TWEETKEEPING: WHY POLITICALLY-INTERESTED IRANIAN ACCOUNTS GET SUSPENDED ON TWITTER

Mehdi Zamani, S12364150
Thesis Research Master Communication Science,
University of Amsterdam

Supervisor: prof. dr. Magdalena Wojcieszak
Amsterdam School of Communication Research

Today, social media platforms such as Twitter and Facebook are entangled in global power struggles. In the wake of the alleged Russian interference in the 2016 US presidential elections, Twitter and Facebook have increasingly ratcheted up their war on malicious activities on their platforms. As a result, they have suspended millions of accounts, including accounts linked to foreign governments engaged in a geopolitical rivalry with the United States. Regulatory measures of this scope raise pressing questions about the impartiality of social media companies in the context of freedom of political speech on their platforms. However, a comprehensive understanding of the conditions under which private Internet companies curb freedom of speech is lacking. I address this gap in knowledge by examining the conditions under which Twitter suspends user accounts interested in Iranian politics. Based on a sample of 601,940 user accounts, this paper unpacks the extent to which the following factors potentially influence likelihood of account suspension by Twitter: (a) number of posted political messages, (b) whether or not an account is automated, (c) user's ideology, (d) whether or not an account takes a favourable stance on the Iranian government, (e) whether or not an account engages in hateful behaviour, and (f) whether or not an account is verified. The paper puts forward a set of innovative methods to test hypotheses regarding this pressing issue as well as a first set of illuminating findings. I find that Twitter is more likely to suspend coordinated and automated accounts, and those that use hateful language. But more importantly, I find that Twitter is more likely to suppress freedom of speech for certain political views: those with a conservative ideology (Principlists) and those taking a favourable stance on the Iranian government.

Keywords: Censorship, Social Media, Political Speech, Iran, Machine Learning

Introduction

Censorship[1] of political speech by social media companies is an increasingly relevant, yet an understudied area of research in political communication. Today, private Internet companies, including social media companies, play a major role in defining and enforcing the "practical conditions of speech" in the digital space while nation-states tend to co-opt and influence such companies for their own interests (p. 1153 Balkin, 2017). Providing some of the most popular social media platforms in the world, since 2015, US-based corporations such as Twitter, Facebook, and Google have been increasingly pressured by the US government and the public to

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

I would like to thank Andreu Casas Salleras for his great supervision throughout the project, Dirck de Kleer for his assistance in the manual annotation process, and Mariken A.C.G. van der Velden for her continued support.

[1] In this paper, censorship is defined as "actions that a public or private governing entity takes on a selective basis to delete expressive content or to prevent speakers from engaging in further expression" (Langvardt, 2017). My specific focus on prevention of engagement in political discussion through suspension of user accounts.

apply censorial measures to counter the promotion of terrorism on their platforms (VanLandingham, 2017). Such measures include but are not limited to suspension of user accounts, fact-checking, flagging and removing contents, in addition to turning over user information to governments. According to media reports, regulatory measures have affected a wide array of social media users from world leaders (Inc., 2019b; Inc, 2021) to media outlets (of Journalists, 2020; teleSUR, 2019), journalists (Eye, 2017; teleSUR, 2019), activists (Dellinger, 2020), academics (Quintana, 2018), and ordinary citizens (Reilly, 2015) based in different countries around the world. These regulatory measures, however, stand in contrast to the claims that social media would provide more equal opportunities for free expression of political views than traditional media (Balkin, 2017). Despite the negative implications of such actions for freedom of speech, scholars have paid little attention to regulation of political speech by social media platforms.

American social media's crackdown on politically interested users, however, has disproportionately targeted users and political views that are not in line with (or even perceived as threatening0 to US security interests. For example, Twitter and Facebook have announced they have removed user accounts based in China, Russia, Iran, Turkey, and Venezuela on multiple occasions in the past few years (Gleicher, 2019; Roth, 2019; @TwitterSafety, 2020). These platforms have stated in their reports that the suspended users were motivated to advance the interests of foreign actors (Roth, 2019), and that they carried out the suspensions in compliance with US law. At times, major American social media companies have adopted regulatory measures due to governmental coercion. For example, US congress(wo)men have unanimously pressured the social media companies to close down the space for foreign terrorist organizations (FTOs) on their platforms, or otherwise, face criminal prosecution (VanLandingham, 2017). As a result, Twitter, Facebook, and Instagram have outlawed publishing contents purported to promote organizations designated as terrorist entities by the US government on their platforms. For example, after Instagram removed posts and suspended accounts that expressed support for the slain Islamic Revolutionary Guard Corps (IRGC) commander, general Qassem Soleimani, a Facebook company spokesperson said: "We operate under U.S. sanctions laws, including those related to the U.S. government's designation of the IRGC and its leadership" (Cockerell, 2020). Another similar instance of crackdown was reported after Instagram targeted accounts and removed posts supporting Mohsen Fakhrizadeh, an Iranian scientist who was assassinated in November 2020 (Agency, 2020; O'Sullivan & Moshtaghian, 2020). Adoption and enforcement of content regulation policies by American social media companies can be due to governmental coercion and intended to serve the US government's interests.

Twitter is an important case to study social media censorship of political content for the implications of its regulatory measures for freedom of speech. With 353 million active monthly users (DataReportal, 2021), Twitter has become a primary source of political information for many around the world (Barberá & Rivero, 2015). Soon after its establishment in 2006, Twitter adopted a policy that would not allow policing user content on its platform except under special circumstances (Klonick, 2017). Twitter's firm belief in the protection of user's right to free speech even led the company to brand itself as "the free speech wing of the free speech party" (Halliday, 2012). Whereas Google and Facebook's tie to the US government's surveillance program was revealed in 2013, Twitter had refused to join the agency's surveillance program and fought for its users' rights in two court cases (Jeong, 2016). This would put the micro-blogging platform in a different league of social media platforms given that unlike Facebook and YouTube, regulation of user content was not enshrined in Twitter's terms of service (Klonick, 2017). However, since 2015, Twitter has adopted an increasingly aggressive approach towards user content creation on its platform, curbing its users' freedom of publishing political content (Jeong, 2016). Since then, Twitter has taken down millions of user accounts in the past few years, including 70 million suspensions only within the span of three months in 2018 (Timberg & Dwoskin, 2018). It is in this context that the study at hand aims to examine the conditions under which Twitter regulates political speech in the form of suspension of accounts belonging to politically interested users.

For two main reasons, I focus my attention on politically interested Iranian users to examine the conditions under which online political speech is suppressed by social media platforms. First, news media reports

as well as reports released by social media companies themselves show that politically interested Iranian users are among the most frequently targeted users on American social media platforms. Second, given the geopolitical rivalry between Iran and the United States, studying politically interested Iranian users helps us understand the ways in which the said rivalry affects an American social media platform's regulation of political speech as far as Iranian users are concerned.

Neutrality and viability of Twitter as a platform for free expression of political views for Iranian users is disputed; on the one hand, the website has been portrayed as a revolutionary tool for social change in Iran, despite the fact the website is blocked inside the country (Khazraee, 2019), and on the other, the platform has removed thousands of user accounts based in Iran in recent years (Roth, 2019). In 2019, Twitter announced the purge of 4,799 user accounts, claiming the users disseminated news with a bias in favour of the Iranian government (Roth, 2019). In October 2020, 104 accounts based in Iran were suspended for artificial intensification of discussions on sensitive topics (@TwitterSafety, 2020). The crackdown of social media companies in line with policies of the Iranian government has negative implications for neutrality of such websites as platforms for political discussions. According to the platform's breakdown of the purge, the removed accounts published global news from perspectives "that benefited the diplomatic and geo-strategic views of the Iranian state", targeted conversations about Iranian politics, or were engaged in discussions about Israel (Roth, 2019). With the purge of accounts based in countries such as Iran, Russia, and China, Twitter may be creating a biased environment against certain voices. Purge of accounts by social media websites suggests that such platforms have started to follow a trend that they were expected to undermine; suppression of the freedom of expression. The take-down of accounts has taken place while top Iranian officials have used Twitter to make their statements heard by a global audience and communicate with their counterparts. Iranians have long used Twitter to engage in political debates, express their political views, and share information about the Iranian politics (Howard, 2010; Mottahedeh, 2015). An analysis on the extent to which accounts are subject to suspension as well as a detailed analysis of the reasons and the (un)intended consequences for political discussion on the platform are lacking. Therefore, this analysis aims to answer the following question: *Under what conditions does Twitter suspend accounts interested in Iranian politics?*

Traditionally, discussions related to Iranian politics are dominated by voices that oppose the Iranian government. An analysis of the Persian Twittersphere concerning the 2009 Iranian presidential election shows that voices that oppose the Iranian government (70% of the nodes) outnumber voices supportive of the government (24% of the nodes) (Khonsari, Nayeri, Fathalian, & Fathalian, 2010). Khazraee's analysis of the Iranian Twittersphere 2019 showed that tweets by eight media sources[2] outside Iran such as BBC News in Persian (`@bbcpersian`) and the official Persian channel of the US Secretary of State (`@USAdarFarsi`) are the most visible sources of messages in the Persian language on Twitter. The account attributed to BBC News in Persian was identified as the most influential journalistic source of messages in the Persian Twittersphere (Khazraee, 2019). Given the negative bias of these outlets against the Iranian government in their reporting, political discussions on Iran seem to be mainly shaped by anti-Iranian government users on Twitter.

