Political Communication

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/upcp20

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

Andreu Casas

To cite this article: Andreu Casas (2024) The Geopolitics of Deplatforming: A Study of Suspensions of Politically-Interested Iranian Accounts on Twitter, Political Communication, 41:3, 413-434, DOI: 10.1080/10584609.2024.2306503

To link to this article: https://doi.org/10.1080/10584609.2024.2306503

POLITICAL COMMUNICATION
2024, VOL. 41, NO. 3, 413–434
https://doi.org/10.1080/10584609.2024.2306503

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

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

Andreu Casas*

Department of Politics, International Relations and Philosophy, Royal Holloway University of London, London, United Kingdom

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. This study contributes to this ongoing debate by looking at content moderation from a geopolitical perspective. The starting premise is that US-based social media companies may be 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. I tracked 600,000 users who follow Iranian elites on Twitter. After accounting for alternative explanations, the results show that Principlist (conservative) users and those supportive of the Iranian government are significantly more likely to be suspended. Further analyses uncover the types of discussions that are being suppressed as a result of these suspensions. Although the exact mechanism at hand cannot be decisively isolated, this paper contributes to building a better understanding of how governments can influence conversations of geopolitical relevance, and how social media suspensions shape political conversations online.

KEYWORDS

Social media; content moderation; deplatforming; geopolitics; Iran

Introduction

Today, private social media companies play a crucial role in moderating freedom of speech (Balkin, 2017; Gillespie, 2018). People around the world increasingly rely on social media to consume news (Shearer & 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 (Bastos, 2021; Bay & Fredheim, 2019; DeNardis & Hackl, 2015). In addition, in recent years many have claimed that widely-used

CONTACT Andreu Casas andreu.casas@rhul.ac.uk Department of Politics, International Relations and Philosophy, Royal Holloway University of London, (McCrea 1-06), Egham, Surrey TW20 0EX, United Kingdom

*Replication data and code available at https://github.com/CasAndreu/twitter-iran-moderation.

Supplemental data for this article can be accessed on the publisher's website at https://doi.org/10.1080/10584609.2024.2306503.

platforms such as Facebook, YouTube and Twitter suspend accounts for political reasons, allegedly targeting conservatives in US politics (Davalos & Brody, 2020) as well as voices supportive of governments involved in a geopolitical rivalry with the West, such as China, Russia, Venezuela and Iran (Cartwright, 2020; O'Sullivan & Moshtaghian, 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.

This study focuses 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 politics at home and abroad. Several studies have explored how governments leverage social media to constrain political speech at home, such as China (King et al., 2013, 2014) and Saudi Arabia (Pan & Siegel, 2020). Some other scholars have researched the ways in which non-Western governments (e.g. Russia) leverage social media communications to influence public opinion abroad (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 the interests of Western governments. For example, in a recent review of digital repression tools, Earl et al. (2022) argue that "although autocrats certainly draw on many forms of digital repression, our review clearly shows that democracies engage in almost all forms of digital repression too" (p.9). The United States is of particular relevance in this context, as some of the most popular and globally used social media platforms are based in the country.

Governments can leverage social media for various geopolitical purposes, such as conducting foreign surveillance (Gray, 2021), promoting their own narratives (Barrie & Siegel, 2021; Stukal et al., 2022; Golovchenko et al., 2020), and suppressing opposing viewpoints (Golovchenko, 2022). In this way, social media can serve as a powerful tool for governments to advance their geopolitical interests. This study focuses on the latter, and discusses how the US may condition US-based social media platforms to deplatform opposing geopolitical views. In particular, the study looks at 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 is of particular relevance because it has been a significant focus of geopolitical conflict for many years and has implications for many other relevant countries such as Russia, China, and the UK. Although Twitter is blocked in Iran, millions of Iranian citizens, including members of Parliament and top government officials, use VPNs and other methods to access and actively use the platform, 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 et al. (2022).

When a social media platform with a global reach is based in a particular country, that government can potentially use the legal system to condition the platform to implement certain content moderation policies with the goal of shaping political conversations abroad – and/or shape conversations of geopolitical interest (Balkin, 2017; Cartwright, 2020; Crasnic et al., 2017; 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. Several 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 Iran's Quds Force, an elite branch of the IRGC. According to a Meta spokesperson, in order to comply with US sanction laws, Instagram and Facebook suspended accounts of users that condemned the assassination or simply covered the story (O'Sullivan & Moshtaghian, 2020): "we operate under US sanctions laws, including those related to the US government's designation of the IRGC and its leadership." 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 (transparent) information available on the scope of these account suspensions and their overall impact on political discussions related to Iran on the platform.

In March 2020, I identified 601,940 users who followed Iranian elites on Twitter, and for a six-month period, periodically collected the messages they posted in the platform and checked whether they had been suspended. Most of the accounts remained *active* after the period of analysis, yet many were (at least temporarily) *suspended* ($N = 3{,}737$). I 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, the results show many toxic behaviors (e.g. using hateful language, spreading misinformation, and bot-like behavior) to be predictive of suspension. More importantly, conservative users and those supportive of the Iranian government are also more likely to be suspended. An analysis of the content more often discussed by non-suspended (v. suspended) users reveals that accounts engaging with more progressive discussions (e.g. criticizing certain actions and policies of the Iranian government) and networks (e.g. private media) are suspended at lower rates, whereas accounts criticizing the killing of General Soleimani and asking for a stronger position of Iran in the international arena are suspended at higher rates.

Unfortunately the nature of the data does not allow to clearly isolate the exact mechanism at play. Anecdotal evidence, such as the above-mentioned statement by a Meta spokesperson (O'Sullivan & Moshtaghian, 2020), or Facebook's Community Standards,[2] point to US-based platforms indeed suspending some Iranian accounts in compliance of US sanction laws. However, it is hard to disentangle whether companies do so based on their own interpretation of these legal prerogatives (using Balkin (2017)'s words, they rather "err on the side of caution"), or whether the US government pushes the platforms to interpret the sanctions as also affecting those praising or engaging (in any way) with sanctioned individuals/organizations on social media. In addition, other behaviors could potentially (at least partially) account for the ideological suspension biases observed in this study. For example, human moderators working for US-based platforms may be less lenient toward particular content (Bergman & Diab, 2022), biasing in turn the content moderation algorithms from these platforms.[3] The contribution of the study is four-fold. First, it contributes to the literature on social media and political content moderation by discussing potential geopolitical motivations and strategies behind existing moderation practices. Second, it contributes to the literature on social media, public diplomacy, geopolitics, and digital repression, by emphasizing that all countries – non-Western countries such as Russia and China, but also Western ones such as the US – can (to a different extent) use or condition social media platforms for their geopolitical interests. Third, the study puts forward a research design and a set of computational techniques that can foster further explorations of the determinants and consequences of political content moderation on social media. Finally, the study concludes with empirical evidence on suspension patterns in the Iranian Twittersphere and how these shape politically-relevant discussions on the platform.

