The Geopolitics of Deplatforming: A Study of Suspensions of Politically-Interested Iranian Accounts on Twitter

Mehdi Zamani* Andreu Casas†

Abstract

Social media companies increasingly play a role in regulating freedom of speech. Debates over ideological motivations behind suspension policies of major platforms are on the rise. We contribute to this ongoing debate by looking at content moderation from a geopolitical perspective. We start from the premise that US-based social media companies are inclined to moderate content on their platforms in compliance with US sanctions laws, especially those concerned with the Specially Designated Nationals and Blocked Persons List. Despite the release of transparency reports by social media companies, we know little about the scope of the problem and the impact of suspensions on political conversations. We tracked 600,000 Twitter users interested in Iranian politics. After accounting for alternative explanations, we find that Principlist (conservative) users and those supportive of the Iranian government are significantly more likely to be suspended. We also uncover the types of discussions that are being suppressed as a result of these suspensions. This paper contributes to building a better understanding of how governments can influence political conversations abroad, and how social media suspensions shape political conversations online.

*Vrije Universiteit Amsterdam
†Vrije Universiteit Amsterdam: a.casassalleras@vu.nl

1 Introduction

Today, private social media companies play a crucial role in moderating freedom of speech (DeNardis and Hackl, 2015; Balkin, 2017; Gillespie, 2018). People around the world increasingly rely on social media to consume news (Shearer and Mitchell, 2021), learn and talk about politics (Barberá et al., 2019), and coordinate political actions (González-Bailón et al., 2011). Despite many initial positive views about the role of social media for enhancing more inclusive, equal and free political conversations, the platforms are increasingly suspending accounts (a phenomenon commonly known as "deplatforming") to address concerns about incivility, hateful behaviors, bots, misinformation, rumors, and conspiracies (DeNardis and Hackl, 2015; Bay and Fredheim, 2019; Bastos, 2021). In addition, in recent years many have claimed that widely-used platforms such as Facebook, YouTube and Twitter suspend accounts for political reasons, allegedly targeting conservatives in US politics (Davalos and Brody, 2020) as well as voices supportive of governments involved in a geopolitical rivalry with the West, such as China, Russia, Venezuela and Iran (O'Sullivan and Moshtaghian, 2020; Cartwright, 2020). Studying the potential suspension biases on social media and their effects on politically-relevant conversations is crucial for theorizing and assessing the role of social media platforms in moderating online speech.

In this paper, we focus on the geopolitical aspect of social media suspensions. Social media platforms are currently at the center of many geopolitical disputes (Cartwright, 2020; Gray, 2021), yet, we lack a clear understanding of the conditions under which platforms can shape the conversation about international politics and its implications. Several studies have explored how governments leverage social media to constrain political speech at home, such as China (King, Pan, and Roberts, 2013, 2014) and Saudi Arabia (Pan and Siegel, 2020). Some other scholars have researched the ways in which non-Western governments leverage social media communications to influence public opinion abroad. The election-interference

operations by the Russian Internet Research Agency (IRA) since the 2016 US election are well known (Golovchenko et al., 2020; Lukito, 2020). Other works have discussed how non-Western countries (e.g. China) can leverage state-controlled platforms (e.g. TikTok) for foreign surveillance (Gray, 2021). However, little is known about how social media platforms can advance interests of Western governments through enforcing their content moderation measures. The United States is of particular relevance in this context, as some of the most popular social media platforms are based in this country.

Governments can leverage social media for various geopolitical purposes, such as conducting foreign surveillance (Gray, 2021), promoting their own narratives (Golovchenko et al., 2020; Barrie and Siegel, 2021; Stukal et al., 2022), and suppressing opposing viewpoints (Golovchenko, 2022). In this way, social media can serve as a powerful tool for governments to advance their geopolitical interests. In this paper we focus on the latter, and explore how the US can push US-based social media platforms to suppress competitive views abroad. In particular, we study suspensions of users interested in the politics of a geopolitical rival of the US, namely Iran, on a US-based platform, Twitter. The relationship between Iran and the US has been a significant focus of geopolitical conflict for many years and has implications for other relevant countries such as Russia, China, and the UK. Studying this case is therefore important and provides valuable insights. Despite being blocked in Iran, millions of Iranian citizens, including members of Parliament and top government officials, use VPNs to access and actively use Twitter, where they frequently discuss political topics. While it may not be the most popular platform in the country, Twitter remains a crucial platform for political discourse in Iran. (see Hashemi, Wilson, and Sanhueza (2022)).

When a social media platform with a global reach is based in a particular country, that government can use the legal system to exert pressure on the platform to implement certain content moderation policies with the goal of shaping political conversations abroad (Crasnic, Kalyanpur, and Newman, 2017; Balkin, 2017; Cartwright, 2020; Golovchenko, 2022). The

US government maintains a list of individuals and organizations (SDN: the *Specially Designated Nationals And Blocked Persons List*) whose assets are blocked, and, by law, US citizens and organizations are prohibited from dealing with. Thousands of Iranian individuals, many of whom being state officials, and organizations are on the SDN list, including the *Islamic Revolutionary Guard Corps* (IRGC) – the official military organization in charge of defending Iran's territorial borders. On January 3, 2020, a US drone strike killed General Qassem Soleimani, the commander of Irans Quds Force, an elite branch of the IRGC. According to a Meta spokesperson, Instagram suspended accounts of users that condemned the assassination or simply covered the story in compliance with US sanctions laws (Cockerell, 2020). While these companies often release reports on the suspension of user accounts for their involvement in state-backed information campaigns (e.g. Twitter),[1] there is limited information available on the scope of these account suspensions and their impact on political discussions related to Iran on the platform.

In March 2020, we identified 601,940 Twitter users who followed Iranian politics, and for a six-month period we periodically collected the messages they posted in the platform and checked whether they had been suspended. Most of the accounts (N=594,852) remained *active* after the period of analysis, yet many were (at least temporarily) *suspended* by Twitter (N=3,737), or *deleted* by either the user or Twitter (N=3,351). We use state-of-the-art computational methods to assess potential ideological differences between the *active* and *suspended* users, after controlling for several confounders; and explore the types of conversations that in turn were to some extent repressed *vs.* amplified as a result of such suspensions. As one would expect, we find many toxic behaviors (e.g. using hateful language and bot-like behavior) to be predictive of suspension. We also find conservative users and those supportive of the Iranian government to also be more likely to be suspended. An analysis of the content more often

[1]See for example information released here: https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-

discussed by non-suspended (v. suspended) users also reveals that accounts engaging with progressive discussions and networks are suspended at lower rates.

The contribution of the paper is four-fold. First, we contribute to the literature on social media and political content moderation by discussing the geopolitical motivations and strategies behind existing moderation practices. Second, we contribute to the literature on social media, public diplomacy and geopolitics, by emphasizing that all countries – non-Western countries such as Russia and China, but also Western ones such as the US – can to some extent, use social media platforms for their geopolitical interest. Third, we put forward a research design and a wide-range of computational techniques that we hope can foster further explorations of the determinants and consequences of political content moderation on social media. We conclude with empirical evidence on suspension patterns in the Iranian Twitter sphere and how these shape politically-relevant discussions on the platform.

2 Political Content Moderation and Deplatforming on Social Media

Social media platforms were originally created with the ideal of enhancing participation, engagement and social connection around the globe (Gillespie, 2018). However, the platforms, the general public, and academics soon became aware of the negative aspects of such an open and free form of communication when some users began to use social media in ways that were considered undesirable from a normative perspective (Gillespie, 2018). Certain behaviors, such as hateful speech, and misinformation, are particularly relevant to the study of politics and freedom of political speech online as they can undermine the development of healthy and democratic conversations on platforms. These issues are the focus of this study. Other controversial behaviors, such as nudity and pornography, are also relevant for understanding the relationship

between platform governance and free speech, but will not be specifically addressed here.

Researchers have identified several social media behaviors, such as expressing extreme views, that can disrupt political conversations online and undermine politics more broadly. For example, a study of Twitter accounts associated with the alt-right movement in the United States Klein (2019) found that the movement used the mainstream social media platform to disseminate messages targeting immigrants, Muslims, and the Black Lives Matter movement prior to the 2017 *Unite the Right* rally in Charlottesville.

Overall, numerous studies have found that mainstream social media platforms often facilitate the dissemination of uncivil and hateful content. A study of Twitter messages mentioning members of the United States Congress in 2017-2018, Theocharis et al. (2020) found that 18% of the tweets contained uncivil language,[2] with spikes of 20-25% on days surrounding significant political events. A study of Twitter messages posted during the 2016 Presidential campaign in the United States, Siegel et al. (2021) found that, while at a lower rate, more extreme hate speech[3] was also present on the platform, with less than 1% of tweets mentioning Trump and Clinton and a few days spiking to around 1-3%.

Also, there is a growing concern regarding the spread of false information on major social media platforms. Numerous studies have documented the circulation of fake news on platforms such as Twitter and Facebook. Grinberg et al. (2019) found that fake news accounted for 6% of the news consumption on Twitter during the 2016 election cycle in the United States. In a similar vein, Guess, Nagler, and Tucker (2019) found that 8.5% of Facebook users shared at least one article containing fake news during the same election cycle.

Some scholars have documented that some political actors have deployed automated bots and manually-controlled social media operations to pursue their often undemocratic goals. As

[2]Defined as "*disrespectful discourse that silences or derogates alternative views*", (Theocharis et al., 2020, 3).

[3]Defined as "*bias-motivated, hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics, especially when the group or individual are unnecessarily labeled*" (Siegel et al., 2021).

an example, Stukal et al. (2022) found that Twitter bots supportive of the Russian government are more likely to tweet and retweet a more diverse set of accounts at times when there are street protests, or when opposition activists are more active on the platform. In addition, research documenting the US-election-interference efforts from the Russian Internet Research Agency (IRA) shows a high level of coordination among IRA accounts: they were likely to post similar messages on the same topics (Green, 2018) and to follow similar temporal posting patterns – Lukito (2020) for example found that in 2015-2017 the IRA used to first try out a wide range of messages on Reddit to later focus only on promoting on larger platforms (such as Twitter), and in a coordinated fashion, those that had achieved higher engagement levels.

Social media platforms have responded to these many politically-relevant threats by implementing a wide range of moderation policies, and in turn, depending on the context, by labeling content with different kinds of warnings, or even removing content and accounts (a practice known as *deplatforming*) that engage in the aforementioned 'toxic' activities (in addition to other moderation policies not necessarily linked to politics *per se*). For example, Facebook's Community Standards[4] states that the platform monitors threats such as violence and incitement, dangerous individuals and organizations, hate speech, account integrity and authentic identity, and misinformation, among others. In a similar fashion, the Twitter Rules[5] mention: violence and extremism, hateful conduct, platform manipulation and spam, civic integrity, and synthetic and manipulated media.