Having said this, the presence of Iran-based news media outlets and journalists on Twitter has grown as top Iranian officials such as Iran's president, Hassan Rouhani, and the minister of Foreign Affairs, Mohammad Javad Zarif have established an active presence on the platform since 2013 (Khazraee, 2019). This has probably changed the balance of influence over public opinion between pro- and anti-government voices given that the high-ranking Iranian officials exert an influence over setting the conversation. Yet, it is not clear that all user accounts interested in Iranian politics enjoy an equal opportunity to freely express their opinions on Twitter given recent purges of accounts by the website, a potential bias that is the subject of the research

[2]These media sources are BBC news in Persian (`@bbcpersian`); *Kaleme*, a non-official news source of the Green Movement in Iran (`@kaleme`); *Deutsche Welle radio* in Persian (`@dw_persian`); *Manoto Persian TV* from London (`@ManotoNews`); the official Persian channel of the US Secretary of State (`@USAdarFarsi`); *Mardomak*, a non-official news source of opposition in Iran (`@mardomak`); and *Radio Farda, Radio Free Europe in Persian* (`@RadioFarda_`) (Khazraee, 2019).

at hand. Moreover, suspension on Twitter means denial of access to information and participation in political discussions.

Twitter's purge of user accounts that are supportive of the Iranian government would only exacerbate the problem of biased information about political issues related to Iran in the current media landscape. Scarcity of accurate information about Iran in mass media has long shaped the perception of citizens of Western democratic countries (Detmer, 1995). For example, the US media coverage of news about Iran have often withheld accurate information from the American public by falling in line with their government's account of major events in the conflicted history of Tehran and Washington. These events include but are not limited to CIA's role in the 1953 coup against the then democratically elected Prime Minister Mohammad Mosaddeq, human rights abuses of Iran's last monarch Mohammad Reza Pahlavi till his ouster in 1979, and the shooting down of an Iranian passenger plane in July 1988, and the Iran-Contra affair (Detmer, 1995; Landon-Murray, Mujkic, & Nussbaum, 2019). Biased coverage of political news about Iran seems to be more a product of the economics of news than an inherent anti-Iran bias (Detmer, 1995). In contrast, social media are widely believed to democratize access to as well as generation of information by freeing information of economic factors that have traditionally hindered its flow (e.g. Effing, van Hillegersberg, & Huibers, 2011; Shirky, 2011). It is in this context that a systemic Twitter bias against voices supportive of the Iranian government would more likely do harm to a free flow of information regarding Iranian politics than guaranteeing a safe environment for public debate.

Given the heavier presence of Twitter users that oppose the Iranian government relative to those that support the government, the platform's recent purges of pro-Iran user accounts may have already marginalized voices supportive of the Iranian government. An imbalanced Persian Twittersphere means that one group, that is, anti-Iran voices, would determine dominant political narratives about Iran, which in turn, would lead to a biased representation of Iranian politics on the platform.

For the first time, this study addresses the empirical gap in our knowledge about the factors that lead to systemic censorship of political speech on a major social media platform. I use state-of-the-art machine learning techniques to analyze Twitter posts in a large-N sample of politically interested Iranians in terms of their political/non-political content in addition to sentiment expressed towards a state actor, while measuring user ideology, hateful user behaviour, coordinated platform manipulation, malicious automated activity, and account verifiability. Thereby, I disentangle potential effects of these factors, and subsequently, compare their magnitude of effects. Both from a methodological and empirical perspective, this study opens doors for researchers to further investigate systemic inequality on social media platforms and its implications for freedom of speech.

Theoretical Background and Hypotheses

Previous works on censorship on social media have looked at the topic from a variety of angles, yet, empirical research on the topic is scarce. Badawy, Ferrara, and Lerman (2018) have focused on malicious political behaviour on Twitter. Langvardt (2017) and Van-Landingham (2017) look at the issue from a legal perspective, contextualizing social media censorship and its implications for free speech. Helberger, Kleinen-von Königslöw, and Van Der Noll (2015) have looked at the subject as a gatekeeping problem, and argue that unlike the traditional gatekeeping which controls access to information, gatekeeping on social media platforms directly targets the user. This conclusion is relevant to the approach taken in this paper in which I view account suspension as a form of regulation of speech. Cobbe (2020) looks at the structural context of algorithmic censorship on social media, arguing that this form of communication control is driven by commercial interests, and leads to more aggressive regulation of both public and private online communication. The closest work to my study was conducted by Chowdhury, Allen, Yousuf, and Mueen (2020). They analyzed multiple forms of malicious behaviour that led to the purge of millions of accounts in a period of three months in 2018 2020. However, their focus is suspension due to malicious behaviour, and not on censorship of politically interested users. My paper, however, intends to address the lack of knowledge about the scenarios that lead to censorship of political speech through user suspension.

Prior research suggests that Twitter users that engage in political discussions are more likely to be suspended.

Based on a globally representative sample, Chowdhury et al. (2020) observed that a great proportion of purged user accounts had consistently engaged in political conversations. Twitter also banned political advertising on its platform in 2019 (Inc., n.d.). Therefore, I test the following hypothesis:

Hypothesis 1 (H1): *Users that post more political messages on Twitter are more likely to be suspended by the platform.*

0.1 Bias towards particular voices of the Persian Twittersphere Iran

The United States government often portrays social media platforms as a tool for democratic change around the globe (e.g. Clinton, 2010). The US government has supported opposition forces against the Iranian government through online technologies. Since the advent of social media, US leaders have stressed the importance of such websites for "advancing democracy" in countries such as Iran as part of a US government's strategy (e.g. Clinton, 2010). In June 2009, at the request of officials in Washington, Twitter rescheduled a network update to delay interruption in the use of its platform by those Iranians protesting the result of the 2009 presidential election (Grossman, 2009). US officials have even been blatant about their government's alignment with private social media companies in furthering their strategic agendas under the banner of advancing democracy abroad. In fact, in a 2010 address to a group of executives of Silicon Valley-based technology companies, the then Secretary of State Hillary Clinton said that the US State Department wishes to put new communication technologies "in the hands of people who will use them to advance democracy and human rights" (e.g. Clinton, 2010). In contrast to how American social media platforms are used as a tool for meddling in internal affairs of other countries, weeks in advance of the 2020 US presidential election, Twitter announced its plan to impose restrictions on retweets of premature victory claims for its users (Elections, 2020), in a move that was intended to protect the election results from interference. Given this history, I expect Twitter to favour political voices against the Iranian government at the expense of voices that support the Iranian government. Therefore, I put forward the following hypothesis:

Hypothesis 2 (H2): *Twitter is more likely to suspend accounts posting messages in favor of the Iranian government.*

he political sphere in Iran is mainly divided into *Reformists* who aim for a more liberal-democratic society, and *Principlists* who want tradition and religion to be the cornerstone of the political realm – with *independents* and *centrists* somewhat in the middle. The current government lies in the middle of this left-right spectrum, making ideology and support/opposition of the government two distinct relevant dimensions. In terms of ideology, there is very little research as to which viewpoints may benefit (or be harmed) by Twitter's suspensions. Anecdotal evidence from other contexts however point to ideology as a relevant predictor. I put forward the following research questions in order to shed new light into whether people's ideological views are predictive of suspension.