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; Gregory, 2008). 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 a nation's geopolitical interests. These can generally be divided into three categories. One way is for governments to promote favorable geopolitical narratives (Miskimmon et al., 2014) on these platforms. These narratives can seek to discredit the narratives of other geopolitical actors, or to promote the nation's views. Sometimes these strategies seek to influence foreign audiences: e.g. Russian operations to undermine democratic processes in Western countries (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). In 2022 for example, Twitter and Facebook identified several bogus accounts, allegedly run by the US military,[4] that "consistently advanced narratives promoting the interests of the United States and its allies while opposing countries including Russia, China, and Iran" (Graphika & Internet Observatory, 2022). On other occasions, 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 Russian Internet Research Agency (IRA) (Stukal et al., 2022), uses social media to influence national debates on international issues such as Crimea (Golovchenko, 2020). Barrie and Siegel (2021) also find that accounts coordinated by the Saudi government often message about international politics (e.g. discussions around Qatar and Iran), and that local audiences engage with these messages at substantive rates.

Governments can also use social media platforms for surveillance. Research shows that governments sometimes track social media communications to silence dissenting voices at home (Pan & Siegel, 2020). The events in recent years regarding TikTok operations in the US illustrate concerns regarding the use 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).[5] 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 (Gray, 2021). In a letter to the Director of National Intelligence, Senators Schumer and Cotton stated that "TikTok is a potential counterintelligence threat we cannot ignore" (Schumer & Cotton, 2019, p. 1), and TikTok's CEO, Shou Zi Chew, had to testify in front of the House Energy and Committee about “TikToks potential threats to data privacy, national security, and childrens online safety” (Busch, 2023, p. 1).

Finally, governments can also leverage social media for their geopolitical interests by suppressing voices on the platforms. The particular strategy will highly depend on whether the platform is based within or outside of the country taking action (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 opposing (geo)political views. For example, access to several Western social media platforms including Twitter is restricted in countries such as China, Russia, and Iran. VKontakte and other platforms controlled by the Russian government are banned in Ukraine (Golovchenko, 2022).

However, a government can leverage the legal system to condition the content moderation policy of platforms based in the country. For example, in this context, in March 2022 the Kremlin passed new legislation to ban and prevent the spread of “fake” news critical of the Russian military operations abroad. Russian social media platforms such as VKontakte and Odnoklassniki are expected to incorporate these directives into their content moderation policy.[6] Around the same time, the Russian government also imposed international sanctions on many top US officials, including President Biden.[7]

This study focuses on the last of the three strategies. It contributes to a better understanding of the geopolitical role of social media by exploring how governments (the US) can advance their geopolitical interests by conditioning content moderation policies (on Twitter) in a way that undermine opposing geopolitical views abroad (Iran) – or about a geopolitical rival more generally, independently of the location of the users. Most existing work on the geopolitical use of social media platforms focuses on non-Western countries such as Russia (Stukal et al., 2022; Golovchenko et al., 2020; Lukito, 2020) 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 from dealing with them. For example, US banks must freeze any account or money transfer 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 & Moshtaghian, 2020) – although it is often unclear what constitutes a form of relevant engagement. This is a good reflection of what Balkin (2017) describes 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 is also a good example of what, in the context of digital repression, Earl et al. (2022) describe as “information channeling:” through international sanctions, governments can condition platforms and users to behave in their preferred way. It can also be seen as “information coercion” (Earl et al., 2022), if the companies indeed act accordingly and take down accounts seen as undesirable, limiting access and information available on the platforms. 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 err 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 can sometimes be 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 Instagram accounts of at least 15 Iranian journalists covering the event (and their posts) had been suspended (IFJ, 2020).

Based on the aforementioned information regarding how US sanction laws can condition the content moderation policies of social media platforms based in the US, I expect the political views of the users in the study to be predictive of suspension. First, I measure the ideology of the users who follow Iranian elites on Twitter in a reformist-principlist (left-right) continuum. Principlist and reformists are the two main ideological groups in Iranian politics. Principlists hold more conservative views and support a stronger foreign policy in regards to Western countries, whereas reformists hold more progressive views and are more open to negotiate with Western countries. In addition, I also measure how supportive the Twitter users in the sample are of the Iranian government. I put forward the following two hypotheses.

H_1 Higher **principlist (conservative)** scores will be predictive of suspension.

H_2 Higher levels of **support for the Iranian government** will be predictive of suspension.

Controlling for Other Predictors of Suspension

The content moderation policies of social media platforms such as Twitter consider many additional behaviors that can lead to the removal of an account. These confounders need to be taken into account in order to accurately explore any potential ideological bias in the suspension of accounts that follow Iranian elites on Twitter. As elaborated below, it is of particular relevance to control for the use of hateful language, the dissemination of misinformation, automatic accounts (bots), as well as coordinated behavior.

Numerous studies find mainstream social media platforms to often facilitate the dissemination of uncivil and hateful content. Theocharis et al. (2020) found 18% of tweets mentioning members of the US Congress in 2017–2018 to contain uncivil language, and Siegel et al. (2021) found about 1% of tweets mentioning Trump and Clinton in 2016 to contain extreme hate speech. There is also a growing concern regarding the spread of false information on major social media platforms, which for example accounted for 6% (Grinberg et al., 2019) and 8.5% (Guess et al., 2019) of the news consumption on Twitter and Facebook, respectively, during the 2016 US election. Some have also documented that certain political actors (e.g. Russian Internet Research

Agency, IRA) have deployed automated bots and manually-controlled social media operations to pursue their political goals (Stukal et al., 2022). Research documenting the US-election-interference efforts from IRA also shows a high level of coordination among their accounts: posting similar messages and on the same topics (Green, 2018; Lukito, 2020).