According to the growing body of research on political content moderation by social media companies, various types of 'toxic' behaviors have been found to be reliable predictors of suspension. A study of Twitter users who posted messages related to the 2020 US presidential election found that approximately 2% of the 21 million users analyzed were suspended a few months following the election Chowdhury et al. (2021). In their comparison of the content

[4]Consulted on August 22nd, 2022: https://transparency.fb.com/en-gb/policies/community-standards/
[5]Consulted on August 22nd, 2022: https://help.twitter.com/en/rules-and-policies/twitter-rules

posted by suspended and active users, the researchers found that suspended users were twice as likely to post offensive tweets and use hate speech, and more likely to share news from fake news websites. Another study that tracked Twitter users during the same election cycle, Yang et al. (2022) found that approximately 4% of the 9,000 partisan users they monitored were suspended a few months after the election. The researchers also discovered that sharing news from fake news websites was a strong predictor of a user being suspended. In a recent study of a related phenomenon, shadowbanning on Twitter in the US (Jaidka, Mukerjee, and Lelkes, Forthcoming), the authors analyzed 25,000 American users to find that bot-like behavior, offensive language, and political engagement were predictive of messages being downgraded by the platform. In a study of approximately 4.5 million Twitter users who posted messages about the 2017 French, UK, and German elections (Majo-Vazquez et al., 2021), a suspension rate of around 5% was observed. The authors found that suspended users were more likely to share news in general, particularly from legacy media as well as from right-wing digital-born outlets, but not necessarily from fake news websites. The researchers also noted some level of coordination among suspended accounts, as they were found to frequently retweet and amplify similar political figures and content. Furthermore, the study's results suggest that suspended users may be more likely to engage in uncivil behavior and use hateful speech, particularly targeting established centrist political figures.

Based on previous research and the Twitter Rules outlining the types of behaviors that can result in suspension, we present a set of hypotheses about the potential predictors of suspension among the Twitter users interested in Iranian politics that we track in this study:

H_1 Using **hateful language** will be predictive of suspension.

H_2 Posting messages that contain **misinformation** will be predictive of suspension.

H_3 Posting patterns associated to **bot** behavior will be predictive of suspension.

H_4 Posting content similar to the content posted by other accounts (**coordination**) will be

predictive of suspension.

3 The Geopolitics of Deplatforming

Governments pursue various forms of foreign policy and public diplomacy in order to safeguard and promote their interests both domestically and internationally (Baldwin, 2000). With the growing influence of social media in politics, online platforms have become a key arena for geopolitical competition (Cartwright, 2020; Gray, 2021).

There are numerous ways in which social media can be utilized to advance geopolitical interests. These can generally be divided into three categories. One way is for governments to promote favorable geopolitical narratives (Miskimmon, O'loughlin, and Roselle, 2014) on these platforms. These narratives can seek to discredit the narratives of other geopolitical actors, or to promote one's own narrative. Sometimes these strategies seek to influence foreign audiences. For example, since 2016, the Russian IRA has engaged in information operations on Western platforms (e.g. Twitter, Facebook, YouTube), with the intention of influencing public opinion abroad by undermining the democratic process in the US and other Western democracies (Golovchenko et al., 2020; Lukito, 2020). Since Hillary Clinton's tenure as Secretary of State, the US has also made numerous efforts through public diplomacy on social media to promote liberal values in different countries (Tsvetkova et al., 2020). Recent evidence for example points to the Department of Defense controlling several Twitter accounts with the objective of pushing favorable content in countries such as Yemen, Syria, Iraq, and Kuwait.[6] In other cases, information campaigns seek to shape geopolitical narratives within a country. For example, research has shown that the Kremlin, either through accounts from state-owned media (Golovchenko, 2020) or through bots and trolls controlled by the IRA (Stukal et al., 2022), uses social media to influence national debates on international issues such as Crimea

[6] https://theintercept.com/2022/12/20/twitter-dod-us-military-accounts/

(Golovchenko, 2020). As another example, in a study of accounts flagged by Twitter as being coordinated by the Saudi government, Barrie and Siegel (2021) find that a substantive part of the messages are about international politics (e.g. discussions around Qatar and Iran); and that local audiences engage with these messages at substantive rates: at least as often as they engage with messages from news organizations.

Governments can also use social media platforms for surveillance. Research shows that governments sometimes track social media communications to crackdown dissenting voices at home (Pan and Siegel, 2020). Yet the events in recent years regarding TikTok operations in the US clearly illustrate the importance of social media for foreign surveillance. TikTok, developed by the Chinese company ByteDance Ltd (although currently based in the Cayman Islands), is today used by millions of US citizens, particularly younger publics (e.g. 67% of teens between 13-17).[7] Since the 2017 China's National Intelligence Law – which states that all organizations and citizens have to cooperate with national intelligence efforts – there are growing concerns among US officials regarding the possibility that TikTok may share private information from US citizens with the Chinese Communist Party (CCP); including information from top government employees and family members who may be on the platform. In turn, the Trump and Biden administrations have passed legislation to constrain TikTok operations in the country (Gray, 2021).

Finally, governments can also leverage social media for their geopolitical interests by suppressing voices on the platforms. The particular strategy at place will mainly depend on whether the platform at hand is based in one's country (Cartwright, 2020) – and so whether a government has any power to regulate its activity. When this is not the case, governments often need to turn to drastic tactics in order to avoid the dissemination of antagonistic (geopolitical) views, such as banning a given platform from being accessed in the country. Access to several Western social media platforms including Twitter is restricted in countries such as

[7] https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

China, Russia, and Iran. VKontakte and other platforms controlled by the Russian government are banned in Ukraine (Golovchenko, 2022). Nevertheless, in all these countries citizens often use VPNs to stay active on the platforms (Hobbs and Roberts, 2018; Hashemi, Wilson, and Sanhueza, 2022; Golovchenko, 2022).

However, when a platform is based in one's country, a government can leverage the legal system to impose particular content moderation policies and to suppress antagonistic views of geopolitical relevance. For example, in this context, in March 2022 the Kremlin passed new legislation to ban and prevent the spread of "fake" news that are critical of the Russian military operations abroad. Russian social media platforms such as VKontakte and Odnoklasniki are expected to incorporate these directives into their content moderation policy.[8] Around the same time, the Russian government also imposed international sanctions on many US top officials, including President Biden.[9]

In this paper we focus on the last of the three strategies that we have discussed. We contribute to a better understanding of the geopolitical role of social media by exploring how governments (the US) can advance their geopolitical interests by pushing for content moderation policies (on Twitter) that suppress antagonistic geopolitical views abroad (in Iran). Most existing work on the geopolitical use of social media platforms focuses on non-Western countries such as Russia (Golovchenko et al., 2020; Lukito, 2020; Stukal et al., 2022) and China (Cartwright, 2020; Gray, 2021); and little is known about a World power such as the US, where most mainstream social media companies such as Twitter, Facebook, or YouTube are based.

Through executive orders, the US government can pass international sanctions designating individuals and organizations to be added to the SDN list. In turn, the assets of these individuals/organizations are to be blocked, and US citizens or organizations are prohibited to dealing with them. For example, US banks are to freeze any account or money transfer

[8](a) https://www.politico.eu/article/russia-expand-laws-criminalize-fake-news/;
(b) https://www.wired.co.uk/article/vk-russia-democracy

[9]https://edition.cnn.com/2022/03/15/politics/biden-us-officials-russia-sanctions/index.html

involving these individuals/organizations. Social media companies based in US soil are not only expected to delete the accounts of those in the SDN list, but also to suspend any account who engage with these users (O'Sullivan and Moshtaghian, 2020) – although it is often unclear what constitutes a form of relevant engagement. This is a good reflection of what Balkin (2017) describers as the "new school of speech regulation". Contrary to the "old" model, where governments were directly involved, mostly through their judiciary branch, in censoring publishers and speakers; in this "new" public-private model, governments "seek to coax the infrastructure provider into helping the state in various ways" (Balkin, 2017, p.1179). This new speech regulation paradigm raises many normative and democratic concerns. For example, as Balkin (2017) points out, from a First Amendment perspective, it raises many legal concerns, as the "enforcement of community norms [by e.g. social media companies] often lacks notice, due process, and transparency" (p.1997). In addition, it also promotes "collateral censorship", as companies rather error on the side of caution and suspend accounts who could be potentially violating a government mandate, even if they are not certain. Anecdotal evidence suggests that this is often the case. For example, right after the killing of General Qassem Soleimani by a US-drone strike, the International Federation of Journalists reported that the accounts of many Iranian journalists covering the event (and their posts) had been suspended (IFJ, 2020).

Based on the aforementioned information regarding the different kinds of pressure the US governments exerts over US-based social media companies for suspending accounts who may be involved in promoting the views of a geopolitical rival country, and given that social media companies are known for erring on the side of caution (Balkin, 2017), in this paper we expect the political views of the Twitter users interested in Iranian politics that we study, to be predictive of suspension. We measure the ideology of the users in a reformist-principlist (left-right) continuum, as well as their level of support for the Iranian government (which claims to be neutral), and put forward the following hypotheses:

H_5 Higher **principlist (conservative)** scores will be predictive of suspension.

H_6 Higher levels of **support for the Iranian government** will be predictive of suspension.

4 Data and Methods

There are many challenges to the study of deplatforming biases (Rogers, 2020). First, some platforms (e.g. Facebook) do not allow independent researchers to collect and analyze user-level data for ordinary users, making it impossible to study deplatforming beyond the suspension of a few salient users/groups. Second, even when looking at platforms that do allow for the study of ordinary accounts (e.g. Twitter), suspensions are likely to be rare, and so a large sample of interest needs to be drawn in order to be able to detect meaningful variations. In addition, behavioral traces for the users of interest need to be collected in a continuous fashion, as data becomes unavailable when a given user is suspended. Finally, accounts may be suspended for many reasons, such as those described in $\mathbf{H}_{1-4}$. Hence, researchers interested in exploring potential political suspension biases need to find ways to control for many additional confounders.

4.1 Sampling

In order to assess the effects of deplatforming on political conversations related to Iranian politics, we needed to find a set of politically-interested users to study. We followed a procedure similar to Barberá et al. (2019). First, we identified the accounts of a group of Iranian elites on Twitter: the Iranian Supreme Leader, all members of Iran's 10th Parliament ($N = 136$), cabinet members of the Rouhani administration ($N = 20$), Iranian news media outlets ($N = 19$), for a total of 176 elite accounts. Then, we pulled the list of followers for each of these elite accounts (a total of 2,410,543 unique followers). To make sure these followers were indeed

interested in politics, we sampled users that followed at least 3 of the 176 elite accounts for our analysis (601,940 users in total).

4.2 Data collection

We tracked the activity of these users between March 11th and September 10th, 2020. We collected all the tweets these users published in 2020 (a total of 65,120,890), and we tracked which accounts became inactive ($N = 7,088$). On October 22nd 2020, we manually checked every inactive account to examine whether they were: (a) *deleted* ($N = 3,351$), (b) *suspended* ($N = 2,491$), or (c) active again ($N = 1,246$, *temporary suspensions*). We dropped the *deleted* accounts from the analysis as we did not know whether those had been suspended by Twitter or by the users themselves. In addition, given that we study suspensions that took place in 2020, we focus on analyzing the messaging behavior of the users during that year, disregarding those users who did not tweet in 2020; for a final analytical sample of 2,151 suspended and 168,940 non-suspended user (171,091 in total).

