The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion

Karsten Müller* Carlo Schwarz†

December 7, 2022

We study the effects of online content moderation on user behavior in the context of a prominent case study: the deletion of President Donald Trump's Twitter account on January 8th, 2021. We provide four key findings. First, the toxicity of tweets sent by Trump followers relative to a representative sample of US Twitter users dropped by around 25% after the account deletion. Second, this effect is larger for pro-Trump tweets and Republican users. Third, Trump's suspension reduced the total number of tweets, suggesting a drop in engagement. Fourth, we find effects on individuals who did not follow Trump directly but followed somebody that did, suggesting network spillovers. This evidence suggests that removing a prominent, polarizing individual can reduce the toxicity of online discourse.

*National University of Singapore, NUS Business School, kmueller@nus.edu.sg.

†Università Bocconi, Department of Economics and IGIER, and PERICLES, CEPR, CAGE, carlo.schwarz@unibocconi.it.

1 Introduction

There is widespread concern about the proliferation of toxic content on social media. Because hateful messaging on social media has been linked to offline harm (*1, 2, 3*), addressing it is a key priority for social networks.[1] Social media companies like Facebook and Twitter are grappling with the contentious issue of how to deal with such content. In October 2018, for example, Facebook removed several accounts associated with the far-right group Proud Boys in an effort to "take action against hate speech and hate organizations to help keep our community safe" (*5*). However, there is relatively little evidence on the effects of such content moderation efforts on the civility of online dialogue. Such evidence is an important input for both public policy and platform providers to assess the desirability and effectiveness of content moderation.

This paper uses a particularly salient case study to investigate the effects of online content moderation on user behavior: the deletion of then-President Donald Trump's Twitter account in the aftermath of the January 6th, 2021 Capitol attack. Twitter suspended Trump's account on January 8th "due to the risk of further incitement of violence" (*6*).[2] The widely-reported account deletion provides us with a natural experiment to test whether removing a prominent, highly influential user who spreads polarizing content can affect the civility of online discourse.

To test this conjecture, we use a difference-in-differences methodology that compares the toxicity of tweets posted by people who followed Trump before his account was suspended with those of Twitter users who did not. This analysis is made possible by a unique dataset we collected during 2020 and 2021 that combines all tweets and the profile information of a representative sample of Twitter users with information on who followed @realDonaldTrump

[1] In 2021, for example, Twitter CEO Jack Dorsey stated that "[o]ffline harm as a result of online speech is demonstrably real, and what drives our policy and enforcement above all" (*4*).

[2] While the suspension was originally said to be permanent, Trump's account was reinstated under new CEO Elon Musk on November 19, 2022 (*7*).

before the account was removed. This dataset allows us to estimate how Trump's account deletion affected the behavior of users that followed him vis-à-vis other users.

Our main analysis studies how Trump's account suspension influenced the online behavior of his followers. The difference-in-differences methodology we use abstracts from differences in the average toxicity of users across time. We find that the number of toxic tweets sent by Trump followers relative to other Twitter users declined by around 22-26% following the account deletion. Consistent with a causal effect of the policy, this drop in toxicity occurs precisely after the account deletion, and we find no differences in the trends of toxic tweets sent by Trump followers compared to other users before. Using the 2016 election as a placebo test, we find no similar change in toxicity of the tweets sent by Trump followers. This makes it unlikely that we are picking up electoral dynamics on Twitter. The evidence is most consistent with the interpretation that deleting a central, widely-followed actor from a social media platform can have a permanent causal effect on the toxicity of online discourse.

The account deletion also had a broader effect on online discourse on Twitter. We find that it reduced the total number of tweets sent by Trump followers, although less than the number of toxic tweets, leading to a decrease in overall toxicity. We also document an effect on the number of tweets mentioning Trump and topics related to his campaign and policies. Importantly, we also find spillovers through his follower network. While the change in user behavior after Trump's account suspension is largest among his followers, we also find a knock-on effect on people who did not *directly* follow him but followed someone that did. This effect on second-degree linkages highlights the importance of network effects in propagating hateful sentiments online. It also suggests that content moderation may have wide-reaching effects on such sentiments.

Several pieces of additional evidence support the idea that deleting Trump's account reduced hateful tweets sent by people sympathetic to his views. First, the effect on toxicity is predominantly driven by accounts we classify as Republicans based on other political accounts

they follow; we find a smaller effect on independents, and no effect on Democrats. This is despite the fact that Republicans make up a relatively small share of Twitter users, a finding we confirm in our data (*8*). Second, the reduction in toxicity is larger for tweets classified as "pro-Trump" based on the hashtags they use. As such, the evidence suggests that removing a prominent, polarizing individual reduces the toxicity of online discourse throughout a social network.

Existing evidence on online content moderation is limited. There is experimental evidence that hiding toxic content with a browser extension reduces engagement with social media platforms (*9*), and that reporting posts as hateful does not affect the behavior of the posters (*10*). Evidence from a "natural experiment" suggests that a law mandating social media platforms to take down hate speech in Germany reduced online toxicity and real-life hate crimes (*11*). Deleting hateful sub-Reddits appears to have affected the posting behavior of users active there (*12*). We add to this existing work by using the unique case study of Trump's account deletion to estimate the effect of online content moderation on toxicity, a key parameter in theoretical models of platform decision-making (*13, 14, 10*). Our study also adds to a broader literature studying the effects of social media on political outcomes using randomized experiments (*15, 16, 17*), and natural experiments with observational data (*18, 19, 1, 2, 3, 20, 21*).

2 Data

2.1 Data Collection

We combine two datasets based on data collected from Twitter: (1) a dataset on Donald Trump's followers and (2) a representative sample of Twitter users. We outline sources and variable construction here; Appendix A contains more details; Appendix A.4. presents summary statistics.

The first dataset includes the list of all accounts that followed Trump before his account was (permanently) suspended on January 8th, 2021. These data were collected in November 2020. In total, we have information on over 80 million users. To classify the political leaning of Trump followers, we also collected data on which other accounts they follow.

This dataset is unique because it is a comprehensive snapshot of Twitter users following Trump *before* his account was suspended. The timing of the data collection is key. Even after Trump's account was reinstated by Elon Musk on November 19th, 2022, it is impossible to measure who followed Trump *before* the account was removed. Having a list of Trump's followers before the account deletion is crucial for the difference-in-differences methodology we use, because the 2022 election loss, January 6th Capitol attack, and subsequent account deletion likely also affected users' incentives to follow Trump or even leave Twitter, which could lead to a biased sample.

The second dataset includes the profiles and tweets of a representative sample of American Twitter users. We use the representative sample of US Twitter users from (*22*), which aims to be representative of the US voting population. We repeatedly collected the tweets of these accounts between 2020 and 2022. In total, we have 395 million tweets these users sent until the end of 2021.

Appendix Table A.3 provides summary statistics for the main variables in our analysis. The panel dataset used in our statistical models has 46,511 users who followed Trump at the time his account was deleted and 452,390 users who did not. Because we are mainly interested in content generated by the users themselves, we exclude retweets (including those of Trump) from the baseline analysis.[3] The average number of tweets per month is 3.92.

[3] As we show in robustness checks, this has no bearing on our results.

2.2 Measuring Toxicity

To measure the toxicity of tweets, we use the Perspective API, a machine learning-based technology for identifying toxic content in text jointly produced by Jigsaw and Google's Counter Abuse Technology team. Perspective API is widely used by commercial clients such as Reddit, The New York Times, Financial Times, or Disqus to automatically moderate online content at scale. It has also been widely used in academic research (*23, 11, 10*)

We measure the toxicity of all tweets in our estimation sample using the Perspective API. Toxicity ranges from 0 to 1. A commonly-used cutoff for identifying toxic text is a toxicity score of 0.8. To get a sense of what these numbers imply: "The Party of Lincoln Comes to Terms with Trump" has a toxicity of 0.1 (not toxic), and "@RandPaul You are the dumbest motherfucker" a toxicity of 0.99 (toxic). 8% of the tweets in our sample have a toxicity score of above 0.8 (see Appendix A.1. for more details).