RQ1 *Is Twitter more likely to suspend user accounts that express conservative Iranian views relative to reformist views?*

RQ2 *Is Twitter more likely to suspend user accounts that are more conservative and post messages in favor of the Iranian government?*

In addition, there are a few rival explanations as to why Twitter suspends user accounts interested in Iranian politics. Therefore, I have controlled for the following potential explanations in my analysis.[3]

0.2 Malicious automated activity

With the growth of Twitter's popularity, development of automated tools for content generation has grown as well (Thomas, Grier, Song, & Paxson, 2011). In the context of online political debate, use of automated accounts, also known as bots, for artificial manipulation of political conversations on social media goes at least back to the 2010 US midterm elections (Bessi & Ferrara, 2016). Although social media bots are not always used for malicious goals, in recent years, social media platforms such as Twitter and Facebook have suffered from malignant bot activities. Automated accounts that

[3] It is noted that the set of control variables in this research is not exhaustive.

spread false information, hate speech, or artificially polarize political debate can compromise the integrity of social media platforms that are expected to promote healthy political debate (Bessi & Ferrara, 2016). To maintain a safe and malice-free environment, Twitter removes automated accounts used for malicious purposes such as manipulation of public opinion or publishing spam content (Chen, Yeo, Lau, & Lee, 2017; Chowdhury et al., 2020). Therefore, I test the following hypothesis:

Hypothesis 3 (H3): *Automated accounts are more likely to be suspended by Twitter.*

0.3 Coordinated platform manipulation

In addition to the use of automated accounts, coordinated behaviour is another common form of platform manipulation on social media. Influence campaigns adopt coordination tactics to sway public opinion about topical subjects such as Covid-19 (Gruzd & Mai, 2020) as well as national elections (Sharma, Ferrara, & Liu, 2020). Coordinated campaigns tend to rely on trolls and/or automated accounts to achieve their goals (Zannettou et al., 2019). Coordination among social media users can be detected through content similarity, spatio-temporal markers, account specifications, or a combination of two or more of these indicators (Pacheco et al., 2020).

Major social media platforms have developed a set of policies to tackle platform manipulation. In this regard, Twitter enforces its policy on platform manipulation to counter what it refers to as "coordinated harmful activity" (manipulation & spam policy, 2020). All else being equal, I expect a higher probability of account suspension among users involved in coordinated behaviour. Therefore, I test the following hypothesis:

Hypothesis 4 (H4): *Accounts involved in coordinated actions are more likely to be suspended by Twitter.*

0.4 Hateful conduct

Twitter enforces its policy on hateful conduct when user accounts make "violent threats against an identifiable target", "incite fear about a protected group", "wish, hop, or call for serious harm on a person or group of people", or "references to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims" (conduct policy, n.d.). Other "big data" studies on hateful conduct on Twitter shows different approaches have been adopted to measure this construct. Burnap and Williams (2015) take a text-based supervised machine learning to detect hostile reactions to the murder of a musician on Twitter, focusing on race, faith and ethnic background. Ribeiro, Calais, Santos, Almeida, and Meira Jr (2018) adopt a user-based approach to detect hateful behaviour on Twitter, going beyond the content of tweets and looking at user profiles and their connections to measure this construct. Therefore, it is expected that users that show signs of hateful behaviour run a higher risk of account suspension by Twitter. To control for this feature, I have developed a supervised machine learning classifier that detects whether or not the text content of the tweets contain markers of hateful behaviour. Therefore, the following hypothesis is tested:

Hypothesis 5 (H5): *Accounts that engage in hateful behaviours are more likely to be suspended by Twitter relative to accounts that do not engage in hateful behaviours.*

0.5 Verified accounts

Twitter may grant a verified status to accounts of public interest in order for other users to know that these accounts are authentic (Accounts, n.d.). Such accounts usually belong to figures active in politics, business, media and journalism etc. Verified accounts on Twitter may be viewed as more credible, which, in turn, can increase one's online visibility on the platform, and make them qualify for more protection against trolling by the platform (Chávez, 2019). A group of user accounts that enjoy special treatment by the platform is world leaders with a verified status. Twitter claims that "the accounts of world leaders are not above our policies entirely" (Inc., 2019b). This means that Twitter is more lenient on accounts of world leaders relative to those of ordinary users unless certain red lines are overstepped, for example, if a high-ranking official promotes terrorism (Inc., 2019b). However, this exception, known as "notice of public interest exception", only applies to verified accounts of incumbent state officials or those running for public office who have 100,000 or more followers on the platform (Inc., 2019a). Given that some of the user accounts in the sample used for this study belong to high-ranking officials, I have controlled for the

potential effect of public interest on the likelihood of account suspension. Therefore, the following hypothesis is tested:

Hypothesis 6 (H6): *Verified accounts are less likely to be suspended by Twitter.*

Data

The overarching objective of this research is to address the following question: *Under what conditions does Twitter suspend user accounts that follow Iranian politics?* First, I identified a group of elite accounts in order to build a sample of users interested in Iranian politics. An elite account, in the context of this research, is an account belonging to any of the following figures: the Iranian Supreme Leader, a member of the tenth Iranian Parliament ($N = 136$), a cabinet member of the Rouhani administration ($N = 20$), an Iranian news media outlet ($N = 19$), for a total of 175 elite accounts. Then, I connected to Twitter's `REST API` to pull the list of followers for each elite account (a total of 2,410,543 unique followers), and basic metadata from each of these followers (e.g. date creation of the account, number of followers and friends, location, etc.). To make sure the follower accounts were interested in Iranian politics, I narrowed my focus down to those users that at least followed three elite accounts. This criterion resulted in 601,940 user accounts. Then, I looked at a range of behavioural markers (see H1-H5) and account properties (see H6) of the follower users within a span of six months, from March 11th 2020 to September 10th 2020. On the starting date of the data collection, I retrieved the last 3,200 tweets posted by the followers prior to this date. Subsequently, every other week[4] I collected every new tweet posted by a follower since the starting date of the data collection; a total of 65,120,890 tweets. In addition, I monitored the activity status of the followers within his time period to record which accounts stopped being active during this time period ($N = 7,088$). However, accounts could have stopped being active either because Twitter suspended or deleted the account, or because the user decided to delete it. For this reason, on October 22nd 2020, I manually checked each of the inactive accounts for whether they were: (a) deleted ($N = 3,351$), (b) suspended ($N = 2,491$), or (c) active again ($N = 1,246$). I dropped the deleted ones from the analysis as these could have either been suspended by Twitter or deleted by the user, and in this analysis I focus on comparing the ones that had been at some point suspended ($N = 3,737$), to the remaining still-active accounts.

Some of the computational methods used in the analysis are very computationally intensive and unfortunately analyzing the full sample of users and tweets in this paper turned out to be unfeasible. For this reason I randomly sampled 30,000 users (out of the ones that had not been deleted nor suspended) to be compared to the 3,737 suspended accounts. Finally, I dropped from this sample of 33,737 users those that did not send any tweet in 2020, given that most of the analytical hypotheses rely on the content of the users' tweets (and so on the users being somewhat active), and I also wanted to rule out that accounts had been suspended/deleted for being inactive. The final sample I use for analysis is composed of 11,859 users (2,340 suspended accounts and 9,519 non-suspended) and the 5,932,777 tweets they sent in 2020.

In the next section I describe the methods I use in order to build user level variables measuring the proportion of tweets users sent about politics, how supportive the political tweets were of the Iranian government, the proportion of tweets containing hateful language, the ideology of the users, and the extent to which users coordinated with other accounts (whether they sent the same content as other accounts). Along with some additional variables (whether an account is verified, and whether they sent an abnormal number of tweets – indicative of the account being automated), I use all these measures to test the hypotheses I previously described.

Methods

0.6 Coding Process

First, a random sample of 1,172 messages published by these user accounts on Twitter was drawn for manual coding in order to build a training set used for training supervised machine learning classifiers predicting: political tweets (*political content*), messages in favor of

[4] I did not automatically set up a script to regularly pull the data, but instead relied on manually executing the script. I did so the following days: 2020-04-16, 2020-04-20, 2020-04-27, 2020-05-06, 2020-05-08, 2020-05-14, 2020-06-08, 2020-06-18, 2020-07-05, 2020-07-10, 2020-07-16, 2020-07-20, 2020-07-31, 2020-08-17, 2020-09-03, and 2020-09-10.

the Iranian government (*political sentiment*), and tweets containing hateful language (*hateful conduct*). I only annotated those tweets written in Persian, Arabic, and English. In the following section, I will explain the process of manual coding for these three variables.

(1) Political content. Tweets were labelled as political if they (a) mentioned to 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 Table 1 for two examples of political versus apolitical messages from the data set.

(2) Political sentiment. I coded the tweets for whether they were positive, neutral (e.g., an opinion about the Iranian government was given without supporting or criticizing it), or negative towards the Iranian government. I only coded for this sentiment dimension those that had previously been coded as political. See Table 2 for a few examples of messages coded as positive, neutral and negative towards the Iranian government.