4.3 Ideology

Our main objective was to assess ideological biases in the suspension of these Twitter accounts. We measured two key ideological dimensions in Iranian politics: where do users fall in the left-right (Reformist-Principlist) spectrum, and how supportive of the Iranian government the users are (which claims to not align with the stances of the different Reformist-Principlist factions in the Parliament).

To measure the ideology of the users in the Reformits-Principlist spectrum, we adapted to the Iranian context a validated and widely used method (*Bayesian Spatial Following model*) for measuring the ideology of elite and ordinary Twitter users in a single left-right dimension (Barberá, 2015). We use these user-level ideology scores to test $\mathbf{H}_5$. The model has been

validated and found to produce accurate ideology estimates for Twitter users in the US context. We provide further details and validate the method in Appendix A, where we show that the resulting ideology scores do a good job at distinguishing between known left-leaning (Reformist) and right-leaning (Principlist) elite accounts in our dataset.

We used a text-based machine learning method to measure the extent to which the accounts were supportive of the Iranian government. We trained a binary BERT multilingual model to distinguish political from non-political tweets, and then another binary BERT multilingual model to distinguish between political messages that expressed support for the Iranian government from messages that expressed a criticism of the government. Finally, we used these model predictions to generate two user-level variables, namely, the amount of political tweets sent in 2020, and the average predicted support for the Iranian government expressed in the politically-relevant tweets (average probability between 0-1). We use the latter to test $\mathbf{H}_6$.

In Table 1 we report the performance of these models (*Political* and *Pro-Iran*), based on three-fold cross-validation on an untouched held-out validation set. The classifiers are highly precise as they correctly predict political and pro-Iran messages 8/10 and 2/3 of the time, respectively; and they also do a good job at detecting most of the political and pro-Iran messages in the dataset (83% and around 70% recall). In Appendix C we provide further information about the manual annotation of the training dataset, as well as the training of the BERT models.

Table 1: Out-of-sample performance of 3 BERT-multilingual models predicting political, hateful, and pro-Iran tweets.

	Labeled	**True Negative**	**True Positive**	**Epochs**	**Accuracy**	**Precision**	**Recall**	**F-Score**
Political	2,893	64%	44%	8	83%	81%	83%	82%
Pro-Iran	607	61%	39%	11	72%	63%	69%	66%
Hateful	1,228	90%	10%	11	93%	66%	46%	53%

4.4 Controls

We also measured a set of user-level controls, in order to test $\mathbf{H}_{1-4}$ and to control for the possibility that accounts are suspended for using hateful language, for engaging in coordinated actions and automated behavior, and for spreading misinformation (Chowdhury et al., 2021; Mackey et al., 2021; Yang et al., 2022; Jaidka, Mukerjee, and Lelkes, Forthcoming).

4.4.1 Hateful content

We fine-tuned another BERT multilingual model to build a binary text classifier predicting whether a message used hateful language. We used the model to create a user-level variable measuring the number of hateful tweets sent by each user in 2020, and use this variable to test $\mathbf{H}_1$. In Table 1 we report the performance of this machine learning classifier. The model is only able to capture about half of the hateful messages in our dataset (about 50% recall) – which means that we will be running conservative tests for $\mathbf{H}_1$. However, although hateful language is very rare in our training data (only 10% of the manually annotated tweets), the model correctly predicts hateful messages 2/3 of the times. We provide further details about the training of this BERT model in Appendix C.

4.4.2 Coordination and Bots

Building on the premise that coordinated accounts post/share very similar (if not the same) content (Green, 2018; Lukito, 2020), we developed a four-step protocol to measure the similarity between the content (tweet text) posted by all possible pairs of users (see details in Appendix E), and created a user-level variable that ranges between 0 and 1 to measure the *average* content similarity (and so likely coordination) between a given user and all the others users in the dataset; and use this measure to test $\mathbf{H}_4$.

In addition, we controlled for automation of accounts in our dataset. Unfortunately, widely

used off-the-shelf tools for bot detection (e.g. *Botometer*) have been recently shown to underperform, particularly in non-English contexts (Rauchfleisch and Kaiser, 2020). Hence, rather than using an off-the-shelf bot-detection model, in our analysis (and to test $\mathbf{H}_3$) we include a set of user-level variables that previous studies have found to be effective at distinguishing bot *v.* human accounts. In particular, we include a set of user-level variables that Bastos and Mercea (2019), Majo-Vazquez et al. (2021) and/or Stukal et al. (2022) have found to be predictive of an account being a *bot*: number of tweets sent by the user (and also whether the user is in the 90th percentile of our s ample in terms of tweeting volume), average daily tweets sent by the user since the creation of the account, the ratio of the number of followers over the number of friends, and the proportion of tweets sent in 2020 that are retweets. We also include a set of variables that this previous literature has found to be predictive of an account being *human*: number of days since the creation of the account, the entropy of the software used for tweeting in 2020, the proportion of tweets sent in 2020 that contain at least one #hashtag, the proportion that are directed at another @user, and whether the user has sent at least one geo-located tweet. And finally, one variable for which existing literature reports mix-findings, some showing that is predictive of an account being a bot (Bastos and Mercea, 2019) and others finding that is predictive of an account being a human (Stukal et al., 2022): whether a user has sent at least one tweet through the web client API.

4.4.3 Misinformation

Given that during the period of our research the platforms were mainly concerned about the spread of misinformation related to COVID-19, in order to test $\mathbf{H}_2$, we focused on identifying users in our data that engaged in spreading misinformation on COVID-19. In particular, we created a user-level variable measuring the number of tweets posted in 2020 that contained one or more hashtags from a set of hashtags that we had previously identified as related to COVID-19 misinformation (see Appendix D for further details).

4.5 Limitations

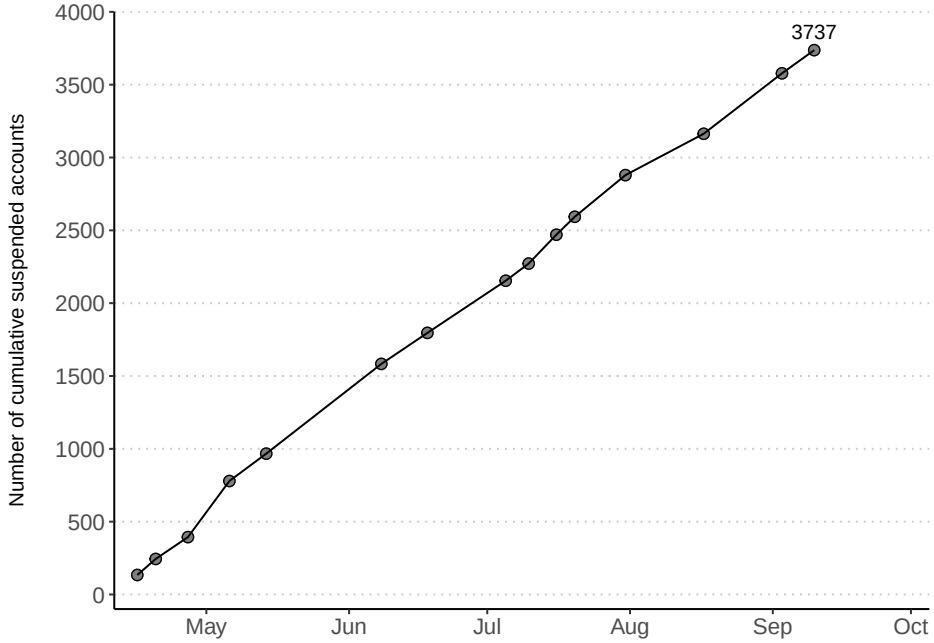
Before presenting the results, we acknowledge that this study is subject to some limitations. First, our analysis is based on one platform (Twitter) and one country (Iran), and so further research is needed to assess whether the patterns uncovered here hold in other contexts. However, we expect similar suspension patterns when it comes the regulation of content related to geopolitical rivals on US-based platforms, as they are all expected to comply with US sanctioning plans. Second, given the observational nature of the study, omitted variable bias is always a concern. Nevertheless, we have developed many measures that allow us to control for the alternative explanations put forward by previous literature. In addition, in Appendix B we show that the key results are robust to different model specifications. Finally, we are not able to clearly distinguish the extent to which (geo)political suspension biases are due to Twitter simply complying with US law and sanctions, or whether the company is erring on the side of caution by suspending any account who may be potentially violating the government mandate. We hope that future research will be able to disentangle more clearly the particular mechanism at hand. However, we believe that the research presented here represents an important step towards building a better understanding of the geopolitical relevance of social media communications, and political content moderation more broadly.

5 Results

In Figure 1 we show the number of cumulative suspensions detected among the 601,940 users that we tracked, a total of 3,737. Each dot corresponds to a moment in time when we checked whether the accounts were still active. About 0.7% of the users were suspended during the period of our analysis, which represents a non-trivial amount of accounts. In addition, the clear linear trend in Figure 1 suggests that Twitter assesses historical data and suspends accounts incrementally in batches, and that we would have most likely uncovered a larger number of

suspensions if we had tracked the accounts for a longer period of time.

Figure 1: Cumulative number of accounts that we tracked and were suspended during the period of analysis.



Clear differences emerge already when simply comparing the suspended and non-suspended users on many relevant descriptives (see Table 2). First, at the top of Table 2 we show the results for the variables that existing literature finds useful for distinguishing bot from human accounts (Bastos and Mercea, 2019; Majo-Vazquez et al., 2021; Stukal et al., 2022). We observe some patterns that are consistent with this existing literature and that suggest that some of the accounts were indeed suspended for engaging in bot-like activity (as predicted by $\mathbf{H}_3$). On average, suspended users had been in the platform for a shorter period of time (1.067 days *v.* 1.337 for non-suspended users), they posted at a much higher rate in 2020 (1.514 tweets *v.* 396), a higher proportion of suspended users were in the 90th percentile in terms of tweeting volume in 2020 (39% *v.* 10%), they had sent a higher number of daily posts since the creation of the accounts (7.12 *v.* 1.32), they had a larger number of followers compared to

Table 2: Descriptive statistics (with 95% confidence interval) for Suspended and Non-Suspended users. The gray cells indicate statistically significant differences at the 0.05 level, based on t-tests.