2.3 Analysing the Content of Tweets

To study changes in the content of tweets with Trump's account deletion, we create two outcome measures. First, we construct variables for the number or share of tweets a user sends that contain the word "Trump." Second, we use a topic modeling approach to identify what users talk about. Topic models like Latent Dirichlet Allocation (*24*) have been widely used in social science research to extract topics from unstructured text data. Given the short text contained in one tweet, we use a Biterm Topic Model (*25*), which has been shown to perform well in short texts (see Appendix A.2. for more details).

2.4 Measuring Political Leaning

We create several measures of political alignment with Trump. First, we classify users as Democrats, Independents, or Republicans based on the Twitter accounts they follow. In particular,

we classify users as Republican if they follow more Republican than Democratic Members of Congress, and vice versa. Users following no party are classified as Independents (see Appendix A).

Second, we measure whether individual tweets are "pro-Trump" or "anti-Trump" based on the hashtags used in the tweet. To create these measures, we create a dictionary of nearly 18,000 pro-Trump and anti-Trump hashtags based on all hashtags that contain the string "trump." We then code any tweet containing a pro-Trump hashtag as pro-Trump, and any tweet containing an anti-Trump hashtag as anti-Trump(see Appendix A.3. for details).

3 Empirical Strategy

On November 3, 2020, President Trump lost the presidential election to Joe Biden against the backdrop of the COVID-19 pandemic. In the following weeks, Trump continued to use his social media channels–most prominently his Twitter account–to make increasingly unsubstantiated claims about widespread voter fraud, and refused to concede his election loss. These statements on social media were accompanied by a large-scale legal effort seeking to overturn the election results in what many have described as a coup attempt (*26*).

These events culminated in the January 6th attack on the capitol. Following a call for action, Trump had broadcasted on Twitter, Trump supporters started gathering in Washington, DC on January 5th to protest the certification of Electoral College votes scheduled for the next day. After an inflammatory speech by Trump calling Biden an "illegitimate president," he encouraged the protesters to march to the capitol. The ensuing attack has been associated with five deaths, numerous injuries, and millions in property damage. On January 8th, Twitter took the unprecedented step to suspend his account entirely, citing a violation of their policies.

Trump's account deletion creates a "natural experiment" to estimate the effect of online content moderation on user behavior. We use a difference-in-differences design to estimate the causal effect of Trump's suspension by comparing tweets sent by accounts that followed Trump before January 8th, who were plausibly more exposed to his rhetoric, compared to other Twitter users. The estimating equation is

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot Trump\ follower_i \times Account\ deletion_t + \theta \mathbf{X}_{it} + \varepsilon_{it}, \tag{1}$$

where Y_{it} is a measure of user behavior by account i in month m. $Trump\ follower_i$ is an indicator variable equal to 1 for users who followed Trump before his account was deleted, and 0 otherwise. $Account\ deletion_t$ is an indicator equal to 1 after January 2021; we also allow for an event study specification to investigate the precise timing of the effect. α_i and γ_t are a full set of user and month fixed effects, which abstract from average differences in toxicity across users and time. $\mathbf{X}_{it}$ is a set of additional control variables, like user-specific linear or cubic time trends, or additional fixed effects.

In our baseline estimation, Y_{it} is the number of toxic tweets sent by a user in a given month, where tweets are defined as toxic if any of the Perspective API toxicity scores is above 0.8. This dependent variable captures the arguably most policy-relevant issue: the total amount of toxic content online.

To interpret β in Equation (1) as the causal effect of Trump's account deletion, we require the assumption that Twitter user behavior would have followed similar trends to the preceding months in absence of the account deletion. While this assumption is inherently untestable, we provide two pieces of evidence in its support.

First, we investigate the link between user behavior and Trump followers in an event study specification. This allows us to see whether the toxicity of Trump followers compared to other Twitter users followed similar trends before the account suspension. The results of this exercise

show that there were no differential trends in outcomes before the deletion, making it less likely that the trends would have diverged in absence of the deletion.

Second, we conduct a placebo test using the 2016 presidential election. This addresses the concern that some changes in Twitter activity between Trump followers and others may be driven by the election cycle itself rather than the account deletion. To address this concern, we re-estimate our regressions in the exact same time window and the same sample of users around the 2016 presidential election. Strikingly, we find no similar pattern, suggesting that our findings are not driven by the election itself. In an alternative specification, we also consider a triple-difference design that explicitly compares the changes in the number of toxic tweets for the same person around the 2016 and 2020 elections:

$$\begin{aligned} Y_{it} = \alpha_i + \gamma_t + \beta \cdot Trump\ follower_i \times Account\ deletion_t \times \mathbf{I}_t^{2021} \\ + \delta_1 Trump\ follower_i \times \mathbf{I}_t^{2021} + \delta_2 Trump\ follower_i \times Account\ deletion_t \\ + \theta \mathbf{X}_{it} + \varepsilon_{it}, \end{aligned} \tag{2}$$

where the sample is extended to include both the 2020-21 and 2016-17 period. The coefficient β in Equation (2) picks up the additional difference between these two periods on top of the baseline differences in outcomes captured by the δ_1 and δ_2 parameters. In Table A.4, we provide summary statistics for the differences in the Twitter behavior between Trump followers and the representative sample of Twitter users before the account deletion.

4 Results

Effects on toxicity Figure 1 presents event study specifications of Equation (1) with the number of toxic tweets as outcome variable. Panel A shows there were no systematic differences in toxicity between Trump followers and other Twitter users before the account deletion in January.

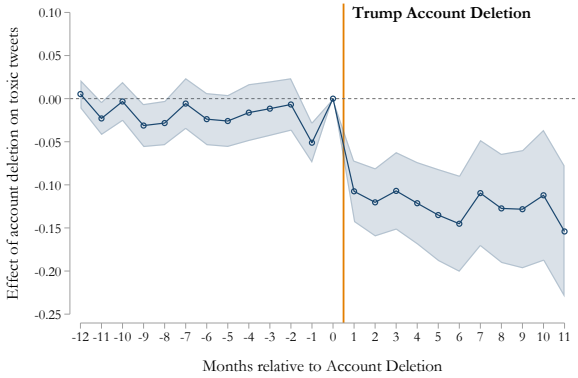
Following the suspension, the number of toxic tweets dropped immediately and permanently, consistent with a causal effect of the policy. Panel B shows a placebo test based on the exact same time frame four years earlier following the 2016 election. We find no similar reduction in toxic tweets, suggesting that the 2021 account deletion indeed reduced hateful posts by Trump followers. Panel C estimates the triple-difference specification in Equation (2) that effectively takes the difference between the estimates in Panel A and B. Again, we find no pre-existing trends in toxic tweets before Trump's account suspension and an immediate and permanent drop afterwards.

Table 1 presents the corresponding difference-in-differences estimates. Depending on the set of control variables, we find point estimates between -0.040 and -0.048 ($p < 0.01$). These imply a reduction in the number of toxic tweets by 22-26% relative to the mean. The effect we find is robust to the inclusion of $User \times Month - of - year$ fixed effects in column 2 and 4. These specifications effectively compare the number of toxic tweets sent by the *same user* in the *same month* in 2021 relative to 2020, depending on whether they followed Trump on Twitter or not. As such, it rules out alternative explanations of our findings, such as differences in user behavior in different months of the year. The triple-difference estimate of -0.051 ($p < 0.01$) in column 5 suggests a 31% reduction in the number of toxic tweets.

