

Open or Blocked Skies? Community Moderation Practices in Bluesky

Saidu Sokoto, Leonhard Balduf, Onur Ascigil, Gareth Tyson, Ignacio Castro, Björn Scheuermann, Andrea Baronchelli, & Michał Król



Our Journey in the Blue Skies

Previous Work:

- IMC '24: Looking at the Blue Skies of Bluesky
- ICWSM '25: Bootstrapping social networks: Lessons from Bluesky starter packs

Today's Focus: community moderation (WWW '26).

Decentralization of Social Media

Centralized



Unlimited power over:

- Moderation
- Access

Federated



Power per instance

- Centralization

Decomposed



Community-contributed functionalities

- Moderation (labels, blocks, blocklists,...)
- Feeds

How do Blocks work

Flatbush Cats

@flatbushcats.bsky.social

5.4K followers 2 following 1 post

501(c)(3) nonprofit focused on reducing the outdoor cat population

Posts

Replies

Media

+ Follow

...

Copy link to profile

Search posts

Add to starter packs

Add to lists

Mute account

Block account

Report account

MAGA

Moderation list by @skywatch.blue

Username, description, and / or handle which includes MAGA, MAHA, TRUMP, or TRUMP 2024, or posts with #MAGA and #TRUMP2024 hashtags or clear indicators of support for Trump. Also includes accounts named after or claiming to be Trumpworld figures.

Subscribe

...

Content Scrapers

Moderation list by @contentscrapers.bsky.social

Scrapers are accounts that steal, sometimes monetize, generic "interesting content" (animals! space! nature!), often with misleading captions, without permission from the creators, to gain clout.

This list primarily focuses on science/art-related content, including photography. Report via DM.

Subscribe

...

The Great AI - NFT - CRYPTO Cull

Moderation list by @rayday.bsky.social

If you prefer to avoid - AI - NFT - CRYPTO content. This lists blocks all three things. Use at your own will.

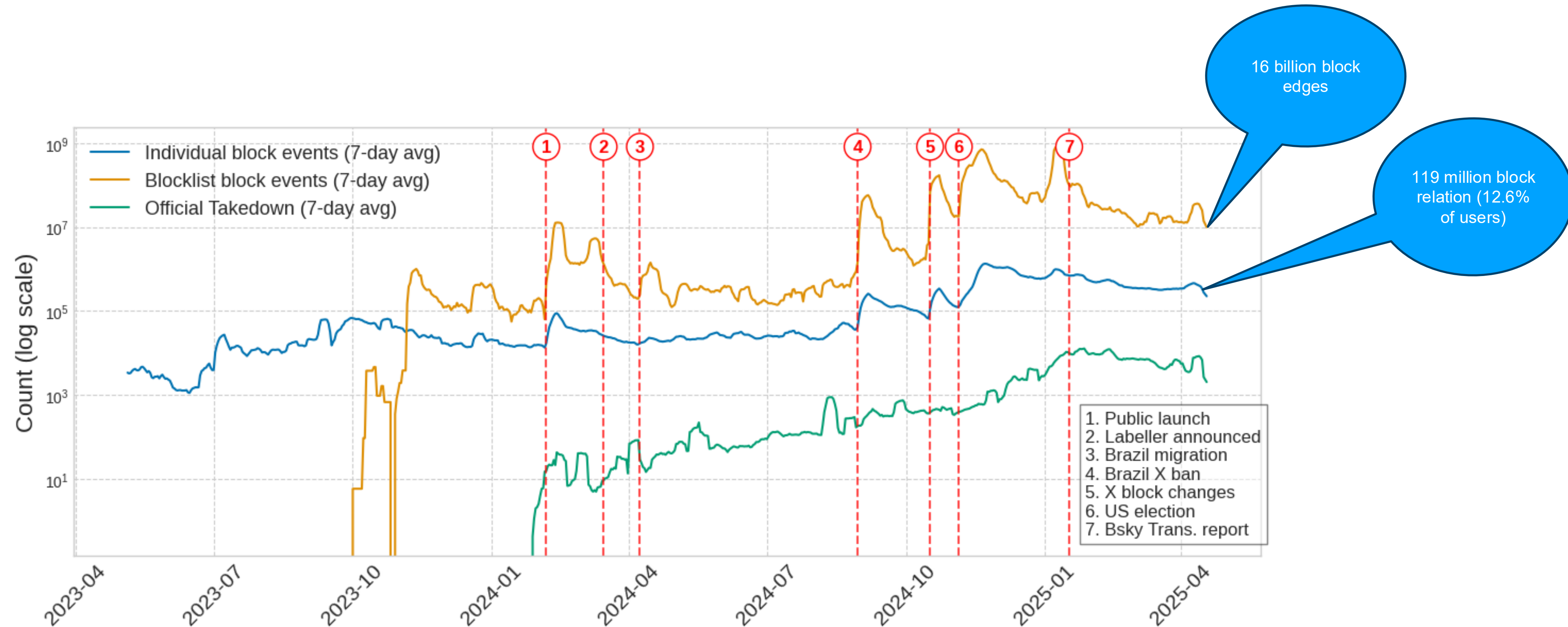
Inclusion on list for: USING - POSTING - REPOSTING - PRO AI POSTS - HARASSING CREATIVES - Includes all GenAI images /memes /music, tech and so on.