(3) Hateful conduct. A tweet was coded as containing hateful language if the message any of the following markers: (1) Making violent threats against an identifiable group, (2) inciting fear about a protected group, (3) wishing, hoping, or calling for serious harm on an individual or group, and (4) making references to violent events. All tweets (political and non-political) were coded for this dimension. For a few examples of messages coded as hateful versus non-hateful, see Table 3.

0.7 Machine Learning Classifiers

I used the manually labeled text to train a total of three machine learning classifiers predicting whether the tweets are (a) political, (b) hateful, and for the political ones, the extent to which they are supportive of the (c) Iranian government. The users in the data set tweet in multiple languages, mostly in Farsi, Arabic, English and Turkish. Rather than training separate models for the tweets on different languages (which would require having a separate large-enough training set for each language), I used the tweets written in Farsi, Arabic and English that had been labeled, and, for each of the three classifiers, I fine-tuned a deep multilingual language model (the *BERT multilingual base model*)[6] that can be trained and used for generating predictions for text data (e.g. tweets) in more than 100 languages.

This model has been self-trained with a large corpora of unlabeled text from Wikipedia, learning deep vocabulary representations that can later be used to accurately perform in many language tasks, such as text classification (the purpose used here) and making next-sentence predictions. For each of the three classifiers, I fine-tuned this pre-trained model by adding a new final layer predicting binary outcomes,[7] and by training this new model architecture for a few more epochs with the tweets I had previously labeled. In Table 4, for each classifier, I report the number of labeled tweets used for training, the percentage of negative and positive labeled cases, the additional epochs used for fine-tuning, as well as the accuracy, precision, recall and f-score of each of the classifiers based on five-fold cross-validation on held-out sets.

I implemented an active learning strategy in order to make sure there were enough labeled tweets, both true positives and true negatives, for each classifier. As previously mentioned, first, I randomly sampled 1,172 tweets and I coded them for each of the three dimensions (political, hateful, and pro-Iran). Then, I used these tweets to train a first version of each of the classifiers. At this stage only the political classifier yielded sufficiently accurate results, this was in part due to having very unbalanced data for the other classifiers – given that only tweets labeled as political were also coded for the pro-Iran, and that tweets using hateful language were very rare (less than 5% of the first batch of 1,172 coded tweets). Hence, I implemented the following approach to increase the size and balance of the training sets for the two remaining classifiers. I used the first rough classifiers to generate hateful and pro-Iran predictions for a set of 30,000 tweets selected at random.

[5] Eshaq Jahangiri is the incumbent Iranian Vice President.

[6] For the code, see `https://github.com/google-research/bert/blob/master/multilingual.md`

[7] For simplification, the pro-Iran classifier was trained only with tweets labeled as being against the Iranian government (`-1`) and tweets being in favor (`1`), disregarding those coded as neutral (`0`). Hence, a predicted probability close to `0` indicates being against the government, probabilities close to `1` indicate being in favor of the government, and probabilities close to `0.5` should indicate tweets that are neutral.

Table 1: Examples of political messages versus non-political messages

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.
Rouhani's tone very defensive today. Said he wrote a letter & warned Europeans there would be "grave consequences" if the US tried to extend the UNSC arms embargo on Iran. #Iran
Non-Political
اثر قطرات اب بر بروی سنگ https://t.co/4RbTFnq1NV *Effect of water drops on stone https://t.co/4RbTFnq1NV*
مناطق طبیعی شمال ایران *Natural areas in the north of Iran*

Table 2: Examples of expressed sentiment towards the Iranian government

Message labeled as showing *negative* sentiment towards the Iranian government
این چه وضع برنامه ریزیه بعد پنج ساعت معطلی در صف اجازه رای ندادن[5] @Eshaq_jahangiri *@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 *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 labeled as showing neutral sentiment towards the Iranian government
اقای دکتر شما نگران نباش ما مجلس داریم اونا رسیدگی میکنن @mah_sadeghi *@mah_sadeghi Please do not worry, sir. We have a parliament which will investigate*
چرا از تجربه کنترل #کرونا برای #اقتصاد استفاده نکنیم؟ تصور کنید روزانه سخنگویی چنین گزارش @drjahanpur @3eyedamir دهد: بیکاران/ شاغلان امروز کارخانجات جدید و تعطیل میزان رشد تولید صادرات واردات و... در پایان از مردم خواهش کند #کالای_ایرانی مصرف و از خرید کالای خارجی پرهیز کنند *Why do we not use our experience with controlling corona for the economy? Just imagine a spokesperson daily releases the following report: Today's number of unemployed/employed, Newly opened factories and closed ones, Growth of export and import etc. In the end, the spokesperson asks the people to consume Iranian products and avoid buying foreign-made products.@3eyedamir @drjahanpur*
Message labeled as showing positive sentiment towards the Iranian government
پای رای مان هستیم #روحانی_تنها_ نیست *We stand by our vote. #Rouhani is not alone*
حالا یک سیلی ای دیشب در #عین_الأسد به اینها زده شد، این مسئله ی دیگری است، آنچه که در مقام مقابله مهم است -این کارهای نظامی به این شکل کفایت آن قضیه را نمیکن -- این است که بایستی حضور فساد برانگیز آمریکا در این منطقه تمام بشود. #انتقام_سخت *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*

For the hateful predictions, I manually labeled those predicted to be hateful and then used those additional labels to train the classifier. I repeated this step a second time, building a training set that yielded satisfactory results. I followed a very similar approach for the pro-Iran classifier, with one key distinction. Instead of only selecting for additional labeling of the positively predicted cases, I selected those that the model confidently predicted ($Pr > 0.85$) to be either against or in favor of the Iran government; also repeating the process a second time.

In Table 4, I show that the resulting classifiers are sufficiently accurate to move forward with the analysis and to learn about the conditions under which accounts in the data set were suspended. The overall accuracy of the political classifier is 82% based on a completely balanced data set (with equal number of true positives and negatives). Moreover, the precision and recall are

Table 3: Examples of hateful vs non-hateful messages

Message labeled as *hateful*
`@realDonaldTrump` Shut your capacious mouth `#mother fucker`
`@McRibFucker69` `@Lubnaneon` I almost can't believe how much of a dumbass you are. Do Americans still have the audacity to talk about democracy in 2020?
Message labeled as *non-hateful*
سردار باغیرت درود به وجدان بیدار و آگاه شما که همچو مرد اشتباه را پذیرفتی `@Hajizadeh_org`
`@Hajizadeh_org` *Praise upon your awake conscience, commander since you accepted responsibility for the mistake. Praise upon you*
کاش دولته بیخیال حذف صفر از پول ملی بشه. میخواد ابرو درست کنه میزنه چشم شو هم کور می کنه
I hope the administration does not remove zeros from the national currency. It wants to solve a problem but instead it would exacerbate it

Table 4: Performance of 3 binary multi-language classifiers predicting tweets about politics, using hateful language, and for those predicted to be about politics, whether they are in favor (v. against) the Iranian government

	Labeled (N)	**Negative (%)**	**Positive (%)**	**Epochs**	**Accuracy**	**Precision**	**Recall**	**F-Score**
Political	1,516	50.3%	49.7%	6	82%	82%	80%	81%
Hateful	1,151	90%	10%	10	93%	74%	36%	47%
Pro-Iran	513	60.8%	39.2%	8	70%	70%	52%	59%

very similar (82% and 80%), meaning that when the classifier gets it wrong (18% of the time only), it is equally likely to misclassify political and non-political tweets, simply adding a bit of noise but not biasing the political measure. The hateful classifier does not perform as well (it misses 2/3 of hateful tweets, underestimating the presence of hateful messages in the data set) but it still helps pick up many hateful tweets with high precision, making correct predictions around 75% of the time. The pro-Iran classifier is also not as precise, and slightly unbalanced (does a better job at detecting tweets in favor rather than against the Iranian government), but overall it is also useful to pick up and analyze the kind of signal I am interested in.

Ideology. I used a *Bayesian Spatial Following* model to estimate the ideology of the followers of the Iranian elites that I study here. This is a widely used strategy in political analyses of social media data (Barberá, 2015; Eady, Zilinsky, Nagler, & Tucker, 2018). The model relies on the homophily assumption that ordinary users follow politicians that better reflect their views, and it has been validated and found to produce accurate ideology estimates for Twitter users in the U.S. context. Below I perform an additional validation to ensure that the method also works in the Iranian context.