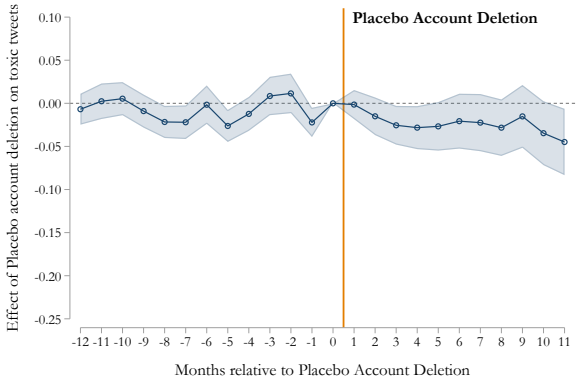
Effects on other outcomes Figure 2 reports estimates for β in Equation (1) using a range of additional outcomes. The point estimates are scaled over the standard deviation of the dependent variable to make them comparable. Appendix Table B.1 contains the unscaled regression estimates and also reports the results from triple-difference regressions as in Equation (2). Trump's account suspension reduced the number of tweets sent by his followers by 21% ($p < 0.01$). We also find a 160% reduction of tweets mentioning "Trump" relative to the mean ($p < 0.01$) as well as a 185% drop in tweets mentioning topics associated with him ($p < 0.01$).

Figure 1: The Effect of Trump's Account Deletion on Toxicity

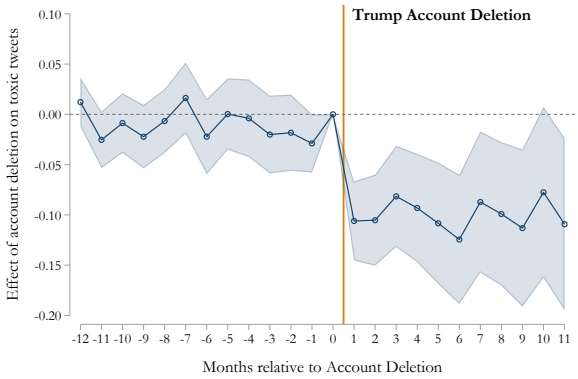
(a) Account Deletion



(b) Placebo Deletion



(c) Triple Difference



Notes: Panels A and B plot the estimates of an event study version of Equation (1). Panel C plots the point estimates of Equation (2). We plot 99% confidence intervals based on standard errors clustered by user.

Table 1: Estimates of Trump Account Deletion on Toxicity

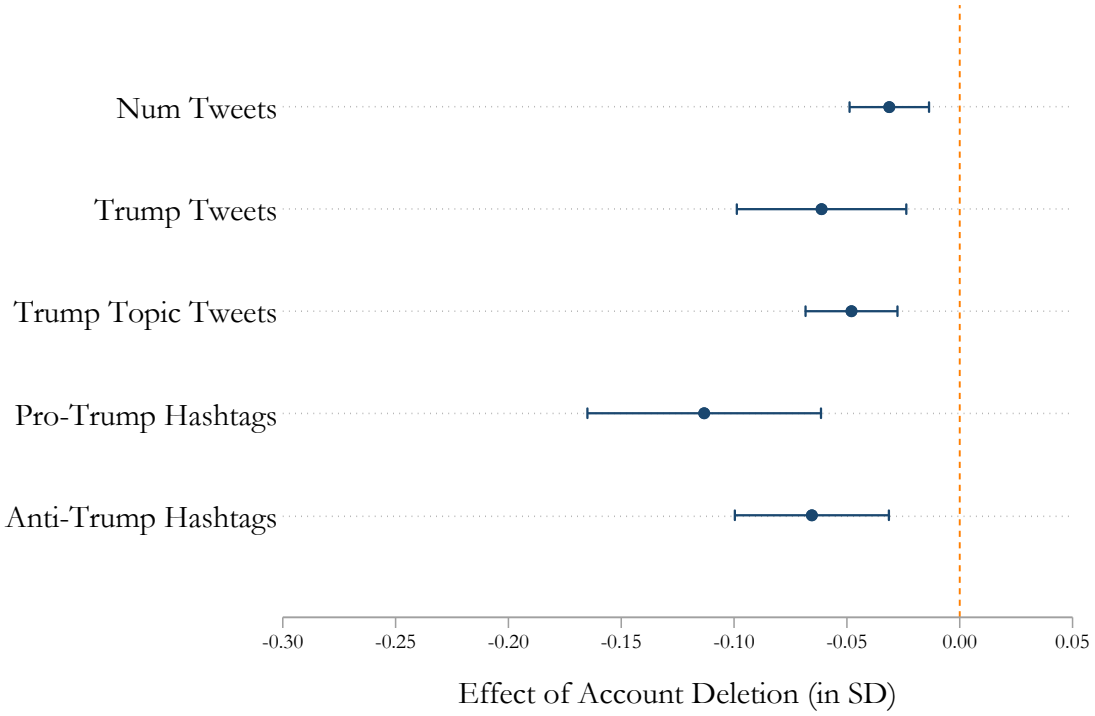
	Diff-in-Diff				DDD
	(1)	(2)	(3)	(4)	(5)
Trump Follower × Account Deletion	-0.048***	-0.041***	-0.046***	-0.040***	
	(0.011)	(0.014)	(0.011)	(0.014)	
Trump Follower × Account Deletion × I[2020]					-0.051***
					(0.013)
User FE	Yes	Yes	Yes	Yes	
Month FE	Yes	Yes	Yes	Yes	
User Linear Time Trend	Yes	Yes	Yes	Yes	
User × Month of Year FE		Yes		Yes	
User Quadratic Time Trend			Yes	Yes	
User × Election FE					Yes
User × Treatment Time FE					Yes
Treatment Time × Election FE					Yes
User × Election Linear Time Trend					Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	24,945,050
Pre-Period Mean of DV	0.183	0.183	0.183	0.183	0.163
R^2	0.52	0.72	0.59	0.72	0.73

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

We also look at heterogeneous effects depending on the alignment of tweets with Trump. The account suspension reduced the number of pro-Trump hashtags by a factor of 7 relative to the mean ($p < 0.01$). The point estimates for anti-Trump hashtags suggest a smaller negative effect of 250% ($p < 0.01$).

Effects by political alignment Table 2 tests for heterogeneous effects by the topic of tweets and a user's party affiliation. In columns 1 and 2, we find reductions in the number of toxic tweets both for tweets that do and do not mention Trump. In columns 3-5, we show that the effect of the account deletion on toxicity is most pronounced for Republicans, where we find a reduction of 34% relative to the mean ($p < 0.05$). The effect is smaller for Independents, which see a 25% drop in toxicity ($p < 0.05$). We find a positive but far from statistically significant effect on Democrats. Taken together, this evidence suggests that the drop in toxicity following

Figure 2: Impact of Account Deletion on Other Outcomes



Notes: This figure plots the point estimates for β based on Equation (1) for different outcome variables Y. We plot 99% confidence intervals based on standard errors clustered by user.

Trump's account deletion is explained by individuals sympathetic to his views, rather than his critics.