DMs ONLY FOR TAKEDOWN/SUBMISSION

Subscribe

...

Scale of Community Moderation



Dataset: 34 million users, 1.2 billion posts, and ~40K blocklists

Reach: 90% of content on the platform

Characterizing the “Blocked”

Activity: 73x more active and receive 450x more likes

Toxicity: There is a clear gradient; blocklisted users show significantly higher toxicity.

Topic bias: Politics is the most over-represented topic for blocked users (+2.44 percentage points > baseline).

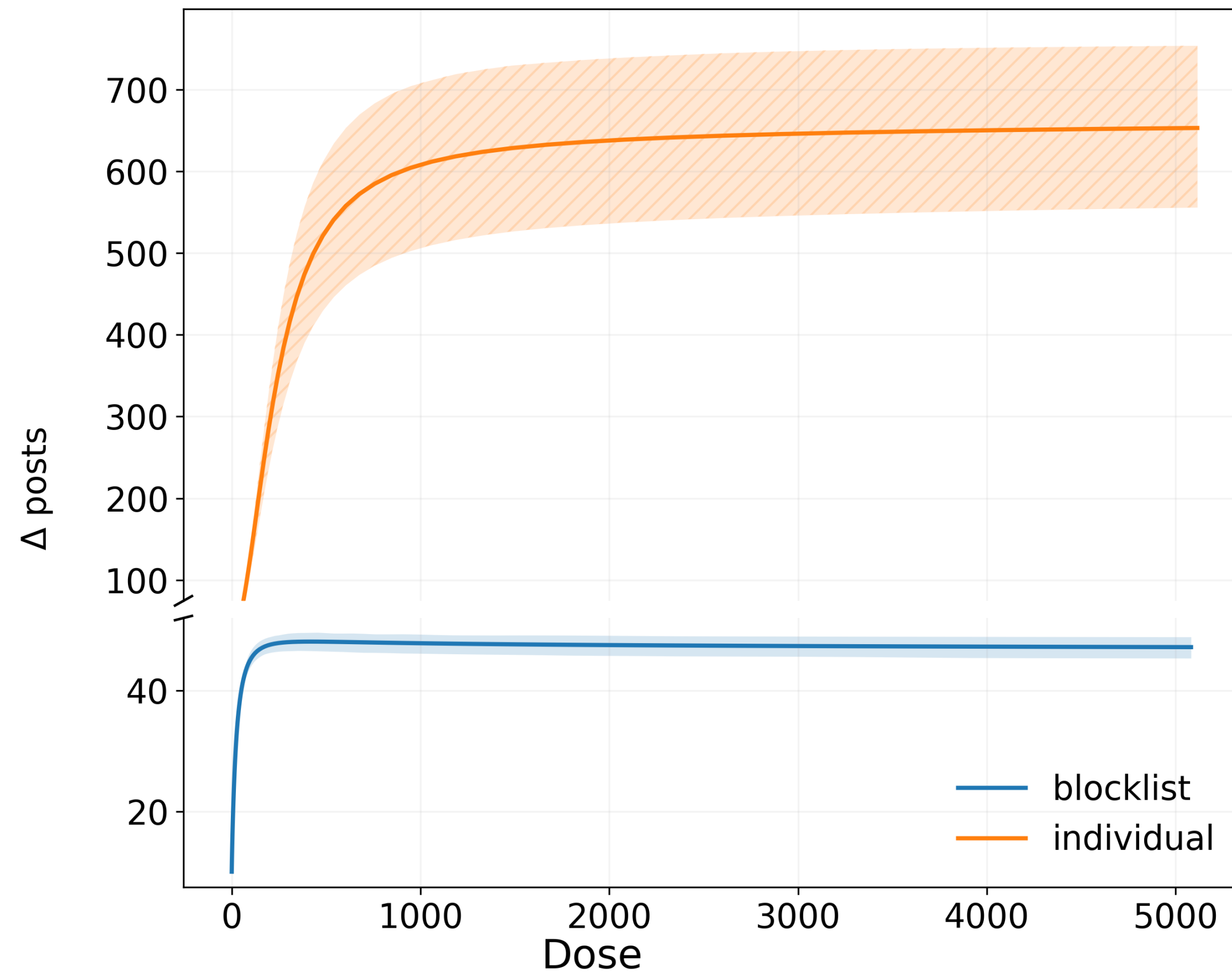
Decentralization or Re-centralization?