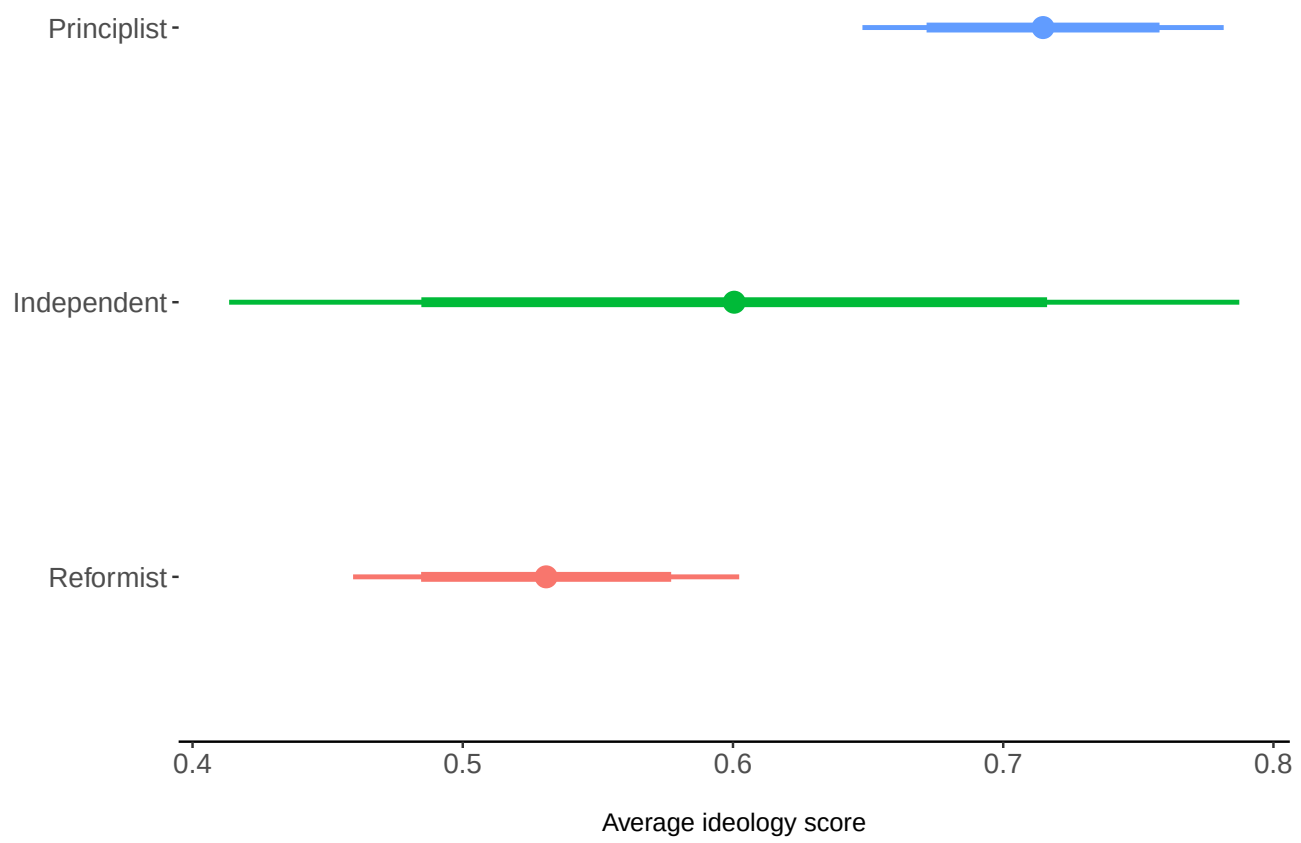
In order to estimate the model, first, I built a bipartite network graph with information about which of the 167 elite accounts each of the 601,940 users in the full sample followed. Then, in order to be able to compute the model,[8] I randomly sampled 5,000 users. These randomly selected users followed most of the elite accounts in the list (152 out of 167, so 91%). Then, I used the `mediascores` package (Eady et al., 2018) to fit the model, obtaining ideology scores (in the same dimension) for the 5,000 users and the 152 elite accounts. Finally, I used the trained model (and so the scores given to the elite accounts) to estimate the ideology of the remaining ordinary users.

In Figure 1 I show that this bayesian spatial model does a good job at distinguishing relevant political dimensions in Iranian politics. I labeled 112 politicians (present in the list of 152 political and media elites with an ideology score) for their political affiliation: *Independent* ($N = 12$), *Principlist* ($N = 41$), or *Reformist* ($N = 59$). Labelling of political affiliation of each elite figure was done according to the available information on the political affiliation of the members of the Parliament (*æsami-e 290*, 2016), and the cabinet members of Rouhani administration (*Gerajesh-e sijasi-e kabine*, 2019).

Principlists, known as *Osul-garajan* in Iran, are defined by their belief in strictly following the doc-

[8] This type of model is very computationally intensive, so one needs to reduce the size of the input matrix in order to be able to estimate the model in a reasonable time frame.

Figure 1: Average ideology scores given by the Bayesian spatial following model to Iranian politicians from the three major affiliations



trine of the Islamic Revolution of Iran while being open to reform (Sadeghi Rad & Mousavi, 2020). Reformists, known as *Eslah-talaban* in Iran, are characterized by their advocacy for civil society and democratization of power in the current political establishment ((Sadeghi Rad & Mousavi, 2020). Those politicians that do not belong to either faction are considered to be independent (including the Supreme Leader).

Figure 1 shows the average standardized [0-1] ideology scores (with 95% confidence intervals) for politicians in each group. A clear (statistically significant) distinction emerges between Reformists and Principlists, indicating that the method does a good job at capturing relevant ideological dimensions in Iranian politics. Also reassuring, the Independents have an ideology score in between the Principalists and the Reformists. The confidence interval for the Independents can be in part explained for the size of the group (4 to 5 times smaller than the other two) but also because politicians in this group remain neutral yet come from a wider ideological background.

Coordination Measure. I used the following empirical strategy to identify coordination among the 11,969 Twitter users I study in detail in this paper. Building on the premise that coordinated accounts share very similar if not the same content, I developed a three-step protocol to measure messaging similarity between all possible pairs of users. First, I selected all the tweets these users sent in 2020. Then, I used the same BERT model that I fine-tuned for building the four machine learning classifiers to generate (768-size) tweet-level embeddings by passing the tweets through the

pre-trained BERT architecture and pulling the output of the second-to-last (fully connected) layer. Then, for each user I generated (768-size) user-level embeddings by averaging the indexed embedding values of all the user's tweets. Finally, I used the resulting $(11,969x768)$ matrix to obtain another $(11,969x11,969)$ matrix measuring the cosine similarity between all possible pairs of author embeddings.

I use the information in this cosine similarity matrix to generate 3 different user-level coordination measures: the (1) maximum, (2) average, and (3) median cosine similarity between the embeddings of a user and the embedding of any other user in this data set. These are all valid ways of capturing the same coordination concept, and I will explore the predictive power of each of them when testing the coordination hypothesis.

Results

In order to address the research questions and test the hypotheses mentioned above, I estimated eight logistic regression models to predict whether accounts were suspended or not as a binary outcome variable. These models differ from each other in that each includes a unique set of terms. The variable that measures stance on the Iranian government (i.e., sentiment variable) is only available for those users who sent at least one political tweet in 2020. Therefore, I excluded the sentiment variable when estimating a first set of models (i.e., Models 1-3). I made this decision to assess the effect of all the other variables without dropping any observation and introducing a potential bias to the data. Exclusion of the sentiment variable was also meant to test the robustness of the findings when estimating another set of models that included additional terms (i.e., Model 4-7). The key difference between the first three models (i.e., Model 1, 2, and 3) is the way in which coordinated behaviour is measured. In Model 1, I measure coordinated behaviour using the average similarity between message contents of each user and contents of messages posted by other users in the data set. In Model 2, rather than taking the average, I measure coordination by looking at the maximum content similarity score, and in Model 3 I use the median of all the scores. Across the three models, the direction of the relationship between coordinated behaviour and the outcome variable did not change regardless of the way in which coordination was operationalized. Therefore, I estimated Model 4, 5, and 6 using only one measurement of coordinated behaviour, that is, average content similarity. These three models (Models 4-6 presented) account for sentiment towards the Iranian government, however, each of them uses a different operationalization of this construct. Model 4 uses average sentiment scores, Model 5 uses median sentiment scores, and Model 6 uses the maximum. Regardless of the way in which sentiment towards the Iranian government was measured, I observed robust results in terms of the direction of marginal effects for all the predictors across these three models. Ultimately, to better analyze the potential effect of automated activity on the likelihood of suspension, I built on the continuous "`Number of tweets 2020`" variable to create a binary measure (i.e., "`Number of tweets 90th percentile`"), which indicates whether or not an account is in the 90th most active percentile. I estimated Model 7 and Model 8 using this binary measure. Model 8, in order to explore the stated RQ2, includes a term that represents the interaction between users' ideology and their sentiment score.

