Five Hundred Days of Farsi Twitter: An overview of what Farsi Twitter looks like, what we know about it, and why it matters

LAYLA M. HASHEMI[1]
George Mason University, USA

STEVEN L. WILSON
Brandeis University, USA

CONSTANZA SANHUEZA PETRARCA
WZB, Germany

International media was quick to dub the Iranian Green Movement a "Twitter revolution" when it erupted in the summer of 2009. State violence against protestors was captured in real time and broadcast worldwide on social media, providing an early example of a regime's helplessness at locking down a narrative in the face of ubiquitous smart phones. Over a decade later, nearly all foreign social media remain officially blocked in Iran, yet Iranians evade state suppression and remain connected to the global community. This article introduces a new dataset of all Farsi-language tweets since September 2019. To date, this amounts to the full text and associated metadata of over 500 million tweets and the evidence shows that the overwhelming majority of this content originates from within the borders of Iran. The study describes the scope of Iran's continued connection to the global community via Twitter, descriptively explores the content of that social media, evaluates what this means for Iranian politics and society, and explores its broader implications for researchers in the age of social media. In particular, we argue that the demonstrated ability to collect the voices of citizens, even from one of the most repressive digital regimes in the world, provides an invaluable framework for scholars with even minimal resources to undertake large-scale digital ethnography.

Keywords: social media, Twitter, Iran, social identity, content analysis

[1] Hashemi: laymay.19@gmail.com
Wilson: stevenwilson@brandeis.edu
Date submitted: 2021-04-18

1. The Internet in Iran

The Iranian regime is one of the strictest regimes in the world in terms of internet surveillance and filtering. The Digital Society Project (DSP) ranks Iran 8th worst in the world in censorship of social media, and among the most capable authoritarian regimes in the world in terms of regime cyber security capacity, internet shutdown capabilities, online monitoring, arrests for online political posting, and the domestic deployment of disinformation online (Mechkova, et al. 2020).[2]

Despite this overwhelming state cyber capacity arrayed against them, the Iranian population widely uses the internet and social media, remaining well connected and tech savvy. Nearly 70% of the over 81 million Iranians have access to the internet.[3] Furthermore, a 2017 survey shows that 24% of Iranians use Twitter every day and that 63% believe Twitter is a trusted source for news[4], proportions that are similar to those found in consolidated democracies around the world.

Iranians have been active online since the dawn of the internet era. Farsi blogs were an early and popular mode of information dissemination prior to the Green Movement in 2009, serving as fora to discuss socio-political issues ranging from human rights to sports and politics.

The early adoption of the internet by Iranian users resulted in a large base of digitally capable youth who successfully used social media platforms for mobilization to contest the 2009 presidential election of Mahmoud Ahmadinejad. Although the government responded to the Green Movement's usage of social media by blocking nearly all foreign social media platforms (including Facebook and Twitter), the homegrown technical capacity of Iranians ensured technical workarounds such as virtual private networks (VPNs), and other such tactics that helped them stay connected to international social media. In fact, Iran ranks fifth among all authoritarian regimes in

[2] Mechkova, Valeriya, Daniel Pemstein, Brigitte Seim, Steven Wilson. 2020. *Digital Society Project Dataset v2.* See indicator v2smgovsmcenprc, v2smgovcapsec, v2smgovshutcap, v2smgovsmmon, v2smarrest, and v2smgovdom in particular.

[3] "Iran Internet Stats and Telecommunications Reports." https://www.internetworldstats.com/me/ir.htm (March 22, 2020).

[4] Jafari, Hamed. 2017. "Infographic: Twitter Usage Statistics in Iran." *TechRasa*. http://techrasa.com/2017/08/02/infographic-twitter-usage-statistics-iran/ (March 22, 2020). While Instagram is also very popular, Twitter public accessibility allows for Iranians who aren't members to obtain Twitter information indirectly. For example, the Telegram (one of Iran's most popular social media sites) group 'Farsi Twitter' has over 100,000 members.

the frequency of average people's use of social media to organize offline political action, a staggeringly high level given the regime's repressive capacity in the cyber realm (Mechkova, et al., 2020).[5]Moreover, Iranians often use alternative platforms to organize and coordinate political action. For example, during the November 2019 blackout, other GPS enabled technologies such as Waze were reportedly used to coordinate car strikes in response to the spike in gas prices (Sanchez 2019). Even when Iranians cannot directly access Twitter, there exist popular alternative dissemination accounts like Twitter Farsi, a Telegram channel with over 530,000 subscribers, which compiles and shares content from the Persian Twittersphere (Official Persian Twitter 2021).