Ownership: Active blocklists are managed by just **0.043%** of users.

Influence: The top 100 lists cover over 66% of all subscribers created by 79 individuals.

The Risk: While the mechanism is decentralized, the governance is concentrated in a tiny group of curators, making their errors (e.g., blocking 31.7% of users who have never even posted) highly impactful.

Community Moderation Reshapes Visibility, Not Participation



Conclusion

1. Blocking is widespread and shapes 90% of content visibility.
2. It targets the most active and polarizing users.
3. It doesn't "silence".
4. A small set of influential curators to exert influence.

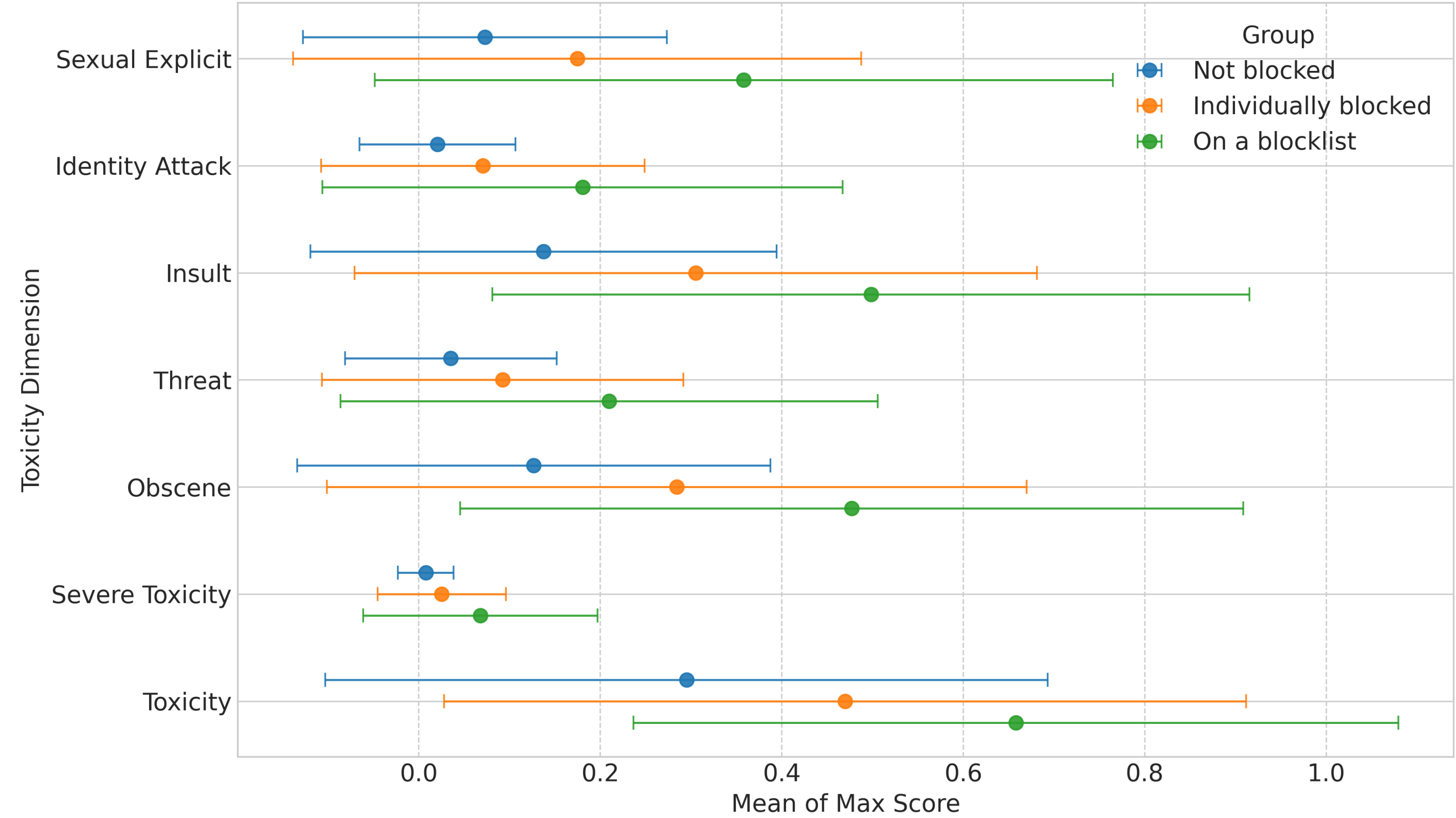


Open Questions

- 1) Does this “curate-your-own-reality” model effectively solve the moderation problem, or is it just automating segregation?
- 2) Are there different interpretations for why blocked users become *more* active?

Characterizing the “Blocked”