Table 2: Heterogeneous Effects of Trump's Account Deletion

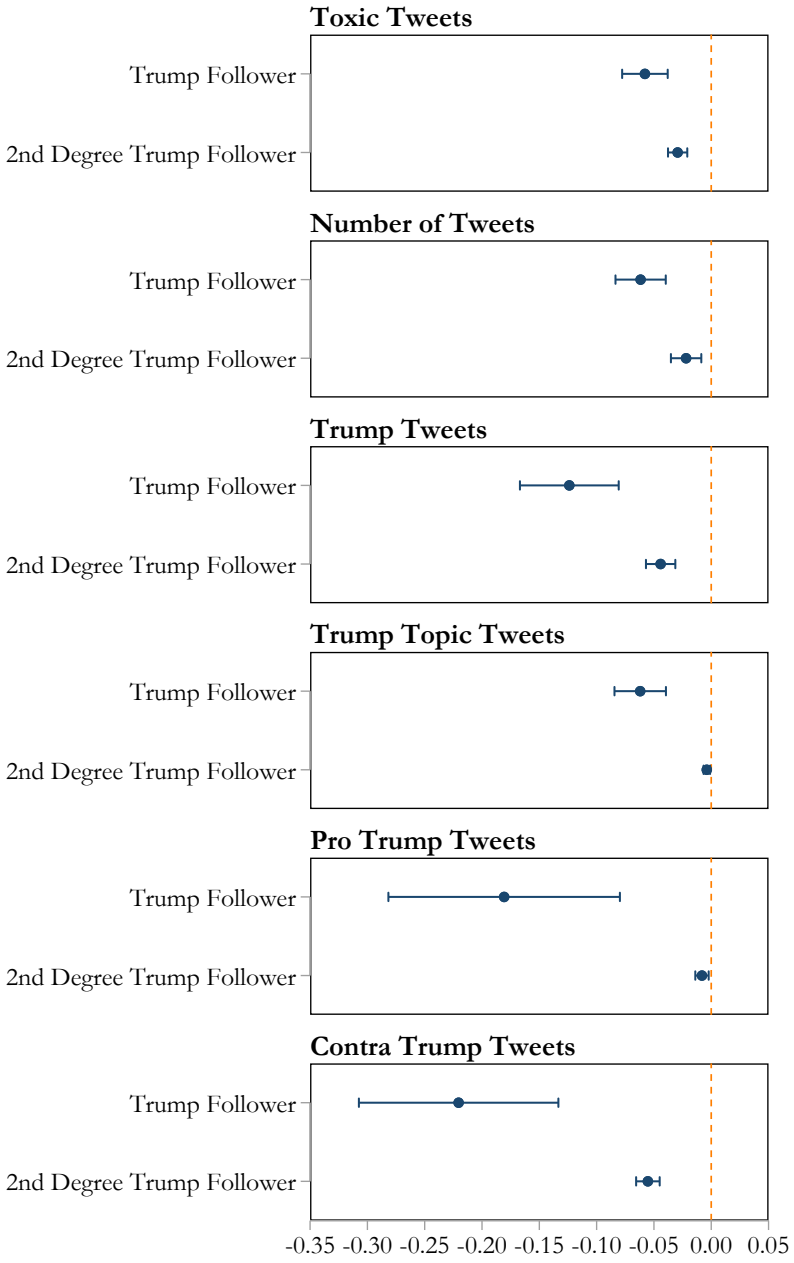
	Dep. Var: Toxicity				
	Trump Tweets (1)	No-Trump Tweets (2)	Democrats (3)	Independent (4)	Republicans (5)
Trump Follower × Account Deletion	-0.001***	-0.047***	0.023	-0.030**	-0.090**
	(0.000)	(0.011)	(0.020)	(0.015)	(0.037)
User FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	2,496,825	9,649,050	326,650
Pre-Period Mean of DV	0.001	0.182	0.408	0.122	0.268
R^2	0.23	0.52	0.50	0.53	0.48

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variables are different samples of toxic tweets. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Network spillover effects Figure 3 shows the results of a modified version of estimating Equation (1) that adds an additional interaction term of the account deletion with an indicator variable equal to 1 for people who do *not* follow Trump directly but follow somebody that does (*2nd Degree Trump follower*). This tests for an indirect effect of Trump's account deletion through the network of his followers. The point estimates suggest an important role for such spillover effects. For all outcomes, we find that the suspension reduced user activity of second-degree linkages.[4] As one would expect, the second-degree effect is always smaller than the direct effect on Trump followers. For the number of toxic tweets, we find that the effect is around half that of the direct effect.

[4]The exception is the number of anti-Trump tweets, for which we continue to find a quantitatively small and statistically insignificant effect.

Figure 3: Network Effects of Trump's Account Deletion



Notes: This figure plots the point estimates from a regression specification akin to Equation (1) for different outcome variables Y. We include interactions of *Trump Follower* and *2nd Degree Trump Follower* with *Account Deletion* as independent variables in the same regression. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Robustness In the Appendix, we conduct extensive robustness checks. Our results are similar when we consider the share rather than the number of tweets. We also consider placebo estimates for the additional outcomes, alternative toxicity measures, and alternative variable transformations. These tests are described in greater detail in Appendix B.

5 Discussion

Whether and how social media companies should regulate the content on their platforms is a contentious issue. Supporters of more comprehensive regulation argue that hateful content poisons public debate, may affect the mental health of people at the receiving end of toxic posts, and may even lead to spikes in offline expressions of hatred. Critics, on the other hand, fear censorship of ideas and an infringement on freedom of speech.

This paper provides empirical evidence informing this debate by showing that account suspensions of polarizing, influential individuals can be an effective tool for decreasing online toxicity. Taken together, our results are most consistent with the interpretation that Trump's account deletion reduced toxic online messaging by his supporters in a quantitatively important way, which led to additional indirect effects through his follower network. The account suspension also led to a decrease in engagement, as measured by the total number of tweets. These results confirm experimental evidence that platforms face a potential trade-off between reducing toxicity and keeping users engaged (*9*).

To be clear, the findings of our study should not be taken as a blanket endorsement of account suspension policies. Rather, we interpret the results here as evidence that content moderation policies can "work" in taming hateful content, which has to be weighed against a potential infringement of freedom of speech to evaluate welfare effects. Such a welfare analysis is beyond the scope of our work.

References

1. K. Müller, C. Schwarz, Fanning the Flames of Hate: Social Media and Hate Crime, *Journal of the European Economic Association* **19**, 2131 (2021).

2. K. Müller, C. Schwarz, From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment, *American Economic Journal: Applied Economics* (2022).

3. L. Bursztyn, G. Egorov, R. Enikolopov, M. Petrova, Social Media and Xenophobia: Evidence from Russia, *Working Paper 26567*, National Bureau of Economic Research (2019).

4. NBC News, Twitter CEO Defends Trump Ban, Cites Threats to Physical Safety, by David Ingram, https://www.nbcnews.com/tech/social-media/twitter-ceo-defends-trump-ban-cites-threats-physical-safety-n1254213 (accessed October 20th, 2022) (2021).

5. NBC News, Facebook Removes Pages Belonging to Far-Right Group 'Proud Boys' , by David Ingram, https://www.nbcnews.com/tech/social-media/facebook-removes-pages-belonging-far-right-group-proud-boys-n926506 (accessed October 20th, 2022) (2018).

6. Twitter, Permanent Suspension of @realDonaldTrump (2021).

7. The Guardian, Elon Musk Reinstates Donald Trump's Twitter Account After Taking Poll, by Dan Milmo (2022).

8. Pew Research Center, Sizing Up Twitter Users, *Tech. rep.* (2019).

9. G. Beknazar-Yuzbashev, R. Jiménez-Durán, J. McCrosky, M. Stalinski, Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment (2022).

10. R. Jiménez Durán, The Economics of Content Moderation: Theory and Experimental Evidence From Hate Speech on Twitter, *Available at SSRN* (2022).

11. R. Jiménez Durán, K. Müller, C. Schwarz, The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG, *Available at SSRN* (2022).

12. E. Chandrasekharan, *et al.*, You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech, *Proc. ACM Hum.-Comput. Interact.* **1** (2017).

13. D. Acemoglu, A. Ozdaglar, J. Siderius, A Model of Online Misinformation, *Working Paper 28884*, National Bureau of Economic Research (2021).