The capacity of Iranians to work around restrictions is not unknown to the Iranian government, which itself extensively uses social media to communicate with the population. Surprisingly, Iranian government officials are active users of the same services they condemn and disparage: former President Hassan Rouhani has over 1.1 million Twitter followers and the Supreme Leader, Ayatollah Khamenei Twitter account @khamenei_ir has nearly 1 million followers. One of the most prominent conservative officials on the platform, Supreme Leader Khameini has accounts in English, Farsi, and Arabic in order to reach a large and diverse audience. While Hassan Rouhani was an active tweeter in both Farsi and English during his presidency, his English account (@HassanRouhani) activity dropped significantly after tweeting several times on September 22, 2020. The majority of his English language tweets during his last year in office addressed the COVID-19 pandemic and championed his administration's efforts to curb the spread of the virus.

Given these developments, the Iranian Twittersphere provides a particularly valuable source into the political discourse of citizens and elites alike in an autocratic context.

2. A New Dataset of Farsi Twitter

Using Twitter's streaming keyword API, we collected all Farsi language tweets worldwide from September 19, 2019 to January 28, 2021. While the streamer continues to collect data (and we plan to collect data indefinitely for future research), this particular article's data ends at that day both for the timing of its writing, and for the convenience of being exactly 500 complete days of data.

[5] Mechkova, Valeriya, Daniel Pemstein, Brigitte Seim, Steven Wilson. 2020. *Digital Society Project Dataset v2.* See indicator v2smorgavgact in particular.

This original dataset of Farsi language tweets provides a wealth of insight into both elite and public discourse of Iranians both inside and outside of the country. Exploring parts of this vast dataset should be a boon to any researcher in the social sciences or humanities with research interests that intersect with Iran in general, and content analysis of that world in particular. Fields that benefit from text analysis of public conversations include sociology, computer science, computational social science, political science, international relations, media and communication studies, ethnography, anthropology, linguistics, literature, musicology, and of course those in area studies with a focus on Iran.

A new Farsi Twitter dataset would allow researchers to examine a wide variety of topics. This includes but is not limited to the study of social media use in repressive contents such as analyzing public discourse of socio-political issues, technology policy implementation, network infrastructure and access, public engagement, and discussion during and across elections, use of media for activism and entertainment and other issues involving Farsi Twitter and Iranian public opinion within and across the various disciplines outlined above.

While the Twitter terms of service disallow directly making the dataset public, accepted norms exist for sharing data internally among the members of a research team. As such, we invite any researchers across disciplines to reach out to the authors for collaboration on this exciting and extensive dataset. Where that is not possible, we can arrange for the public sharing of compilations of unique tweet ids (in accordance with Twitter's terms of service) on subsets of the data of interest to particular researchers in order to rehydrate those subsets via Twitter's API on their own systems.

2.1 Data Collection and Robustness Checks

We accomplished this data collection by designing the stream to return all tweets classified as being in the Farsi language and matching a list of specified keywords. Twitter's API does not have a minimum word length for searching and returns partial word matches (i.e. searching for "a" would return any tweet with the word "bat", or any other instance of the letter "a"). Therefore, we constructed a keyword search list of the 32 letters of the Farsi alphabet, plus a selection of 332 stop words from Google's Farsi language stop word list, for robustness. This would in principle return *all* tweets in the Farsi language in real time since it is asking for all tweets that are in Farsi and contain a letter from the Farsi alphabet, provided the total number of returned tweets was less than one percent of the total Twitter stream (the cap for streamers).

The result has been over 500 million tweets to date, averaging just under 1.1 million tweets per day (a rate which is increasing over time). As a robustness check, we compared our figures to the totals of Farsi language tweets from the Twitter sample stream, which has a guaranteed 1% random sample of all tweets worldwide. If our stream was indeed capturing the full Farsi stream, our totals should have been approximately one hundred times those values. With expected minor variation because of the randomness of the process (and a small number of tweets with a Farsi language tag, but no actual Farsi letters), this was the case, leading us to have strong confidence we were indeed capturing all Farsi language tweets.

