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

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. 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.

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

*Royal Holloway University of London. Department of Politics, International Relations and Philosophy: andreu.casas@rhul.ac.uk. This research has received funding from a VENI grant from NWO (VI.Veni.211R.052, PI: Andreu Casas).

1 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 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.

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, Pan, and Roberts, 2013, 2014) and Saudi Arabia (Pan and 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, Maher, and Pan (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 (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. 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, Wilson, and Sanhueza (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 (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. 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 Irans 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 and 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

[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.

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 and 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 towards particular content (Bergman and Diab, 2022), biasing in turn the content moderation algorithms from these platforms.[3]

[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/

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.

2 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 nations 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 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 and Stanford 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 and 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

[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/

"TikTok is a potential counterintelligence threat we cannot ignore" (Schumer and 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 Odnoklasniki 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

[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

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 (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 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 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 is also a good example of what, in the context of digital repression, Earl, Maher, and Pan (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, Maher, and Pan, 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.

3 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, Nagler, and Tucker, 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, Yang et al. (2022) find suspended users (4% of 9,000 partisan users) to share fake news at higher rates. In a recent

[8]Consulted on August 22nd, 2022: https://help.twitter.com/en/rules-and-policies/twitter-rules

study on shadowbanning on Twitter in the US, Jaidka, Mukerjee, and Lelkes (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.

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

4.1 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, Wilson, and Sanhueza, 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 towards having more

[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.

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 towards the government, for example, by demanding to stop the imprisonment and execution of dissidents.

4.2 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).

4.3 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

[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.

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 (Barbera 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 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 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*

[11]Note that, as recommended when training classifiers for unbalanced classes, I used an active learning approach

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%

4.4 Controls

4.4.1 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]

(Miller, Linder, and Mebane, 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).

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), 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 and Kaiser, 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 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 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).

4.4.4 Additional Controls

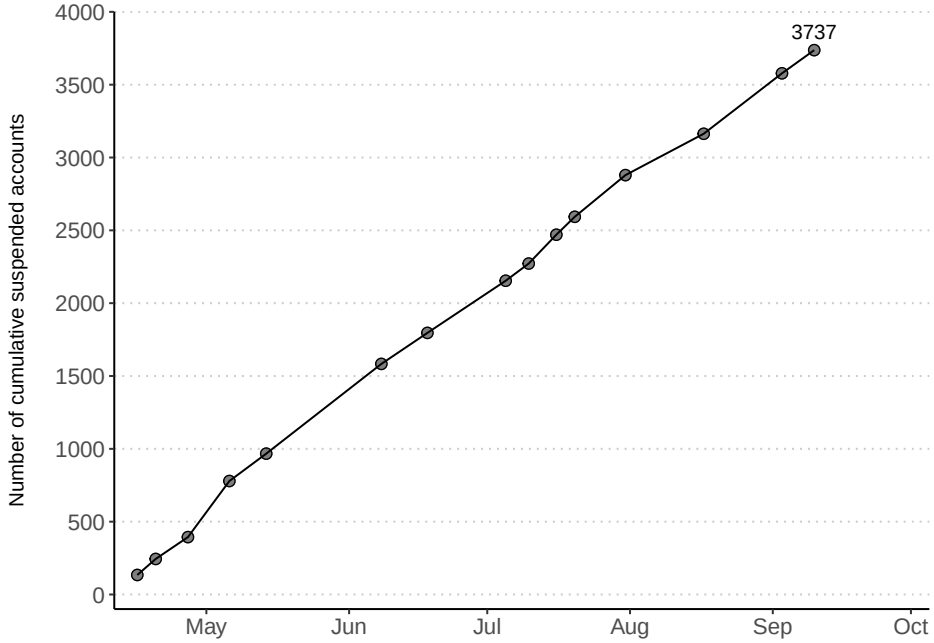
Three additional controls are included in the analyses. First, a control accounting for the possibility of verified accounts to be less likely to be suspended (as Twitter may want to avoid public controversies surrounding the suspension of salient accounts). Second, a control accounting for the language 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).

5 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.

Figure 1: Cumulative number of accounts that I 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, the top of Table 2 shows 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). 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 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

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	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]

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

[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.

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

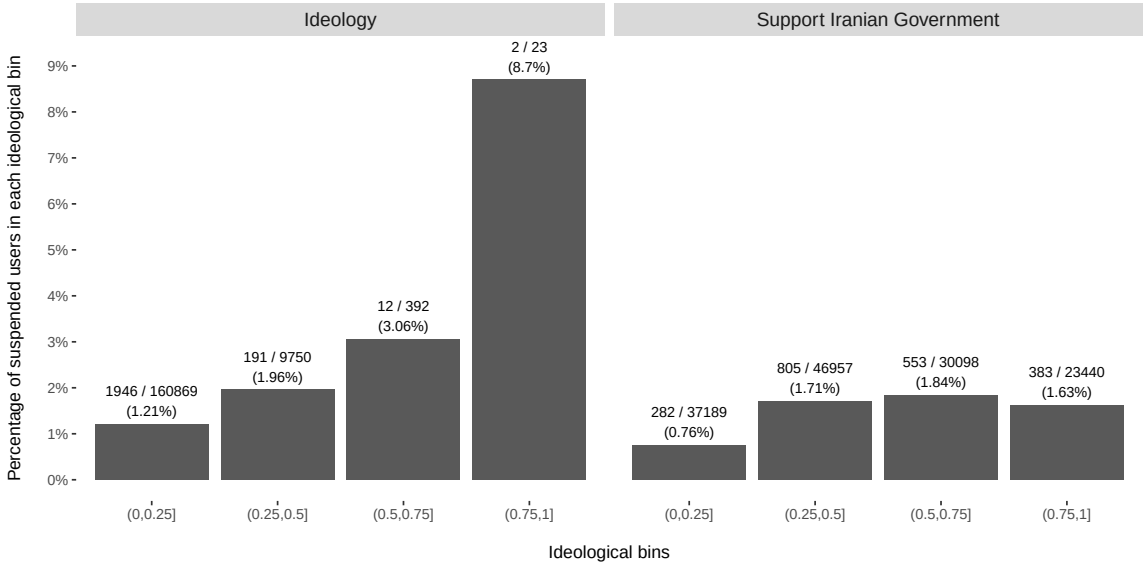
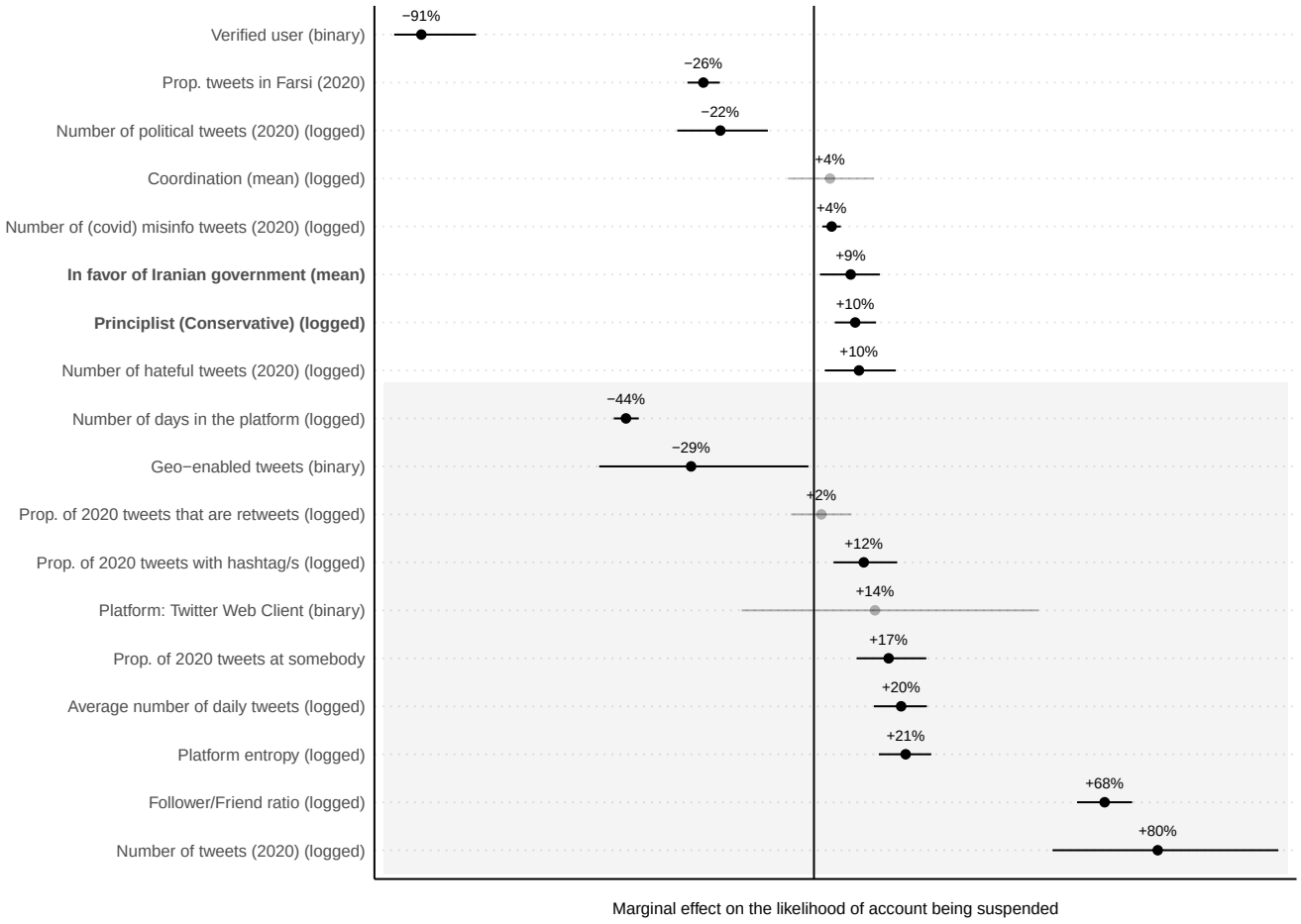


Figure 3 provides more stringent evidence for these differences, which shows the results of a multivariate logistic regression predicting suspensions. Skewed variables have been log-transformed (see distribution of all numeric/continuous variables in Appendix B), but the key findings remain the same 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.