14. Y. Liu, P. Yildirim, Z. J. Zhang, Social Media, Content Moderation, and Technology (2021).

15. C. A. Bail, *et al.*, Exposure to Opposing Views on Social Media Can Increase Political Polarization, *Proceedings of the National Academy of Sciences* **115**, 9216 (2018).

16. R. Levy, Social Media, News Consumption, and Polarization: Evidence from a Field Experiment, *American Economic Review* **111**, 831 (2021).

17. H. Allcott, L. Braghieri, S. Eichmeyer, M. Gentzkow, The Welfare Effects of Social Media, *American Economic Review* **110**, 629 (2020).

18. M. Petrova, A. Sen, P. Yildirim, Social Media and Political Donations: New Technology and Incumbency Advantage in the United States, *Working Paper* (2017).

19. R. Enikolopov, A. Makarin, M. Petrova, Social Media and Protest Participation: Evidence from Russia, *Econometrica* **88**, 1479 (2020).

20. E. Zhuravskaya, M. Petrova, R. Enikolopov, Political Effects of the Internet and Social Media, *Annual Review of Economics* **12** (2020).

21. T. Fujiwara, K. Müller, C. Schwarz, The Effect of Social Media on Elections: Evidence From the United States, *NBER Working Papper* (2021).

22. A. A. Siegel, *et al.*, Trumping Hate on Twitter? Online Hate in the 2016 US Election Campaign and its Aftermath, *Quarterly Journal of Political Science* **16**, 71 (2021).

23. S. Ali, *et al.*, Understanding the Effect of Deplatforming on Social Networks, *13th ACM Web Science Conference 2021* p. 187–195 (2021).

24. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3**, 993 (2003).

25. X. Yan, J. Guo, Y. Lan, X. Cheng, A Biterm Topic Model for Short Texts, *Proceedings of the 22nd International Conference on World Wide Web* pp. 1445–1456 (2013).

26. D. Pion-Berlin, T. Bruneau, R. B. Goetze, The Trump Self-Coup Attempt: Comparisons and Civil–Military Relations, *Government and Opposition* p. 1–18 (2022).

27. P. Barberá, Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data, *Political Analysis* **23**, 76 (2015).

28. E. Wulczyn, N. Thain, L. Dixon, Ex Machina: Personal Attacks Seen at Scale, *Proceedings of the 26th International Conference on World Wide Web* pp. 1391–1399 (2017).

29. L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and Mitigating Unintended Bias in Text Classification, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* pp. 67–73 (2018).

30. C. Schwarz, ldagibbs: A Command for Topic Modeling in Stata Using Latent Dirichlet Allocation, *The Stata Journal* **18**, 101 (2018).

31. P. Barberá, *et al.*, Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data, *American Political Science Review* **113**, 883 (2019).

32. M. F. Porter, An Algorithm for Suffix Stripping, *Program* (1980).

33. M. F. Porter, Snowball: A Language For Stemming Algorithms (2001).

Acknowledgments

We are very grateful to Pablo Barbera for sharing his data with us. We thank Giovanni Moscariello for outstanding research assistance. We also thank seminar participants at Bocconi for their valuable comments. Karsten Müller would like to acknowledge generous funding received as part of a Presidential Young Professorship at National University of Singapore.

Conflict of Interest

The authors are not aware of any relevant conflict of interests.

Supplemental Materials
The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion

by Karsten Müller and Carlo Schwarz

Correspondence to: kmueller@nus.edu.sg or carlo.schwarz@unibocconi.it

This appendix presents further details on data construction and additional robustness exercises:

- Appendix A provides additional details on the data.
- Appendix B shows additional results.

A Materials and Methods

This Appendix provides additional details on data collection and the construction of our main outcome measures. The starting point for our data collection is the list of 498,901 randomly-selected American Twitter users collected in 2015 by (*22*). Using this sample of Twitter users has two main advantages. First, it allows us to capture a representative picture of overall Twitter activity in the United States. Secondly, these users were already active on Twitter at the time of Trump's first presidential run. This allows us to do a within-person comparison of their behavior around the 2016 and 2020 presidential elections. The first data collection step was to download the tweets sent by each user in the (*22*) sample. In total, we obtained 395 Million tweets during the time period 2014- the start of 2022, as well as the user profile information.

The unique dataset enabling our analysis is information on the universe of Twitter followers of Donald Trump as of November 2020. This time period is ideal because it is in immediate proximity to the 2020 election but pre-dates Trump's removal from the platform. This list of

Trump followers allows us to identify which Twitter users were more exposed to the account deletion.

We further make use of a list of the followers of all US Congress members from the 110th to 115th US congress, which we collected in the run-up to the 2020 election (*21*). We use these follower lists to identify the political leaning of each Twitter user. More specifically, we classify a user as Republican if they follow more Republicans than Democrats on Twitter, and vice versa. Users who do not follow any Republican or Democratic Member of Congress on Twitter are classified as an independent.[5]

To study network spillovers, we are also interested in Twitter users who do not directly follow Trump but follow someone that does. We call these accounts second-degree Trump followers. To identify them, we scraped the list of followed accounts for a random subsample of our Twitter users. In total, we obtained the list of followed accounts for 268,215 Twitter users. Based on this list, we can then identify if Donald Trump is among the followed accounts of an account or not. We code users as second-degree Trump followers only if they do not directly follow Trump on Twitter.

We create several dependent and independent variables of interest based on this dataset by employing natural language processing algorithms. In particular, we 1) classify the toxicity of tweets using Google's Perspectives API, 2) train a topic model to identify shifts in the topics users are talking about, and 3) identify pro- and anti-Trump tweets based on the hashtags used. We describe each of these algorithms in the following sections.

A.1. Additional Details on Toxicity Scores

Our approach to measuring the toxicity of tweets is based on Google's Perspective API (*28, 29*). The Perspectives API has been widely used as a measure of hate speech and is known to achieve

[5]The motivation for this approach comes from (*27*), who also classifies the political leaning of Twitter users based on who they follow.

state-of-the-art performance in classifying hateful text. We sent all tweets in our data to the API, which then returns a score for the toxicity of each tweet. The resulting toxicity score ranges from 0 and 1 for the following six dimensions: toxicity, severe toxicity, identity attack, insult, profanity, and threat. The Perspectives API is able to assign toxicity scores for tweets in one of the following languages: English, Spanish, French, German, Portuguese, Italian, and Russian. These languages make up the vast majority of tweets in our data, with English by far being the most prominent. Most tweets for which the API does not provide a toxicity score do not contain any text (e.g., because they only contain a link).

Following the existing literature, we use a cutoff of 0.8 to identify highly toxic tweets. Table A.1 presents the text of ten tweets with very high toxicity scores. Overall, the Perspectives API appears to be accurate in capturing toxic tweets. Our main measure is based on a composite score which is the maximum of the six toxicity dimensions. As such, a tweet in our data will be coded as toxic if it is above 0.8 in any one of the toxicity dimensions. As we show in robustness checks, the exact cutoff or measure of toxicity have no bearing on our results.

Table A.1: Examples of Toxic Tweets

Tweet	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
@RepAbaham You are an asshole too. Eat Trump's shit	0.99	0.87	0.62	0.99	0.98	0.25
RT @MasonWattles: Trump and Hillary are both shit, but trump is the better shit	0.96	0.81	0.64	0.95	0.97	0.27
@RandPaul You are the dumbest mother-fucker	0.99	0.926	0.56	0.99	0.99	0.235
@laurenboebert Lol, goddamn are you a fucking imbecile	0.99	0.92	0.71	0.99	0.99	0.231
@Jim_Jordan Gym Jordan you're an asshole #gymjordanisanasshole	0.994	0.89	0.48	0.98	0.98	0.18
@LindseyGrahamSC Get fucked you piece of shit.	0.99	0.92	0.59	0.95	0.99	0.82

A.2. Additional Details on the Topic Model

To classify the topics Twitter users talk about, we use a machine learning topic model. Topic models such as Latent Dirichlet Allocation (*24*) have been widely used in social science research to extract topics from unstructured text data.[6] For example, (*31*) uses a topic model to study which topics politicians and citizens talk about. The topic model will describe each document (here, a tweet) as a probabilistic mixture of topics, and each topic as a probabilistic mixture of words. This allows to more systematically analyze clusters of related tweets.