	Non-Suspended	**Suspended**
Potential predictors of bot or human accounts		
Avg. Number of days since account creation	1337 [1332-1342]	1067 [1023-1111]
Avg. daily posts (since creation)	1.32 [1.28-1.35]	7.12 [6.36-7.87]
Avg. Follower/Friend ratio	2.3 [1.7-2.9]	111.7 [55.66-167.74]
Avg. Entropy of platform use (2020)	0.2 [0.19-0.2]	0.2 [0.19-0.22]
Avg. Proportion of tweets at somebody (2020)	0.48 [0.47-0.48]	0.46 [0.45-0.47]
Prop. of Geo-enabled accounts (2020)	0.03	0.02
Avg. Proportion of tweets with a hashtag (2020)	0.21 [0.21-0.21]	0.23 [0.22-0.24]
Avg. Proportion of retweets (2020)	0.23 [0.23-0.23]	0.27 [0.26-0.29]
Prop. using Twitter Web Client platform (2020)	0.04	0.02
Avg. Number of tweets (2020)	396 [390-402]	1514 [1421-1606]
Prop. in the 90th most active percentile (2020)	0.10	0.39
Other covariates of interest		
Prop. of verified users	0.003	0.001
Avg. Number of political tweets (2020)	153 [151-156]	562 [522-603]
Avg. Prop. of political tweets (2020)	0.35 [0.35-0.35]	0.37 [0.36-0.38]
Avg. Number of hateful tweets (2020)	1 [1-1]	7 [6-8]
Avg. Prop. of hateful tweets (2020)	0.008 [0.008-0.008]	0.006 [0.005-0.008]
Avg. Number of Covid-Misinfo tweets (2020)	0 [0-0]	1 [1-2]
Avg. Coordination score {0-1}	0.947 [0.947-0.948]	0.974 [0.972-0.975]
Avg. Prop. In favor of Iranian government {0-1}	0.425 [0.424-0.426]	0.455 [0.448-0.462]
Avg. Principlist (Conservative) score {0-1}	0.164 [0.164-0.165]	0.18 [0.175-0.185]

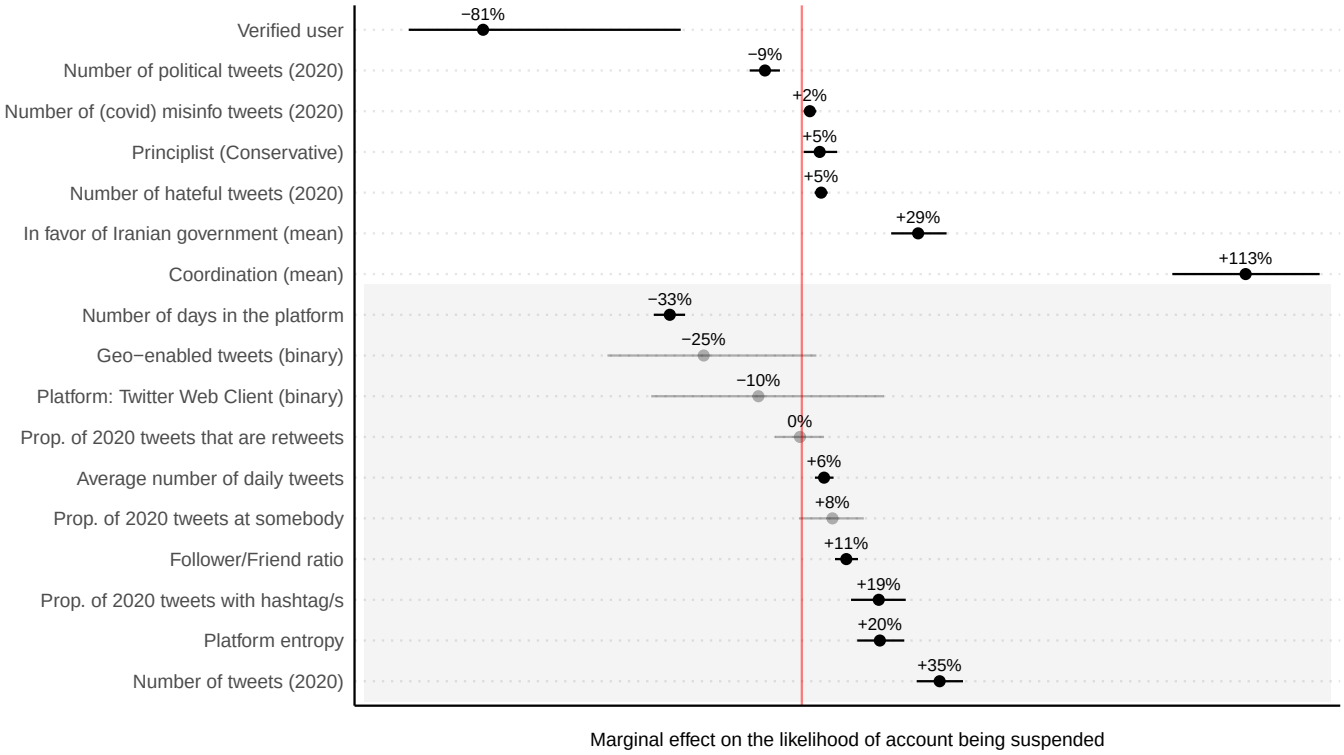
friends (111.7 follower/friend ratio *v.* 2.3), a lower proportion tweeted at least one geo-located message (2% *v.* 3%), they sent a lower proportion of tweets at somebody (46% *v.* 48%), a higher proportion of retweets (27% *v.* 23%), and a lower proportion sent at least one tweet using the Twitter Web Client platform (2% *v.* 4%). There are only two findings regarding these potential predictors that are not consistent with existing research: Stukal et al. (2022) found the proportion of tweets with hashtags to be predictive of human accounts (but we find higher proportion among non-suspended users) and we do not find any difference between suspended and non-suspended accounts in terms of the entropy of platforms used for posting messages.

The comparisons for the remaining covariates of interest are consistent with our initial expectations. In line with $\mathbf{H}_1$, we observe that suspended users sent a larger number of tweets containing hateful language in 2020 (7 *v.* 1). Although this is in part explained by the fact that they also sent many more tweets: the average proportion of tweets that were hateful was actually similar for both groups (between 6 and 8%). We also observe that suspended users sent more tweets containing the COVID-related misinformation hashtags that we had identified (1 *v.* 0), which aligns with our $\mathbf{H}_2$. In line with $\mathbf{H}_4$ we also observe higher coordination scores among suspended users (0.98 *v.* 0.95).

More importantly, these comparisons also reveal substantive ideological differences. In the last two rows of Table 2 we observe suspended users to be substantially more supportive of the Iranian government ($\mathbf{H}_6$) as well as more ideologically conservative ($\mathbf{H}_5$). On average, we observe for example that 46% of the political tweets posted by suspended users expressed support for the Iranian government, compared to 43% for non-suspended users; and suspended users to be more conservative on average (0.18 in a 0-1 index where higher values indicate higher conservatism; *v.* 0.16 for non-suspended users).

We provide more stringent evidence for these differences in Figure 2, where we show the results of a multivariate logistic regression predicting suspensions. In particular, we show the marginal effect (expressed as changes in the likelihood of suspension) of a one standard deviation change for numeric variables; and of being a verified, geo-locating at least one tweet in 2020, using the Twitter Web Client platform at least once, and so forth, for the remaining binary variables in the model. In line with Table 2, $\mathbf{H}_3$, and the aforementioned literature on social media bots, we find several of the potential identifiers of (human) bot behavior to be predictive of an account (not) being suspended. For example, having been in the platform for longer negatively predicts suspension (-33%), and a larger tweeting volume (measured as the average number of daily tweets, +6%, as well as the number of tweets sent in 2020, +35%) and a higher follower/friend ratio (+11%) predict suspension.

Figure 2: Logistic regression predicting whether an account was suspended. Marginal effects expressed in percentual change (%). Note: *The variables at the bottom of the figure, in the gray area, are potential predictors of bot (v. human) activity.*



The model results are also supportive of all the other hypotheses. In line with $\mathbf{H}_1$, we observe a one standard deviation increase in hateful tweets to be predictive of a 5% increase in the likelihood of suspension. As expected in $\mathbf{H}_2$, a similar increase in the number of tweets containing COVID-related misinformation is also predictive of an increase in the likelihood of suspension (+2%). The same applies to those with higher levels of coordination ($\mathbf{H}_4$): a one standard deviation increase in the average similarity of the textual tweet content posted by a user, and the content posted by the other users in the dataset, is predictive of a +113% increase in the likelihood of suspension. This is actually the stronger predictor of suspension in the model, denoting that Twitter is particularly concerned about politically-engaged accounts posting content in a coordinated fashion (remember that these are all accounts that follow at

least three Iranian political elites); potentially as a way to uncover information operations.

More importantly, we also find that, after controlling for the many confounders in the model, the two ideological measures of interest (conservatism and support for the Iranian government) are also predictive of suspension; findings that are robust to many model specifications (see Appendix B). A one standard deviation increase in conservatism (Principlism) is correlated with a 5% increase in the likelihood of suspension. The same increase in support for the Iranian government is also predictive of a 29% increase in the chances of being suspended. Overall, the model results show that first, accounts are in part suspended to reduce toxic and malicious behavior and improve the health of the platform. However, the findings also show some clear political biases in the suspension of users. These political biases are likely to be the result of Twitter suspending accounts that mention, engage with, praise, etc., people and organizations included in the SDN list (e.g. IRGC or Qassem Soleimani). In Table 2 and Figure 2 we observe that by doing so, Twitter is influencing which ideological views get to have a voice on the platform; advancing to some extent the geopolitical interests of the United States. The Principlists (conservatives), as well as those supportive of the Iranian government, particularly support a tougher Iranian foreign policy at the international arena, specially *vis-a-vis* the United States.

To shed more light on these ideological biases, in Figure 3.A we analyze the content of the tweets and explore the hashtags most often used by suspended *vs.* non-suspended users. For each hashtag used by any of the users under analysis, we first calculated the proportion of unique suspended and non-suspended users who used the hashtag in any of their tweets in 2020, and then calculated the difference between the suspended and non-suspended proportions. In Figure 3.B we analyze their networks and use the same procedure (comparing the proportion that follows each elite) to explore which elite accounts are most often followed by suspended *vs.* non-suspended users. The positive (and red) bars are hashtags and elite accounts most often used/followed-by suspended, and the green ones are most often used/followed-by non-

