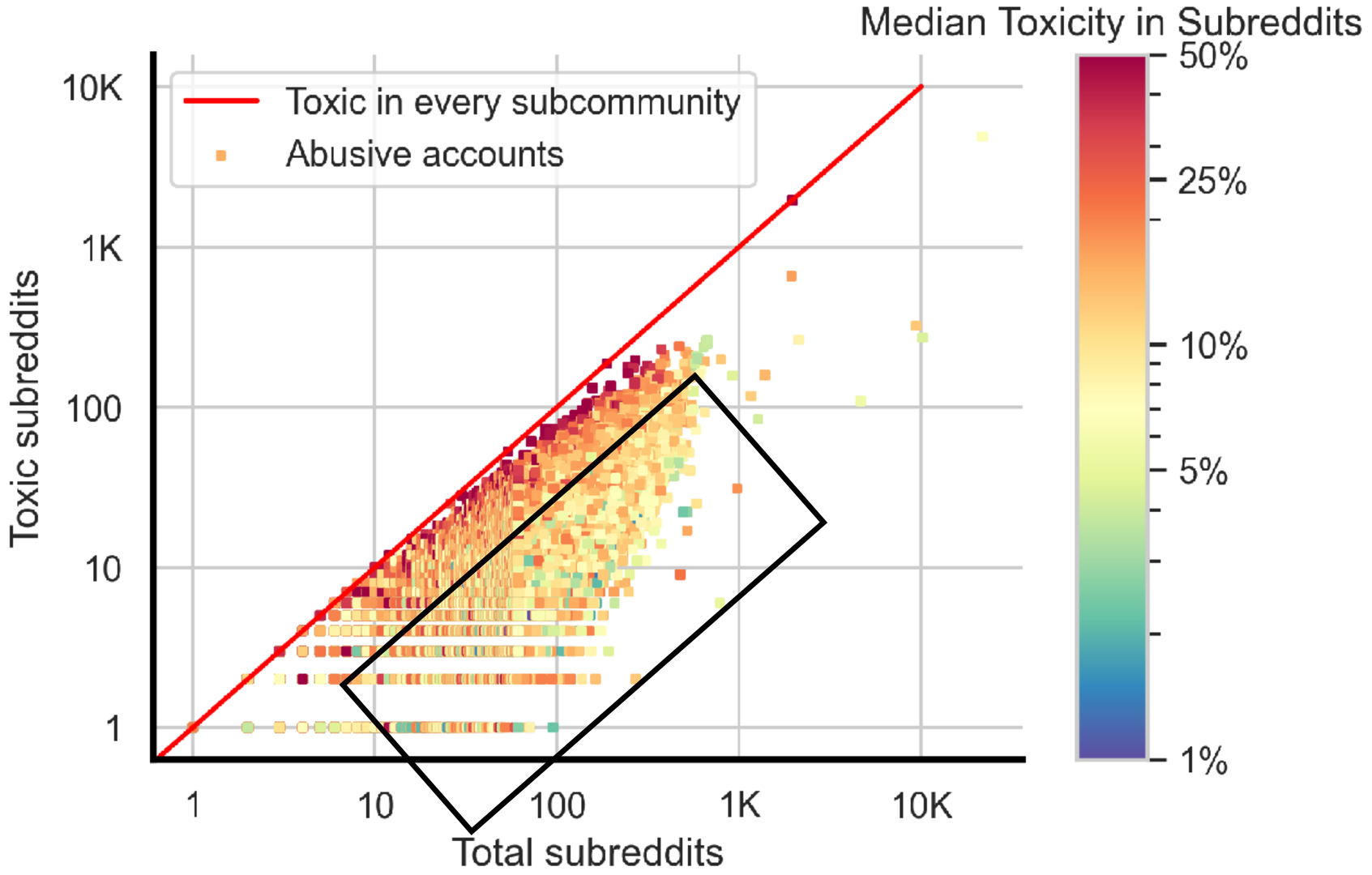
Abusive Account Toxicity



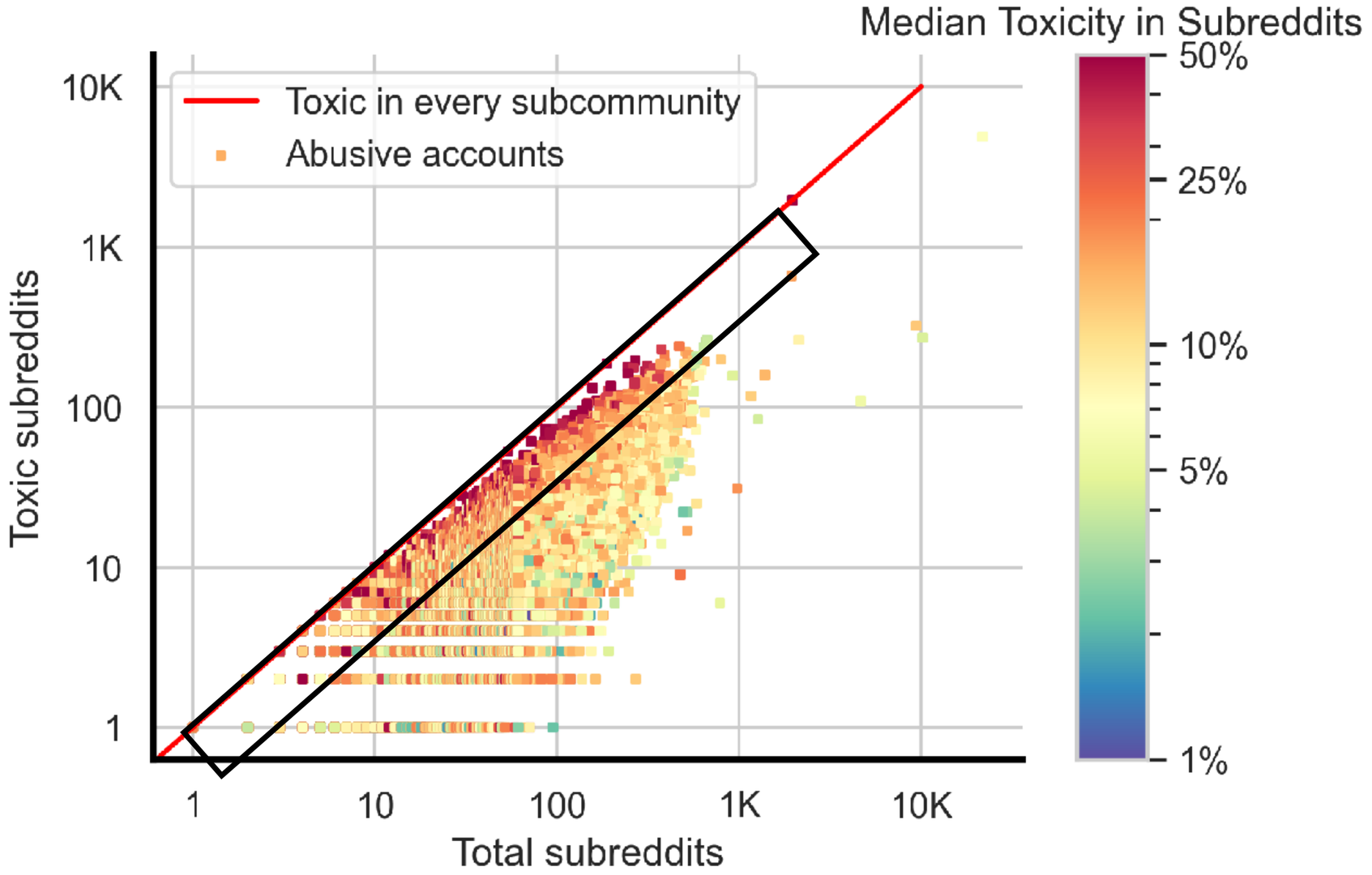
Abuser Behaviors Vary by Subreddit



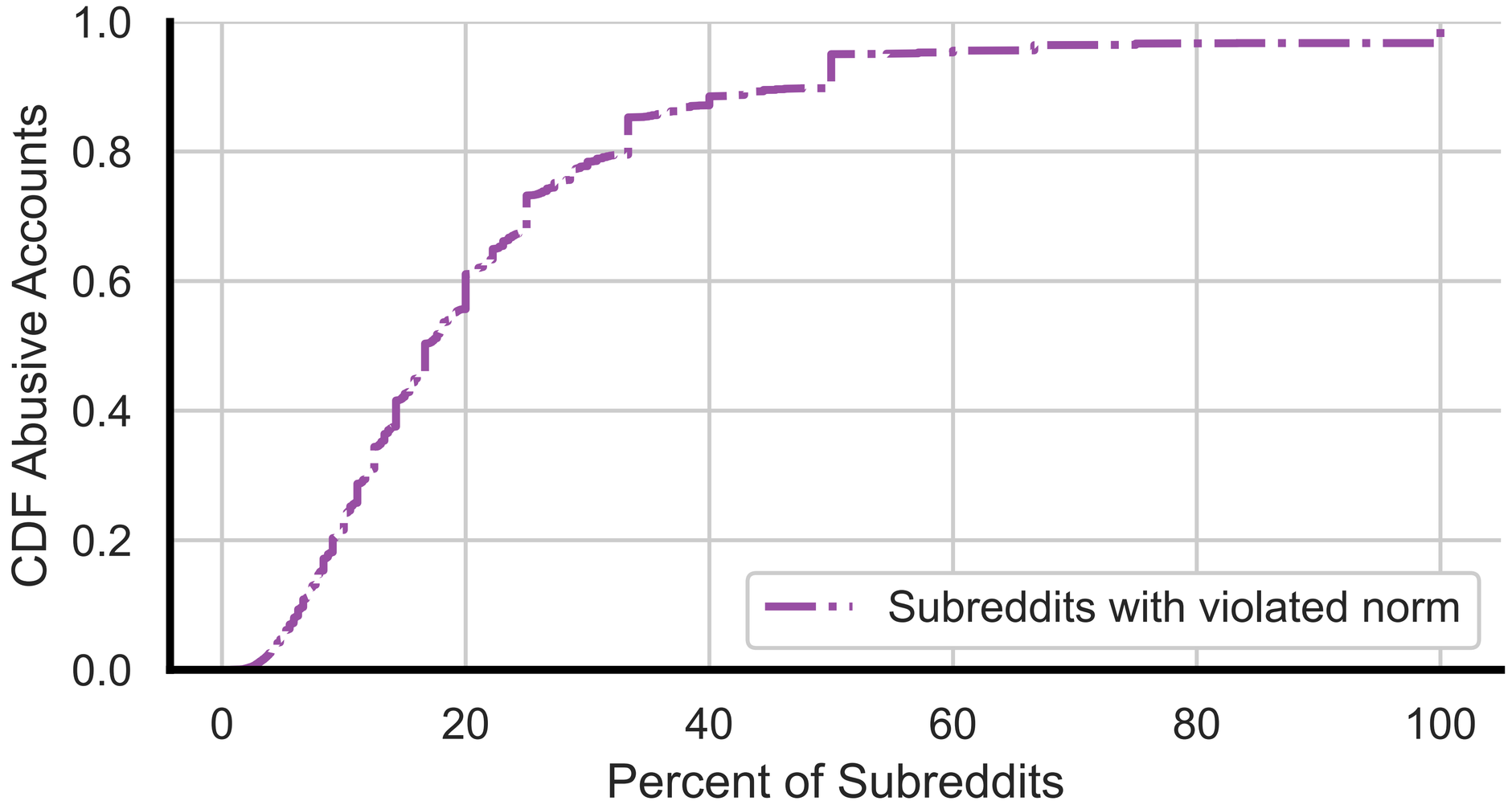
Abuser Behaviors Vary by Subreddit