Social media platforms have responded to these threats by implementing a wide range of moderation policies, and removing content and accounts. For example, the Twitter Rules[8] state that accounts can be suspended for engaging in violence and extremism, hateful conduct, platform manipulation and spam, undermining civic integrity, and using synthetic and manipulated media.

According to the growing body of research on political content moderation by social media companies, these types of "toxic" behaviors have been found to be reliable predictors of suspension. In a study of Twitter users messaging about the 2020 US presidential election, Chowdhury et al. (2021) find suspended users (2% of 21 million) to be twice as likely to post offensive tweets and use hate speech, and more likely to share news from fake news websites. In another study tracking Twitter users during the same election cycle, Mohse et al. (2024) find suspended users (4% of 9,000 partisan users) to share fake news at higher rates. In a recent study on shadowbanning on Twitter in the US, Jaidka et al. (2023) find that bot-like behavior, offensive language, and political engagement were predictive of messages being downgraded by the platform. In a study of Twitter users who posted messages about the 2017 French, UK, and German elections, Majo-Vazquez et al. (2021) find suspended users (5% of 4.5 million) to be more likely to be coordinated, use hateful language, and share news in general, although not necessarily from fake news websites.

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 the previous section. Hence, researchers interested in exploring potential political suspension biases need to find ways to control for many additional confounders.

Sampling

The study relies on a sample of politically-interested users to assess the effect of deplatforming on political conversations related to Iranian politics on Twitter.

There are different approaches to building such sample, each with their strengths and weaknesses. Some studies rely on a pre-defined set of politically-relevant hashtags/keywords to identify a sample of interest (e.g. Jost et al. (2018); Casas and Webb Williams (2018)). This is particularly useful when aiming to study a clearly defined set of users (e.g. those engaging with a particular protest movement). However, this approach is not necessarily useful when aiming at identifying a broader population of users who engage in a constantly-changing set of political topics that is unknown *ex ante*. A second option could have been to track all users messaging in a given language (e.g. Farsi, (Hashemi et al., 2022)). However, this would have yielded large numbers of non politically-interested users, exponentially complicating an already arduous process of data collection, processing and analysis. In addition, users who follow and engage in Iranian politics may also post in other languages (e.g. Arabic, English, etc.).

In the end, I opted for a network-based procedure similar to Barberá et al. (2019) and looked for users who follow Iranian elites on Twitter. First, I identified the accounts of a group of elites: the Iranian Supreme Leader (Ayatollah Khamenei), all members of Iran's 10th Parliament ($N = 136$), cabinet members of the Rouhani administration ($N = 20$), and state-owned as well as independent Iranian news media outlets ($N = 19$), for a total of 176 elite accounts.[9] Then, I 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, I sampled users that followed at least 3 of the 176 elite accounts for the analysis (601,940 users in total).

A clear advantage of this procedure is that it yielded a large (yet manageable) sample of users who are interested in Iranian politics, independently of their language, and their political topics of interests. As key limitation, although some elite private media accounts that are sometimes critical of the government were included, most of the seed accounts were government elites. In turn, the resulting sample is likely to be biased toward having more pro-(Iranian)government users than the average user of interest in this study. However, I argue that this actually means that the hypotheses will be submitted to a hard test: there will be fewer chances to compare suspensions among staunch critics of the government (which are expected to be suspended at lower rates) *vs* clear government supporters (which are expected to be suspended at higher rates), and will have to rely more heavily in comparing moderate opponents/supporters, to more clear and outspoken supporters. That being said, as subsequent analyses show, many accounts in the sample openly voice (hard) criticism toward the government, for example, by demanding to stop the imprisonment and execution of dissidents.

Data Collection

I tracked the users in the sample between March 11th and September 10th, 2020, collecting all the tweets they published in 2020 (a total of 65,120,890), as well as information about which accounts became inactive ($N = 7,088$) and when. On October 22nd 2020, the inactive accounts were manually checked for whether they had been: (a) *deleted* ($N = 3,351$), (b) *suspended* ($N = 2,491$), or (c) were active again ($N = 1,246$, *temporary suspensions*).[10] The *deleted* accounts are not included in the analysis as it is unclear whether they had been suspended by Twitter or by the users themselves. In addition, given that the study focuses on suspensions that took place in 2020, users who did not tweet in 2020 are excluded, for a final analytical sample of 2,151 suspended and 168,936 non-suspended users (171,087 in total).

Ideology

The main objective is to assess ideological biases in the suspension of these Twitter accounts. Two key ideological dimensions in Iranian politics are used for this purpose: 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 Reformist-Principlist spectrum, I adapted to the Iranian context a validated and widely used method (*Correspondance Analysis*) for measuring the ideology of elite and ordinary Twitter users in a single left-right dimension (Barberá et al., 2015), and use these user-level ideology scores to test $\mathbf{H}_1$. The model has been validated and found to produce accurate ideology estimates for Twitter users in the US context. Further details regarding the validation of the method in the Iranian context are available in Appendix A, which shows that the resulting ideology scores do a good job at distinguishing between known left-leaning (Reformist) and right-leaning (Principlist) elite accounts in the dataset (members of the 10th Parliament).

A text-based machine learning method is used to measure the extent to which the accounts were supportive of the Iranian government. I trained a binary BERT multilingual model to distinguish political from nonpolitical tweets, and then another binary BERT multilingual model to distinguish between political messages that expressed support for the Iranian government from messages that expressed criticism of the government. Finally, these model predictions are used 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). The latter is used to test $\mathbf{H}_2$.

Table 1 shows the performance of these models (*Political* and *Pro-IranGov*), based on five-fold cross-validation on an untouched held-out validation set. The *Labeled* column indicates how many tweets were manually annotated to train and validate the classifiers, and the *Negative* and *Positive* columns indicate the percentage of the annotated messages that were coded as (not) being political, and as (not) being in favor of the government.[11] The *Epochs* column indicates the number of training/fine-tuning iterations for these classifiers. Finally, the remaining columns provide information about common performance metrics used in machine learning: *Accuracy*, *Precision*, *Recall* and *F1-Score*.