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.*



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 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 4.A 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 4.B 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. 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 4.A 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 4.A 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-thegreat). Some other hashtags at the top represent some of the common Principlist narratives, such as ایران قوی (*strong Iran*) and جهش تولید *(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 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.

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

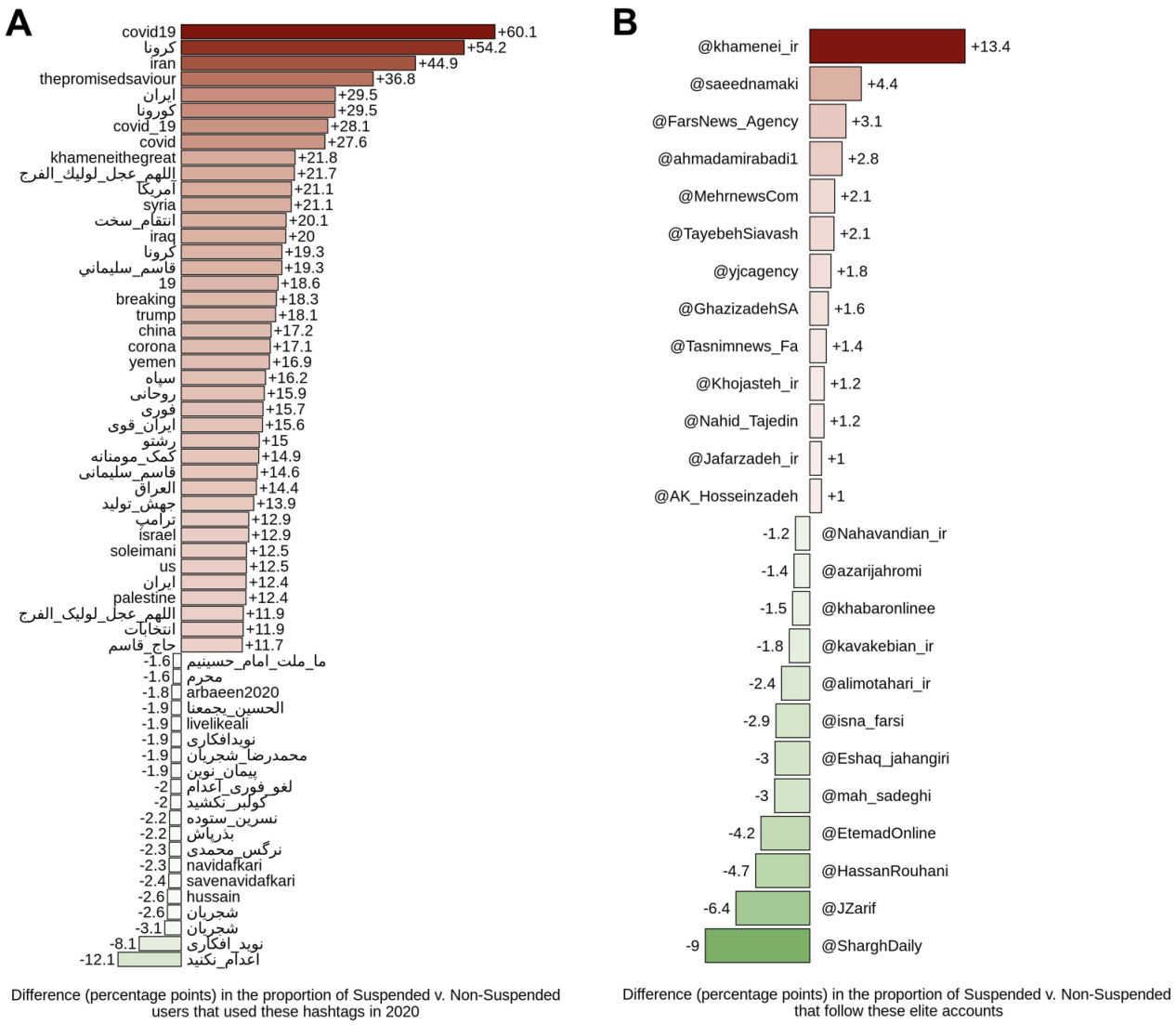


Figure 4.B shows similar ideological biases. Among the most-followed elite accounts by the suspended users, there is the Supreme Leader of Iran (Ayatollah Seyyed Ali Khamenei) as well as some conservative media outlets, including *Tasnim News* and *Fars News Agency*. On the contrary, among the most-followed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

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 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 towards 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, Maher, and Pan (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]

[15] https://www.washingtonpost.com/national-security/2022/09/19/pentagon-psychological-operations-facebook-twitter/

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.

Barbera, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26(10): 1531–1542.

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.

Bergman, A Stevie, and Mona T Diab. 2022. "Towards Responsible Natural Language Annotation for the Varieties of Arabic." *arXiv preprint arXiv:2203.09597* .

Busch, Kristen E. 2023. "TikTok: Recent Data Privacy and National Security Concerns." *Congressional Research Service Report No. IN12131: https://crsreports.congress.gov/product/pdf/IN/IN12131* .

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

Casas, Andreu, and Nora Webb Williams. 2018. "Images that Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72(2): 360–375.

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.

Chowdhury, Farhan Asif, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. 2020. On Twitter Purge: A Retrospective Analysis of Suspended Users. In *Companion Proceedings of the Web Conference 2020*. pp. 371–378.

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.

Earl, Jennifer, Thomas V. Maher, and Jennifer Pan. 2022. "The digital repression of social movements, protest, and activism: A synthetic review." *Science Advances* 8(10): eabl8198.

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.

Graphika, and Stanford Internet Observatory. 2022. "Unheard Voice: Evaluating five years of pro-Western covert influence operations." *Report available at: https://cyber.fsi.stanford.edu/io/publication/unheard-voice-evaluating-five-years-pro-western-covert-influence-operations-takedown* .

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/* .

Gregory, Bruce. 2008. "Public diplomacy: Sunrise of an academic field." *The ANNALS of the American academy of political and social science* 616(1): 274–290.

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.

IFJ. 2020. "Iran: Journalists demand end to censorship of Iranian media on Instagram." Accessed on Aug. 30, 2022: 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, Kokil, Subhayan Mukerjee, and Yphtach Lelkes. 2023. "Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse." *Journal of Communication* 73(01): 163–178.

Jost, John T, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A Tucker. 2018. "How social media facilitates political protest: Information, motivation, and social networks." *Political psychology* 39: 85–118.

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).

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.

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.

Schumer, Charles E., and Tom Cotton. 2019. "[Letter from Senators Charles E. Schumer and Tom Cotton to the Acting Director of National Intelligence Joseph Maguire]." *Retrieved from https://www.democrats.senate.gov/imo/media/doc/10232019%20TikTok%20Letter%20-%20FINAL%20PDF.pdf* .

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.

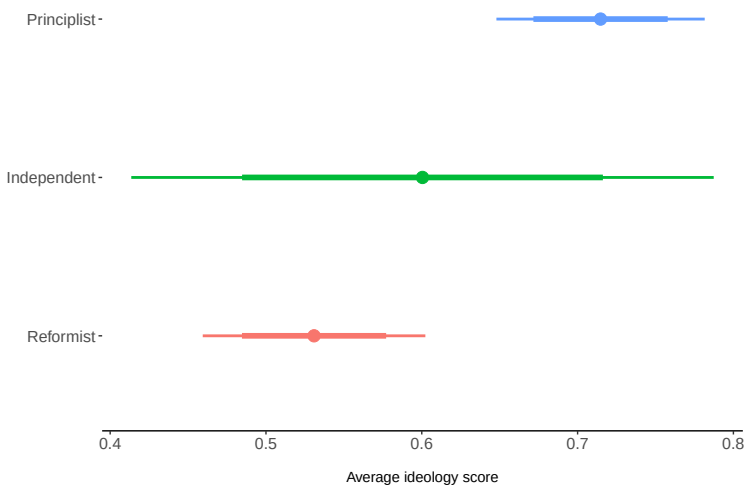
Building on Barbera et al. (2015), and the code available in the replication repository for the paper,[16] I used correspondence analysis (CA) to implement the method developed by Barberá (2015) for estimating the ideology of users and elite accounts (in this case, from Iran) in the same ideological continuum. As a robustness check, I also used the original method in Barberá (2015) to estimate ideology scores for elite accounts and a random set of 5,000 users, which yielded extremely similar results.

To estimate the model that I used in the paper for generating ideology scores for the users under analysis, I 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, I randomly sampled 5,000 users, who followed most of the elite accounts in the list (140 out of 176, so 80%). Then I fit a CA model to the data, obtaining ideology scores (in the same dimension) for the 5,000 users and the 140 elite accounts. Finally I used the trained model 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).