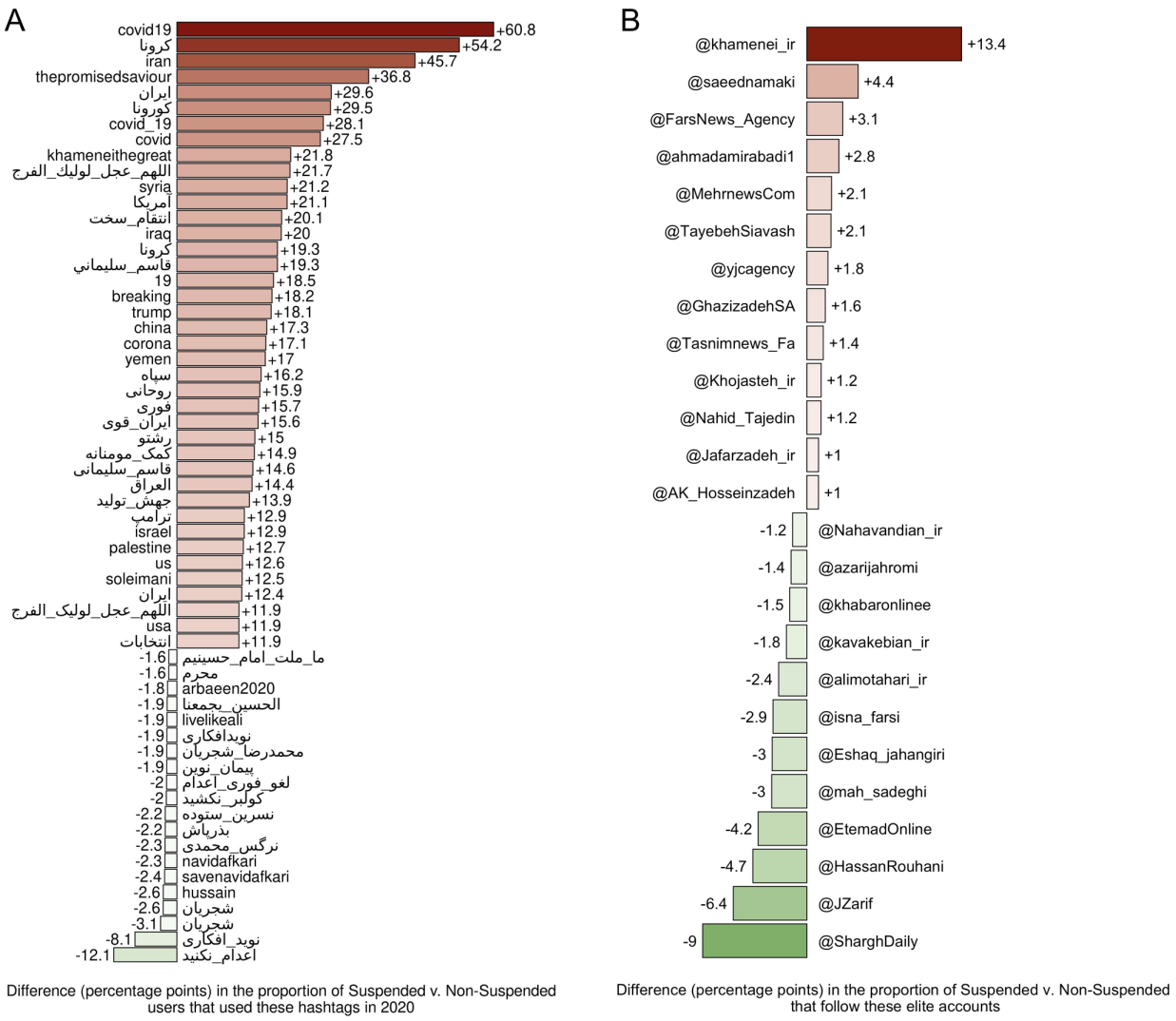
suspended users.

Figure 3.A illustrates the type of content that was to some extent repressed *vs.* emphasized as a result of the suspensions. First, we observe that (at least some) suspended users posted about COVID-19 at a much higher rate than non-suspended users. Many of the hashtags at the top of Figure 3.A are related to coronavirus, such as کرونا, covid, and covid19. In line with Table 2 and Figure 2, we believe this reassures the idea that some of the accounts were suspended for spreading misinformation on this topic.

Also, we observe many relevant political and ideological differences. Among the hashtags most often used by the suspended users, some are about General Qassem Soleimani (e.g. قاسم■سلیمانی) and some priase the Supreme Leader of Iran (khameneithegreat). We also observe other hashtags representing some of the common Principlist narratives, such as ایران■قوی (*strong Iran*) and جهش■تولید *(production growth)*. On the contrary, we can see that many hashtags that indicated opposition to the Iranian government were disproportionally used by non-suspended users, which were amplified to some extent as a result of the suspension of pro-Iranian government accounts. For example, hashtags against the execution of Navid Akfari, who was executed in 2020 for murdering a security guard in 2018, such as اعدام■نکنند (*do not execute*), نوید■افکاری (*Navid Afkari*), savenavidafkari, and navidafkari.

We find similar ideological biases in Figure 3.B. Among the most-followed elite accounts by the suspended users, we find the Supreme Leader of Iran (Ayatollah Seyyed Ali Khamenei) as well as some conservative media outlets, including *Tasnim News* and *Fars News Agency.* On

Figure 3: Differences in hashtag usage (A), and elite following (B), between suspended and non-suspended users.



the contrary, among the most-followed elite accounts by the non-suspended users, we identify Reformist media outlets (e.g., *Shargh Daily*) and figures such as Iran's former President Hassan Rouhani and some of his cabinet members, including Mohammad Javad Zarif, the former Iranian Foreign Minister, who was the chief diplomat in the negotiations over Iran's nuclear program between 2013 and 2015. Generally speaking, what distinguishes Principlists from

Reformists in terms of foreign policy is that whereas Iranian Reformists seek closer ties with the West, and the US in particular, Principlists seek to promote a tougher and sovereigntist foreign policy approach, especially with regards to Iran's defense and nuclear program.

6 Conclusion

Social media platforms are increasingly becoming important for politics: an increasing number of citizens around the world use such platforms to consume news, learn about politics, and engage in politics. To combat malicious behavior, the platforms suspend accounts that e.g. use hateful language and/or spread misinformation. In recent years, however, accusations of politically-motivated censorship have been leveled at Western social media platforms, such as Facebook and Twitter. We addressed this question from a geopolitical perspective. Although there has been much research on how non-Western countries (ab)use social media for (geo)political reasons in relation to Russia and China, little is known about how a Western country such as the United States can leverage its international sanctioning plans to push US-based social media companies to suspend accounts (and influence political conversations) to advance its geopolitical interests.

For a six-month period in 2020, we tracked about 600,000 Twitter accounts interested in Iranian politics. About 4,000 of them had been suspended after the period of analysis. We find two overarching patterns when comparing suspended and non-suspended accounts, and when using multivariate regressions to model suspension. First, accounts that engaged in different kinds of toxic/malicious behavior (e.g. used uncivil and hateful language, spread misinformation, and are suspected to be automated bots) were more likely to be suspended. Yet, after accounting for these confounders, we found clear ideological suspension biases: Principlists (conservative) accounts and those supportive of the Iranian government were also more likely to be suspended. An analysis of the content and networks of suspended (vs. non-suspended)

users indicated that these suspensions may contribute to advance the geopolitical interests of the US, amplifying voices critical of the Iranian government to the detriment of voices supportive of the government and a strong stance against the US in the international arena.

This research makes many relevant contributions to the emergent literature on political deplatforming. First, by emphasizing its geopolitical role, it provides (and illustrates) a clear theoretical framework and expectations about the conditions under which accounts may be suspended. The Russian social-media information operations in the last few US elections, and the social media bans from Western countries and Russia as a result of the Ukraine crisis, highlight the relevance of social media for public diplomacy in the current digital environment. This paper advances our understanding of the size of the problem, and the extent to which geopolitically-motivated suspensions shape political conversations in the platform. Second, the paper contributes crucial empirical evidence to the theoretical and normative debate on new forms of (political) speech regulation, or as Balkin (2017) describes it, the "new school of speech regulation". Whereas in the past governments were directly involved in censoring publishers and speakers (in most cases with the judiciary branch playing a key role), this new private-public model of speech regulation raises many legal and normative concerns. We expect the findings presented here to spearhead further debates in this area. Finally, the paper puts forward a research design that not only allows for clear comparisons between suspended and non-suspended accounts, but that it also does not rely on data made available by the platforms, which we never fully know how it was curated. We hope for future research to leverage similar designs to explore potential political-suspension biases (or lack thereof) in many additional contexts and platforms, in order to build a better understanding of the conditions under which social media suspensions may shape political conversations around the globe.

References

Baldwin, David A. 2000. "Success and failure in foreign policy." *Annual Review of Political Science* 3(1): 167–182.

Balkin, Jack M. 2017. "Free speech in the algorithmic society: big data, private governance, and new school speech regulation." *UCDL Rev.* 51: 1149–1210.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political analysis* 23(1): 76–91.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. "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(4): 883–901.

Barrie, Christopher, and Alexandra A Siegel. 2021. "Kingdom of trolls? Influence operations in the Saudi Twittersphere." *Journal of Quantitative Description* 1: 1–41.

Bastos, Marco. 2021. "This Account Doesnt Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate." *American Behavioral Scientist* 65(5): 757–773.

Bastos, Marco T., and Dan Mercea. 2019. "The Brexit Botnet and User-Generated Hyperpartisan News." *Social Science Computer Review* 37(1): 38–54.

Bay, Sebastian, and Rolf Fredheim. 2019. *Falling Behind: How Social Media Companies Are Failing to Combat Inauthentic Behaviour Online.* NATO StratCom COE.

Cartwright, Madison. 2020. "Internationalising state power through the internet: Google, Huawei and geopolitical struggle." *Internet Policy Review* 9(3): 1–18.

Chowdhury, Farhan Asif, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha, and Abdullah Mueen. 2021. Examining Factors Associated with Twitter Account Suspension Following the 2020 U.S. Presidential Election. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* ASONAM '21 New York, NY, USA: Association for Computing Machinery p. 607612.

Cockerell, Isobel. 2020. "Instagram shuts down Iranian accounts after Soleimanis death.".

Crasnic, Loriana, Nikhil Kalyanpur, and Abraham Newman. 2017. "Networked liabilities: Transnational authority in a world of transnational business." *European Journal of International Relations* 23(4): 906–929.

Davalos, J., and B. Brody. 2020. "Facebook, Twitter CEOs Sought by Senate Over N.Y. Post Story." *Bloomberg:*
https://www.bloomberg.com/news/articles/2020-10-15/facebook-twitter-chided-anew-by-republicans-over-ny-post-story .

DeNardis, L., and A.M. Hackl. 2015. "Internet governance by social media platforms." *Telecommunications Policy* 39(9): 761–770.

Eady, Gregory, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2020. "News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians." *OSF Preprint: osf.io/ch8gj* (Nov).

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Golovchenko, Yevgeniy. 2020. "Measuring the scope of pro-Kremlin disinformation on Twitter." *Humanities and Social Sciences Communications* 7(1): 1–11.

Golovchenko, Yevgeniy. 2022. "Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media." *The Journal of Politics* 84(2): 639–654.

Golovchenko, Yevgeniy, Cody Buntain, Gregory Eady, Megan A Brown, and Joshua A Tucker. 2020. "Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election." *The International Journal of Press/Politics* 25(3): 357–389.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The dynamics of protest recruitment through an online network." *Scientific reports* 1(1): 1–7.

Gray, Joanne Elizabeth. 2021. "The geopolitics of" platforms": The TikTok challenge." *Internet policy review* 10(2): 1–26.

Green, JJ. 2018. "Tale of a Troll: Inside the Internet Research Agencyin Russia." *Washington's Top News: https://wtop.com/j-j-green-national/2018/09/tale-of-a-troll-inside-the-internet-research-agency-in-russia/* .

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363(6425): 374–378.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science advances* 5(1): eaau4586.

Hashemi, Layla, Steven Wilson, and Constanza Sanhueza. 2022. "Five Hundred Days of Farsi Twitter: An overview of what Farsi Twitter looks like, what we know about it, and why it matters." *Journal of Quantitative Description: Digital Media* 2.