The classifiers are highly precise as they correctly predict political and pro-Iranian -Government messages about > 80% of the time, and they also do a good job at detecting most of the political and pro-Iranian-Government messages in the dataset (83% and around 77% recall). Appendix C provides further information about the manual annotation of the training dataset, as well as the training of the BERT models.[12]

Table 1. Cross-validated out-of-sample performance of 3 BERT-multilingual models predicting political, hateful, and pro-Iranian-government tweets.

	Labeled	Negative	Positive	Epochs	Accuracy	Precision	Recall	F1-Score
Political	2,893	56%	44%	7	83%	81%	83%	82%
Hateful	1,998	79%	21%	2	88%	76%	66%	70%
Pro-IranGov	1,294	50%	50%	4	81%	77%	77%	76%

Controls

Hateful Content

I fine-tuned another BERT multilingual model to build a binary text classifier predicting whether a message used hateful language. The model is used to create a user-level variable measuring the number of hateful tweets sent by each user in 2020. Table 1 also reports the performance of this machine learning classifier. The model is able to capture 2/3 of the hateful messages in the dataset (about 66% recall), and it correctly predicts hateful tweets 76% of the time. Appendix C also provides further details about the training of this BERT model.[13]

Coordination and Bots

Building on the premise that coordinated accounts post/share very similar (if not the same) content (Green, 2018; Lukito, 2020), I 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.

In addition, I controlled for automation of accounts in the 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 et al., 2020). Hence, rather than using an off-the-shelf bot-detection model, in the analysis I include a set of user-level controls that previous studies have found to be effective at distinguishing bot *v*. human accounts. In particular, I 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, 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. I 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 & 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.

Misinformation

Given that during the period of this research the platforms were mainly concerned about the spread of misinformation related to COVID-19, in order to control for misinformation, I focused on identifying users in the data that engaged in spreading misinformation on COVID-19. In particular, I created a user-level variable to measure the number of tweets posted in 2020 that contained one or more hashtags from a set of hashtags that had been previously identified as related to COVID-19 misinformation (see Appendix D for further details).

Additional Controls

Three additional controls are included in the analyses. First, a control accounting for the possibility of verified accounts to be less likely to be suspended (as Twitter may want to avoid public controversies surrounding the suspension of salient accounts). Second, a control accounting for the language used by the users in the dataset (Prop. of tweets in Farsi), as automatic content moderation tools by Twitter may not perform equally well across languages. Finally, a control for the amount of political messages posted by the users, as some previous research finds higher suspension rates for accounts posting about politics (Chowdhury et al., 2020).

Results

Figure 1 shows the number of cumulative suspensions detected among the 601,940 users tracked in the study, a total of 3,737. Each dot corresponds to a moment in time when the accounts were checked for whether they were still active. About 0.6% of the users were suspended during the period of analysis, which represents a non-trivial amount. The clear linear trend in Figure 1 suggests that Twitter assesses historical data and suspends accounts incrementally in batches, and that a larger number of suspensions would have been found if the accounts had been tracked for a longer period of time.

Clear differences emerge already when simply comparing the suspended and non-suspended users on many relevant descriptives (see Table 2). First, the top of Table 2 shows the results for the variables that existing literature finds useful for distinguishing bot from human accounts (Bastos & Mercea, 2019; Stukal et al., 2022; Majo-Vazquez et al., 2021). Most patterns are consistent with this existing literature and suggest that some of the accounts were most likely suspended for engaging in bot-like activity. On average, suspended users had been in the platform for a shorter period of time (1,067 days *v.* 1,337 for

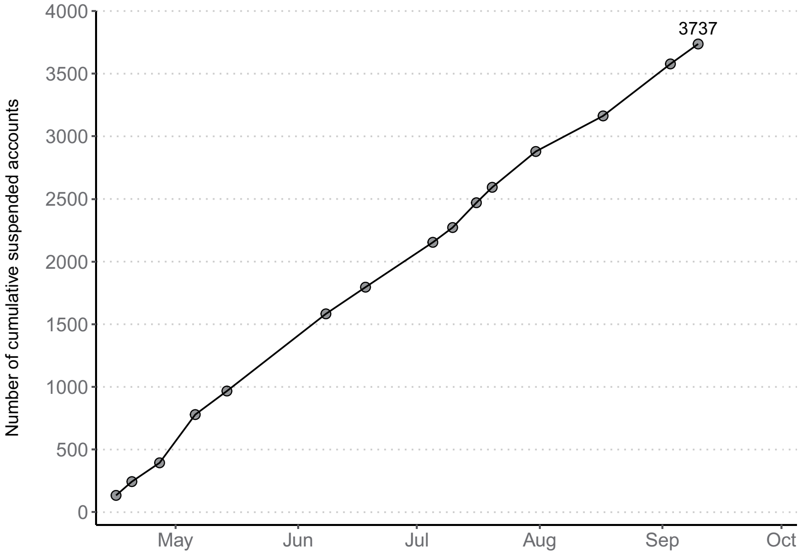

Figure 1. Cumulative number of accounts that I tracked and were suspended during the period of analysis.

Table 2. Descriptive statistics (with 95% confidence interval) for suspended and non-suspended users. The asterisks 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	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	0.2 [0.19–0.2]	0.2 [0.19–0.22]
Prop. of Geo-enabled accounts	0.03*	0.02*
Avg. Proportion of tweets with a hashtag	0.21 [0.21–0.21]*	0.23 [0.22–0.24]*
Avg. Proportion of tweets at somebody	0.48 [0.47–0.48]*	0.46 [0.45–0.47]*
Avg. Proportion of retweets	0.23 [0.23–0.23]*	0.27 [0.26–0.29]*
Prop. using Twitter Web Client platform	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)	7 [7–7]*	33 [30–36]*
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. tweets in Farsi (2020)	0.611 [0.609–0.613]*	0.559 [0.542–0.575]*
Avg. Prop. tweets if English (2020)	0.136 [0.135–0.138]*	0.146 [0.135–0.157]*
Avg. Prop. tweets in Arabic (2020)	0.07 [0.069–0.071]*	0.112 [0.101–0.122]*
Avg. Principlist (Conservative) score {0–1}	0.112 [0.111–0.112]*	0.126 [0.122–0.13]*
Avg. Prop. In favor of Iranian government {0–1}	0.429 [0.428–0.431]*	0.49 [0.479–0.501]*

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 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%).[14]

As one would expect, suspended users sent a larger number of tweets containing hateful language in 2020 (33 *v.* 7). 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%, no statistically significant difference), which is to some extent surprising. This could be a function of conducting simple bivariate analyses between accounts that also differ on many additional dimensions. In a subsequent analysis (Figure 3), where suspension are modeled as a function of all these covariates together in the same model, the results show hateful tweets to be predictive of suspension. Table 2 also shows that suspended users sent more tweets containing COVID-related misinformation hashtags (1 *v.* 0), and higher coordination scores among suspended users (0.98 *v.* 0.95).