I conducted the following validation exercise to make sure that the model adapted well to the Iranian context. I checked the average ideology score given by the model to the members of Parliament that are known to be affiliated to either the Reformist (N = 59), Independent (N = 12), and Principlist (N = 41) factions in the chamber, 112 in total. I obtained the information regarding their political affiliation from this source. I report the average ideology scores for these three groups in Figure A1 (with 95% and 80% confidence intervals), where I 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 Independents in the chamber and so in this validation group (N = 12), but also because these Independents are a less homogeneous group in terms of their ideological leaning.

[16] https://github.com/pablobarbera/twitter_ideology/tree/master/2020-update

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 3 of the paper I 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 Tables B1 and B2 of this appendix, I provide the coefficient table for that model (*Model 3*), plus additional logistic regressions predicting suspensions, with the goal of assessing the robustness of the findings presented in Figure 3.

Table B1: Coefficient tables for 6 logistic regression predicting whether an account was (temporarily) suspended during the period of analysis.
Note: *The asterisks (*) indicate findings that are statistically significant at the 0.05 level or below*

Variable	Model 1	Model 2	Model 3
(Intercept)	0.0784 (0.2665)	0.0302 (0.2521)	-0.0138 (0.3223)
Principlist (Conservative) (logged)	0.134 (0.0354)*	0.1335 (0.0354)*	0.1418 (0.0369)*
In favor of Iranian government (mean)			0.2778 (0.116)*
Coordination (mean) (logged)	1.0845 (1.691)		2.1014 (2.7128)
Coordination (max.) (logged)		1.2713 (3.9603)	
Verified user (binary)	-2.6537 (0.5888)*	-2.6553 (0.5888)*	-2.5917 (0.5886)*
Number of tweets (2020) (logged)	0.2293 (0.0316)*	0.2327 (0.0321)*	0.2682 (0.0353)*
Number of political tweets (2020) (logged)	-0.0904 (0.0307)*	-0.0907 (0.0307)*	-0.1275 (0.034)*
Number of hateful tweets (2020) (logged)	0.1027 (0.0269)*	0.102 (0.027)*	0.0759 (0.0299)*
Number of (covid) misinfo tweets (2020) (logged)	0.1382 (0.0427)*	0.1376 (0.0426)*	0.1521 (0.043)*
Number of days in the platform (logged)	-0.5932 (0.0279)*	-0.5936 (0.0279)*	-0.593 (0.0293)*
Average number of daily tweets (logged)	0.2333 (0.0356)*	0.2305 (0.0355)*	0.2546 (0.0367)*
Follower/Friend ratio (logged)	0.3839 (0.0149)*	0.384 (0.0149)*	0.3728 (0.0155)*
Platform entropy (logged)	0.1014 (0.014)*	0.1014 (0.014)*	0.1047 (0.0145)*
Prop. of 2020 tweets at somebody	0.369 (0.1027)*	0.3776 (0.1016)*	0.4902 (0.1147)*
Geo-enabled tweets (binary)	-0.3134 (0.1738)	-0.3119 (0.1738)	-0.356 (0.1812)*
Prop. of 2020 tweets with hashtag/s (logged)	0.059 (0.0248)*	0.0599 (0.0248)*	0.0807 (0.0279)*
Prop. of 2020 tweets that are retweets (logged)	0.0051 (0.0209)	0.0066 (0.0207)	0.01 (0.0227)
Platform: Twitter Web Client (binary)	0.1027 (0.1663)	0.104 (0.1663)	0.1227 (0.1726)
Prop. tweets in Farsi (2020)	-0.7922 (0.0658)*	-0.7844 (0.0643)*	-0.8091 (0.073)*
N	171087	171087	137680
AIC	19932.84	19933.16	18288.65

I 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. I 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 I 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 I predicted to be about politics (see Appendix C for further information on the machine learning model used to generate the political predictions).

Table B2: Continuation of Table B1

Variable	Model 4	Model 5	Model 6
(Intercept)	-0.6792 (0.4318)	-0.3589 (0.3515)	-47.66 (3.0415)*
Principlist (Conservative)			1.1929 (0.2622)*
Principlist (Conservative) (logged)	0.1018 (0.0466)*	0.141 (0.0369)*	
In favor of Iranian government (mean)	0.3678 (0.1648)*	0.834 (0.2672)*	0.3025 (0.0978)*
In fav. of Iranian gov. X *Coord. (mean) (logged)*		13.8919 (6.0436)*	
Coordination (mean)			45.0509 (3.1608)*
Coordination (mean) (logged)	-1.2035 (4.767)	-4.6829 (3.7281)	
Verified user (binary)		-2.5957 (0.5886)*	-1.8067 (0.7986)*
Number of tweets (2020)			0.0002 (0)*
Number of tweets (2020) (logged)	0.3219 (0.049)*	0.2758 (0.0353)*	
Number of political tweets (2020)			-0.0002 (0)*
Number of political tweets (2020) (logged)	-0.1645 (0.0481)*	-0.1277 (0.034)*	
Number of hateful tweets (2020)			0.0035 (0.0004)*
Number of hateful tweets (2020) (logged)	0.0478 (0.0434)	0.0646 (0.0302)*	
Number of (covid) misinfo tweets (2020)			0.0043 (0.0017)*
Number of (covid) misinfo tweets (2020) (logged)	0.1201 (0.0542)*	0.1575 (0.0431)*	
Number of days in the platform			-0.0003 (0)*
Number of days in the platform (logged)	-0.6224 (0.0409)*	-0.5887 (0.0294)*	
Average number of daily tweets			0.008 (0.0014)*
Average number of daily tweets (logged)	0.2758 (0.0513)*	0.2562 (0.0368)*	
Follower/Friend ratio			0.0005 (0.0001)*
Follower/Friend ratio (logged)	0.439 (0.0214)*	0.3717 (0.0155)*	
Platform entropy			0.5582 (0.0789)*
Platform entropy (logged)	0.1139 (0.0191)*	0.1052 (0.0145)*	
Prop. of 2020 tweets at somebody	0.2497 (0.1626)	0.4961 (0.1147)*	0.2652 (0.1222)*
Geo-enabled tweets (binary)	-0.4195 (0.2994)	-0.3557 (0.1811)*	-0.3427 (0.1822)
Prop. of 2020 tweets with hashtag/s			0.8838 (0.136)*
Prop. of 2020 tweets with hashtag/s (logged)	0.0689 (0.041)	0.0786 (0.0279)*	
Prop. of 2020 tweets that are retweets			-0.0704 (0.1125)
Prop. of 2020 tweets that are retweets (logged)	0.0112 (0.0319)	0.0073 (0.0228)	
Platform: Twitter Web Client (binary)	-0.0515 (0.3147)	0.12 (0.1726)	-0.2131 (0.1769)
Prop. tweets in Farsi (2020)		-0.7876 (0.0735)*	-0.8965 (0.0695)*
N	88890	137680	137680
AIC	10670.41	18285.25	19565.37

- *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 I 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).
- *Prop. tweets in Farsi (2020)*: a continuous user-level variable, ranging from 0 to 1, measuring the proportion of tweets sent in 2020 by a given user that have been labeled as being in Farsi by twitter.

In addition, I 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 I 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.

In the main model in Figure 3 (Model 3 in Table B1), as well as in many alternative specifications that I describe below, I applied a log transformation to numeric/continuous variables that I identified as being skewed. You can find a list and the distribution of the ones I log-transformed in Figure B1, and of the ones I did not transform in Figure B2. However, as I describe in more detail below, the key findings of the paper (regarding the observed ideological suspension biases) hold even when I do not apply any log-transformation.

In the additional model specifications shown Tables B1 and B2 I assess the impact of the following modeling choices. First, given that I created two variables capturing different ideological dimensions (reformist-principlist position, and support for the Iranian government), I wanted to assess whether including only one of these (only ideology, *Principlist (Conservative)*, in Model 1 and 2) into the model would yield different results than including both. In addition, given that I do not have a measure of *In favor of Iranian government* for users who did not send any politically-relevant message in 2020, I also wanted to see if findings hold before adding this variable into the model and so having to drop some observations (see larger N in Model 1 and 2, compared to the other models). The robustness of these estimates (*Principlist (Conservative)* & *In favor of Iranian government*) across the models reveals that these two dimensions have a distinguishable and robust effect on suspension.

I also wanted to assess the robustness of the findings to an alternative way to represent coordination. In the main model shown in Figure 3 of the paper, Model 3 in B1, I measure 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 I model coordintion as the maximum content/text similiarity between the tweets sent by a given user and any other user in the dataset. Similar to what I observe in the remaining models (where I use the average coordination score), I do not see this version of the variable to be predictive of suspension. Although as I detail below, I do find in Model 6 the non-logged version of this variable to be a strong predictor of suspension.