In addition, in our close reading of many thousands of tweets as part of this analysis, it was clear that with very few exceptions the tweets were clearly in Farsi, and that the Twitter language identifier was not systematically falsely identifying non-Farsi tweets as Farsi (for instance, we kept an eye out for Arabic language tweets falsely misclassified as Farsi due to the similar alphabet). Finally, we manually tweeted several test tweets in Farsi from other Twitter accounts and were able to document that our streamer captured them. As such, we are confident that we are capturing close to all Farsi language tweets and have found no evidence of systematic false negatives or false positives on that scraping process.

Geocoded tweets represent about 1.5% of all tweets worldwide, however the proportion of Farsi language tweets that are geocoded are a third that at 0.5%.[6] When GIS analysis is done to identify the country of origin on the basis of geocodes, only about half of that tiny number of geocoded Farsi tweets originate from within Iran. Furthermore, only 3.9% of Farsi language Twitter users *ever* use geocoding in *any* of their tweets, compared to 17% of Russian language Twitter users, and 17% of Swedish language Twitter users (to draw from two comparable sets of language data collections). Since any usage of VPNs or similar technology would preclude geocoding, this makes reasonable sense.

We also performed robustness checks for bot activity in the Farsi language stream, using version three of Indiana University's Botometer software on a random sample of 5,000 user accounts from our stream (Yang, et al, 2019). It assessed only 1.6% of the users as likely to be bots at a 95% confidence level. We used the language neutral version of the Botometer algorithm, which is designed to detect inauthentic behavior based on a training set of known bots, based on the social network structure, friend and retweet networks, and time-based patterns in behavior.

[6] For summary statistics on geocoded tweets globally, see: Wilson, Steven L. *Social Media as Social Science Data*. Forthcoming.

In addition, we can provide striking evidence that the bulk of this Twitter activity is taking place within Iran itself, despite the government blocks of foreign social media sites. Ironically, we can do so thanks to the government's periodic shutdowns of the *entire* internet within Iran at key points. These create a natural experiment-like setting: while Farsi language Twitter activity originating from outside of Iran would not be impacted by these total shutdowns, domestic activity ceases. As such, by mapping our traffic temporally onto internal shutdowns, we can estimate the proportion of Farsi language Twitter users who are using technical workarounds to avoid government's censorship to post from within Iran itself.

According to Netblocks, an independent, non-partisan civil society group working at the intersection of digital rights, cyber-security and internet governance, Iran's traffic dropped to 5% of normal activity beginning on November 16, 2019 and only began to be restored late November 21 (Sapra, 2021). The country's limited mobile broadband connectivity, distributed by three national providers, experienced similar disruptions. The return of internet traffic to normal happened gradually over several days, which could have several explanations. It is possible that Iranians began self-censoring once the internet was restored, fearing the same fate as the hundreds who were arrested and killed during the protests (Amnesty International 2019). An alternative explanation could be that connections were only partially restored with limited speed and access in the first days which impacted the activity.

Figures 1 and 2 highlight the starkness of the drop in Farsi language Twitter activity during the five day internet shutdown, including the two weeks before and after. The total number of Farsi tweets dropped by 63% (from an average of 770,000 per day before and after to 288,000 during the shutdown) and the total distinct user accounts per day dropped by 62% (from an average of 85,000 per day before and after to 32,000 per day during). This evidence convincingly demonstrates that around two-thirds of Farsi language Twitter activity is originating from within Iran, rather than from the global Farsi-speaking diaspora.