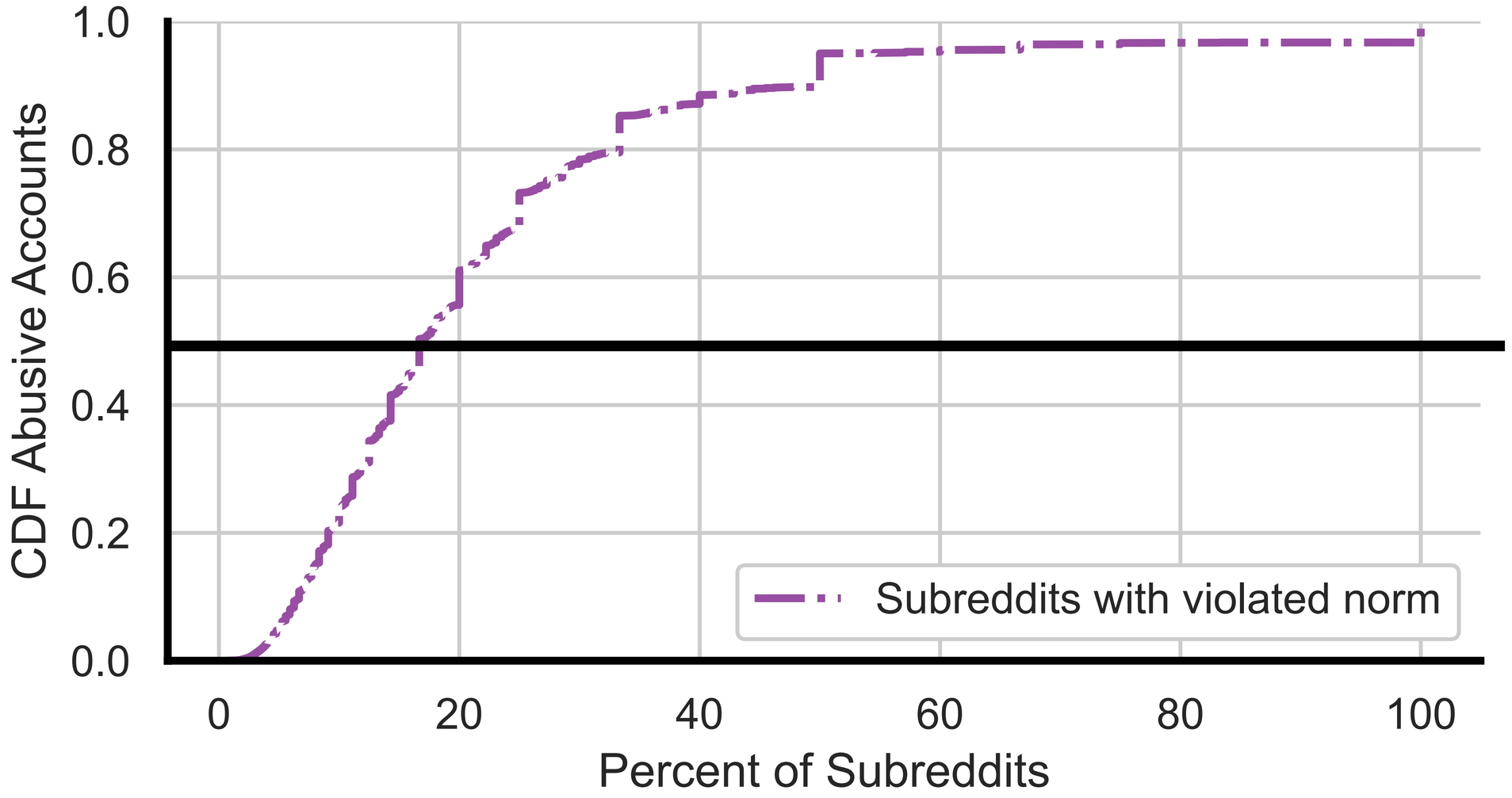
Abuser Behaviors Vary by Subreddit



Abusers Violate Subreddit Norms



Abusers Violate Subreddit Norms



Building Abuser Personas

- Varied abuser toxicity behaviors imply *classes, or personas* of abusive accounts that can inform defenses
- Clustered (with PCA and K-Means) abusive accounts based on four features:
  - Fraction of abuser comments that are toxic
  - Fraction of subreddits that abusers post a toxic comment in
  - Median fraction of toxic comments for each subreddit the abuse participates in
  - The fraction of sub communities that each abusive account violates a norm in
- Identified **3 classes of abusive accounts**