Variable	Model 1	Model 2	Model 3
(Intercept)	-12.2525 (0.9642)*	-108.2734 (7.9932)*	-12.6842 (1.006)*
Coordination (mean)	11.0224 (1.0188)*		
Coordination (median)			11.2842 (1.0448)*
Coordination (max.)		106.8219 (8.0192)*	
Verified user	-1.7999 (0.7352)*	-1.8137 (0.7248)*	-1.7975 (0.734)*
Principlist (Conservative)	1.395 (0.2598)*	1.2543 (0.2606)*	1.3663 (0.26)*
Number of tweets (2020)	0.0003 (0)*	0.0003 (0)*	0.0003 (0)*
Number of tweets 90th percentile (2020)			
Number of political tweets (2020)	0.0001 (0.0001)	0 (0.0001)	0 (0.0001)
Number of hateful tweets (2020)	0.0197 (0.0039)*	0.019 (0.0038)*	0.0198 (0.0039)*
In favor of Iranian government (mean)			
In favor of Iranian government (median)			
In favor of Iranian government (max.)			
Principlist x In favor Iran gov. (mean)			
N	11859	11859	11859
AIC	10969.49	10864.93	10966.34

Variable	Model 4	Model 5	Model 6	Model 7	Model 8
(Intercept)	-14.3487 (1.3731)*	-14.1995 (1.3704)*	-6.9422 (1.1315)*	-13.9258 (1.3688)*	-13.8123 (1.3757)*
Coordination (mean)	12.4018 (1.4237)*	12.4104 (1.4233)*	3.2473 (1.2139)*	11.9341 (1.4192)*	11.9519 (1.4205)*
Coordination (median)					
Coordination (max.)					
Verified user	-1.9226 (0.7201)*	-1.9134 (0.7204)*	-1.6168 (0.6485)*	-1.6804 (0.6213)*	-1.6752 (0.6214)*
Principlist (Conservative)	1.547 (0.2684)*	1.5511 (0.2686)*	0.6889 (0.2718)*	1.6081 (0.2719)*	0.7233 (1.06)
Number of tweets (2020)	0.0003 (0)*	0.0003 (0)*	0.0002 (0)*	0 (0)	0 (0)
Number of tweets 90th percentile (2020)				1.5046 (0.11)*	1.5054 (0.11)*
Number of political tweets (2020)	0 (0.0001)	0 (0.0001)	0 (0.0001)	-0.0001 (0.0001)	-0.0001 (0.0001)
Number of hateful tweets (2020)	0.0209 (0.004)*	0.0211 (0.004)*	0.0116 (0.0033)*	0.0127 (0.0037)*	0.0127 (0.0037)*
In favor of Iranian government (mean)	2.3322 (0.2551)*			2.303 (0.2577)*	1.9481 (0.4849)*
In favor of Iranian government (median)		1.9473 (0.2299)*			
In favor of Iranian government (max.)			3.7115 (0.1941)*		
Principlist x In favor Iran gov. (mean)					2.4495 (2.8347)
N	9898	9898	9898	9898	9898
AIC	9821.07	9833.08	9485.1	9643.75	9645

In this section, I report the effect of the statistically significant predictors of account suspension while rank-ordering them by the magnitude of their effect. Tables on page 13 provides the coefficients for all the models. However, in order to more clearly interpret the results, in Figure 2 I use the model parameters to report standardized marginal effects, which are changes in the likelihood of account suspension due to a one standard deviation increase in each predictor while keeping the remaining predictors constant.[9]

First, I find that accounts that show bot-like behaviour run the highest risk of suspension by Twitter. In particular, accounts in the 90th most active percentile in terms of messaging are 350% more likely to be taken down by the platform. These findings show that there is an extremely high risk of account suspension associated with hyper-messaging (i.e., posting messages in abnormally large numbers). Hyper-messaging is a common behavioural trait among automated accounts (Alfifi & Caverlee, 2017), and these findings suggest that Twitter is more sensitive to automated activity than any other factor measured here. As for the continuous measure of automated activity, I also observe that a one standard deviation increase in the number of messages posted in 2020 increases the likelihood of suspension at least by 25% and at most by 50%, depending on the model's specifications. Despite the smaller effect size of the continuous measure of automation, these results solidify the claim that Twitter is intolerant to bot-like behaviour, confirming my hypothesis (H3). Given that US lawmakers and citizens have increased the pressure on social media companies in recent years to take tougher measures against malicious automated activities on their platforms, it stands to reason that Twitter substantially punishes automated accounts with suspension (Gorwa & Guilbeault, 2020). This observation is consistent with findings of previous studies that have established a positive relationship between automated behaviour and probability of account suspension (e.g. Ferrara, 2020).

Coordinated behaviour, measured as a maximum content similarity, is the second strongest predictor of the likelihood of account suspension. A one standard deviation increase in this predictor doubles the likelihood of account suspension. This finding confirms my hypothesis (H4). The marginal effect remains very large even when measuring coordination using the average rather than the maximum similarity score between the content of a user's tweets and the messages sent by any other users in the data set. Notably, the large effect sizes of automated nature and coordinated behaviour shows that both variables have a strong association with account suspension. This finding may be due to the close link between coordinated behaviour and use of automated accounts. Previous studies have shown coordinated social media campaigns, at times, deploy automated accounts for malign purposes including artificial manipulation of public opinion and incitement of hatred (e.g. Gruzd & Mai, 2020).

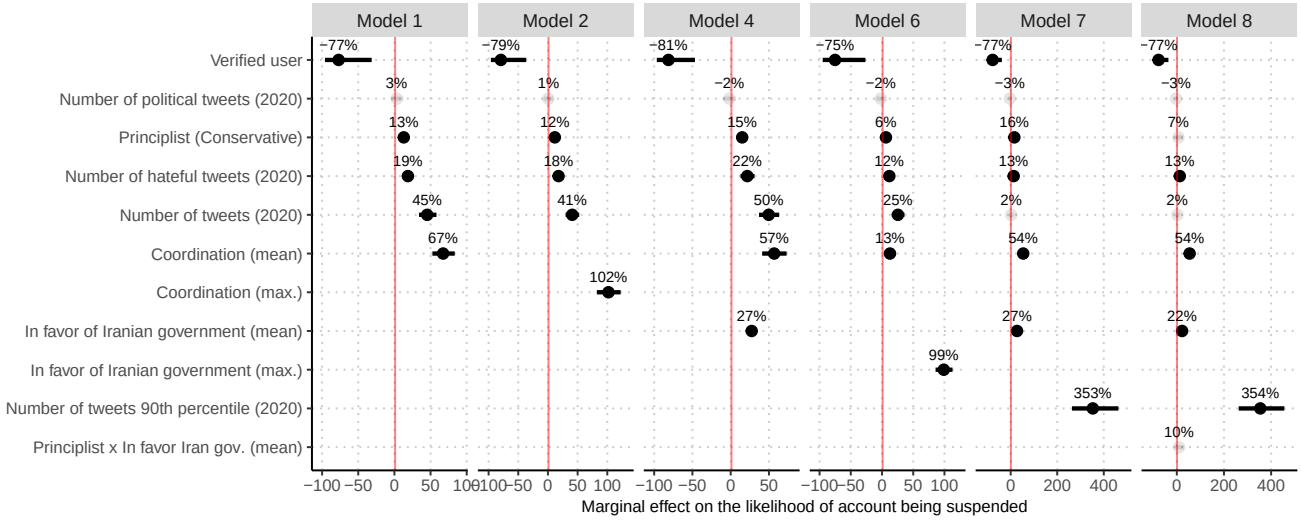
Interestingly, expression of positive sentiment towards the Iranian government, operationalized as the maximum sentiment score, is as strong a predictor of account suspension as coordinated behaviour. A one standard deviation increase in this variable leads to a 99% increase in the likelihood of account suspension. When the sentiment variable was measured as average sentiment score, I observed a weaker positive association between this predictor and account suspension (22% - 27%). These observations confirm my hypothesis (H2), indicating that Twitter is more likely to suppress voices that support the Iranian government or users whose political views are in line with the stance of the Iranian government. Moreover, these findings suggest that Twitter does not provide its Iranian users with a neutral platform for engagement in political discussions.

Verified accounts are at least 75% and at most 81% less likely to be suspended by Twitter compared to unverified ones, confirming my hypothesis (H6). There are at least two explanations for the substantially higher probability of suspension among unverified users relative to verified ones. First, verified users on Twitter are assigned a public interest value, as a result of which, the platform is more lenient with them compared to unverified ones even if the former violates some of the platform's terms of service in the same manner as the latter (Paul, Khattar, Kumaraguru, Gupta, & Chopra, 2019). Second, account verification authenticates the user administering the account, indicating that messages are posted by a genuine user.

