Are shadowbans effective in checking the spread of misinformation?

Kokil Jaidka[1], Svetlana Churina[1], Subhayan Mukerjee[1], Yphtach Lelkes[2]
[1]National University of Singapore, [2]University of Pennsylvania

Work in Progress

Introduction

Objective: This study investigates *shadowbans*, a form of algorithmic censorship on Twitter, through a large-scale multi-year audit of over 25,000 U.S.-based accounts to examine *who* gets shadowbanned and *why*.

Dataset and Method

Sample: We sampled 30 Twitter users from each of 1,607 U.S. counties, using geotagged tweets from the 1% Twitter firehose.

Feature Extraction through NLP APIs:

a) Profile features: account age, verified status, bot-likeness
b) Content features: offensiveness, political hashtags
c) Social influence: follower count, engagement metrics

Validity: Six shadowban audits conducted between June 2020 and June 2021; **Another is in progress in April'25**

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
June 12, 2020	June 19, 2020	June 29, 2020	July 13, 2020	July 24, 2020	June 9, 2021

We queried Shadowban.eu web service with each username to test whether they were facing one of four kinds of shadowbans:

Shadowban Type	Description
Search Suggestion	Account does not appear in search suggestions
Search Ban	Tweets are hidden from all search results
Ghost Ban	Replies are invisible to others
Reply Downtiering	Replies are hidden behind 'Show more replies'

Data Processing

Initial sample	Existing users	Banned users	Users that tweeted
10,107 users in Run 1 to 4, 50,000 users in Run 5 and 6	35,000 users existed out of the sample of users tested (other accounts have been suspended or deleted)	6.2% of the accounts have been banned at least once	28,925 posted a tweet in the ten days before the shadowban data collection

Results

Number of shadowbans detected

Dataset statistics: The number of accounts that incurred at least one shadowban across multiple runs. The numbers in brackets indicate the number of bans applied to 'active accounts' who had tweeted in the ten-day period preceding the shadowban check. (TWITTER, TWEET, RETWEET and the Twitter Bird logo are trademarks of Twitter Inc. or its affiliates).

Run	No. of checks	Existing users	Ideology available	Active users	Tweets	Search suggestion ban	Search ban	Ghost ban	Downtiered replies
1	10,107	7,836	6,885	6,539	371,175	16 (15)	30 (24)	5 (5)	216 (200)
2	10,107	7,657	6,716	6,422	414,468	17 (16)	26 (24)	5 (4)	209 (196)
3	10,107	7,680	6,803	6,579	504,002	26 (21)	36 (31)	7 (6)	204 (199)
4	10,107	7,667	6,746	6,573	584,140	18 (17)	28 (25)	6 (6)	243 (235)
5	50,000	38,921	26,897	26,338	1,802,404	216 (209)	117 (116)	23 (23)	1028 (980)
6	50,000	34,514	24,829	20,404	809,450	519 (226)	246 (86)	1,684 (288)	1,168 (687)