Abuser Personas

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

Abuser Personas – Occasional Abusers

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

Abuser Personas – Moderate Abusers

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

Abuser Personas – Serial Abusers

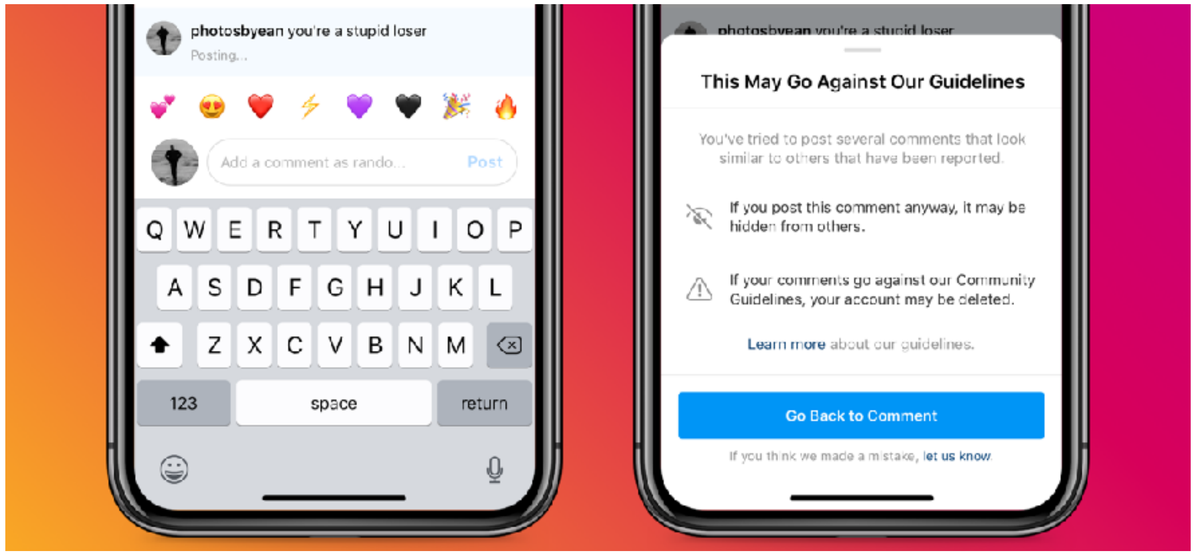
Metric	Sub-Metric	Cluster 1	Cluster 2	Cluster 3
Cluster	Size	52K (60%)	30K (30.8%)	4.8K (5.6%)
Toxicity	Aggregate Toxicity	5.1%	12.1%	20%
	Fraction of subreddits with toxic comments	19.6%	36%	50%
	Norm Violated Subreddits	12.5%	24%	75%

Abusive accounts are not singular, and defenses should follow abuse patterns

Cluster-1 Informed Defenses – Nudges

Reconsidering Tweets:
Intervening During Tweet Creation Decreases Offensive Content

Matthew Katsaros[1], Kathy Yang[2], Lauren Fratamico[2]
[1]Yale Law School
[2]Twitter Inc.
matthew.katsaros@yale.edu
{kathyy, lfratamico}@twitter.com



Cluster-2 Informed Defenses – Community-based Rules

Twitter's new strike system will target prolific COVID-19 fake information spreaders

Twitter says repeat offenders will be booted from the platform.

Strike system

With the new Rule[0] reivision, we'll also be introducing a strike system in an attempt to improve the content quality and encourage people to read and follow the new rule. Authors of posts that will be removed for violating the new revision of Rule[0] will receive **1** strike for every post removed. Please note that the strike system currently only applies to Rule[0]. The following punishments will be given for receiving strikes:

- Strike 1 - 1 day tempban
- Strike 2 - 3 day tempban
- Strike 3 - 7 day tempban
- Strike 4 - 30 day tempban
- Strike 5 - permanent ban

Cluster-3 Informed Defenses – Platform-Wide Bans

Hello,

Your account has been suspended due to multiple or repeat violations of the Twitter Rules: https://twitter.com/rules.

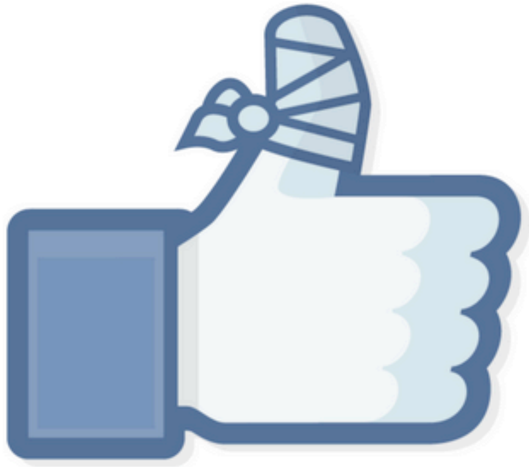
Please do not respond to this email as replies and new appeals for this account will not be monitored.