I observe a positive yet somewhat weaker relation-

[9] With the exception of the binary variable "Verified user". In this case I report the marginal effect of a user being verified.

Figure 2: Marginal effect of different predictors on the likelihood of account suspension for five models



ship between engagement in hateful conduct and account suspension. A one standard deviation change in this variable increases the probability of account suspension at least by 12% and at most by 22%. Users that posted more hateful messages during the period of study suffered a higher probability of account takedown. This finding confirms my hypothesis (H5) that Twitter is more likely to suspend user accounts that posted hateful messages. The weak association can be attributed to challenges that Twitter and other social media companies face regarding the detection of hateful conduct, mainly due to heavy reliance of such platforms on manual methods of identification and suspension of abusive users (Zhang & Luo, 2019), and likely difficulties associated with dealing with this issue in a multilanguage fashion.

Finally, I find that the weakest (although still substantive) predictor of account suspension in this sample is a user's ideology. A one standard deviation change in this variable increases the probability of account suspension at least by 6% and at most by 16%. Answering my first research question (RQ1), this finding indicates that Twitter is biased against conservative Iranian voices compared to reformist Iranian voices on its platform. Given that a large-N observation of political ideology bias on Twitter does not have precedence in the relevant literature, this finding, for the first time, sheds new light on political ideology bias on the platform. Some have argued that conservative Americans run a higher risk of suspension than liberal Americans on the platform (Hanania, 2018), but the findings presented here can hardly be generalized to the US context, as the geopolitical aspect in the Iranian context is likely to play a crucial role.

Also, I analyzed the relationship between engagement in political discussions and account suspension due to the growing relevance of social media for citizens, politicians and political institutions in the context of spreading political information and shaping public opinion (Stieglitz & Dang-Xuan, 2013). I examined the number of political tweets that each user in the sample posted in 2020 to examine the marginal effect of this variable on the likelihood of account suspension. Across the models, I observe that posting political messages has a small and statistically insignificant marginal effect on the probability of suspension. Therefore, the data does not support my first hypothesis (H1). These findings suggest that despite existing concerns and preliminary evidence, Twitter does not necessarily suspend accounts for engaging in political discussions on its platform. Even though prior research (e.g. Chowdhury et al., 2020) found a great number of politically interested users to be part of a set of accounts purged by Twitter, I did not find a systemic pattern in this regard while controlling for other relevant covariates.

As for my second research question (RQ2), I do not find evidence that the interaction between user ideology, and support of the Iranian government, influences the likelihood of account suspension. Ideology (Reformist-Principlist continuum) and opposi-

tion/support of the government are two distinct analytical dimensions. From an ideological standpoint, government officials tend to be moderate or independent, and support of the government usually mostly captures how supportive users are of specific government policies and its position vis-a-vis other countries. I find that users who are Principlists and also supportive of the government are not more likely to be suspended than those who are not supportive of the government.

Discussion

In the twenty-first century, social media platforms have established themselves as the new public square on a promise of democratization of communicative infrastructure. At the same time, the clash between adherence to ideals of free speech on one side and confronting practical matters including different forms of abuse on their platforms in addition to resisting governmental pressures on the other puts social media platforms in a double bind. What is alarming about the current dynamic is the fast-paced deviation of Western social media platforms from the democratic ideal of free speech that these technologies promised to promote with their advent. This problem can be exacerbated due to challenges related to rapidly evolving techniques of social media abuse and further entanglement of social media platforms in the interests of nation-states and priorities defined by them. Moreover, social media platforms are virtually assigned to regulation of online speech, including censoring content and suppressing voices (Balkin, 2018). A key issue with social media platforms' regulatory roles is lack of transparency about them since it is not clear what the conditions under which private social media companies suppress free speech on their platforms are. This lack of transparency poses a threat to democratic norms and accountability, and in the long run may affect the platforms' ability to maintain an impartial and inclusive environment for expression of political views.

In this paper, I analyzed a set of account features and online behaviour of 601,940 users that follow Iranian politics over a period of six months in 2020, and compared differences between a group of suspended accounts to a group of active accounts. I used state-of-the-art machine learning techniques to predict which messages were about politics, supportive of the Iranian government, and contained hateful content in addition to measuring coordinated behaviour and user ideology. After controlling for many potential confounders, unsurprisingly I found that accounts that are likely to be bots as well as accounts that are likely to participate in malicious coordinated activities, are much more likely to be suspended by Twitter. But more importantly, I also find that conservative (Principlist) accounts as well as those supportive of the Iranian government are also more likely to be suspended. The paper hence not only brings a first set of illuminating findings on the role of private social media companies in regulating freedom of speech, but it also makes a substantial contribution to our understanding of the role that these companies play in ongoing geopolitical struggles. Furthermore, the paper also makes a clear methodological contribution, laying out a set of innovative computational methods that can be used by many in the future to address further questions about freedom of speech in social media platforms.

The findings of the study are robust across many statistical specifications, but also have some limitations. First, despite growing concerns about misinformation on Twitter (Allcott, Gentzkow, & Yu, 2019), I did not control for user engagement in dissemination of misinformation. Future research should find a way to reliably measure misinformation dissemination in the Persian Twittersphere assess whether the findings presented here hold after controlling for this additional confounder. Second, I only looked at Twitter users following formal Iranian elites (e.g. media accounts and members of the General Assembly), not taking into account many other users interested in Iranian politics despite not following the group of elites studied here. However, preliminary anecdotal evidence suggested that supporters of the Iranian government were particularly vulnerable to account suspensions, therefore I focused on this group of elites. Finally, generalization of the findings to other countries currently in a geopolitical rivalry with the United States (such as Russia or Venezuela) is unknown. Twitter has suspended many users based in countries that are a US allies (e.g., Saudi Arabia). Further research is needed in order to uncover a more complete picture of regulation of political speech on Twitter's platform, in general, and the conditions under which it allows free political expression for users in countries with a geopolitical rivalry with Western countries, in particular.

In conclusion, this paper is a first of a kind large-scale research on systemic censorship of political speech on Twitter. My analysis showed that political speech on Twitter comes at a cost for Iranians, and possibly for users from countries in a rivalry with the US. This work intends to help research on political censorship by social media move forward through its research design and empirical findings, encouraging future research from many disciplines, including political and communication science, public policy, journalism, legal studies, and any other field concerned with freedom of political speech on the Internet.

1 References

Accounts, A. V. (n.d.). *About verified accounts.* Retrieved 2020-01-09, from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Agency, F. N. (2020). *Instagram removes posts supporting martyred iranian n. scientist.* Retrieved 2020-11-28, from https://www.farsnews.ir/en/news/13990908000750/Insagram-Remves-Pss-Sppring-Maryred-Iranian-N-Scienis

Alfifi, M., & Caverlee, J. (2017). Badly evolved? exploring long-surviving suspicious users on twitter. In *International conference on social informatics* (pp. 218–233). https://doi.org/10.1007/978-3-319-67217-5_14

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 2053168019848554. https://doi.org/10.1177/2053168019848554

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)* (pp. 258–265). https://doi.org/10.1109/ASONAM.2018.8508646

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL Rev.*, *51*, 1149.

Balkin, J. M. (2018). Free speech is a triangle. *Colum. L. Rev.*, *118*, 2011.

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

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, *33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, *21*(11-7).

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy

and decision making. *Policy & internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Chávez, A. (2019). *Twitter tips the scale toward incumbents by refusing to verify primary challengers.* Retrieved 2020-02-07, from https://theintercept.com/2019/10/11/twitter-verification-2020-candidates-incumbents/

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2017). A study on real-time low-quality content detection on twitter from the users' perspective. *PloS one*, *12*(8), e0182487. https://doi.org/10.1371/journal.pone.0182487

Chowdhury, F. A., Allen, L., Yousuf, M., & Mueen, A. (2020). On twitter purge: A retrospective analysis of suspended users. In *Companion proceedings of the web conference 2020* (pp. 371–378). https://doi.org/10.1145/3366424.3383298

Clinton, H. (2010). *Statement: Hillary clinton on internet freedom.* Retrieved 2020-10-10, from https://www.ft.com/content/f0c3bf8c-06bd-11df-b426-00144feabdc0

Cobbe, J. (2020). Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 1–28. https://doi.org/10.1007/s13347-020-00429-0

Cockerell, I. (2020). *Instagram shuts down iranian accounts after soleimani's death.* Retrieved 2021-01-24, from https://www.codastory.com/authoritarian-tech/instagram-iran-soleimani/

conduct policy, H. (n.d.). *Hateful conduct policy.* Retrieved 2020-01-09, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

DataReportal. (2021). *Global social media overview.* Retrieved 2021-01-24, from https://datareportal.com/social-media-users

Dellinger, A. (2020). *Facebook suspended the accounts of environmental activists trying to plan a protest.* Retrieved 2021-01-24, from https://www.mic.com/p/facebook-suspended-the-accounts-of-environmental-activists-trying-to-plan-a-protest-34580215

Detmer, D. (1995). Covering up iran: Why vital information is routinely excluded from us mass media news accounts. *The US media and the Middle East: Image and perception*, 91–101.