In Model 4 I was interested in exploring whether one would observe different patterns, particularly regarding the ideological suspension biases observed in the main model in Figure 3 (Model 3 in B1), if I only looked at users we can be confident that were messaging from inside Iran. Although in the current global and social media world, the US should be interested in shaping political conversations regarding Iranian politics generally, independently of the location of users, one could argue that they should be particularly interested in influencing conversations from inside Iran. I leveraged data from Hashemi, Wilson, and Sanhueza (2022) for this purpose. These authors collected all tweets sent in Farsi conciding with the same period of anlaysis. In addition, they leveraged a one-week internet shutdown that took place in Iran to identify which users were messaging from abroad *vs.* inside Iran using a VPN (Twitter is banned in Iran and it can only be accessed through a VPN). First, I only kept in the dataset those users who used Farsi in the majority of their tweets (>50% of the tweets – although the pattern holds when using higher thresholds: e.g. 70%, 80%). And then, I dropped from the resulting sample those users that Hashemi, Wilson, and Sanhueza (2022) had identified as tweeting from outside Iran: ending with a final sample for estimating Model 4 of 88,890 users (1,164 suspended; 87,726 non-suspended) I am highly confident that were tweeting from inside Iran (because they tweeted in Farsi and so included in Hashemi, Wilson, and Sanhueza (2022)'s dataset, and they did not tweet during the internet shutdown according to Hashemi, Wilson, and Sanhueza (2022)). I also observe in this model that Principlist and those in favor of the Iranian government to be more likely to be suspended, and the effect size to be of comparable magnitude to those shown in Model 3.

In Model 5 I was interested in exploring an interaction between the coordination measure and the support for the Iranian government (and weather the main findings would hold after accounting for this potential interaction). Twitter has openly reported in the past to have suspended accounts that they suspected were being coordinated by the Iranian government to spread "diplomatic and geostrategic views of the Iranian state".[17]. They usually also make available for researchers at large, datasets with account and tweet-level information for the suspended accounts, and/or for particular research teams (e.g. the Stanford Internet Observatory) for a more detailed analysis. However, it is hard to tell exactly how this dataset was curated, and how these suspended accounts compare to others they could have suspended but did not. So I wanted to check whether the main effects reported in Figure 3 (Model 3 in B1) were simply a mere reflection of this phenomenon. The results in Model 5 show that, although coordination *per se* is not a relevant predictor in any of the models (with the exception of its non-logged version in Model 6), coordinated accounts that showed high support for the Iranian governments were indeed statistically and substantively suspended at higher rates. However, the findings for the key variables of interest remain significant and of similar magnitude, indicating that after controlling for this interaction, I still observe support for the government in general (independently of the level of coordination) to be predictive of suspension.

Finally, in Model 6 I wanted to assess if the key findings regarding the ideological suspension biases observed in the other models hold when not applying non-linear transformations to the numeric/continuous skewed variables in the dataset, and I do find the results to hold.

[17] https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter

Figure B1: Distribution of predictor variables that I log transformed in Models 1-5 in Tables B1 and B2.

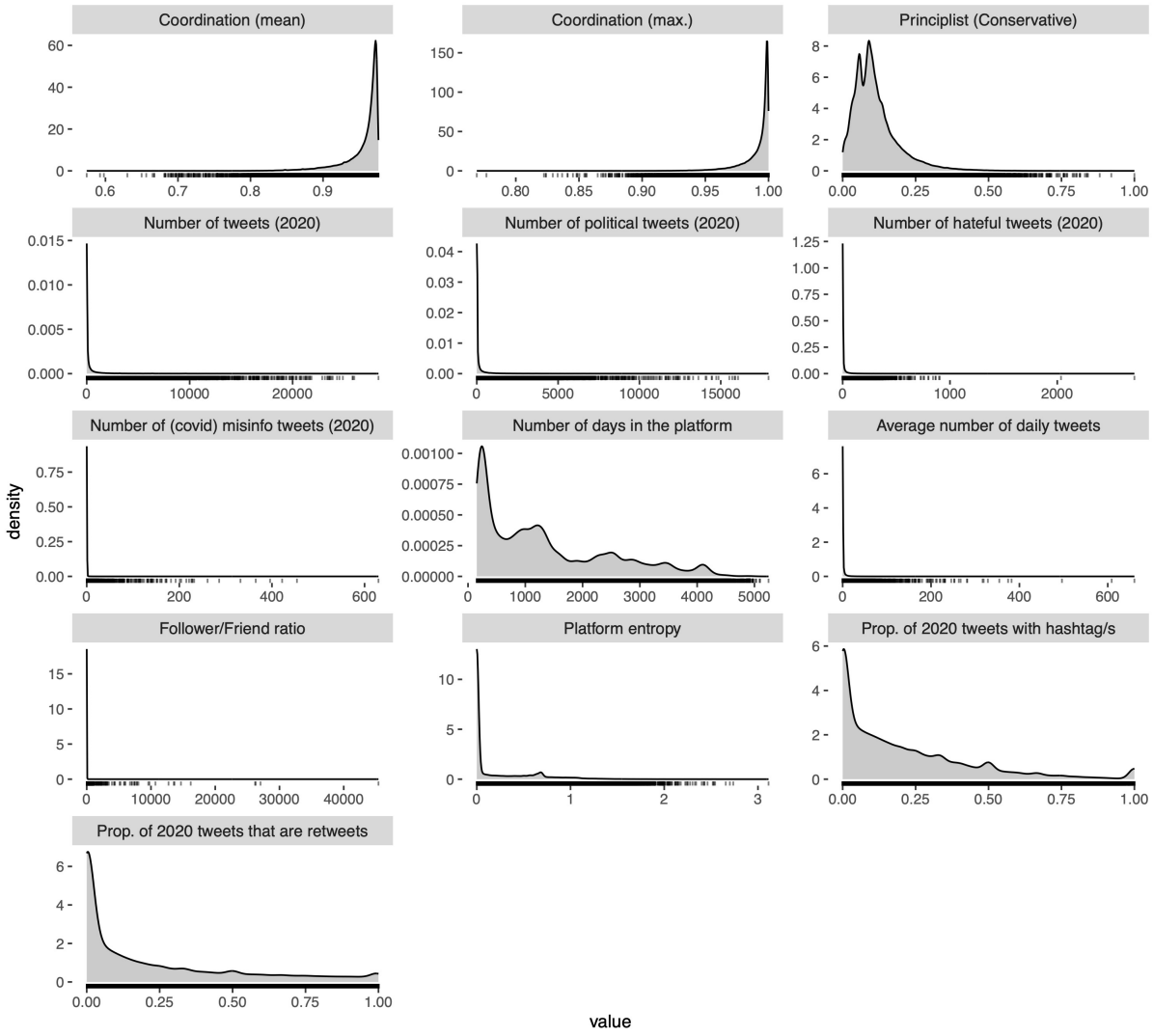
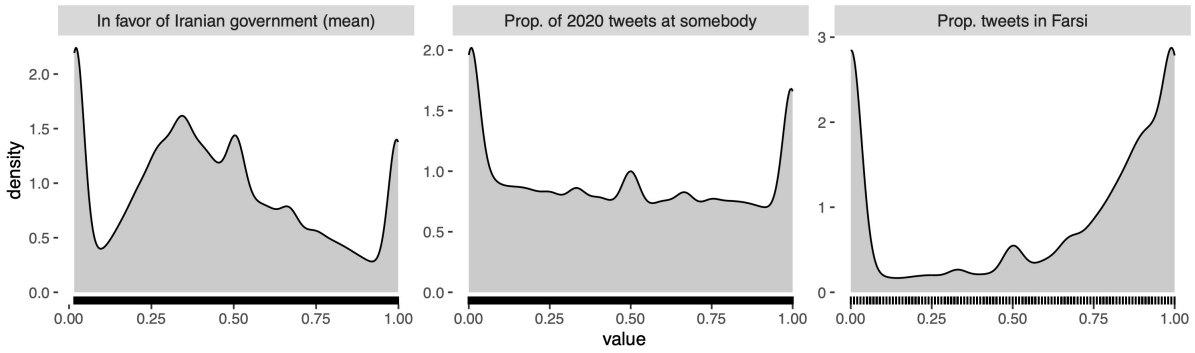


Figure B2: Distribution of predictor variables that I did not logged in Models 1-5 in Tables B1 and B2.



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

I 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. I used the following procedure to train each of these models.

Table C1: 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 باغيرت درود به وجدان بيدار و آگاه شما كه همچو مرد اشتباه را پذيرفتي @Hajizadeh_org
(EN translation) *@Hajizadeh_org Praise upon your awake conscience, commander since you accepted responsibility for the mistake. Praise upon you*
كاش دولته بدخال حذف صفر از پول ملی بشه. مخواد ابرو درست كنه منزنه چشم شو هم كور می كنه
(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 I tracked, I sampled some tweets at random (plus some others using an active-learning procedure (Miller, Linder, and Mebane, 2020)): 1,998 for the hateful coding, 2,893 for the political coding, and I selected 1,294 of the political tweets for the coding of the support of the Iranian government. Following Twitter's definition of hateful language,[18] 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 C1). 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 C2).

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

Table C2: 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 تسلحات 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 طبيعي شمال ايران {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 C3). 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 I 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 I split the data into a train-test-validation split. I saved 20% to assess performance on a final fully-untoched validation set, and split the remaining data 80%/20% for training/testing when training the model. I trained each model until the test loss stop improving: 7 iterations for the political model, 2 iterations for the hateful model, and 4 iterations for the pro-Iran government one. I evaluated the performance of the model using 3-fold cross-validation based on the untouched validation set. I report the performance of the models in Table C4. The *Labeled* column provides information about the total number of messages I 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 indeed predicted to be positive (recall), and the average of both (f-score). Overall *Accuracy* is high for all models (>80%), as well as the *Precision*: although true positives are rare in the training data, e.g. 21% for the hateful classifier, the models make correct predictions >75% of the time. *Recall* is also high for the political and pro-Iran classifiers (83% and 77%, respectively, and only slightly lower for the hateful model, 66% (also resulting on a slightly lower f-score for this model, of 70%). This means that, as it happens with any machine learning model, I'm measuring the quanitites of interest with some noise. However, I don't have any reason to believe that there is a correlation between the error of the model and the outcome of interest.