More importantly, these comparisons also reveal substantive ideological differences. The last two rows of Table 2 show suspended users to be more ideologically conservative ($\mathbf{H}_1$) and to be substantially more supportive of the Iranian government ($\mathbf{H}_2$). On average, for example 49% 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 (0.13 in a 0–1 index where higher values indicate higher conservatism; *v.* 0.11 for non-suspended users).

Figure 2 shows these bivariate ideological differences in suspensions in more detail. Users are clustered into different ideological bins (left-panel) and bins representing different levels of support for the Iranian government (right-panel), with higher values, and so bars on the right in each panel, indicating the rates for more conservative users, and higher support for the government. When looking at the ideology measure, there is a suspension rate of 1.21% for the least conservative (Principlist) users but a suspension rate of 3.06% and 8.7% for the most conservative ones. Regarding the measure of support for the Iranian government, the lowest rate is for the users who supported the government the least in their Twitter communications (0.76% for those who were supportive in 0–25% of their political messages), compared to suspensions rates that are more than twice as large (>1.63%) for those who supported the government in more than 25% of their political tweets.

Figure 3 provides more stringent evidence for these differences, which shows the results of a multivariate logistic regression predicting suspensions. Skewed variables have been log-transformed (see distribution of all numeric/continuous variables in Appendix B), but the key findings remain the same when not applying these non-linear transformations (see Model 6 in Table B2, Appendix B). In particular, Figure 3 shows 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, and the aforementioned literature on social media bots, it shows 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 (−44%), and a larger tweeting volume (measured as the average number of daily tweets, +20%, as well as the number of tweets sent in 2020, +80%) and a higher follower/friend ratio (+68%) positively predict suspension.

In regards to the other controls in the model, the results also align with what one would expect. A one standard deviation increase in hateful tweets is predictive of a 10% increase in the likelihood of suspension. A similar increase in the number of tweets containing COVID-related misinformation is predictive of a 4% increase in the likelihood of suspension. On the contrary, verified users are predicted to be suspended at lower rates (91% less likely). Contrary to the findings by Chowdhury et al. (2020) in the US context, accounts messaging about politics are suspended at lower rates. Accounts messaging in Farsi are also less likely to be suspended. Contrary to the expectations, I find a null effect for the coordination variable, although a model where the coordination variable is interacted with support for the Iranian government shows that coordinated accounts that are supportive of the government are statistically and substantially much more likely to be suspended compared to supportive accounts that are not coordinated (see Model 5 in B2, Appendix B).

More importantly, in line with $\mathbf{H}_1$ and $\mathbf{H}_2$, I 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, including when only focusing on accounts that are likely to tweet from inside Iran). A one standard

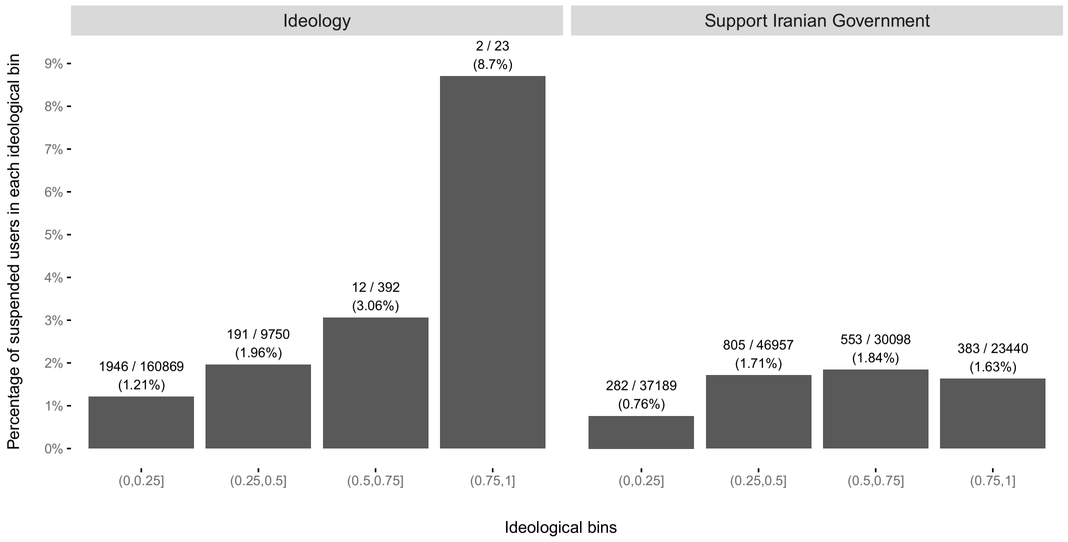

Figure 2. Percentage of suspended users by their ideology, and by how supportive they are of the Iranian government in their tweets.