A challenge for applying topic modeling to tweets is that, given Twitter's limit of 280 characters, they are relatively short. As such, they provide relatively little context for the topic model to work with. To alleviate this problem, we use a Biterm Topic Model (*25*) (BTM henceforth). BTM makes use of the global co-occurrences of words in the entire corpus to identify more coherent topics. To prepare the data for estimating the BTM, we remove stopwords, links, and non-alphanumerical characters from each tweet. Then, we reduce words to their morphological roots using a stemmer (*32, 33*). To avoid less meaningful clusters, we only fit the topic model on the subset of tweets in English.[7]. Finally, we train the BTM on a random subset of 20 million tweets and restrict the vocabulary to the 75,000 most frequent tokens. We use the trained model to obtain topics for all English tweets in our data.

As with any clustering algorithm, we need to pre-specify the number of topics (clusters) we want to create. Since we are interested in relatively coarse topics, we set the number to 25. In unreported results, we have confirmed that the findings are almost identical if we allow for 50, 75, or 100 topics. In Table A.2, we show the 10 most frequent words for each of the 25 topics we obtained using BTM. For our analysis, we use topic 6, which groups tweets related to Trump and politics.

[6] (*30*) provides a Stata implementation for Latent Dirichlet Allocation.

[7] The results are virtually unchanged if we fit the topic model to all tweets. The only difference is that some of the topics simply group together non-English words.

Table A.2: Topics from BTM

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
get	like	game	food	state	like	trump	follow	help	get
like	get	year	new	new	day	peopl	one	health	like
day	shit	tonight	free	year	get	say	new	use	got
today	fuck	play	today	school	eat	presid	today	new	time
make	lol	night	get	today	good	vote	via	need	watch
feel	know	last	home	citi	make	media	unfollow	busi	look
time	nigga	day	day	nation	love	like	win	get	wear
love	say	first	check	first	drink	right	person	peopl	game
work	got	win	make	counti	one	would	stat	work	one
good	bitch	time	use	team	today	american	day	data	day
Topic 10	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**
get	like	peopl	like	get	like	day	play	win	new
time	get	god	get	new	get	join	game	game	check
like	love	life	love	time	peopl	today	get	team	get
one	lol	today	girl	like	know	come	like	play	app
day	know	may	look	day	make	week	time	year	use
year	good	one	got	work	want	tomorrow	watch	season	video
good	day	make	lol	one	feel	new	year	time	post
level	one	love	day	make	think	open	back	get	free
make	time	thing	know	look	thing	friday	one	first	today
love	got	feel	one	today	someon	event	see	day	great
Topic 20	**Topic 21**	**Topic 22**	**Topic 23**	**Topic 24**					
new	love	love	year	job					
use	like	day	time	work					
art	get	thank	one	get					
design	make	happi	day	need					
work	song	time	today	help					
look	know	one	get	year					
get	see	great	new	vote					
learn	thank	life	need	time					
like	time	year	like	make					
make	much	today	peopl	student					

A.3. Additional Details on Pro-/Contra Trump Hashtags

To classify whether tweets contain messages in support or in opposition to Trump, we classify hashtags. As a starting point, we extract all hashtags in our data set that contain the term "trump". This provides us with a list of 17,845 hashtags. We hand-code these hashtags into the three categories pro-Trump, contra-Trump, and neutral. Examples of pro-Trump hashtags are "#fightfortrump", "#trumpsecondterm", or "#trumpwillwin." Contra-Trump hashtags are "#removetrumpnow", "#trumptreason", or "#trumpcoupattempt." Hashtags that are unclear, such as "#lovetrumps", "#trumpplaza", or "#donalddtrump" were coded as neutral.

We extend this list of hashtags with other frequently-used hashtags. We add "#MAGA", "#StopTheSteal", "#VoterFraud", and "#RiggedElection" to the list of pro-Trump hashtags and "#25thAmendmentNow", "#11MAGATerrorism", and "#CapitolRiot" to the list of anti-Trump hashtags. Note that this modification has no bearing on our results.

Equipped with this list, we code tweets as pro-Trump or anti-Trump based on which hashtags they contain. Tweets that do not contain any of the coded hashtags are coded as neutral.

A.4. Descriptive Statistics

Equipped with the combined tweet-level dataset and user profile information, we create a user-month-level panel data set, which allows us to analyze the Twitter activity of the 498,901 users in our data. Table A.3 presents summary statistics for the most important variables. In Table A.4, Table A.5, and Figure A.1, we also show a comparison between Trump followers and the representative sample of Twitter users excluding Trump followers in the period before the account deletion.

Compared to other Twitter users, Trump followers send more tweets (5.8 compared to 3.3 per month) and produce a much higher share of toxic content (0.3 compared to 0.17 toxic tweets per month). Appendix Figure A.1 visualizes these patterns. However, Table A.5 shows

that, apart from these baseline differences, the user profiles of Trump followers appear to be remarkably similar to those of randomly-picked Twitter users. The rank correlation between the frequency of first names and terms used in user bios is 0.92 and 0.94, respectively. The 10 most common names and terms mentioned in user bios are very similar.

Table A.3: Summary Statistics

Variable	Mean	SD	p50	Min	Max	N
Toxicity Variables						
Any Toxicity> 0.8	0.20	2.06	0.00	0.00	397.00	12,472,525
Toxicity> 0.8	0.08	1.11	0.00	0.00	303.00	12,472,525
Severe Toxicity> 0.8	0.04	0.57	0.00	0.00	195.00	12,472,525
Identity Attack> 0.8	0.02	0.44	0.00	0.00	247.00	12,472,525
Insult> 0.8	0.07	0.92	0.00	0.00	227.00	12,472,525
Profanity> 0.8	0.14	1.59	0.00	0.00	328.00	12,472,525
Threat> 0.8	0.03	0.35	0.00	0.00	85.00	12,472,525
Other Outcomes Variables						
Number of Tweets	3.92	24.39	0.00	0.00	3249.00	12,472,525
Number of Trump Tweets	0.00	0.20	0.00	0.00	156.00	12,472,525
Tweet Trump Topic	0.01	0.29	0.00	0.00	238.00	12,472,525
Number of Pro-Trump Hashtags	0.00	0.16	0.00	0.00	262.00	12,472,525
Number of Anti-Trump Hashtags	0.02	0.88	0.00	0.00	1587.00	12,472,525
User Variables						
Trump Follower	0.09	0.29	0.00	0.00	1.00	12,472,525
2nd Degree Trump Foll.	0.81	0.40	0.00	1.00	1.00	6,705,375

Notes: This table presents the mean, standard deviation, median, minimum, maximum, and number of observations of our main outcome variables, main variables of interest, and control variables for the full estimation sample.