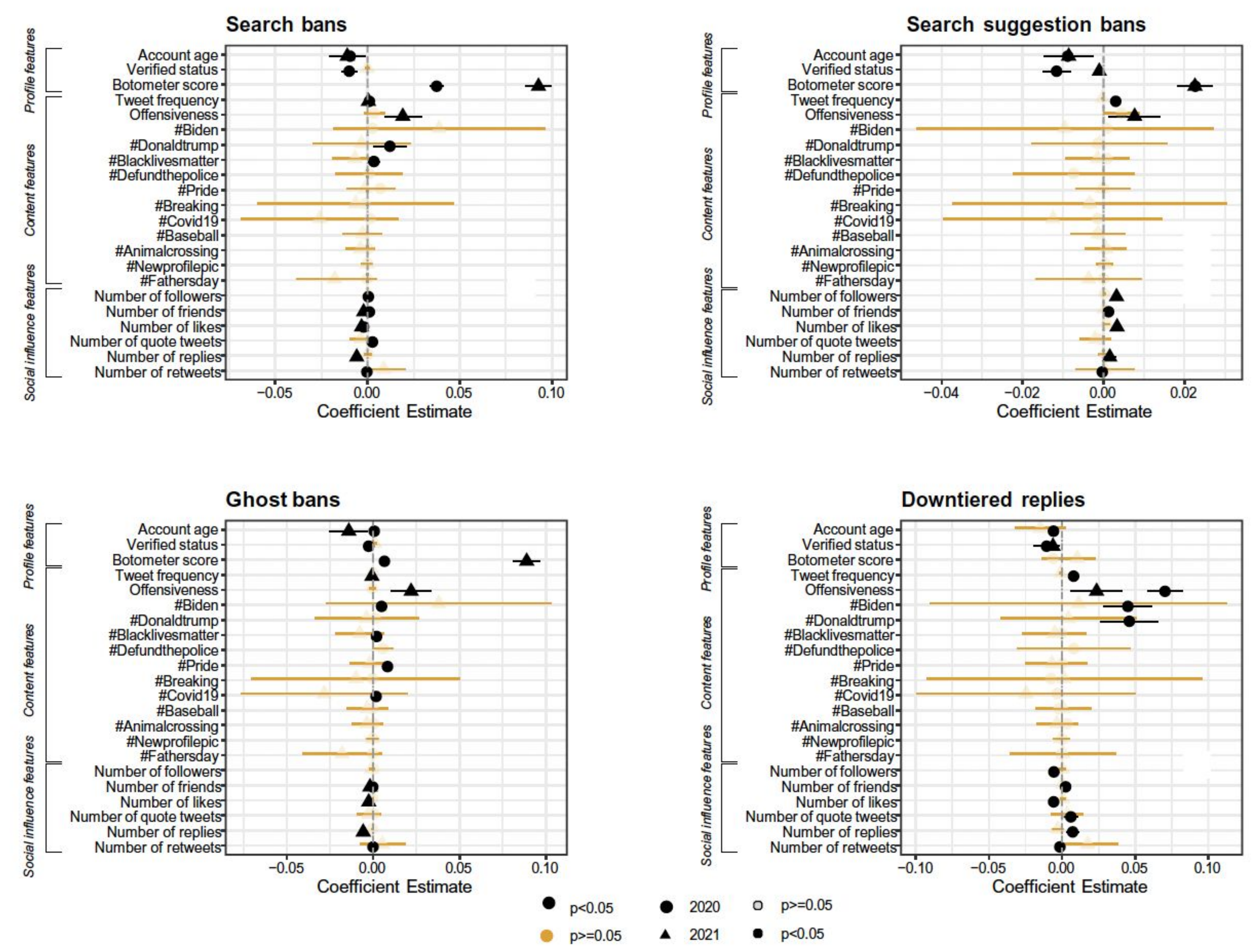

Figure 1. *Effect sizes of the rescaled independent variables on whether the accounts are shadowbanned or not, using ridge regression. The effects are reported as a percentage (x 10^2)*

Probing the Effects on Misinformation

Table 2. *We cross-referenced the URLs (5M+ links) and mentions (530K+ handles) in shadowbanned tweets with credibility labels (Robertson et al., 2018; Mukerjee et al., 2022). The Table shows the most frequent domains and users rated controversial and credible.*

Domains rated controversial	% of total	Domains rated credible	% of total
worldstarhiphop	1.84%	ktla	6.10%
townhall	1.63%	medium	5.12%
gellerreport	0.39%	time	4.51%
thefederalist	0.33%	hightimes	4.26%
mrctv	0.26%	thehill	3.26%
bipartisanreport	0.19%	buffalonews	2.93%
realfarmacy	0.18%	mashable	2.82%
bigleaguepolitics	0.16%	ew	2.72%
truepundit	0.16%	nymag	2.21%
bongino	0.12%	techcrunch	2.19%

Users rated controversial	% of total	Users rated credible	% of total
nowthisnews	1.15%	nytimes	3.30%
theblaze	0.12%	ap	1.14%
washtimes	0.10%	forbes	1.13%
rt_com	0.10%	jaketapper	0.85%
foxnewssunday	0.07%	mercnews	0.82%
oann	0.06%	taylorswift13	0.76%
americanewsroom	0.06%	reuters	0.72%
thefive	0.05%	thr	0.66%
rt_america	0.05%	thebuffalonews	0.65%
dailymailuk	0.04%	variety	0.60%