Hobbs, William R, and Margaret E Roberts. 2018. "How sudden censorship can increase access to information." *American Political Science Review* 112(3): 621–636.

IFJ. 2020. "Iran: Journalists demand end to censorship of Iranian media on Instagram.".

Jaidka, Kokil, Subhayan Mukerjee, and Yphtach Lelkes. Forthcoming. "Censorship on social media: The gatekeeping function of shadowbans in the American Twitterverse." *Journal of Communication* .

King, Gary, Jennifer Pan, and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American political science Review* 107(2): 326–343.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199).

Klein, Adam. 2019. "From Twitter to Charlottesville: Analyzing the fighting words between the alt-right and Antifa." *International Journal of Communication* 13: 297–318.

Lukito, Josephine. 2020. "Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017." *Political Communication* 37(2): 238–255.

Mackey, Tim K, Vidya Purushothaman, Michael Haupt, Matthew C Nali, and Jiawei Li. 2021. "Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter." *The Lancet Digital Health* 3(2): e72–e75.

Majo-Vazquez, Silvia, Mariluz Congosto, Tom Nicholls, and Rasmus Kleis Nielsen. 2021. "The Role of Suspended Accounts in Political Discussion on Social Media: Analysis of the 2017 French, UK and German Elections." *Social Media + Society* 7(3): 20563051211027202.

Miller, Blake, Fridolin Linder, and Walter R. Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." *Political Analysis* 28(4): 532551.

Miskimmon, Alister, Ben O'loughlin, and Laura Roselle. 2014. *Strategic narratives: Communication power and the new world order*. Routledge.

O'Sullivan, D., and A. Moshtaghian. 2020. "Instagram says it's removing posts supporting Soleimani to comply with US sanctions." *CNN: https://edition.cnn.com/2020/01/10/tech/instagram-iran-soleimani-posts/index.html* .

Pan, Jennifer, and Alexandra A. Siegel. 2020. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review* 114(1): 109125.

Rauchfleisch, Adrian, and Jonas Kaiser. 2020. "The False positive problem of automatic bot detection in social science research." *PLOS ONE* 15(10): 1–20.

Rogers, Richard. 2020. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." *European Journal of Communication* 35(3): 213–229.

Shearer, Elisa, and Amy Mitchell. 2021. "News use across social media platforms in 2020.".

Siegel, Alexandra A., Evgenii Nikitin, Pablo Barber,, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. "Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath." *Quarterly Journal of Political Science* 16(1): 71–104.

Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. 2022. "Why Botter: How Pro-Government Bots Fight Opposition in Russia." *American Political Science Review* 116(3): 843857.

Theocharis, Yannis, Pablo Barber,, Zolt,n Fazekas, and Sebastian Adrian Popa. 2020. "The Dynamics of Political Incivility on Twitter." *SAGE Open* 10(2): 2158244020919447.

Tsvetkova, Natalia, Dmitrii Rushchin, Boris Shiryaev, Grigory Yarygin, and Ivan Tsvetkov. 2020. "Sprawling in Cyberspace: Barack Obamas Legacy in Public Diplomacy and Strategic Communication." *Journal of Political Marketing* pp. 1–13.

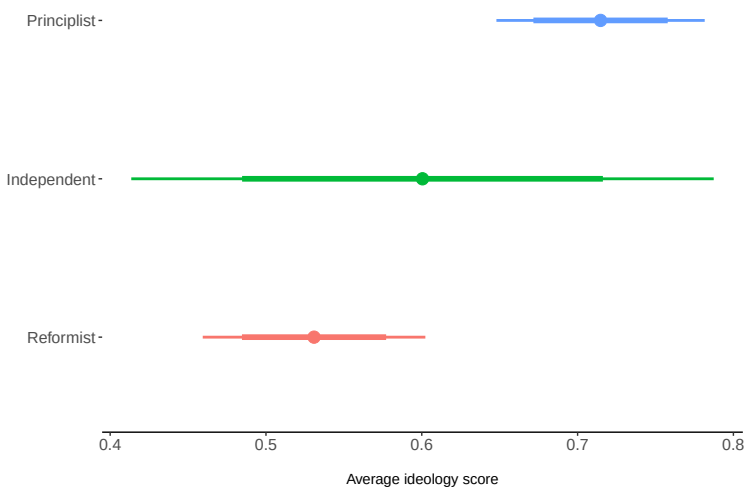
Yang, Qi, Mohsen Mosleh, Tauhid Zaman, and David G Rand. 2022. "Is Twitter biased against conservatives? The challenge of inferring political bias in a hyper-partisan media ecosystem.".

Appendix A Further details and validation of the ideology score.

To estimate the model, we first built a bipartite network graph with information about which of the 176 elite accounts each of the 601,940 users in the full sample followed. Then, in order to be able to estimate this computationally-intensive model, we randomly sampled 5,000 users, who followed most of the elite accounts in the list (152 out of 176, so 86%). Then, we used the `mediascores` package (Eady et al., 2020) to fit the model, obtaining ideology scores (in the same dimension) for the 5,000 users and the 152 elite accounts; finally we used the trained model (and so the scores given to the elite accounts) to estimate the ideology of the remaining ordinary users (based on the elites they followed), and standardized the scores between 0 and 1 (with 0 indicating extreme reformists, and 1 extreme principlists).

We conducted the following validation exercise to make sure that the *Bayesian Spatial Following model* adapted well to the Iranian context. We checked the average ideology score given by the model to the members of Parliament that are known to be affiliated to either the Reformist, Independent, and Principlist factions in the chamber. We report the average ideology scores for these three groups in Figure A1, where we observe the method to perform as expected and to generate ideological scores that do a good job at distinguishing between reformists and principlists. Also as expected, the model estimated independents to have an average ideology between the averages estimated for reformists and principlist. The confidence interval is rather large for the independents in part due to the low number of members in this group.

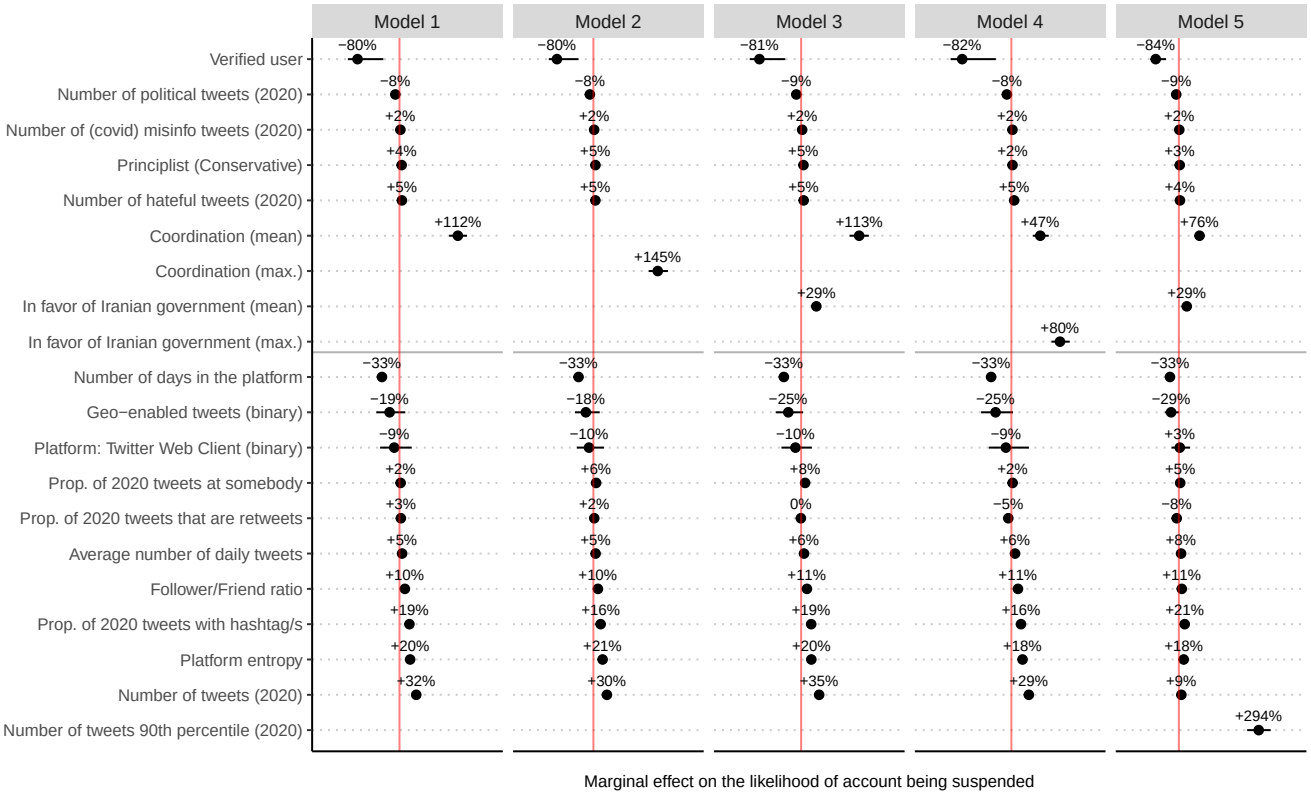
Figure A1: Average ideology score for elite accounts known to be Principlists (conservative), Reformists (liberal) and Independent.



Appendix B Model details and alternative model specifications.

In Figure 2 of the paper we present the results of a logistic regression predicting suspension as a function of the two explanatory variables of interest (ideology and support for the Iranian government) plus a set of confounders. In Figure B1 of this appendix we estimated additional logistic regressions predicting suspensions, with the goal of assessing the robustness of the findings presented in Figure 2. Model 3 in Figure B1 is the same reported in Figure 2, although for simplicity, in the version reported in this appendix we rounded the estimated effects to full integers.

Figure B1: Logistic regressions predicting whether the account was suspended. **Note:** *The variables under the horizontal gray line are potential predictors of bot (v. human) activity.*



We estimated all these logistic regression at the user level, with a binary outcome variable indicating whether a user had been (at least temporary) suspended by the end of data collection. We included the following user-level predictors in the models:

- *Verified user*: whether the user was verified – the blue check mark on Twitter indicating whether the user is a person of interest.