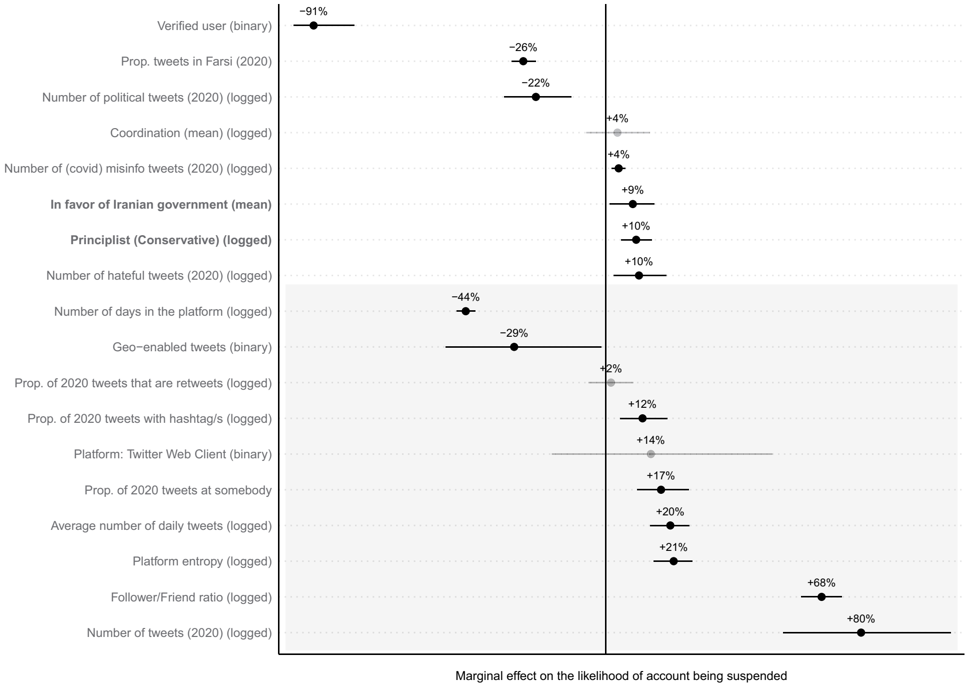

Figure 3. Logistic regression predicting whether an account was suspended. Marginal effects expressed in percentual change (%). Note: *the variables at the bottom of the figure, in the gray area, are potential predictors of bot (or human) activity.*

deviation increase in conservatism (Principlism) is correlated with a 10% increase in the likelihood of suspension. The same increase in support for the Iranian government is also predictive of a 9% 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 to improve the health of the platform. However, the findings also show some clear political biases in the suspension of users, and in turn, that these suspensions have consequences for which ideological views get to have a stronger presence on the platform. 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. Hence, although due to limitations in the data I am unable to definitely pin down the exact mechanism at play, in line with the theoretical framework, these suspension patterns contribute (at least to some extent) to advance the geopolitical interests of the US.

To shed more light on these ideological biases, in Figure 4A. I 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, I 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 4B. I 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.

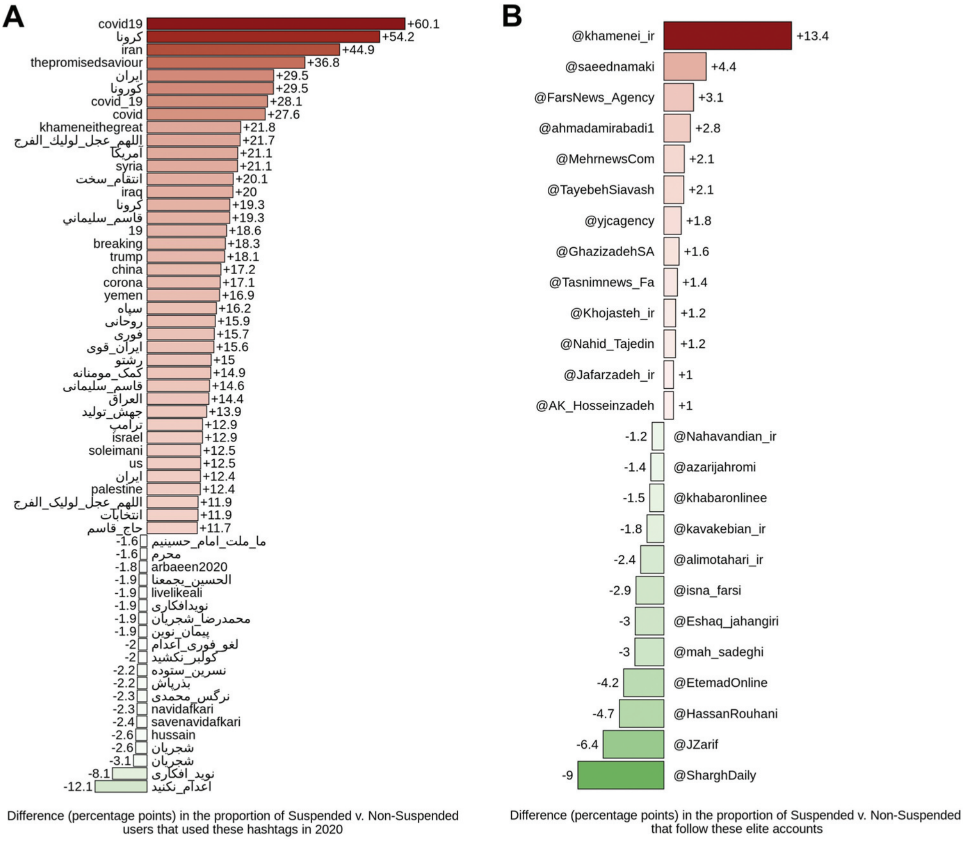

Figure 4. Differences in hashtag usage (A), and elite following (B), between suspended and non-suspended users.

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 4. A illustrates the type of content that was to some extent repressed *vs.* emphasized as a result of the suspensions. First, it shows 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 4A are related to coronavirus, such as کرونا, covid, and covid19. In line with Table 2 and Figure 3, this reassures the idea that some of the accounts were suspended for spreading misinformation on this topic.

Also, Figure 4A shows 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▮سلمانی) and some praise the Supreme Leader of Iran Ayatollah Khamenei (Khamenei the great). Some other hashtags at the top represent some of the common Principlist narratives, such as ارانا▮قوی (*strong Iran*) and جهش▮تولد *(production growth)*. On the contrary, 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 Afkari, 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.

Figure 4B shows similar ideological biases. Among the most-followed elite accounts by the suspended users, there is the Supreme Leader of Iran (Ayatollah Seyyed Ali Khamenei) as well as some conservative media outlets, including *Tasnim News* and *Fars News Agency*. On the contrary, among the most-followed elite accounts by the non-suspended users, there are 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.

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 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. This study addresses 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 condition the content moderation policies of US-based social media companies, and in turn, advance its geopolitical interests.

For a six-month period in 2020, I tracked about 600,000 Twitter accounts interested in Iranian politics. About 4,000 of them had been suspended after the period of analysis. Two overarching patterns emerge 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, the results also show 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.