Table C3: Examples of messages in favor *v.* against the Iranian government

Message coded as being *against* the Iranian government
@Eshaq_jahangiri[19] 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 I used them to generate political, hateful, and pro-Iran predictions for the rest of the unlabeled tweets in the dataset. Finally, I used these machine-labeled tweets to generate 3 user-level variables that I then included in the analyses. First, I counted the number of political tweets sent in 2020 by any of the users in the dataset. Then, I 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, I 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). I then created a user-level variable measuring the average support for the Iranian government (averaging the tweet-level probabilities for a given user).

Table C4: Performance of 3 BERT-multilingual models predicting political, hateful, and pro-Iran 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%

Appendix D Identifying covid-misinformation hashtags.

In the analysis of potential political/ideological suspension among those discussing Iranian politics on Twitter, I 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, I decided to focus on detecting the spread of COVID-19 related misinformation among users in the dataset, given that platforms at that time were particularly concerned about eradicating misinformation on the topic. I developed a four-step protocol to create a list of hashtags that could be clearly linked to the spread of COVID-19 misinformation. First, I generated a list of 39 keywords in Farsi, Arabic and English (the most spoken languages in the dataset) that would help us identify COVID-19 related messages (see Table D1 for a list), and then I manually annotated a random sample of 1,000 messages containing any of these 39 keywords for whether those messages contained misinformation. Next, I 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, I 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, I 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 the analysis of potential political/ideological suspension among those discussing Iranian politics on Twitter, I wanted to control for the alternative explanation that accounts may also be suspended for acting in a coordinated fashion. To accomplish this goal, I developed a method to identify overall content/text similarity between the tweets posted by a given user, and the other users in the dataset. First, I 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 I 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 I generated a (768-size) user-level embeddings by averaging the indexed embedding values of all the users tweets. Third, I calculated the cosine similarity between all possible pairs of user embeddings. And finally, I 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).

Appendix F Most distinctive hashtags for tweets classified using each BERT model.

The goal of this Appendix is to provide further clarity on the three BERT models I fine-tuned in the paper to predict political and hateful tweets, as well as tweets supportive of the Iranian government. Contrary to traditional machine learning models (e.g. logistic regression, decision trees or support vector machines) transformer models are a black box: it is hard to tell exactly what exact textual features are predictive of a given model class. However, in this Appendix I conducted the following exercise in order to provide some intuition as which unigrams and hashtags are likely to be associated with predicted positive and negative messages for each of the three binary classifiers. For a balanced (by classifier and language) random sample of 100,000 tweets in the dataset, I first created a list of all unique unigrams and hashtags in the dataset (I restricted this analysis to tweets in Farsi, English, and Arabic, which represent the vast majority of tweets in the data). Then, for each of the classifiers, I calculated the proportion of predicted positives (e.g. hateful) and the proportion of predicted negatives (e.g. not hateful) tweets that contained each of the unigrams and hashtags. Finally, I calculated the difference in proportions, with positive differences indicating that those textual features appeared at higher rates in tweets predicted to be political, hateful, and in favor of the Iranian government (green rows in all the figures included in this appendix), whereas negative differences indicating that those features showed up at higher rates in tweets predicted to *not* be political, nor hateful and to be contrary to the Iranian government (red rows in the figures included in this appendix). In the figures included below I report the top and bottom 20 unigrams/hashtags, after sorting them by the difference in proportion.

The results of this exercise clearly speak to the face validity of the classifiers. Among the top unigrams and hashtags most associated with messages in English predicted to be political, I observe terms such as *iran*, *israel*, *war*, *trump*, *america*, *palestine*, *regime*, *military*, *#Soleimani*, *#Syria*, *#China*, or *#Afghanistan*. Among the bottom ones, and so the unigrams/hashtags least associated with messages predicted to be political, I observe terms such as *friends*, *person*, *love*, *happy*, *bless*, *AI*, *DataScience*, or *Python*.

Among the most distinctive unigrams/hashtags associated with tweets in English predicted to be in favor of the Iranian Government, I observe *revenge*, *soleimani*, *zionist*, *thepromisedsavior*, *#KhameniTheGreat*, *#US*, and *HardRevenge*. Among the most distinctive unigrams/hashtags associated with tweets predicted to be against the Iranian Government, I observe *court*, *workers*, *protesters*, *police*, *#IranProtests*, *#FreeNazanin*, *#SaveNavidAfkari*, and *#StopExecutionsInIran*.

And among the most distinctive unigrams/hashtags associated with tweets in English predicted to be hateful, I observe *death*, *die*, *isarel*, *america*, *revenge*, *zionist*, *kill*, *#HardRevenge*, *#Hard_revenge*, *#Jihad*, *#Soleimani*, and *#TerroristTrump*. Among the most distinctive unigrams/hashtags associated with non-hateful tweets, I see *coronavirus*, *covid*, *media*, *news*, *health*, *rights*, *#covid19*, *#coronavirus*, and *#BREAKING*.

Figure F1: **Political** classifier: most distinctive **unigrams**.

Model	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion
political	en	iran	0.05677844271	ar	العراق	0.05642574979	fa	ایران	0.1034009869
political	en	people	0.03596444833	ar	كورونا	0.02953912007	fa	آمریکا	0.05970850364
political	en	israel	0.03441200427	ar	ايران	0.0291814443	fa	کشور	0.03443608278
political	en	president	0.03155702639	ar	اليمن	0.02808639254	fa	اسلامی	0.0331771733
political	en	trump	0.03083861541	ar	العالم	0.0244711237	fa	انتقامسخت	0.03212830635
political	en	war	0.02708687253	ar	الشعب	0.02363270739	fa	انتقام	0.03015267936
political	en	iranian	0.0260675481	ar	السعودية	0.02227576723	fa	دولت	0.02924570672
political	en	government	0.02196094711	ar	إيران	0.0217690467	fa	ترامپ	0.02686594106
political	en	america	0.0219093149	ar	لبنان	0.02091882706	fa	جمهوری	0.02594171456
political	en	palestine	0.02167964317	ar	سليماني	0.0197421646	fa	سردار	0.02577120614
political	en	country	0.02123646828	ar	رئيس	0.01910633524	fa	انقلاب	0.02549084984
political	en	american	0.01905433702	ar	عاجل	0.0188058724	fa	سلیمانی	0.02537411487
political	en	regime	0.01851931194	ar	حزب	0.018616671	fa	قاسم	0.02436483851
political	en	india	0.01681400622	ar	أمريكا	0.01852246585	fa	رژیم	0.02420826459
political	en	police	0.01667052382	ar	المقاومة	0.01848389143	fa	مجلس	0.02379438464
political	en	imam	0.01665476222	ar	امريكا	0.01775385692	fa	مرگ	0.02158442964
political	en	military	0.01640235455	ar	فلسطين	0.01752161924	fa	امام	0.02155258422
political	en	china	0.01534231264	ar	السيد	0.01749643036	fa	سال	0.02085345699
political	en	pakistan	0.01529815425	ar	ترامب	0.01625075163	fa	جنگ	0.02063738545
political	en	killed	0.01505797477	ar	علي	0.01597259804	fa	سپاه	0.0203989207
political	en	mask	-0.003234292368	ar	شاد	-0.00366276696	fa	منو	-0.00421537482
political	en	share	-0.003355038839	ar	رب	-0.00416056379	fa	شب	-0.00430549178
political	en	sun	-0.003389595972	ar	شاءالله	-0.00416948749	fa	دیدم	-0.00459789031
political	en	friends	-0.003483666112	ar	صل	-0.00446995033	fa	عشق	-0.00476206487
political	en	stay	-0.003527417515	ar	رو	-0.00461165360	fa	چیه	-0.00483064669
political	en	person	-0.003691915388	ar	ربنا	-0.00467174617	fa	خوبه	-0.00486779179
political	en	heart	-0.003909765924	ar	إلا	-0.00507454674	fa	برم	-0.00488616788
political	en	morning	-0.003950890392	ar	واقعا	-0.00511837413	fa	دل	-0.00498079870
political	en	birthday	-0.004691537813	ar	آره	-0.00572822350	fa	بابا	-0.00523407182
political	en	check	-0.004775100217	ar	اره	-0.00578831606	fa	داره	-0.00539195322
political	en	nice	-0.005320633032	ar	تو	-0.00612735333	fa	دوست	-0.00558324253
political	en	allah	-0.005872326193	ar	بله	-0.00661701756	fa	چی	-0.00614520612
political	en	checked	-0.006065220854	ar	احسنت	-0.00684033154	fa	میکنم	-0.00649423229
political	en	beautiful	-0.006114226125	ar	نه	-0.00776770004	fa	بچه	-0.00663275076
political	en	automatically	-0.006300146461	ar	خدا	-0.00803405030	fa	باشه	-0.00706126297
political	en	happy	-0.006833451083	ar	يا	-0.00809018733	fa	خونه	-0.00713390926
political	en	life	-0.007657253918	ar	والا	-0.00829593871	fa	دلم	-0.00726538032
political	en	god	-0.009305266577	ar	ممنون	-0.00930937976	fa	دیگه	-0.01175679861
political	en	bless	-0.010583973	ar	اللهم	-0.0149507396	fa	منم	-0.01176237823
political	en	love	-0.01083466021	ar	الله	-0.01648134381	fa	خدا	-0.01234919631

Figure F2: **Political** classifier: most distinctive **hashtags**.