- *Number of (COVID-19) misinfo tweets (2020)*: the number of messages the user sent in 2020 that included (at least) one of the hashtags we identified as clearly linked to the spread of COVID-19 related misinformation (see Appendix D for further information).
- *Number of political tweets (2020)*: the number of messages the user sent in 2020 that we predicted to be about politics (see Appendix C for further information on the machine learning model used to generate the political predictions).
- *Principlist (Conservative)*: a standardized continuous score between 0 and 1 indicating the ideology of the user in a Reformist-Principlist (left-right) continuum, where higher scores indicate more Principlist/conservative users (see Appendix A for further information about the ideology scores).
- *Number of hateful tweets (2020)*: the number of tweets the user sent in 2020 that we predicted to contain hateful language (see Appendix C for further information on the the machine learning model used to generate the hateful predictions).
- *Coordination (mean)*: a continuous user-level variable, ranging between 0 and 1, measuring the *average* content/textual similarity between the tweets sent by a given user, and all the other users in the dataset (see Appendix E for further details on how this coordination score is calculated).
- *Coordination (max.)*: a continuous user-level variable, ranging between 0 and 1, measuring the *maximum* content/textual similarity between the tweets sent by a given user, and any other user in the dataset (see also Appendix E for further details).
- *In favor of the Iranian government (mean)*: a continuous user-level variable, ranging between 0 and 1, measuring the *average* predicted support for the Iranian government in *all* the political tweets sent by the user in 2020 (see Appendix C for further information on the model used to predict the probability of a political tweet to be supportive of the Iranian government). This variable is NA for users who did not send any politically-relevant tweet in 2020 (about 19.2% of the users in the dataset).
- *In favor of the Iranian government (max.)*: a continuous user-level variable, ranging between 0 and 1, measuring the *maximum* predicted support for the Iranian government in *any* of the political tweets sent by the user in 2020. This variable is NA for users who did not send any politically-relevant tweet in 2020 (about 19.2% of the users in the dataset).

In addition, we added to the models the following set of variables that previous literature has found to be predictive of bot (*v.* human) behavior (Stukal et al., 2022; Bastos and Mercea, 2019):

- *Number of tweets (2020)*: the number of messages the user sent in 2020.
- *Number of tweets 90th percentile (2020)*: whether the user is in the 90th percentile in terms of numbers of tweets sent in 2020.
- *Number of days in the platform*: number of days between the creation of the account and the day we started data collection.

- *Geo-enabled tweets (binary)*: whether the user sent at least 1 geolocated tweet in 2020.
- *Platform: Twitter Web Client (binary)*: whether the user sent at least 1 tweet through the web client API in 2020.
- *Prop. of 2020 tweets at somebody*: proportion of the tweets the user sent in 2020 that were directed at another @user.
- *Prop. of 2020 tweets that are retweets:* proportion of the tweets the user sent in 2020 that were retweets, instead of original messages.
- *Average number of daily tweets:* average number of tweets/day the user sent in 2020.
- *Follower/Friend ratio:* a ratio measuring the number of followers over the number of friends for a given user.
- *Prop. of 2020 tweets with hashtag/s:* proportion of messages the user sent in 2020 that contained at least 1 #hashtag.
- *Platform entropy:* entropy of the software platform used for tweeting in 2020 for a given user.

Table B1: Coefficient tables for the five logistic regressions presented in Fig. 2 and Fig. B1. Note: *The asterisks indicate findings that are statistically significant at least at the 0.05 level.*

Variable	Model 1	Model 2	Model 3
(Intercept)	-31.5144 (1.7906)*	-76.6569 (4.5327)*	-47.4234 (2.9689)*
Coordination (mean)	28.1086 (1.8707)*		43.9004 (3.0584)*
Coordination (max.)		72.5554 (4.5654)*	
Verified user (binary)	-1.9894 (0.8641)*	-2.0475 (0.8677)*	-2.0321 (0.8486)*
Principlist (Conservative)	0.4839 (0.232)*	0.5299 (0.2304)*	0.5043 (0.2384)*
Number of tweets (2020)	0.0002 (0)*	0.0002 (0)*	0.0002 (0)*
Number of tweets 90th percentile, binary (2020)			
Number of political tweets (2020)	-0.0001 (0)*	-0.0001 (0)*	-0.0001 (0)*
Number of hateful tweets (2020)	0.0044 (0.0007)*	0.0043 (0.0007)*	0.0042 (0.0007)*
Number of (covid) misinfo tweets (2020)	0.004 (0.0017)*	0.0041 (0.0017)*	0.0042 (0.0017)*
In favor of Iranian government (mean)			1.3388 (0.1497)*
In favor of Iranian government (max.)			
Number of days in the platform	-0.0004 (0)*	-0.0004 (0)*	-0.0004 (0)*
Average number of daily tweets	0.0069 (0.0014)*	0.0069 (0.0014)*	0.0071 (0.0014)*
Follower/Friend ratio	0.0005 (0.0001)*	0.0005 (0.0001)*	0.0005 (0.0001)*
Platform entropy	0.5482 (0.0762)*	0.5636 (0.0761)*	0.5338 (0.0789)*
Prop. of 2020 tweets at somebody	0.0581 (0.1081)	0.1675 (0.106)	0.2273 (0.1222)
Geo-enabled tweets (binary)	-0.2276 (0.1746)	-0.2141 (0.1746)	-0.3043 (0.1821)
Prop. of 2020 tweets with hashtag/s	0.7401 (0.1212)*	0.6373 (0.119)*	0.8113 (0.1332)*
Prop. of 2020 tweets that are retweets	0.0853 (0.1026)	0.0665 (0.1014)	-0.0217 (0.111)
Platform: Twitter Web Client (binary)	-0.1162 (0.1704)	-0.1135 (0.1703)	-0.1279 (0.1767)
N	171087	171087	138252
AIC	21609.67	21543.56	19761.43

Variable	Model 4	Model 5
(Intercept)	-27.9326 (2.6627)*	-37.0156 (2.823)*
Coordination (mean)	22.5693 (2.8112)*	33.0279 (2.9098)*
Coordination (max.)		
Verified user (binary)	-2.0488 (0.8528)*	-2.2104 (0.8426)*
Principlist (Conservative)	0.2104 (0.2376)	0.3444 (0.2375)
Number of tweets (2020)	0.0002 (0)*	0.0001 (0)*
Number of tweets 90th percentile, binary (2020)		1.3704 (0.0638)*
Number of political tweets (2020)	-0.0001 (0)*	-0.0002 (0)*
Number of hateful tweets (2020)	0.004 (0.0007)*	0.0035 (0.0007)*
Number of (covid) misinfo tweets (2020)	0.0038 (0.0017)*	0.0033 (0.0017)
In favor of Iranian government (mean)		1.3403 (0.1536)*
In favor of Iranian government (max.)	2.33 (0.1913)*	
Number of days in the platform	-0.0004 (0)*	-0.0004 (0)*
Average number of daily tweets	0.0076 (0.0014)*	0.0102 (0.0015)*
Follower/Friend ratio	0.0005 (0.0001)*	0.0005 (0.0001)*
Platform entropy	0.4984 (0.0789)*	0.4878 (0.0791)*
Prop. of 2020 tweets at somebody	0.0633 (0.1208)	0.1476 (0.1226)
Geo-enabled tweets (binary)	-0.307 (0.1823)	-0.3472 (0.1825)
Prop. of 2020 tweets with hashtag/s	0.6694 (0.1344)*	0.885 (0.134)*
Prop. of 2020 tweets that are retweets	-0.1818 (0.1115)	-0.2845 (0.1129)*
Platform: Twitter Web Client (binary)	-0.1071 (0.1766)	0.0165 (0.1765)
N	138252	138252
AIC	19640	19337.49

In the additional model specifications shown Figure B1 we assess the impact of the following modeling choices (coefficient tables are available at Table B1). First, given that we created two variables capturing different ideological dimensions (reformist-principlist position, and support for the Iranian government), we wanted to assess whether including only one of these into the model would yield different results than including both. The robustness of these estimates across the models reveals that these two dimensions have a separate and distinguishable effect on suspension.

We also wanted to assess the robustness of the findings to different ways in which we could have represented the following covariates of the model. We decided for the main model in Figure 2 to represent coordination as the average content/text similarity between the tweets sent by a given user, and all the other users in the dataset. However, one could also argue that what matters for suspension is to have high levels of coordination with simply one other account, and for this reason in Model 2 we model coordintion as the maximum content/text similiarity between the tweets sent by a given user and any other user in the dataset. We also see this version of the variable to be a strong (even stronger, +145%. *v.* +112%) predictor of suspension, reassuring the findings presented in Figure 2. Similarly, one could also argue that what matters for suspension is not how supportive one is of the Iranian government on average, but whether a user sent simply a tweet that is very supportive of the government. For this reason in Model 4 we model such support by using the maximum support score in any of the political tweets sent by a given user in 2020. We also see this version of the variable to be very (and even more strongly, +80%. *v.* +29%) predictive of suspension. Finally, we also wondered whether a simple count of the number of tweets sent in 2020 was enough good proxy to capture bot-like behavior. Instead, in Model 5 we use a more extreme version of this variable and use a binary version capturing whether the user is in the 90th percentile in terms of tweet

volume. We observe this binary variable to also be a very strong predictor of suspension. More importantly, the results for the remaining covariates, in particular the two key explanatory variables of interest (*Principlist (Conservative)* and *In favor of Iranian government (mean)*) are robust to these additional modeling choices. In addition, in all models we include many variables that have been found in previous literature (Stukal et al., 2022; Bastos and Mercea, 2019) to be predictive of either bot (v. human) or human (v. bot) behavior on Twitter; to rule out the possibility that the results for the other covariates of interest (i.e. ideology and support for the Iranian government) could be in part explained by the accounts being automatic bots.

Appendix C BERT multilingual models predicting political and uncivil tweets, and support for the Iranian government.

We fine-tuned three BERT multilingual models (`bert-base-multilingual-cased`) predicting whether a given tweet uses hateful language, whether it is political, and if so, whether it is supportive of the Iranian government. We used the following procedure to train each of these models.

Table B2: Examples of hateful *v.* non-hateful messages

Message coded as *hateful*
مرگ بر■خامنه■ای #IranRegimeChange# گه نخورررر صادقسی @mah_sadeghi
(EN translation) *#Death_to_Khamenei #IranRegimeChange Cut the crap Sadeghi @mah_sadeghi*
خدا شاهده همه مردم نفرينتون ميكنن بلاخره يكي از نفرين هاي ملت مظلوم ميگيره @mah_sadeghi جناب صادقي شماها نمايندگان بي عرضه روسرگ بر شما كه فكر خودتونيد فقط
(EN translation) *Swear to God all people curse you. Eventually you, incompetent MPs, will be damned by the curse of the oppressed @mah_sadeghi Mr. Sadeghi. Death to you for you only care about yourselves.*
Message coded as *non-hateful*
سردار باغیرت درود به وجدان بیدار و آگاه شما که همچو مرد اشتباه را پذیرفتی @Hajizadeh_org
(EN translation) *@Hajizadeh_org Praise upon your awake conscience, commander since you accepted responsibility for the mistake. Praise upon you*
کاش دولته بدخال حذف صفر از پول ملی بشه. مخواد ابرو درست کنه منزنه چشم شو هم کور می کنه
(EN translation) *I hope the administration does not remove zeros from the national currency. It wants to solve a problem but instead it would exacerbate it*

First, from all Farsi, Arabic and English tweets sent in 2020 by the users we tracked, we sampled some tweets at random (plus some others using an active-learning procedure (Miller, Linder, and Mebane, 2020)): 1,228 for the hateful coding, 2,893 for the political coding, and we selected 607 of the political tweets for the coding of the support of the Iranian government. Following Twitter's definition of hateful language,[10] messages were considered as being hateful if they: (1) made violent threats against an identifiable group, (2) incited fear about a group/community, (3) wished, hoped, or called for serious harm on an individual or group, and (4) made references to violent events (see some examples in Table B2). Tweets were coded as political if they (a) mentioned a policy topic (e.g., economy, foreign policy, defense, social welfare, etc.), (b) mentioned a political event and/or an institution (e.g., a national election, protest, parliament, etc.), and/or (c) mentioned a member of the political elite (e.g., a politician, military official) in the form of a reply and/or a mention (see some examples in Table B3).