I acknowledge that this study is subject to several limitations. First, the 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, similar suspension patterns are to be expected when it comes to the regulation of content related to geopolitical rivals on US-based platforms, as they are all expected to comply with US sanctions. Second, given the observational nature of the study, omitted variable bias is always a concern. Nevertheless, I have developed many measures that allow to control for the alternative explanations put forward by previous literature. In addition, Appendix B shows that the key results are robust to different model specifications. Finally, I am not able to clearly distinguish the extent to which (geo)political suspension biases are due to Twitter simply complying with US law, whether the company is erring on the side of caution by suspending any account who may be potentially violating the government mandate, or whether the patterns uncovered here can also be the result of other kinds of biases that may emerge during the development of content moderation procedures (e.g. language/cultural/ideological biases in internal manual annotations for content that violates the Twitter Rules). Future research should aim to disentangle more clearly the particular mechanism at hand. However, the research presented here represents an important step toward building a better understanding of the geopolitical relevance of social media communications, and political content moderation more broadly.

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 and geopolitics in the current digital environment. This paper advances our understanding of the size of the problem, and the extent to which geopolitically-motivated suspensions can 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. I 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 curated datasets of suspended accounts made available by the platforms, which are difficult to independently assess. However, this research did rely on access to Twitter data through their public API, which has recently been discontinued – emphasizing the urgency for researchers to be able to access, and independently analyze, data from major social media platforms. Future research can build on the theoretical, methodological, and empirical work presented here to explore potential political-suspension biases (or lack thereof) in many additional contexts and platforms, in order to create a better understanding of the conditions under which social media suspensions may shape political conversations around the globe. In addition, building on the work of Earl et al. (2022), future research can also explore in more detail additional ways through which the US government can leverage communications on US-based platforms to advance their geopolitical interests, by for example deploying accounts promoting content that is beneficial to their geopolitical interests abroad.[15]

Notes

1. Until 2022, Twitter made recurrent public statements regarding sets of accounts the company suspended for being involved in covert information operations. For example, in this statement from 2019 they reported a set of accounts they suspended for being allegedly coordinated by the Iranian government "to support the diplomatic and geostrategic views of the Iranian state:" https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter. They made datasets with account- and tweet-level information for the suspended accounts available to the research community: https://transparency.twitter.com/en/reports/moderation-research.html. However, it is hard to tell exactly how these datasets were curated, and how the suspended accounts compare to others they could have suspended but did not. These publicly-available datasets are restricted to accounts suspended for being linked to state-backed operations, and little is known regarding suspension of ordinary users. Moreover, since 2022, Twitter decided to only share future data with a closed consortium of researchers, making it even harder for researchers at large to independently analyze the political determinants and effects of their content moderation policy.
2. https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/.
3. For example, Facebook's Oversight Board has been recently discussing the conditions under which messages containing the term "shaheed," martyr, should be moderated: https://www.oversightboard.com/news/1299903163922108-oversight-board-announces-a-review-of-meta-s-approach-to-the-term-shaheed/.
4. https://www.washingtonpost.com/national-security/2022/09/19/pentagon-psychological-operations-facebook-twitter/.
5. https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/.
6. (a) https://www.politico.eu/article/russia-expand-laws-criminalize-fake-news/;
   (b) https://www.wired.co.uk/article/vk-russia-democracy.
7. https://edition.cnn.com/2022/03/15/politics/biden-us-officials-russia-sanctions/index.html.
8. Consulted on August 22nd, 2022: https://help.twitter.com/en/rules-and-policies/twitter-rules.
9. Twitter handles were collected for 179 elites, but 3 of them were excluded because they were protected and some crucial information, such as their followers, could not be gathered.
10. This was a very straightforward task. The message provided by Twitter when trying to access suspended/deleted profiles was very clear regarding whether the profile had been suspended by the platform, or deleted (which I do not know if it was done by the platform or the user). An account was determined to have been only temporarily suspended if it was back to being active, and so the timeline was visible.
11. Note that, as recommended when training classifiers for unbalanced classes, I used an active learning approach (Miller et al., 2020) to determine the sample of messages to be annotated. In turn, the *Negative* and *Positive* percentages in Table 1 are not a reflection of the overall presence of these types of messages in the dataset.
12. The inter-rater reliability for the two coders involved in the annotation was 0.89 and 0.83 (Cohen's Kappa) for the political and pro-Iranian-government task, respectively.
13. The inter-rater reliability for the two coders involved in the annotation was 0.72 (Cohen's Kappa).
14. 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 I find higher proportion among non-suspended users) and I do not find any difference between suspended and non-suspended accounts in terms of the entropy of platforms used for posting messages.
15. https://www.washingtonpost.com/national-security/2022/09/19/pentagon-psychological-operations-facebook-twitter/.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research has received funding from a VENI grant from NWO (VI.Veni.211 R.052, PI: Andreu Casas

Notes on contributor

Andreu Casas (PhD, University of Washington) is an Assistant Professor in Political Communication at Royal Holloway University of London, in the Department of Politics, International Relations and Philosophy. His research explores new political communication dynamics in the digital society, the policymaking process more broadly, and the use and development of novel computational methods for the study of politics.

References

Baldwin, D. A. (2000). Success and failure in foreign policy. *Annual Review of Political Science*, *3*(1), 167–182. https://doi.org/10.1146/annurev.polisci.3.1.167

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL Rev*, *51*, 1149–1210. https://doi.org/10.2139/ssrn.3038939

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (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. https://doi.org/10.1017/S0003055419000352

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barrie, C., & Siegel, A. A. (2021). Kingdom of trolls? Influence operations in the Saudi twittersphere. *Journal of Quantitative Description*, *1*, 1–41. https://doi.org/10.51685/jqd.2021.012

Bastos, M. (2021). This account doesn't exist: Tweet decay and the politics of deletion in the brexit debate. *American Behavioral Scientist*, *65*(5), 757–773. https://doi.org/10.1177/0002764221989772

Bastos, M. T., & Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Bay, S., & Fredheim, R. (2019). *Falling behind: How social media companies are failing to combat inauthentic behaviour online*. NATO StratCom COE.