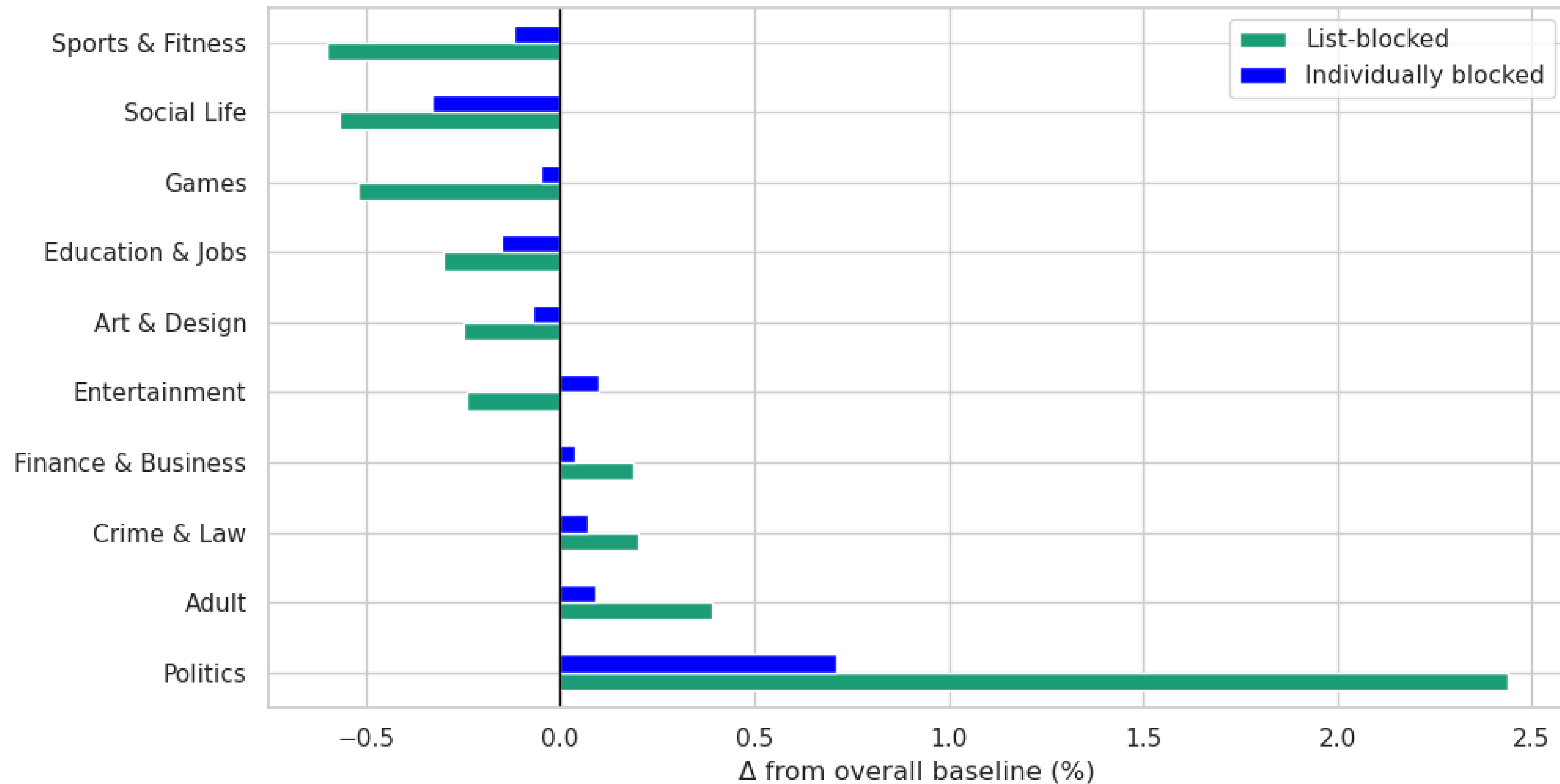
Toxicity



Characterizing the “Blocked”