Table A.4: Comparing Trump Followers to Other Twitter Users (Pre-Period)

	Twitter users			Trump Follower		
Variable	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Any Toxicity> 0.8	0.171	1.741	5,881,070	0.304	2.235	604,643
Toxicity> 0.8	0.070	0.943	5,881,070	0.131	1.246	604,643
Severe Toxicity> 0.8	0.033	0.514	5,881,070	0.048	0.556	604,643
Identity Attack> 0.8	0.018	0.401	5,881,070	0.024	0.374	604,643
Insult> 0.8	0.059	0.770	5,881,070	0.142	1.359	604,643
Profanity> 0.8	0.125	1.389	5,881,070	0.182	1.463	604,643
Threat> 0.8	0.022	0.294	5,881,070	0.040	0.369	604,643
Number of Tweets	3.288	18.888	5,881,070	5.806	24.226	604,643
Number of Trump Tweets	0.004	0.203	5,881,070	0.013	0.333	604,643
Tweet Trump Topic	0.005	0.265	5,881,070	0.032	0.698	604,643
Number of Pro-Trump Hashtags	0.001	0.079	5,881,070	0.014	0.622	604,643
Number of Anti-Trump Hashtags	0.024	0.763	5,881,070	0.113	2.618	604,643

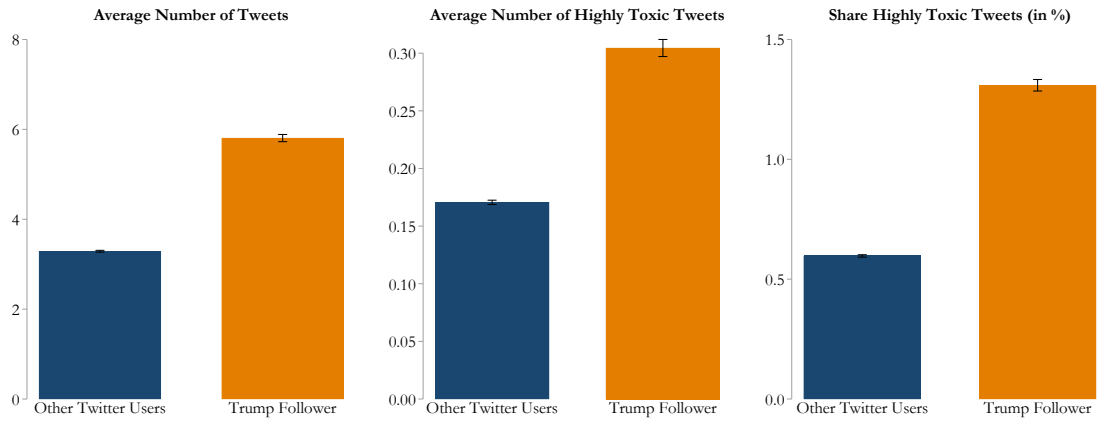
Notes: This table presents the mean, standard deviation, median, minimum, maximum, and number of observations of our main outcome variables, main variables of interest, and control variables for period before Trump's account deletion in January 2021.

Table A.5: Comparison User Profiles of Trump Followers to Other Twitter Users

User first names (Corr. = 0.92)		**Terms used in user bios (Corr. = 0.93)**	
Twitter users	Trump followers	Twitter users	Trump followers
david	michael	love	love
michael	john	de	life
john	david	life	co
chris	chris	co	de
daniel	mike	la	http
alex	brian	http	http co
sarah	jason	http co	fan
mike	james	follow	music
kevin	mark	music	husband
jessica	matt	like	father

Notes: This table compares the individual characteristics of Twitter users who follow "@RealDonaldTrump" to other Twitter users. We plot the ranking of the most common first names and terms used in a Twitter user's "bio."

Figure A.1: Average Toxicity of Trump's Twitter Followers



Notes: This figure plots the average number of tweets, the average number of toxic tweets, and the share of toxic tweets for Twitter users who followed Donald Trump before the deletion of his account and a representative sample of Twitter users who do not follow Trump. The whiskers represent 95% confidence intervals.

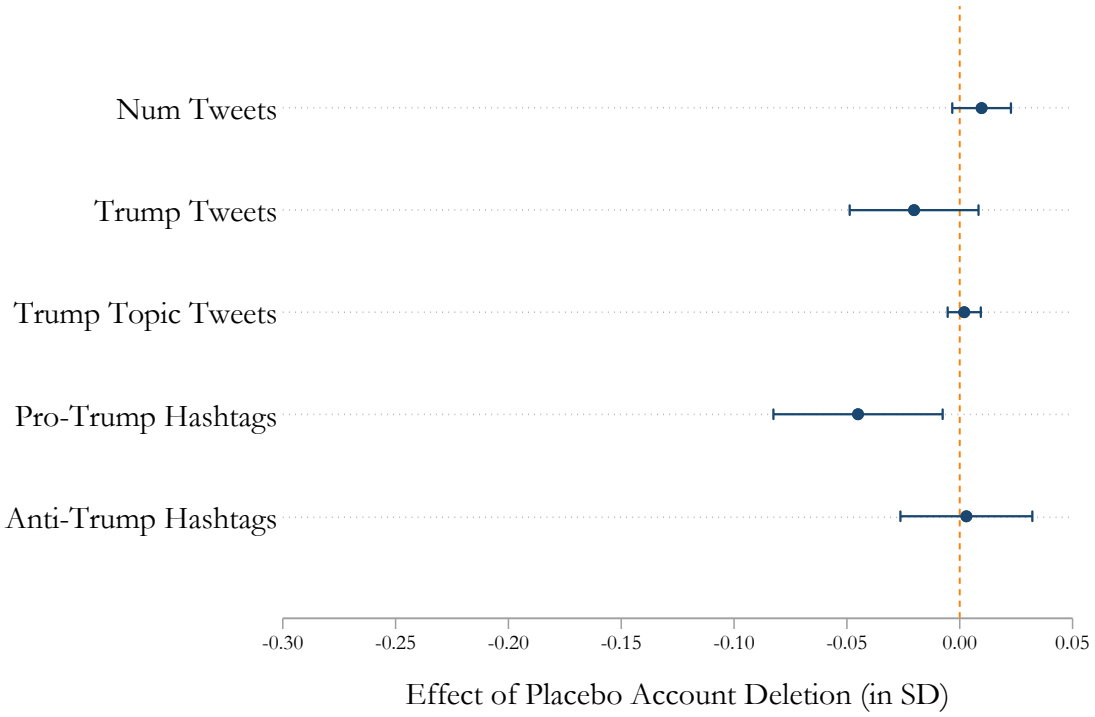
B Additional Results

We conduct several robustness exercises for our main results, which we describe in the following. Appendix Figure B.1 shows the placebo estimates for the 2016 election for each of the additional outcomes. The point estimates are much smaller than those for our baseline results in Figure 2 and mostly statistically insignificant, despite the identical sample size. The only statistically significant placebo estimate for the number of pro-Trump tweets is around one third of the magnitude of the estimate in Figure 2. This again provides support for the hypothesis of a causal effect of the account deletion.

In Table B.2, we use alternative measures of toxicity obtained from the Perspective API as outcomes. We find a negative and statistically significant effect of similar magnitude on most of them. Table B.3 considers alternative cutoffs of the toxicity core for classifying tweets as "toxic," which all yield similar results. Table B.4 considers alternative pre-periods for which the $Account\ deletion_t$ variable is equal to 0. Extending the pre-period back to January 2019, 2018, or 2017 (instead of 2020) leads to highly similar point estimates.

Table B.5 and Table B.6 consider alternative transformations of the toxicity measure and standard errors that again yield similar estimates. In Table B.7, we show that our findings also hold when we scale our outcomes over the number of tweets sent by each user. This table suggests that the account deletion did not only reduce the number of toxic tweets but also their share in overall Twitter content. Table B.8 repeats our baseline exercise but includes retweets. This results in very similar patterns in that we find a reduction of toxicity, the number of tweets, tweets about Trump, pro-Trump tweets, and the similarity of tweets vis-à-vis Trump's. Lastly, Table B.9 shows the results if we restrict our analysis to tweets in English. The language of a tweet again has no bearing on our results.