Model	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion
political	en	#Iran	0.01458681009	ar	#العراق	0.01313695036	fa	#انتقام_سخت	0.03057455285
political	en	#Palestine	0.01201444962	ar	#قاسم_سليماني	0.01285354381	fa	#اعدام_نکنید	0.01157700495
political	en	#QudsDay	0.007045141273	ar	#كورونا	0.009980903918	fa	#Palestine	0.007952264307
political	en	#US	0.005363477957	ar	#اليمن	0.0097271481	fa	#مرگ_بر_کلیت_و_	0.007778348472
political	en	#Israel	0.004385277995	ar	#عاجل	0.008447356791	fa	#KhameneiTheG	0.00710628864
political	en	#Syria	0.00434284006	ar	#ايران	0.007755896714	fa	#کرونا	0.007071211914
political	en	#Iraq	0.004149038921	ar	#السعودية	0.006012421156	fa	#QudsDay	0.006878599346
political	en	#Soleimani	0.003991108383	ar	#حزب_الله	0.00507612896	fa	#القدس_درب_الشهداء	0.006078082956
political	en	#KhameneiTheGreat	0.003367267032	ar	#إيران	0.004487797728	fa	#قاسم_سليماني	0.005219118924
political	en	#BREAKING	0.003247834028	ar	#البحرين	0.004324576316	fa	#ایران	0.005110144915
political	en	#covid1948	0.003209336494	ar	#لبنان	0.004286001887	fa	#قاسم_سلیمانی	0.004907718805
political	en	#Pakistan	0.00262267921	ar	#القدس_درب_الشهداء	0.004204391181	fa	#سردار_سلیمانی	0.004247933791
political	en	#BREAKING:	0.002269634066	ar	#أطلقوا_سجناء_البحري	0.00401965163	fa	#آمریکا	0.003597890051
political	en	#Trump	0.002224569197	ar	#ابو_مهدي_المهندس	0.003877948357	fa	#مرگ_بر_آمریکا	0.003414946489
political	en	#Iranian	0.002192638998	ar	#انتقام_سخت	0.003869815766	fa	#KhameneiVirus	0.003264473839
political	en	#Lebanon	0.002190012064	ar	#فيروس_كورونا	0.003860892067	fa	#حاج_قاسم_سلیمانی	0.002909594762
political	en	#Yemen	0.002107763127	ar	#Palestine	0.003796337651	fa	#ترامپ	0.002818122981
political	en	#China	0.00207057906	ar	#الكويت	0.003736245084	fa	#HardRevenge	0.002578158869
political	en	#Afghanistan	0.001996210925	ar	#سوريا	0.003654634377	fa	#روحانی	0.002378266305
political	en	#Iran's	0.001839593854	ar	#ليبيا	0.003573023671	fa	#هواپیمای_اوکراینی	0.002309524162
political	en	#301DaysOfCode	-0.0004724781479	ar	#:	-0.00028340654	fa	#الحسين_يجمعنا	-0.00019648515
political	en	#Analytics	-0.0004724781479	ar	#LiveLikeAli	-0.00028340654	fa	#پرسپولیس_النصر	-0.00021145383
political	en	#angularjs	-0.0004724781479	ar	#PS752	-0.00028340654	fa	#نهج_البلاغه	-0.00021794801
political	en	#BigData	-0.0004724781479	ar	#WhoisHussein	-0.00028340654	fa	#اللهم_عجل_لولیک_ا	-0.00021921479
political	en	#coder	-0.0004724781479	ar	#أمين_الجوفي	-0.00028340654	fa	#رمضان	-0.00024194443
political	en	#coding	-0.0004724781479	ar	#Hussain	-0.00034349911	fa	#هوشنگ_ابتهاج	-0.00024519152
political	en	#COVID	-0.0005122891492	ar	#بداية_الإنتقام	-0.00034349911	fa	#ریت	-0.00025619956
political	en	#1	-0.0005521001506	ar	#اللهم_عجل_لوليك_الف	-0.00042510982	fa	#الهی_عظم_البلا	-0.00026269375
political	en	#IoT	-0.0005905976849	ar	#سيد_المواقف	-0.00042510982	fa	#استراماچونی_را_برگ	-0.00027694888
political	en	#RStats	-0.0005905976849	ar	#شب_زيارتي	-0.00042510982	fa	#حاج‌قاسم	-0.00029967852
political	en	#vuejs	-0.0005905976849	ar	#صباح_الخ	-0.00042510982	fa	#شیفت_شب	-0.00029967852
political	en	#DataScience	-0.0007087172218	ar	#الحسين_يجمعنا	-0.00048520238	fa	#قرار_عاشقی	-0.00031718075
political	en	#javascript	-0.0007087172218	ar	#فيروس_كورونا_من_	-0.00048520238	fa	#پرسپولیس	-0.00033793007
political	en	#MachineLearning	-0.0007087172218	ar	#كلنا_معك_على_السم	-0.00048520238	fa	#در_خانه_بمانیم	-0.00033793007
political	en	#php	-0.0007087172218	ar	#وارث_الصبر_والنص	-0.00048520238	fa	#شیفت_شب	-0.00035091844
political	en	#Python	-0.0007087172218	ar	#MTVirus	-0.00056681309	fa	#مولانا	-0.00040540544
political	en	#tech	-0.0007087172218	ar	#الموعود	-0.00056681309	fa	#حافظ	-0.00042940185
political	en	#AI	-0.0008268367588	ar	#ياحسين	-0.00056681309	fa	#سعدی	-0.00045015117
political	en	#100DaysOfCode	-0.0009449562958	ar	#ميلاد_القائد	-0.00070851636	fa	#اللهم_عجل_لولیک_الف	-0.00054415650
political	en	#CoronavirusOutbreak	-0.0009847672971	ar	#صباح_الخير	-0.00103049734	fa	#ThePromisedS	-0.00068385369

Figure F3: **Pro Iran Government** classifier: most distinctive **unigrams**.

Model	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion
proiran	en	israel	0.02679511592	ar	الله	0.1258908981	fa	انتقامسخت	0.05279953591
proiran	en	god	0.02672900045	ar	يا	0.03318599515	fa	ایران	0.05136932674
proiran	en	imam	0.02384087836	ar	العراق	0.02224508822	fa	انتقام	0.0501789732
proiran	en	palestine	0.02373553278	ar	محمد	0.02039012347	fa	آمریکا	0.0436884188
proiran	en	thepromisedsavi	0.01818147656	ar	السلام	0.01959126234	fa	امام	0.03071628448
proiran	en	war	0.01726346505	ar	النصر	0.01735695381	fa	سردار	0.02826750294
proiran	en	bless	0.01613561035	ar	اللهم	0.01575620207	fa	قاسم	0.02820080605
proiran	en	allah	0.01473348402	ar	علي	0.01497894247	fa	ترامپ	0.02818016709
proiran	en	peace	0.01465823718	ar	اليمن	0.0147759672	fa	الله	0.02766302209
proiran	en	america	0.01417953156	ar	العالم	0.01423685119	fa	سلیمانی	0.02626285622
proiran	en	zionist	0.0115683857	ar	والله	0.01262556213	fa	حاج	0.02562953364
proiran	en	day	0.01105315458	ar	امريكا	0.01253362395	fa	خدا	0.02527670404
proiran	en	soleimani	0.0108372238	ar	سليماني	0.01247132201	fa	اسرائیل	0.02523506989
proiran	en	syria	0.01037678095	ar	الموت	0.01236831963	fa	شهید	0.02332511225
proiran	en	love	0.01011485409	ar	السيد	0.01227084935	fa	جنگ	0.0199399376
proiran	en	revenge	0.009714947902	ar	ايران	0.01224371574	fa	شهادت	0.01748964733
proiran	en	mahdi	0.009435267072	ar	المقاومة	0.01202717366	fa	قدس	0.01617785725
proiran	en	iraq	0.008860347256	ar	ع	0.01081380042	fa	علی	0.01547020616
proiran	en	qudsday	0.008592841597	ar	أمريكا	0.01063242668	fa	حضرت	0.01458278605
proiran	en	east	0.008394156007	ar	إيران	0.01050782278	fa	جهان	0.01450767146
proiran	en	video	-0.005455476607	ar	آره	-0.004341510563	fa	میگه	-0.005444511928
proiran	en	workers	-0.005548477423	ar	نوش	-0.004647488204	fa	منم	-0.005473290203
proiran	en	black	-0.005690178925	ar	درسته	-0.005029334602	fa	توی	-0.005701841827
proiran	en	checked	-0.005743949625	ar	شد	-0.00505646822	fa	نفر	-0.005793640377
proiran	en	deaths	-0.005753081003	ar	شو	-0.005110735455	fa	مرگبرکلیتوتمامیتجمهور	-0.006408809208
proiran	en	lockdown	-0.005843377218	ar	الصحة	-0.005665920887	fa	چیه	-0.006663971217
proiran	en	court	-0.00586738837	ar	لازم	-0.005798559495	fa	مگه	-0.007122963966
proiran	en	virus	-0.005906788082	ar	احسنت	-0.005936730199	fa	اینا	-0.007236124101
proiran	en	protesters	-0.005957684584	ar	جان	-0.006256274649	fa	نمیشه	-0.007529585501
proiran	en	president	-0.006671601301	ar	رو	-0.006318576597	fa	خونه	-0.007564677057
proiran	en	health	-0.006737207981	ar	از	-0.006844125796	fa	باشه	-0.007638434395
proiran	en	media	-0.006773394304	ar	این	-0.006928029264	fa	اعدام	-0.008917530121
proiran	en	breaking	-0.006990173057	ar	متاسفانه	-0.007290776756	fa	نداره	-0.009009670524
proiran	en	iranian	-0.01016169815	ar	اره	-0.007553551356	fa	اگه	-0.01358357322
proiran	en	government	-0.01020363287	ar	واقعا	-0.00772689039	fa	میشه	-0.01415836277
proiran	en	trump	-0.0107614771	ar	با	-0.007878627903	fa	اعدامنکنید	-0.0144248474
proiran	en	covid	-0.01597565501	ar	والا	-0.008496115283	fa	داره	-0.01525988191
proiran	en	coronavirus	-0.01741075427	ar	بله	-0.01139664634	fa	چی	-0.01763446188
proiran	en	police	-0.01835733816	ar	تو	-0.01259645277	fa	اون	-0.01954943901
proiran	en	people	-0.02505769036	ar	نه	-0.0181809728	fa	دیگه	-0.02695323781