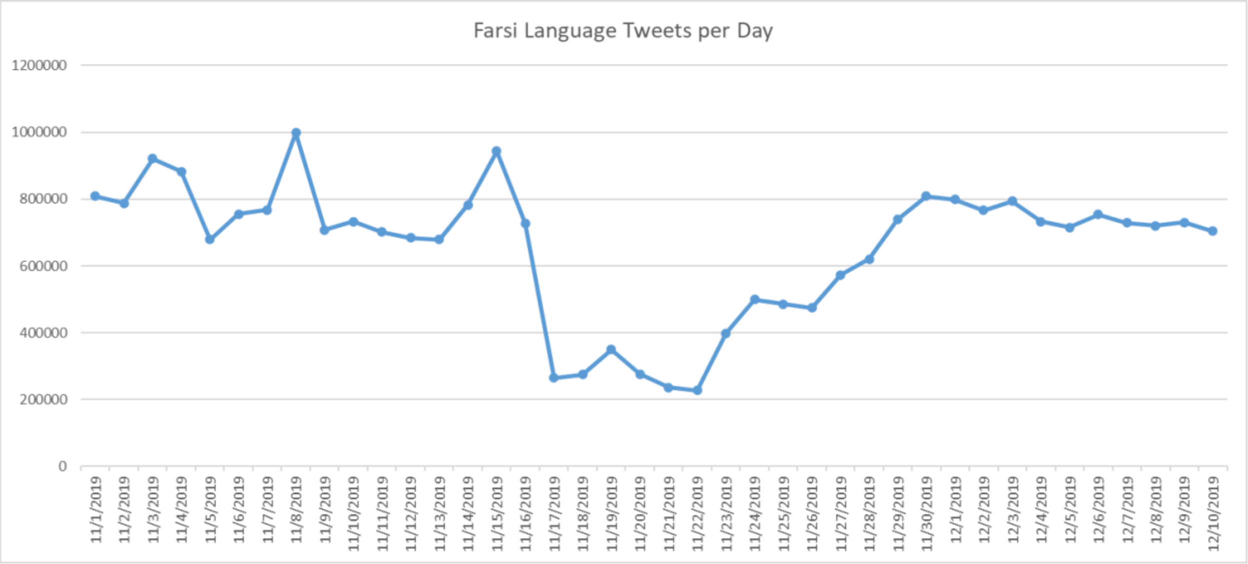

Figure 1. Farsi Language Tweets Per Day (November 2019)

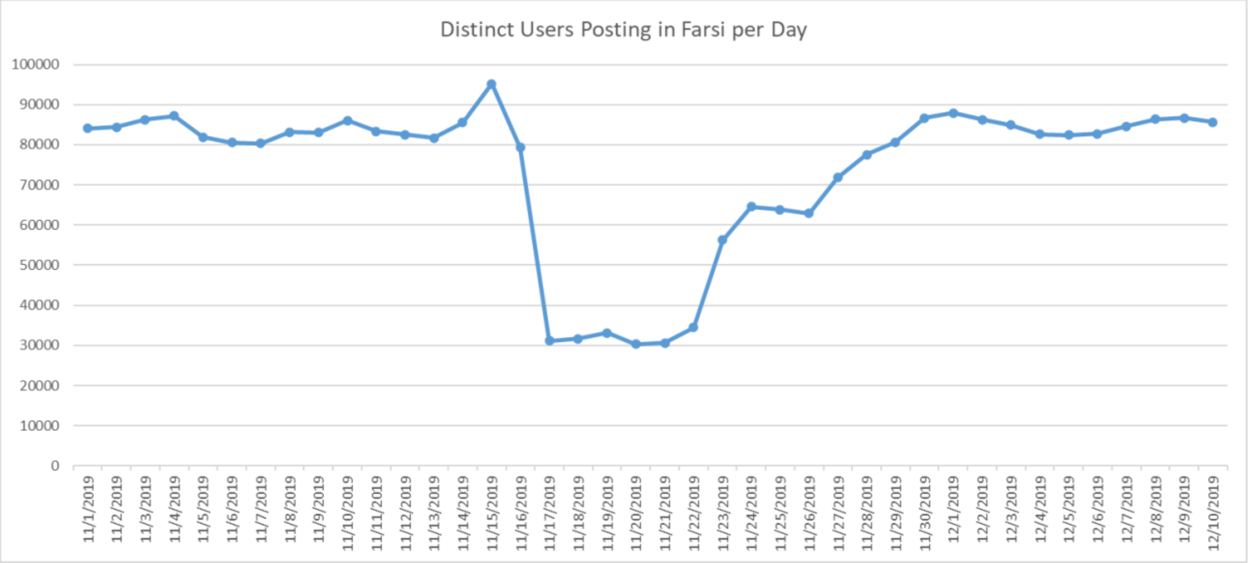

Figure 2. Distinct Farsi Language Users Per Day (November 2019)

2.2 Summary Statistics of Farsi Twitter

Our dataset contains 538 million Farsi language tweets, which amounts to approximately 9 billion words of textual content. The tweets were posted by 3.3 million distinct users, representing about 6% of Iran's internet-connected population. Metadata and non-textual content reveal additional rich layers of communication. Table 1 summarizes the distinct instances of hashtags, mentions, URLs, images, and videos, along with the total number of instances of each metadata item in our dataset.