[10] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Table B3: Examples of political *v.* non-political messages

Message coded as *political*
@WhiteHouse @realDonaldTrump #GhasemSoleimani was the man who swiped out isis with the help of the resistance (Iraq and Syria) a fact that every political analyst knows about, I don't know how can anybody believe what he is saying.
محمد جواد ظریف وزیر امور خارجه ایران آمریکا را بزرگترین فروشنده تسلیحات نامید. ظریف اعلام کرد: آمریکا مدت هاست که در صدر کشورهای هزینه کننده نظامی، صدر فروشندگان سلاح، صدر آغاز کنندگان و تحریک کنندگان جنگ و صدر سودجویان درگیری های جهان است.
(EN translation) *Mohammad Javad Zarif the Iranian foreign minister called the US the biggest supplier of arms. Zarif stated: The US has been the top military spender, arms supplier, and instigator and beneficiary of wars across the world for quite a while.*
Message coded as *non-political*
اثر قطرات اب بر بروی سنگ{annonymized-url}
(EN translation) *Effect of water drops on stone* {*annonymized-url*}
مناطق طبیعي شمال ایران {annonymized-url}
(EN translation) *Natural areas in the north of Iran* {annonymized-url}

Messages were coded as being in favor of the Iranian government if they voiced support for the Supreme leader and his views, the administration or any government agency (including any branch of the military), and/or took a strong stance towards defending Iran against foreign interventions (in line with the views of the current administration). Tweets were coded as being against the Iranian government if they criticized the Supreme leader, the administration or any government agency, and/or any government policy (see some examples in Table B4). A second coder annotated 100 tweets for each of these variables, resulting in a Cohen's Kappa of 0.89 for the political coding, 0.83 for the support of the Iranian government coding, and 0.72 for the coding of hateful messages.

Then we used the labeled data to fine-tuned the three BERT multilingual models (after attaching a final binary prediction layer to the model), using an AdamW optimizer and a learning rate of 1e-5. In each case we split the data into a train-test split (80-20). We trained each model until the test loss stop improving: 8 iterations for the political model, and 11 iterations for the remaining two models. We evaluated the performance of the model using 5-fold cross-validation based on the unseen test set. We report the performance of the models in Table B5. The *Labeled* column provides information about the total number of messages we labeled, and the *Negative* and *Positive* columns about the proportion of true negatives/positives that resulted from the coding exercise. For the three models, the percentage of true negatives was higher than the true positives, and so it should be used to judge the overall *Accuracy* of the model (this is the percentage of cases a naive model attributing negative/positive labels at random would get right). The percentage of positives should be used to judge the *Precision*, *Recall*, and *F-Score*, as these provide information about the percentage of predicted positive labels are indeed positives (precision), the percentage of true positives that we indeed predicted to be positive (recall), and the average of both (f-score). Overall *Accuracy* is high for all mod-

els (>70%), as well as the *Precision*: although true positives are rare, e.g. 10% for the hateful classifier, the models make correct predictions 2/3rds of the time, and perform substantially better than a naive model making random predictions, e.g. 6.6 times better (66%/10%) for the hateful classifier. *Recall* is also high for the political and pro-Iran classifiers, but slightly lower for the hateful model (also resulting on a slightly lower f-score for this model). This means that, as it happens with any machine learning model, we're measuring our quanitites of interest with some noise. However, we don't have any reason to believe that there is a correlation between the error of the model and our outcome of interest, which in practice means that we'll be running harder and conservative tests of our hypotheses.

Table B4: Examples of messages in favor *v.* against the Iranian government

Message coded as being *against* the Iranian government
@Eshaq_jahangiri[11] این چه وضع برنامه ریزیه بعد پنج ساعت معطلی در صف اجازه رای ندادن
(EN translation) *@Eshaq_jahangiri What kind of management is this. After five hours waiting in the line, they did not let voting*
ای کاش لیست قراردادهای منعقده بعد از برجام بدون ضمانت اجرای محکم منتشر میشد تا خسارتی که #دولت_تدبیر به کشور وارد کرد، روشن میشد. در بیشتر این قراردادها، تحریم به عنوان فورس ماژور محسوب شده و عملا طرف در حالی که بخشی امبلغ قرارداد گرفته، بدون هیچ خسارتی، ایران را ترک میکند
(EN translation) *I wish a list of contracts signed after the JCPOA that lack an execution guarantee would be made public so as to illustrate the damages the administration of rationality (i.e., Rouhani administration) has caused the country. In most of these contracts, sanctions are deemed as force majeure, and in practice, while the party to the contract has received part of the agreed payment, without any compensation, leaves Iran.*
Message coded as being *in favor* the Iranian government
پای رای مان هستیم #روحانی_تنها_نیست
(EN translation) *We stand by our vote. #Rouhani is not alone*
حالا یک سیلی ای دیشب در #عین_الأسد به انها زده شد، این مسئله ی دیگری است، آنچه که در مقام مقابله مهم است این کارهای نظامی به این شکل کفایت آن قضیه را نمیکند این است که بایستی حضور فساد برانگیز آمریکا در این منطقه تمام بشود. #انتقام_سخت
(EN translation) *Last night, a slap in the face was delivered in #Ein_Al-Assad, this is a different matter, military actions of this kind will not suffice. The main response involves putting an end to the corrupt presence of America in the region. #Severe_revenge*

Overall, these models performed well enough to continue with the analyses, and so we used them to generate political, hateful, and pro-Iran predictions for the rest of the unlabeled tweets in our dataset. Finally, We used these machine-labeled tweets to generate 3 user-level variables that we then included in our analyses. First, we counted the number of political tweets sent in 2020 by any of the users in our dataset. Then, we counted the number of hateful tweets sent in 2020. Finally, to measure a given user's support for the Iranian government, rather than generating binary predictions, we used the logistic nature of the machine-learning

model to generate probability predictions for each tweet (so the probability of a given tweets to be supportive of the Iranian government, ranging from 0 to 1). We then created two user-level measures: the average support for the Iranian government (averaging the tweet-level probabilities for a given user), as well as the maximum support for the Iranian government (pulling the highest tweet-level probability for a given user). We use the average measure for the main model reported in Figure 2, and the maximum measure in some of the alternative model specifications reported in Appendix B.

Table B5: Performance of 3 BERT-multilingual models predicting political, hateful, and pro-Iran tweets.

	Labeled	**Negative**	**Positive**	**Epochs**	**Accuracy**	**Precision**	**Recall**	**F-Score**
Political	2,893	64%	44%	8	83%	81%	83%	82%
Hateful	1,228	90%	10%	11	93%	66%	46%	53%
Pro-Iran	607	61%	39%	11	72%	63%	69%	66%

Appendix D Identifying covid-misinformation hashtags.

In our analysis of potential political/ideological suspension among those discussing Iranian politics on Twitter, we wanted to control for the alternative explanation that accounts may also be suspended for spreading misinformation. Detecting involvement in the dissemination of misinformation generally turned out to be incredibly challenging (most existing studies in the US context for example rely on existing lists of fake news sites and then explore how often users share links from those sites; but such lists do not exists, and are very hard to develop, for the Iranian context). Hence, we decided to focus on detecting the spread of COVID-19 related misinformation among users in our dataset, given that platforms at that time were particularly concerned about eradicating misinformation on the topic. We developed a four-step protocol to create a list of hashtags that could be clearly linked to the spread of COVID-19 misinformation. First, we generated a list of 39 keywords in Farsi, Arabic and English (the most spoken languages in our dataset) that would help us identify COVID-19 related messages (see Table D1 for a list), and then we manually annotated a random sample of 1,000 messages containing any of these 39 keywords for whether those messages contained misinformation. Next, we selected the unique hashtags in those coded as containing COVID-19 related misinformation, and went back to the full dataset to pull 10 random messages containing each of those hashtags. After coding those 10 messages per hashtag again for whether they contained misinformation, we treated as clear COVID-19 misinformation hashtags those for which at least 8 of the 10 random messages had been coded as containing misinformation. Finally, we generated the a user-level variable measuring number of messages a user sent in 2020 that contained one these 7 COVID-19 misinformation hashtags: chinesevirus, chineseviruscensorship, coronafromusa, islamicrepublicvirus, wuhanvirus, chinesevirus19, محاکمها عاملنا شیوع کرونا در ایران.

Table D1: List of 39 keywords used to generate a first sample of tweets discussing COVID-19.

keyword	language
covid	english
corona	english
virus	english
china_virus	english
chinese_virus	english
chinesevirus	english
chinavirus	english
biologic	english
bio-weapon	english
alcohol	english
کووید	farsi
ویروس ـکرونا	farsi
کرونا	farsi
کووید ـ۹۱	farsi
کووید	farsi
ویروس کرونا	farsi
کرونا	farsi
کووید ۹۱	farsi
#بیولوژیک	farsi
#جنگ ـبیولوژیک	farsi
جنگ ـبیولوژیک	farsi
بیولوژیک	farsi
متانول	farsi
الکل	farsi
كوفيد٩١	arabic
كورونا	arabic
فيروس كورونا	arabic
كوفيد٩١	arabic
كورونا	arabic
فيروس ـكورونا	arabic
الفيروس الصيني	arabic
الفيروس ـالصيني	arabic
بيولوجية	arabic
حرب بيولوجية	arabic
حرب الجرثومية	arabic
بيولوجية	arabic
حرب ـبيولوجية	arabic
حرب ـالجرثومية	arabic
كحول	arabic

Appendix E Identifying coordination.

In our analysis of potential political/ideological suspension among those discussing Iranian politics on Twitter, we wanted to control for the alternative explanation that accounts may also be suspended for acting in a coordinated fashion. To accomplish this goal, we developed a method to identify overall content/text similarity between the tweets posted by a given user, and the other users in our dataset. First, we used the same BERT multilingual model used in Appendix C to generate (768-size) tweet-level embeddings for all messages sent in 2020 by the users we tracked, by passing the tweets through the pre-trained BERT architecture and pulling the output of the second-to-last (fully connected) layer. Second, for each user we generated a (768-size) user-level embeddings by averaging the indexed embedding values of all the users tweets. Third, we calculated the cosine similarity between all possible pairs of user embeddings. And finally, we used these cosine similarities to generate two user-level variables: one measuring the *average* content/text similarity between all tweets sent by a given user, and the tweets sent by the rest of the users in the dataset (so between a given user's embedding, and the user-level embeddings for the other users); and second, the *maximum* cosine similarity between the tweets sent by a given user and the ones sent by any other user in the dataset (so between a given user's embedding, and the user-level embedding of any other user in the dataset).