EasyChair Preprint

№ 6003

An audit of Twitter's shadowban sanctions in the United States

Kokil Jaidka, Subhayan Mukerjee and Yphtach Lelkes

July 3, 2021

7^{th} International Conference on Computational Social Science IC^{2}S^2
July 27-31, 2021, ETH Zürich, Switzerland

An audit of Twitter's shadowban sanctions in the United States

Kokil Jaidka[1], Subhayan Mukerjee[1], and Yphtach Lelkes[2]

[1]National University of Singapore
[2]University of Pennsylvania

Keywords: Political Communication, Social Media, Policy Discussion

Extended Abstract

Concerns have been raised about the algorithms that curate content on social media platforms, which are designed to show users the perspectives that they agree with, cutting us off from information and viewpoints that they don't. Twitter has also adopted a strategy of censorship in their efforts to make their platform more civil, and curb the propagation of misinformation. Twitter blocks or partially blocks a user or their content from the rest of the community. This practice has often led to accusations, particularly from the political right in the United States, of covertly sanctioning controversial accounts (Menegus & McKay, 2018) and right-leaning partisan users, such as senate representatives (Thompson, 2018). This has led to a slew of media coverage and preliminary audits to test (Fosse, n.d.), support (Eordogh, 2018), refute (Feldman, 2018) and criticize (Shadowban.eu, 2020) these claims. In response, Twitter has issued a clarification asserting that any evidence of shadow banning is solely an artifact of their ranking algorithms (Gadde & Beykpour, 2018). While audit studies exist that seek to understand the censorship of social media platforms by states (e.g., by Turkey (Tanash et al., 2015) and China (King et al., 2014)), there is no systematic evidence regarding Twitter's shadow bans.

In this study, we examined the linguistic factors predicting shadow bans on a dataset of over 40,000 Twitter accounts, fetched in March 2020, comprising a stratified random sample of users based on their FIPS codes. To test whether an account is shadow banned, we sent queries to the shadow ban.eu [1] web-service that checks whether a public account on Twitter is facing a (a) **Search ban**, i.e. it doesn't show up in Twitter search; (b) **Search suggestion ban**, i.e. it doesn't show up in the search suggestions while someone is searching for it. (c) **Ghost ban** i.e. replies or tweets posted by the affected user in a thread don't show up for others, or (d) **Reply deboost** i.e. replies by the affected account don't show up naturally for others; they only show up when others click on "show more replies". We coded an account to be shadow banned, if it had at least one of the above restrictions imposed on it, and not, otherwise. Data was repeatedly collected over six different runs during a one-year period. We found that 6.2% of the 41,092 existing accounts in our dataset had been shadow banned at least once during the study period.

Our independent variables included: the level of incivility in the account's tweets (estimated using state-of-the-art natural language classifiers (Davidson et al., 2017)), the likelihood of the account being a bot (using the Botometer API (Yang et al., 2020)), and the account's network size (extracted using the Twitter API). All independent variables were rescaled to [0,1]. We fit

[1]https://shadowban.eu/

fixed effects linear models to our dataset, grouping the accounts based on the time when their shadow ban statuses were fetched, and estimated robust standard errors for the effect sizes. Table 1 reports the results for a general set of users over five runs in 2020 and a validation run in 2021. The Table shows that predictors of the likelihood of shadowbanning among the accounts are incivility, botlike behavior and tweet frequency in the ten days before the shadowban data was collected. Discussion about politics is associated with deboosted replies, but has the opposite effect in search suggestion listing. Many of these results were consistent in a repeated analysis in 2021, as seen in the figures in bold.

To conclude, the algorithms we have used to predict linguistic features and botlike behavior have been evaluated and applied in many recent studies to examine the characteristics of Twitter users. Details on their feature engineering, and a list of the model features and their coefficients have been published by the original authors (Davidson et al., 2017; Yang et al., 2020). However, Twitter has provided no such information about the algorithms underlying its shadow ban sanctions. We recommend that Twitter could follow suit and publish the technical features of its algorithms to engender greater trust amongst its users and advance the cause of open science.

Table 1: Effect sizes of the rescaled independent variables on whether the accounts are shadow banned or not. Estimates are based on fixed effect linear models with clustered standard errors. Figures in bold are also statistically significant in the validation analysis in 2021.

	Dependent variable:			
	Search suggestion ban	Search ban	Ghost ban	Reply deboost
	(1)	(2)	(3)	(4)
Incivility	-0.191	-0.909**	-2.944**	0.641
	(0.529)	(0.403)	(1.280)	(0.403)
Botometer score	**3.144*****	**3.241*****	**2.948*****	-0.699***
	(0.295)	(0.231)	(0.665)	(0.231)
Verified user	**-14.510*****	-12.991***	-14.043***	**-13.999*****
	(0.347)	(0.249)	(0.471)	(0.249)
Number of followers	0.228**	0.089	0.226*	-0.219***
	(0.097)	(0.061)	(0.121)	(0.061)
Number of friends	0.152	0.120*	**-0.332****	0.242***
	(0.104)	(0.069)	(0.139)	(0.069)
Number of retweets	**-0.00004****	0.00000	-0.00000	-0.00000**
	(0.00002)	(0.00000)	(0.00001)	(0.00000)
Number of quotes	-0.520	-1.515	1.120***	0.049
	(0.358)	(1.807)	(0.383)	(1.807)
Number of replies	-0.186*	-0.827	-2.244	-0.019
	(0.102)	(0.532)	(1.873)	(0.532)
Number of likes	**0.006****	-0.012	-0.161	-0.0003
	(0.002)	(0.049)	(0.139)	(0.049)
Tweet frequency	**0.002*****	0.001***	-0.001	0.001***
	(0.0002)	(0.0002)	(0.001)	(0.0002)
Account age	**-0.001*****	-0.0004***	0.0002	-0.0001
	(0.0001)	(0.0001)	(0.0002)	(0.0001)
Politics	-14.058**	-5.486	-24.646	15.614***
	(6.512)	(3.959)	(15.935)	(3.959)

Note: *p<0.1; **p<0.05; ***p<0.01

References

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 512–515.

Eordogh, F. (2018). Why repulicans weren't the only ones shadow banned on twitter. *Forbes*. https://www.forbes.com/sites/fruzsinaeordogh/2018/07/31/why-republicans-werent-the-only-ones-shadow-banned-on-twitter/?sh=2c031dc5434b

Feldman, B. (2018). Twitter is not 'shadow banning' republicans. *The Intelligencer*. https://nymag.com/intelligencer/2018/07/twitter-is-not-shadow-banning-republicans.html

Fosse, K. (n.d.). Twitther shadowban test. *Shadowban.eu Blog*. https://shadowban.eu/

Gadde, V., & Beykpour, K. (2018). Setting the record straight on shadow banning. *Twitter Blog*. https://blog.twitter.com/official/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in china: Randomized experimentation and participant observation. *Science*, *345*(6199).

Menegus, B., & McKay, T. (2018). Twitter may be demoting controversial accounts in search results. *Gizmodo*. https://gizmodo.com/twitter-may-be-demoting-controversial-accounts-in-searc-1827788070

Shadowban.eu. (2020). Twitter thread. *Twitter*. https://threadreaderapp.com/thread/1316544057691140096.html

Tanash, R. S., Chen, Z., Thakur, T., Wallach, D. S., & Subramanian, D. (2015). Known unknowns: An analysis of twitter censorship in turkey. *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, 11–20.

Thompson, A. (2018). Twitter is shadow banning prominent republicans like the rnc chair and trump jr's spokesman. *Vice News*. https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman

Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1096–1103.