Figure B.1: Impact of Placebo Account Deletion on Other Outcomes



Notes: This figure plots the placebo point estimates for β based on Equation (1) for different outcome variables Y estimated in the sample 2016-17, exactly four years before our baseline sample period. We plot 99% confidence intervals based on standard errors clustered by user.

Table B.1: Trump Account Deletion and Other Outcomes

	Number Tweets	Trump Tweets	Trump Topic Tweets	Pro Trump Hashtags	Contra Trump Hashtags
	(1)	(2)	(3)	(4)	(5)
Panel A: Diff-in-Diff					
Trump Follower × Account Deletion	-0.742***	-0.008***	-0.013***	-0.014***	-0.084***
	(0.163)	(0.002)	(0.003)	(0.002)	(0.011)
User FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes
User × Month of Year FE	Yes	Yes	Yes	Yes	Yes
User Quadratic Time Trend	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	3.523	0.005	0.007	0.002	0.033
R^2	0.73	0.61	0.75	0.60	0.61
Panel B: Triple Difference					
Trump Follower × Account Deletion × I[2020]	-0.802***	-0.007***	-0.021***	-0.011**	-0.073***
	(0.153)	(0.002)	(0.003)	(0.004)	(0.014)
User × Election FE	Yes	Yes	Yes	Yes	Yes
User × Treatment Time FE	Yes	Yes	Yes	Yes	Yes
Treatment Time × Election FE	Yes	Yes	Yes	Yes	Yes
User × Election Linear Time Trend	Yes	Yes	Yes	Yes	Yes
Observations	24,945,050	24,945,050	24,945,050	24,945,050	24,945,050
Pre-Period Mean of DV	3.653	0.004	0.005	0.002	0.024
R^2	0.75	0.63	0.62	0.59	0.64

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the outcome measure in the top row. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.2: Robustness: Alternative Toxicity Measures

	Baseline	Toxicity	Sev. Toxicity	Ident. Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trump Follower × Account Deletion	-0.048***	-0.026***	-0.004	0.001	-0.036***	-0.015**	-0.003**
	(0.011)	(0.006)	(0.003)	(0.002)	(0.006)	(0.008)	(0.002)
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.183	0.076	0.034	0.018	0.067	0.130	0.024
R^2	0.52	0.51	0.48	0.46	0.48	0.52	0.41

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month based on different definitions of toxicity available via the Perspective API. In all cases toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.3: Robustness: Alternative Toxicity Thresholds

	Toxicity ≥ 0.5	Toxicity ≥ 0.6	Toxicity ≥ 0.7	Toxicity ≥ 0.8	Toxicity ≥ 0.9	Toxicity ≥ 0.95
	(1)	(2)	(3)	(4)	(5)	(6)
Trump Follower × Account Deletion	-0.120***	-0.099***	-0.075***	-0.048***	-0.026***	-0.009**
	(0.033)	(0.025)	(0.019)	(0.011)	(0.008)	(0.004)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.763	0.564	0.436	0.183	0.186	0.083
R^2	0.49	0.50	0.50	0.52	0.51	0.50

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Toxic tweets are those with a Perspective API aggregate toxicity score larger than the cutoff in the top row. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.4: Robustness: Choice of Pre-Period

	Baseline	Jan 2019	Jan 2018	Jan 2017
	(1)	(2)	(3)	(4)
Trump Follower × Account Deletion	-0.040***	-0.041***	-0.036***	-0.039***
	(0.014)	(0.013)	(0.012)	(0.012)
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes
User × Month of Year FE	Yes	Yes	Yes	Yes
User Quadratic Time Trend	Yes	Yes	Yes	Yes
Observations	12,472,525	17,960,436	23,947,248	29,934,060
Pre-Period Mean of DV	0.183	0.155	0.137	0.129
R^2	0.72	0.68	0.58	0.50

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. We vary the estimation sample to include 2020-21 (column 1), 2019-21 (column 2), 2018-21 (column 3), and 2017-21 (column 4). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.5: Robustness: Variable Transformations

	Baseline	ln(1+Toxic Tweets)	asinh(Toxic Tweets)	I[Toxic Tweet>0]	Share of Toxic Tweets
	(1)	(2)	(3)	(4)	(5)
Trump Follower × Account Deletion	-0.048***	-0.012***	-0.015***	-0.008***	-0.001***
	(0.011)	(0.001)	(0.002)	(0.001)	(0.000)
User FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.183	0.054	0.068	0.040	0.007
R^2	0.52	0.69	0.69	0.60	0.31

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is a measure of toxicity. Column 1 uses our baseline measure, the number of toxic tweets, defined as those with a Perspective API toxicity score of larger than 0.8. Column 2 uses the natural logarithm of the number of toxic tweets (with 1 added inside). Column 3 instead uses the inverse hyperbolic sine transformation, which naturally accommodates zero and negative values. Column 4 considers a dummy dependent variable equal to 1 for users that post at least one toxic tweet, and 0 otherwise. Column 5 uses the share of toxic tweets in all user posts. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6: Robustness: Standard Errors

	User	User × Month	User × Month of Year
	(1)	(3)	(4)
Trump Follower × Account Deletion	-0.048***	-0.048***	-0.048***
	(0.011)	(0.008)	(0.008)
User FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.183	0.183	0.183
R^2	0.52	0.52	0.52

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by the level indicated in the top row. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.7: Effect of Account Deletions in Shares (in %)

	Share Toxic Tweets	Share Trump Tweets	Share Trump Topic Tweets	Share Pro Trump Hashtags	Share Contra Trump Hashtags
	(1)	(2)	(3)	(4)	(5)
Trump Follower × Account Deletion	-0.139***	-0.015***	-0.051***	-0.040***	-0.195***
	(0.026)	(0.005)	(0.009)	(0.005)	(0.016)
User FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.663	0.018	0.029	0.008	0.114
R^2	0.31	0.12	0.18	0.12	0.19

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variables are the shares (in %) of the indicated outcomes for a user in a given month. Toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.8: Effect of Account Deletions Including Retweets

	Toxic Tweets	Number Tweets	Trump Tweets	Trump Topic Tweets	Pro Trump Hashtags	Contra Trump Hashtags
	(1)	(2)	(3)	(4)	(5)	(6)
Trump Follower × Account Deletion	-0.052***	-0.778***	-0.011***	-0.025***	-0.024***	-0.129***
	(0.014)	(0.186)	(0.002)	(0.003)	(0.004)	(0.018)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.319	5.702	0.009	0.011	0.004	0.089
R^2	0.50	0.52	0.27	0.37	0.19	0.29

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. The sample is based on all tweets and retweets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.9: Effect of Account Deletions Excluding Non-English Tweets

	Toxic Tweets	Number Tweets	Trump Tweets	Trump Topic Tweets	Pro Trump Hashtags	Contra Trump Hashtags
	(1)	(2)	(3)	(4)	(5)	(6)
Trump Follower × Account Deletion	-0.047***	-0.753***	-0.009***	-0.025***	-0.018***	-0.115***
	(0.013)	(0.147)	(0.002)	(0.003)	(0.003)	(0.017)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User Linear Time Trend	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525	12,472,525
Pre-Period Mean of DV	0.266	3.662	0.008	0.011	0.003	0.081
R^2	0.50	0.53	0.27	0.37	0.19	0.29

Notes: This table presents the point estimates for β based on Equation (1), where the dependent variable is the number of toxic tweets sent by a user in a given month. Toxic tweets are those with a Perspective API toxicity score of larger than 0.8. $Trump\ follower$ is an indicator variable equal to 1 for Twitter users who followed Trump before his account was suspended, and 0 otherwise. $Account\ Deletion$ is an indicator variable equal to 1 for the months after the account deletion in January 2021. Standard errors are clustered by user. The sample is based on all English tweets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.