Figure F4: **Pro Iran Government** classifier: most distinctive **hashtags**.

Model	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion
proiran	en	#ThePromisedSaviour	0.01739620477	ar	#انتقام_سخت	0.007550521874	fa	#انتقام_سخت	0.05056608372
proiran	en	#Palestine	0.01346088403	ar	#قاسم_سليماني	0.007155108668	fa	#Palestine	0.009700525118
proiran	en	#QudsDay	0.008346133702	ar	#العراق	0.005413683615	fa	#QudsDay	0.008503467359
proiran	en	#انتقام_سخت	0.004537125905	ar	#اليمن	0.005094139165	fa	#KhameneiTheG	0.00837723642
proiran	en	#Soleimani	0.004185072427	ar	#ThePromisedS	0.003268810644	fa	#القدس_درب_الشهداء	0.008124051913
proiran	en	#covid1948	0.003824057166	ar	#القدس_درب_الشهداء	0.002998001332	fa	#ThePromisedS	0.00707050475
proiran	en	#KhameneiTheGreat	0.003751854114	ar	#إيران	0.00277592716	fa	#قاسم_سليماني	0.006119405199
proiran	en	#Syria	0.003468959895	ar	#سننتقم	0.00277592716	fa	#قاسم_سليمانی	0.006000337098
proiran	en	#Israel	0.003180147686	ar	#فيروس_كورونا	0.002540286178	fa	#سردار_سلیمانی	0.004829103631
proiran	en	#US	0.003162054525	ar	#ابو_مهدي_المهندس	0.002109704641	fa	#HardRevenge	0.004382268667
proiran	en	#HardRevenge	0.002951532954	ar	#سوريا	0.001998667555	fa	#حاج_قاسم_سلیمانی	0.003413293754
proiran	en	#Yemen	0.002262475367	ar	#Palestine	0.001887630469	fa	#covid1948	0.003229696654
proiran	en	#Iran	0.002258922787	ar	#اللهم_عجل_لوليك_ال	0.001887630469	fa	#آمریکا	0.003154391669
proiran	en	#Afghanistan	0.001874405161	ar	#بيروت	0.001887630469	fa	#مرگ_بر_آمریکا	0.003128567377
proiran	en	#Lebanon	0.001799158315	ar	#فلسطين	0.001887630469	fa	#ایران	0.002960804675
proiran	en	#Hussain	0.001570543583	ar	#حزب_الله	0.00187406366	fa	#حاج_قاسم	0.002916255466
proiran	en	#القدس_درب_الشهداء	0.001308786319	ar	#ليبيا	0.001776593382	fa	#ترامپ	0.002665238849
proiran	en	#ImamMahdi	0.00129373695	ar	#لبنان	0.001700724626	fa	#FlyTheFlag	0.002528231923
proiran	en	#Iraq	0.001269556202	ar	#الموت_لأمريكا	0.001665556296	fa	#مرگ_بر_اسرائیل	0.002486636697
proiran	en	#قاسم_سليماني	0.001221533898	ar	#أمريكا	0.001603254348	fa	#ایران_قوی	0.002402723614
proiran	en	#BREAKING:	-0.0006349476965	ar	#الأمم_المتحدة_شريك	-0.000679789327	fa	#کولبر_نکشید	-0.0005522365581
proiran	en	#CoronavirusOutbreak	-0.0006439094797	ar	#الحريري	-0.000679789327	fa	#نویدمان_را_کشتند	-0.0005522365581
proiran	en	#FreeNazanin	-0.0006439094797	ar	#البحرين	-0.000706922945	fa	#KingdomWithP	-0.000583055797
proiran	en	#Iran:	-0.00074621127	ar	#الحجر_المنزلي	-0.000790826413	fa	#اعدام	-0.000618933446
proiran	en	#COVID2019	-0.0007913593771	ar	#السعودية:	-0.000790826413	fa	#استراماچونی_را_برد	-0.0006275415433
proiran	en	#NavidAfkari	-0.0008184143223	ar	#مطار_الكويت	-0.000790826413	fa	#نه_به_اعدام	-0.000652643205
proiran	en	#SaveNavidAfkari	-0.0008184143223	ar	#تشريفاتي	-0.000790826413	fa	#جاویدشاه	-0.0007100093657
proiran	en	#StopExecutionsInIran	-0.0008184143223	ar	#تقلبي	-0.000790826413	fa	#مرگ_بر_جمهوری_	-0.0007774288836
proiran	en	#MAGA	-0.0008485130603	ar	#حظر_التجوال	-0.000790826413	fa	#مرگ_بر_خامنه‌ای	-0.0008283548371
proiran	en	#BlackLivesMatter	-0.0009689080125	ar	#حظر_تجول	-0.000790826413	fa	#اعتراضات_سراسری	-0.0009115452895
proiran	en	#Tehran	-0.0009809135886	ar	#حظر_جزئي	-0.000790826413	fa	#StopExecutions	-0.000928761484
proiran	en	#IranProtests	-0.00106816601	ar	#دعم_حسابات_الاحرا	-0.000790826413	fa	#لغو_فوری_اعدام	-0.0009538631457
proiran	en	#Iran's	-0.001128363486	ar	#مجلس_الوزراء	-0.000790826413	fa	#مرگ_بر_خامنه‌ای_	-0.0009868502747
proiran	en	#Iranian	-0.001158462224	ar	#مردم_را_می‌کشند_ز	-0.000790826413	fa	#SaveNavidAfka	-0.001137460245
proiran	en	#Coronavirus	-0.001281900969	ar	#مرگ_بر_کلیت_و_	-0.000790826413	fa	#کرونا	-0.001182947012
proiran	en	#COVID−19	-0.001450318226	ar	#هاااا_وعاجل	-0.000790826413	fa	#نوید_افکاری_را_کش	-0.001271576651
proiran	en	#…	-0.001468411388	ar	#خلك_في_البيت	-0.000853128361	fa	#اعتصابات_سراسری	-0.001346881636
proiran	en	#Covid19	-0.00159472433	ar	#ثم_القبض	-0.001977066034	fa	#نوید_افکاری	-0.001552689877
proiran	en	#اعدام_نکنید	-0.001841432225	ar	#أطلقوا_سجناء_البحري	-0.002052934791	fa	#مرگ_بر_کلیت_و_	-0.006333504223
proiran	en	#COVID19	-0.006878570292	ar	#اعدام_نكنيد	-0.003163305654	fa	#اعدام_نکنید	-0.01379730585

Figure F5: **Hateful** classifier: most distinctive **unigrams**.