Bergman, A. S., & Diab, M. T. (2022). Towards responsible natural language annotation for the varieties of arabic. *arXiv preprint arXiv:2203.09597*. https://doi.org/10.48550/arXiv.2203.09597

Busch, K. E. (2023). TikTok: Recent data privacy and national security concerns. *Congressional Research Service Report No. IN12131*: https://crsreports.congress.gov/product/pdf/IN/IN12131https://crsreports.congress.gov/product/pdf/IN/IN12131

Cartwright, M. (2020). Internationalising state power through the internet: Google, Huawei and geopolitical struggle. *Internet Policy Review*, *9*(3), 1–18. https://doi.org/10.14763/2020.3.1494

Casas, A., & Webb Williams, N. (2018). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, *72*(2), 360–375. https://doi.org/10.1177/1065912918786805

Chowdhury, F. A., Allen, L., Yousuf, M., & Mueen, A. (2020). On twitter purge: A retrospective analysis of suspended users. In *Companion Proceedings of the Web Conference 2020*, Taipei, Taiwan (pp. 371–378).

Chowdhury, F. A., Saha, D., Rashidul Hasan, M., Saha, K., & Mueen, A. (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 (pp. 607612).

Crasnic, L., Kalyanpur, N., & Newman, A. (2017). Networked liabilities: Transnational authority in a world of transnational business. *European Journal of International Relations*, *23*(4), 906–929. https://doi.org/10.1177/1354066116679245

Davalos, J., & Brody, B. (2020). Facebook, twitter CEOs sought by senate over N.Y. Post story. *Bloomberg* Advance online publication.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

Earl, J., Maher, T. V., & Pan, J. (2022). The digital repression of social movements, protest, and activism: A synthetic review. *Science Advances*, *8*(10), eabl8198. https://doi.org/10.1126/sciadv.abl8198

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

Golovchenko, Y. (2020). Measuring the scope of pro-kremlin disinformation on twitter. *Humanities and Social Sciences Communications*, *7*(1), 1–11. https://doi.org/10.1057/s41599-020-00659-9

Golovchenko, Y. (2022). Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media. *The Journal of Politics*, *84*(2), 639–654. https://doi.org/10.1086/716949

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (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. https://doi.org/10.1177/1940161220912682

González-Bailón, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, *1*(1), 1–7. https://doi.org/10.1038/srep00197

Graphika, & Internet Observatory, S. (2022). Unheard voice: Evaluating five years of pro-western covert influence operations. https://cyber.fsi.stanford.edu/io/publication/unheard-voice-evaluating-five-years-pro-western-covert-influence-operations-takedown

Gray, J. E. (2021). The geopolitics of" platforms": The TikTok challenge. *Internet Policy Review*, *10*(2), 1–26. https://doi.org/10.14763/2021.2.1557

Green, J. J. (2018). Tale of a troll: Inside the internet research Agency in 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/

Gregory, B. (2008). Public diplomacy: Sunrise of an academic field. *The ANNALS of the American Academy of Political and Social Science*, *616*(1), 274–290. https://doi.org/10.1177/0002716207311723

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Hashemi, L., Wilson, S., & Sanhueza, C. (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*, *2*. https://doi.org/10.51685/jqd.2022.005

IFJ. (2020). Iran: Journalists demand end to censorship of Iranian media on instagram. Retrieved August 30, 2022, from https://www.ifj.org/media-centre/news/detail/category/press-releases/article/iran-journalists-demand-end-to-censorship-of-iranian-media-on-instagram.html.

Jaidka, K., Mukerjee, S., & Lelkes, Y. (2023). Silenced on social media: The gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication*, *73*(1), 163–178. https://doi.org/10.1093/joc/jqac050

Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., & Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, *39*(S1), 85–118. https://doi.org/10.1111/pops.12478

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343. https://doi.org/10.1017/S0003055413000014

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, *345*(6199). https://doi.org/10.1126/science.1251722

Lukito, J. (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. https://doi.org/10.1080/10584609.2019.1661889

Majo-Vazquez, S., Congosto, M., Nicholls, T., & Kleis Nielsen, R. (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. https://doi.org/10.1177/20563051211027202

Miller, B., Linder, F., & Mebane, W. R. (2020). Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches. *Political Analysis*, *28*(4), 532551. https://doi.org/10.1017/pan.2020.4

Miskimmon, A., O'loughlin, B., & Roselle, L. (2014). *Strategic narratives: Communication power and the new world order*. Routledge.

Mosleh, M., Yang, Q., Zaman, T., Pennycook, G., & Rand, D. G. (2024). Is twitter biased against conservatives? The challenge of inferring political bias in a hyper-partisan media ecosystem. *PsyArxiv Preprint*. Retrieved February 7, 2024, from https://doi.org/10.31234/osf.io/ay9q5

O'Sullivan, D., & Moshtaghian, A. (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, J., & Siegel, A. A. (2020). How Saudi crackdowns fail to silence online dissent. *American Political Science Review*, *114*(1), 109125. https://doi.org/10.1017/S0003055419000650

Rauchfleisch, A., Kaiser, J., & Zollo, F. (2020). The false positive problem of automatic bot detection in social science research. *PLOS ONE*, *15*(10), 1–20. https://doi.org/10.1371/journal.pone.0241045

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Schumer, C. E., & Cotton, T. (2019). [Letter from senators Charles E. Schumer and Tom Cotton to the acting Director of national intelligence Joseph Maguire]. https://www.democrats.senate.gov/imo/media/doc/10232019.

Shearer, E., & Mitchell, A. (2021). News use across social media platforms in 2020. *PEW Research*. Retrieved April 3, 2023, from https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

Siegel, A. A., Nikitin, E., Barber, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (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. https://doi.org/10.1561/100.00019045

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2022). Why Botter: How pro-government bots fight opposition in Russia. *American Political Science Review*, *116*(3), 843857. https://doi.org/10.1017/S0003055421001507

Theocharis, Y., Barber, P., Zolt, N. F., & Adrian Popa, S. (2020). The dynamics of political incivility on twitter. *SAGE Open*, *10*(2), 2158244020919447. https://doi.org/10.1177/2158244020919447

Tsvetkova, N., Rushchin, D., Shiryaev, B., Yarygin, G., & Tsvetkov, I. (2020). Sprawling in cyberspace: Barack Obama's legacy in public diplomacy and strategic communication. *Journal of Political Marketing*, 1–13. https://doi.org/10.1080/15377857.2020.1724425