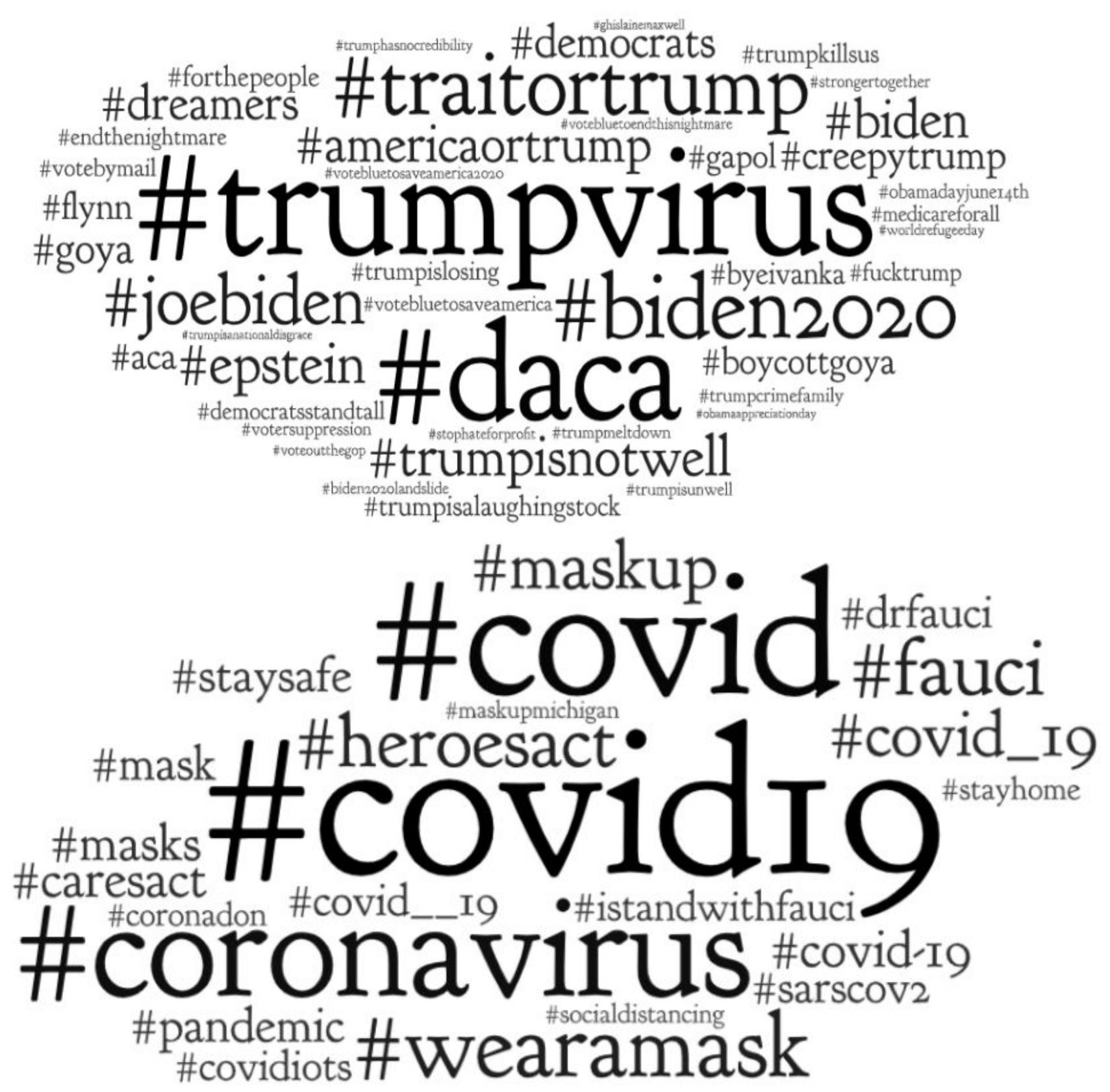

Figure 2. *Hashtag topics are semantically similar hashtags that at least 250 accounts tweeted. The size reflects the frequency of the hashtag in the corpus.*

Key Insights

Shadowbanned users and tweet characteristics

a) Profile features:
- Users exhibiting "botlike" behavior, high tweet frequency, uncivil posts are more likely to be banned

b) Social influence:
- Users with more retweets are less likely to receive search suggestion bans
- Users with more likes are more likely to receive search suggestion bans

Temporal instability

Hashtags triggering bans in 2020 (e.g., #Pride, #BLM) were not banned in 2021

Misinformation

Tweets frequently cite credible news outlets like New York Times and Associated Press, and weekly magazines like Time and New Yorker. Tweets by shadowbanned users are expected to cite controversial accounts such as Now This News, The Blaze, and media sites such as Town Hall and Geller Report.

Conclusion and Future Work

While platform norms shift over time, moderation decisions remain opaque and unevenly enforced.

In analyses with the Twitter Community Notes dataset, we found that verified accounts posted 12% (26k) of misleading tweets.

These accounts are less likely to be shadowbanned, making it easier for misinformation to spread widely.

References

[1] Mukerjee, S., Jaidka, K., & Lelkes, Y. (2022). The political landscape of the US Twitterverse. *Political Communication, 39*(5), 565-588.
[2] Just, N., & Latzer, M. (2017). Governance by algorithms: Reality construction by algorithmic selection on the internet. *Media, Culture & Society, 39*(2), 238–258.
[3] Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within google search. In *Proceedings of the ACM on human-computer interaction, 2*(CSCW), 1-22.