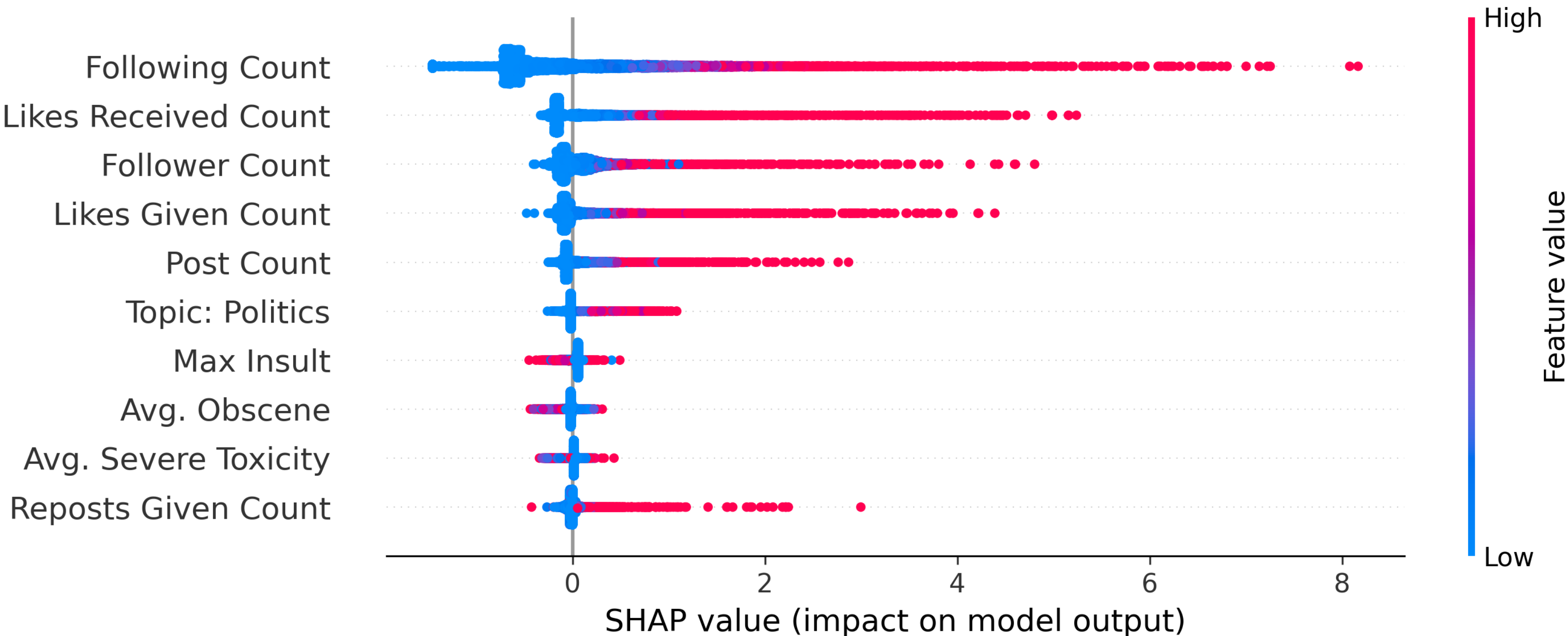
Topic bias

- For politics (+2.44 percentage points > baseline).