Thanks,

Twitter Support

Sorry, this page isn't available

The link you followed may be broken, or the page may have been removed.



Go back to the previous page · Go to the Facebook homepage · Visit the Help Center

Key Takeaways

- Toxic content is distinct from spam and fraud; traditional defensive mechanisms are not as easy to apply without downstream platform health
- Toxic content has a wide reach on Reddit, and impacts a significant number of interactions even if volume is low
- Harassment interventions cannot be “one-size-fits-all”; need to be nuanced to capture varied attacker patterns

Key Takeaways

- Toxic content is distinct from spam and fraud; traditional defensive mechanisms are not as easy to apply without downstream platform health
- Toxic content has a wide reach on Reddit, and impacts a significant number of interactions even if volume is low
- Harassment interventions cannot be “one-size-fits-all”; need to be nuanced to capture varied attacker patterns

Deepak Kumar

kumarde@cs.stanford.edu

@_kumarde

Backup

Identifying Abusive Accounts

- Perspective API has several classifiers trained to look for different types of abuse
  - We looked for the a classifier + threshold combination that maximized precision, as to reduce the number of *false positives*
- Leveraged data collected from Reddit in prior work* to identify an “abusive account” threshold

Perspective Classifier	Max Threshold	Precision
IDENTITY_ATTACK	0.9	0.62
INSULT	0.9	0.53
THREAT	0.9	0.43
TOXICITY	0.9	0.51
SEVERE_TOXICITY	**0.9**	**0.75**

*Designing Toxic Content Classification for a Diversity of Perspectives, SOUPS 2021

Identifying Abusive Accounts

- Perspective API has several classifiers trained to look for different types of abuse
  - We looked for the a classifier + threshold combination that maximized precision, as to reduce the number of *false positives*
- Leveraged data collected from Reddit in prior work* to identify an “abusive account” threshold
- Choosing SEVERE_TOXICITY >= 0.9 threshold offers a high precision signal for whether a comment is toxic; but significantly reduced data volume
  - Only 203K (**0.03%**) comments from 156K accounts meet this threshold

Perspective Classifier	Max Threshold	Precision
IDENTITY_ATTACK	0.9	0.62
INSULT	0.9	0.53
THREAT	0.9	0.43
TOXICITY	0.9	0.51
SEVERE_TOXICITY	**0.9**	**0.75**

51*Designing Toxic Content Classification for a Diversity of Perspectives, SOUPS 2021

Expanding Abusive Comments Set

- Can we expand the set of abusive comments to capture more abusive behaviors?
    - Hypothesis: If an account engages in toxic behavior, other posts made by the same account may be toxic at a *lower threshold*

Expanding Abusive Comments Set

- Can we expand the set of abusive comments to capture more abusive behaviors?
  - Hypothesis: If an account engages in toxic behavior, other posts made by the same account may be toxic at a *lower threshold*
- Randomly sampled 200 comments from abusive accounts and 200 from nonabusive accounts; evaluated precision from comments at each threshold

Threshold	Nonabusive Precision	Abusive Precision	Fraction of Comments
0.5	0.19	0.56	6.6M (1.3%)
0.7	0.22	0.68	2.6M (0.5%)
0.9	NA	0.75	203K (0.05%)

Expanding Abusive Comments Set

- Can we expand the set of abusive comments to capture more abusive behaviors?
  - Hypothesis: If an account engages in toxic behavior, other posts made by the same account may be toxic at a *lower threshold*
- Randomly sampled 200 comments from abusive accounts and 200 from nonabusive accounts; evaluated precision from comments at each threshold
- We could lower the threshold with only a modest drop in precision; strong enough signal to start measuring behaviors at scale

Threshold	Nonabusive Precision	Abusive Precision	Fraction of Comments
0.5	0.19	0.56	6.6M (1.3%)
0.7	**0.22**	**0.68**	**2.6M (0.5%)**
0.9	NA	0.75	203K (0.05%)