Model	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion	Lang	Unigram	Diff. Proportion
hateful	en	death	0.07914297162	ar	الله	0.1277554966	fa	انتقام	0.1208664285
hateful	en	god	0.06636687261	ar	النصر	0.0632437337	fa	انتقامسخت	0.1098421845
hateful	en	bless	0.0612545653	ar	الموت	0.05466903538	fa	مرگ	0.08360329988
hateful	en	israel	0.05508900217	ar	يا	0.04956211431	fa	اسرائیل	0.04123277345
hateful	en	die	0.05437626082	ar	الصهاينة	0.02655457305	fa	خاک	0.03833470973
hateful	en	palestine	0.05027794548	ar	اسرائيل	0.02438093904	fa	آمریکا	0.03591614663
hateful	en	america	0.04640126676	ar	اليمن	0.02320346162	fa	صهیونیستی	0.0348886914
hateful	en	revenge	0.04566306785	ar	سليماني	0.02103608454	fa	رژیم	0.03470937859
hateful	en	allah	0.03726107725	ar	إسرائيل	0.01864467434	fa	خون	0.03251156735
hateful	en	imam	0.03505334864	ar	امريكا	0.0179736178	fa	صهیونیست	0.02993112898
hateful	en	zionist	0.03265576592	ar	المقاومة	0.01747241103	fa	ترامپ	0.0295222283
hateful	en	iran	0.03179159331	ar	الصهيوني	0.01675551373	fa	نابودی	0.02863088164
hateful	en	save	0.02958370701	ar	اللهم	0.01662524777	fa	سخت	0.02703869747
hateful	en	kill	0.02942772038	ar	الحسين	0.01641685699	fa	ایران	0.02411827135
hateful	en	victory	0.02611967541	ar	قتل	0.01630849031	fa	سردار	0.02372933872
hateful	en	soleimani	0.02136794374	ar	الشهداء	0.01602923102	fa	قاسم	0.02293755243
hateful	en	dead	0.0203926024	ar	عليهم	0.01452561073	fa	قدس	0.0222089586
hateful	en	thepromisedsavi	0.02018091751	ar	المجرمين	0.0141963394	fa	کشته	0.02208610406
hateful	en	qudsday	0.01986493201	ar	سعود	0.0134137877	fa	سلیمانی	0.02067496385
hateful	en	zionists	0.01869858019	ar	والله	0.01328977868	fa	نابود	0.01995995061
hateful	en	global	-0.00566219392	ar	ألف	-0.003179176008	fa	نداره	-0.00368962557
hateful	en	international	-0.00578277125	ar	جدا	-0.003236487818	fa	قرنطینه	-0.00373099653
hateful	en	national	-0.00611981662	ar	دولار	-0.003289628335	fa	خائم	-0.00379470593
hateful	en	rights	-0.00625551430	ar	فيروس	-0.003404251954	fa	شرایط	-0.00379698807
hateful	en	political	-0.00668289630	ar	وزير	-0.003573058913	fa	درست	-0.00391233300
hateful	en	minister	-0.00744215507	ar	بفيروس	-0.004014868222	fa	دولت	-0.00393599786
hateful	en	security	-0.00765345450	ar	الناس	-0.004249329573	fa	نظر	-0.00424740597
hateful	en	china	-0.00776006785	ar	آره	-0.004293084685	fa	رای	-0.00432789979
hateful	en	india	-0.00803624636	ar	مجلس	-0.004294127508	fa	ماه	-0.00442105253
hateful	en	virus	-0.00844295396	ar	الصحة	-0.004518160631	fa	دوست	-0.00451030420
hateful	en	health	-0.00902115399	ar	قناة	-0.004572343971	fa	باشه	-0.00457150704
hateful	en	government	-0.00904182840	ar	واقعا	-0.004962055585	fa	دکتر	-0.00460772869
hateful	en	video	-0.00935303076	ar	وزارة	-0.005185045885	fa	داره	-0.00463860056
hateful	en	news	-0.01018310812	ar	بله	-0.005854016784	fa	روحانی	-0.00475921710
hateful	en	media	-0.01105260702	ar	نه	-0.00619058788	fa	رئیس	-0.00478958211
hateful	en	breaking	-0.01106294422	ar	عاجل	-0.006472975635	fa	انتخابات	-0.00547290392
hateful	en	president	-0.01177761285	ar	رئيس	-0.006805375439	fa	آقای	-0.00637344720
hateful	en	covid	-0.01517313859	ar	ممنون	-0.00747121787	fa	اعدامنکنید	-0.00678901399
hateful	en	trump	-0.01841630448	ar	والا	-0.007749434334	fa	مجلس	-0.00812179626
hateful	en	coronavirus	-0.03252947684	ar	كورونا	-0.01178515901	fa	کرونا	-0.0146662372

Figure F6: **Hateful** classifier: most distinctive **hashtags**.

Model	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion	Lang	Hashtag	Diff. Proportion
hateful	en	#Palestine	0.03007635356	ar	#انتقام_سخت	0.02650038971	fa	#انتقام_سخت	0.105117009
hateful	en	#QudsDay	0.02014149598	ar	#قاسم_سليماني	0.01285631195	fa	#مرگ_بر_کلیت_و_	0.01785838371
hateful	en	#ThePromisedSaviour	0.01956252503	ar	#القدس_درب_الشهداء	0.008517386508	fa	#القدس_درب_الشهداء	0.01630841401
hateful	en	#انتقام_سخت	0.01647133354	ar	#اليمن	0.007901556126	fa	#Palestine	0.01413802244
hateful	en	#covid1948	0.009733895352	ar	#سننتقم	0.007236756518	fa	#QudsDay	0.012567221
hateful	en	#HardRevenge	0.009170815666	ar	#الموت_لأمريكا	0.005845674201	fa	#قاسم_سليماني	0.009573463929
hateful	en	#Soleimani	0.007043857404	ar	#حصار_اليمن_جريمة	0.005455962588	fa	#HardRevenge	0.009033423668
hateful	en	#القدس_درب_الشهداء	0.004379277007	ar	#QudsDay	0.005066250974	fa	#مرگ_بر_آمریکا	0.007782641929
hateful	en	#FlyTheFlag	0.003785956623	ar	#انتقامی_سخت	0.004676539361	fa	#قاسم_سلیمانی	0.007110877057
hateful	en	#Israel	0.003599729291	ar	#السيد_حسن_نصرالله	0.003729351998	fa	#سردار_سلیمانی	0.006833330292
hateful	en	#zionist	0.003112251349	ar	#Palestine	0.003617856848	fa	#covid1948	0.005537388748
hateful	en	#Iran	0.003065733938	ar	#فلا_غالب_لكم	0.003395909371	fa	#مرگ_بر_اسرائیل	0.005489542075
hateful	en	#قاسم_سليماني	0.002946312968	ar	#يوم_القدس_العالمي	0.003395909371	fa	#FlyTheFlag	0.005036325091
hateful	en	#TerroristTrump	0.002795494935	ar	#الصهيونية_اخطر_من	0.003117692907	fa	#زندگی_سگی_اسرائیل	0.004637542714
hateful	en	#QassemSoleimani	0.002780374586	ar	#يوم_الأسير_الفلسطيني	0.003117692907	fa	#انتقام	0.004498566965
hateful	en	#Hard_revenge	0.002478738521	ar	#حزب_الله	0.003115607261	fa	#حاج_قاسم_سلیمانی	0.004440636027
hateful	en	#HardReveng	0.002478738521	ar	#لبنان	0.002781121812	fa	#انتقامی_سخت	0.004308248211
hateful	en	#IranAttacks	0.002257487345	ar	#قادة_النصر	0.002727981294	fa	#سننتقم	0.003734200347
hateful	en	#Jihad	0.0022172949	ar	#الحشد_الشعبي	0.002671712307	fa	#نوید_افکاری_را_کش	0.003595224598
hateful	en	#شهيد_القدس	0.0022172949	ar	#الذكرى_100_لمجزر	0.00233826968	fa	#عین_الأسد	0.003431920781
hateful	en	#EU	-0.0006637535262	ar	#روسيا	-0.00055747574	fa	#امیرحسین_مرادی	-0.00036908057
hateful	en	#Turkey	-0.0007091145725	ar	#صفقة_القرن	-0.00055747574	fa	#جهش_تولید	-0.00038104224
hateful	en	#Kashmir	-0.0008046198114	ar	#ابو_مهدي_المهندس_	-0.00066897089	fa	#سیستان_و_بلوچستان	-0.00039918710
hateful	en	#Pakistan	-0.0008750529541	ar	#السعودي	-0.00066897089	fa	#کرونا	-0.00040658450
hateful	en	#coronavirus.	-0.0008850047016	ar	#العربية_عاجل	-0.00066897089	fa	#شارمین_میمندی_نژا	-0.00045671331
hateful	en	#OOTT	-0.0009403174954	ar	#فيروس_كورونا_من_	-0.00066897089	fa	#مجلس_قوی	-0.00047835451
hateful	en	#Taliban	-0.0009403174954	ar	#العراق_ينتفض	-0.00078046604	fa	#جمعیت_امام_علی	-0.00052042272
hateful	en	#Iran's	-0.0009554378442	ar	#قاسم_سليماني_شهيد	-0.00083673503	fa	#StopExecutions	-0.00058372739
hateful	en	#Iran.	-0.0009956302893	ar	#اعدام_نكنيد	-0.00089196119	fa	#بورس	-0.00063816200
hateful	en	#COVID－19	-0.001232001813	ar	#رمضان	-0.00089196119	fa	#تاجگردون	-0.00066249007
hateful	en	#Covid19	-0.001272194259	ar	#بيروت	-0.00089300402	fa	#ویروس_کرونا	-0.00068063493
hateful	en	#CoronaVirus	-0.001327507052	ar	#الصين	-0.00100345635	fa	#قالیباف	-0.00078332095
hateful	en	#BREAKING:	-0.001508565783	ar	#المسيرة	-0.00122644665	fa	#رائفی_پور	-0.00079877895
hateful	en	#India	-0.001659383815	ar	#المسيرة_نت	-0.00122644665	fa	#ایران_قوی	-0.00101342577
hateful	en	#Coronavirus	-0.002036621625	ar	#الكويت	-0.00122748947	fa	#روحانی	-0.00115240152
hateful	en	#China	-0.002061693721	ar	#ايران	-0.00128375845	fa	#KhameneiTheG	-0.00115482991
hateful	en	#BREAKING	-0.00233825769	ar	#مصر	-0.0013379418	fa	#کمک_مومنانه	-0.00120065293
hateful	en	#US	-0.002438931531	ar	#فيروس_كورونا	-0.00206422450	fa	#ThePromisedS	-0.00184433492
hateful	en	#coronavirus	-0.005224860173	ar	#عاجل	-0.00212153631	fa	#کرونا	-0.00414807003
hateful	en	#COVID19	-0.006481934083	ar	#كورونا	-0.00429934162	fa	#اعدام_نکنید	-0.00642302502