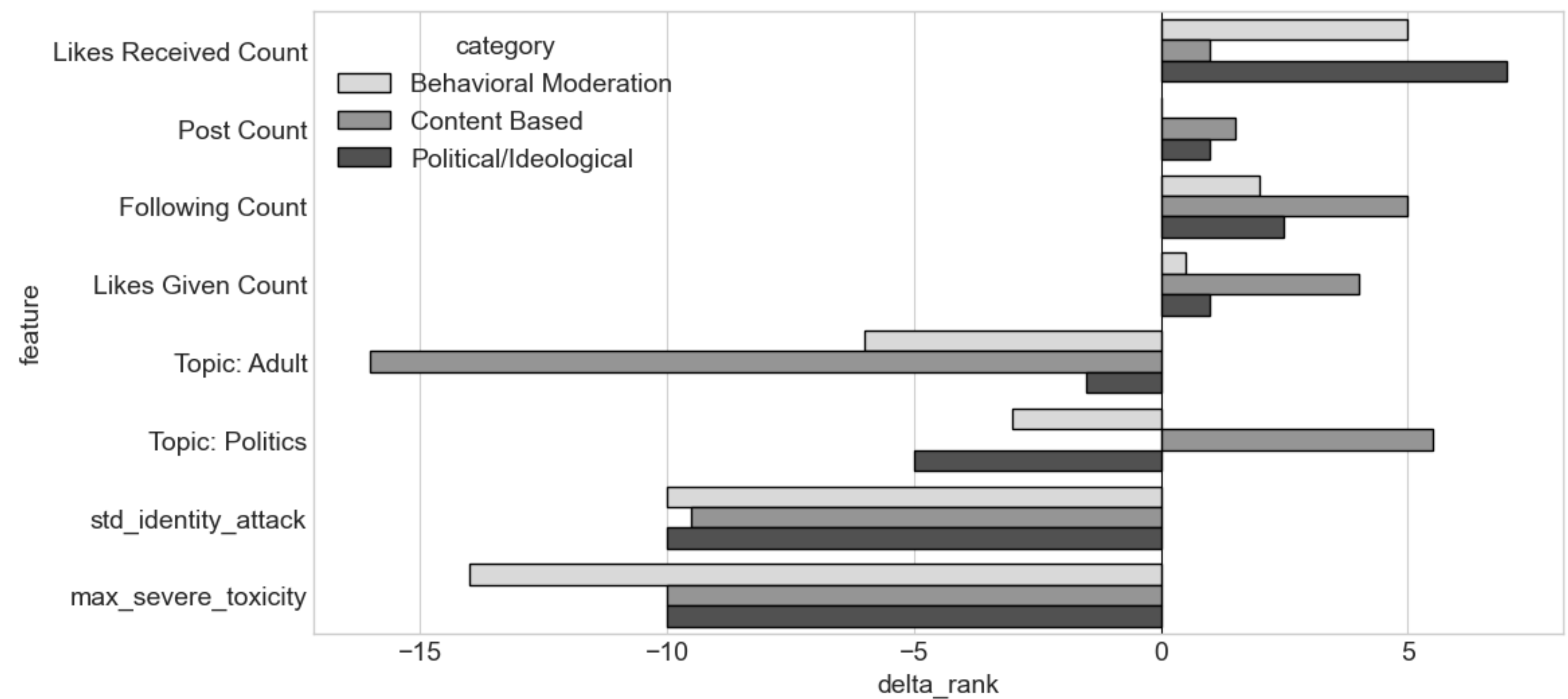
Predicting Blocking

General classifier



Predicting Blocking

Feature ranking



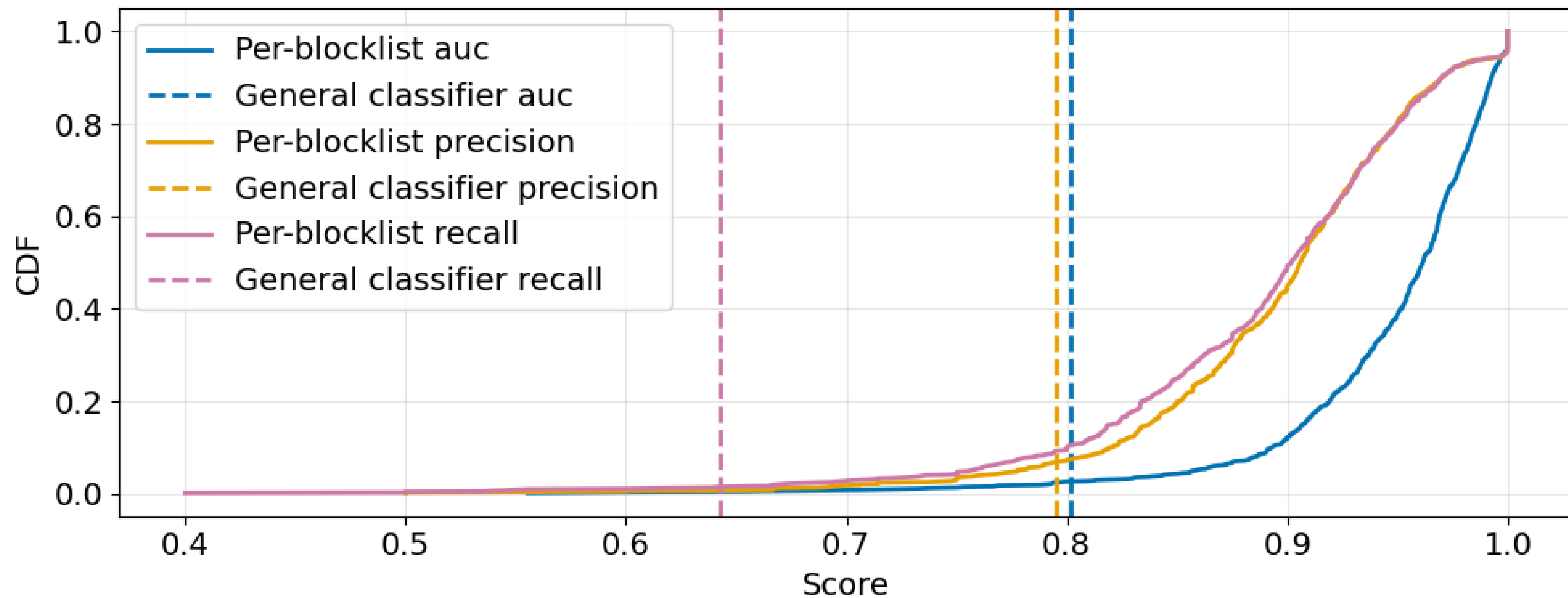
Nearest Neighbor Matching

TABLE D.1: Differences between blocked users and non-blocked users.

Outcome (SMD, SD units)	Indv. Blocked	On Blocklist
<i>Not matched on</i>		
Avg. toxicity	0.108	0.155
Avg. insult	0.124	0.179
Avg. obscene	0.099	0.140
Avg. severe toxicity	0.073	0.113
Avg. identity attack	0.072	0.113
Avg. sexual explicit	0.096	0.162
Political topic share	0.163	0.224
Left stance probability	0.077	0.094
Right stance probability	0.060	0.096
Neutral stance probability	−0.055	−0.110
Matched treated users	405,569 (81.1%)	311,218 (62.2%)
Mean match distance	0.184	0.226

Predicting Blocking

General vs. Per Blocklist classifier



Causality Analysis