Eady, G., Zilinsky, J., Nagler, J., & Tucker, J. (2018). *Methodological supplement.* Retrieved from http://htmlpreview.github.io/?https://github.com/SMAPPNYU/mediascores/blob/master/vignettes/mediascores-vignette.html

Effing, R., van Hillegersberg, J., & Huibers, T. (2011). Social media and political participation: Are facebook, twitter and youtube democratizing our political systems? In E. Tambouris, A. Macintosh, & H. de Bruijn (Eds.), *Electronic participation* (pp. 25–35). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23333-3_3

Elections, U. (2020). *Us election: Twitter tightens rules on retweets and victory claims.* Retrieved 2020-10-09, from https://www.bbc.com/news/technology-54485697

Eye, M. E. (2017). *Twitter under fire after suspending egyptian journalist wael abbas.* Retrieved 2021-01-24, from https://www.middleeasteye.net/news/twitter-criticised-after-suspending-account-egyptian-journalist-wael-abbas

Ferrara, E. (2020). Bots, elections, and social media: a brief overview. *Disinformation, Misinformation, and Fake News in Social Media*, 95–114. https://doi.org/10.1007/978-3-030-42699-6_6

Gleicher, N. (2019). *Removing coordinated inauthentic behavior from china.* Retrieved 2020-06-12, from https://about.fb.com/news/2019/08/removing-cib-china

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, *12*(2), 225–248. https://doi.org/10.1002/poi3.184

Grossman, L. (2009). *Iran protests: Twitter, the medium of the movement.* Retrieved 2020-11-10, from http://content.time.com/time/world/article/0,8599,1905125,00.html

Gruzd, A., & Mai, P. (2020). Going viral: How a single tweet spawned a covid-19 conspiracy theory on twitter. *Big Data & Society*, *7*(2), 2053951720938405. https://doi.org/10.1177/2053951720938405

Halliday, J. (2012). *Twitter's tony wang: "we*

are the free speech wing of the free speech party". Retrieved 2021-01-24, from https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech

Hanania, R. (2018). *It isn't your imagination: Twitter treats conservatives more harshly than liberals.* Retrieved 2021-01-27, from https://quillette.com/2019/02/12/it-isnt-your-imagination-twitter-treats-conservatives-more-harshly-than-liberals/

Helberger, N., Kleinen-von Königslöw, K., & Van Der Noll, R. (2015). Regulating the new information intermediaries as gatekeepers of information diversity. *info*. https://doi.org/10.1108/info-05-2015-0034

Howard, P. N. (2010). *The digital origins of dictatorship and democracy: Information technology and political islam*. Oxford University Press.

Inc., T. (n.d.). *Political content.* Retrieved 2021-01-27, from https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html

Inc., T. (2019a). *Defining public interest on twitter.* Retrieved 2020-01-09, from https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html

Inc., T. (2019b). *World leaders on twitter: principles & approach.* Retrieved 2020-01-09, from https://blog.twitter.com/official/en_us/topics/company/2019/worldleaders2019.html

Inc, T. (2021). *Permanent suspension of @realDonaldTrump.* Retrieved 2021-01-24, from https://blog.twitter.com/en_us/topics/company/2020/suspension.html

Jeong, S. (2016). *The history of twitter's rules.* Retrieved 2021-01-24, from https://www.vice.com/en/article/z43xw3/the-history-of-twitters-rules

Khazraee, E. (2019). Mapping the political landscape of persian twitter: The case of 2013 presidential election. *Big Data & Society*, *6*(1), 2053951719835232. https://doi.org/10.1177%2F2053951719835232

Khonsari, K. K., Nayeri, Z. A., Fathalian, A., & Fathalian, L. (2010). Social network analysis of iran's green movement opposition groups using twitter. In *2010 international conference on advances in social networks analysis and mining* (pp. 414–415). https://doi.org/10.1109/ASONAM.2010.75

Klonick, K. (2017). *The terrifying power of internet censors.* Retrieved 2021-01-24, from https://www.nytimes.com/2017/09/13/opinion/cloudflare-daily-stormer-charlottesville.html

Landon-Murray, M., Mujkic, E., & Nussbaum, B. (2019). Disinformation in contemporary us foreign policy: Impacts and ethics in an era of fake news, social media, and artificial intelligence. *Public Integrity*, *21*(5), 512–522. https://doi.org/10.1080/10999922.2019.1613832

Langvardt, K. (2017). Regulating online content moderation. *Geo. LJ*, *106*, 1353.

manipulation, P., & spam policy. (2020). *Platform manipulation and spam policy.* Retrieved 2020-01-09, from https://help.twitter.com/en/rules-and-policies/platform-manipulation

Mottahedeh, N. (2015). *# iranelection: Hashtag solidarity and the transformation of online life*. Stanford University Press.

of Journalists, I. F. (2020). *Iran: Journalists demand end to censorship of iranian media on instagram.* Retrieved 2021-01-24, from https://www.ifj.org/media-centre/news/detail/category/press-releases/article/iran-journalists-demand-end-to-censorship-of-iranian-media-on-instagram.html

O'Sullivan, D., & Moshtaghian, A. (2020). *Instagram says it's removing posts supporting soleimani to comply with us sanctions.* Retrieved 2020-09-17, from https://edition.cnn.com/2020/01/10/tech/instagram-iran-soleimani-posts/index.html

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2020). Uncovering coordinated networks on social media. *arXiv preprint arXiv:2001.05658*.

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M., & Chopra, S. (2019). Elites tweet? charac-

terizing the twitter verified user network. In *2019 ieee 35th international conference on data engineering workshops (icdew)* (pp. 278–285). https://doi.org/10.1109/ICDEW.2019.00006

Quintana, C. (2018). *Why did these scholars suddenly find their twitter accounts suspended?* Retrieved 2021-01-24, from https://www.chronicle.com/article/why-did-these-scholars-suddenly-find-their-twitter-accounts-suspended/

Reilly, R. J. (2015). *Fbi: When it comes to @ISIS Terror, retweets = endorsements.* Retrieved 2021-01-24, from https://www.huffpost.com/entry/twitter-terrorism-fbi_n_55b7e25de4b0224d8834466e

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *Proceedings of the international aaai conference on web and social media* (Vol. 12).

Roth, Y. (2019). *Information operations on twitter: Principles, process, and disclosure.* Retrieved 2020-11-24, from https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter.html

Sadeghi Rad, H., & Mousavi, S. R. (2020). Ayatollah khamenei and reformism in iran. *Journal of Contemporary Research on Islamic Revolution*, *2*(4), 1–20.

Sharma, K., Ferrara, E., & Liu, Y. (2020). *Identifying coordinated accounts in disinformation campaigns.*

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28–41.

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social network analysis and mining*, *3*(4), 1277–1291. https://doi.org/10.1007/s13278-012-0079-3

teleSUR. (2019). *Twitter suspends accounts of cuba's largest media outlets.* Retrieved 2021-01-24, from https://www.telesurenglish.net/news/Twitter-Suspends-Accounts-of-Cubas-Largest-Media-Outlets-20190912-0004.html

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 acm sigcomm conference on internet measurement conference* (pp. 243–258). https://doi.org/10.1145/2068816.2068840

Timberg, C., & Dwoskin, E. (2018). *Twitter is sweeping out accounts like never before.* Retrieved 2021-01-24, from https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

@TwitterSafety. (2020). *Disclosing networks to our state-linked information operations archive.* Retrieved 2020-10-08, from https://blog.twitter.com/en_us/topics/company/2020/disclosing-removed-networks-to-our-archive-of-state-linked-information.html

VanLandingham, R. E. (2017). Jailing the twitter bird: Social media, material support to terrorism, and muzzling the modern press. *Cardozo L. Rev.*, *39*, 1.

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science* (pp. 353–362). https://doi.org/10.1145/3292522.3326016

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925–945. https://doi.org/10.3233/SW-180338