Table 1. Summary Statistics of Data in Farsi Twitter (Sep 2019 - Jan 2021)

Data Item	**Distinct Instances**	**Total Instances**
Tweets	-	538 million
Users	3,346,946	-
Hashtags	1,927,620	178 million
Mentions	1,870,922	536 million
URLs	2,459,512	6 million
Images	13,876,435	83 million
Videos	2,095,574	11 million

As we will discuss in detail later, the use of imagery, videos, and hashtags reveals a vast and multidimensional array of communication, ranging from political issues to sports to international music and television. This variety of content is apparent even from top level glances at the most frequently posted items in each category. For example, the top three mentions are all of Twitter accounts of Korean pop artists (K-Pop), while the most frequently used hashtag (used three times more often) is #اعدام_نکنید (DoNotExecute), protesting political executions by the state. The most shared video is an animation of a coffin draped in an American flag as a counter for causality figures mounts next to Donald Trump's name, and "severe revenge" is sought for Soleimani's assassination by the United States.

These also provide glimpses into the way that Farsi Twitter users network with both each other and the greater world. Of the 1.9 million distinct user accounts mentioned in our dataset, nearly 1.1 million of them (58%) exist in our set of 3.3 million users who have posted in the Farsi language. This balance of in and out group mentions points to a population that is strongly networked, but not cut off from the rest of the world, engaging in a balance of dialogue both internally and externally to Farsi Twitter.

This is further reinforced by an analysis of the URLs being shared on Farsi Twitter. Only 8.6% of the URLs are for domains within the Iranian country-code (.ir), while 9.1% are links to YouTube videos and 4.1% are links to songs on Spotify. There is also evidence of interaction with the extended ecosphere of social media such as Telegram (5.5%), Instagram (4.2%), and to a much lesser degree Facebook (0.6%).

Patterns in links to news organizations are particularly insightful. The top news website to which Farsi Twitter links is mojahedin.org, an opposition news website that represents at 5.7% the third most linked to domain behind YouTube and Telegram. The next most popular news site is the BBC (3.1%). Conspicuously, not a single American sourced news site even cracks 0.1% of link traffic, while Al Jazeera barely registers at 0.02% and Sputnik (an infamous Russian source of disinformation) represents 0.26% of links -- more than the top three American news websites combined.

These features of the data point to a population that is circumventing governmental controls to engage with their own regime and the global community in complex and nuanced ways: both political and nonpolitical, in support of the government in some contexts and forming sites of resistance in others.

3. A Brief Digital Ethnography of Farsi Twitter

With hundreds of millions of tweets posted by millions of users, Farsi Twitter defies summarization into soundbites. And while its authors cannot be considered representative of the population as a whole, the data reflects a wide diversity of political and social thought that cannot be categorized as an echo chamber of any sort. We argue that rather than using this data as a metric for any particular consensus (for instance, pro-regime or pro-opposition), it is best understood as a site of digital contention over the *content* of Iranian national identity.[7]

That contention can be framed in several dimensions: contestation over information (what Iran and Iranians *are*), conflict over policy (what Iran and Iranians *should be*), transnational collaboration (connecting with both the Iranian diaspora and with culturally close groups), and connectivity with global culture and community.

In the next section, we show how these dimensions emerge latently from the data via a set of bottom-up NLP techniques, and then proceed to ethnographically explore each dimension with a combination of qualitative description and descriptive statistics.

3.1 Emergent Patterns

We extracted a day-clustered 1% random sample totaling five million tweets (i.e., a 1% random sample of *each* day to ensure temporal representation) in order to create a representative sample that was small enough to be tractable for available computing resources. We then used three different approaches to tease out the patterns in the data with a minimum of researcher priors or bias.

First, we applied the ParsBERT dual layer supervised topic identifying neural network to our random sample of tweets (Farahani et al, 2020). The first layer is a model pre-trained on the structure of the Farsi language with a 1.3 billion word corpus from 3.9 million articles. The second layer is a model built on the PersianNews labeled set of 16,438 Farsi language news articles from various news websites, that contains eight classifications: economic, international, political, science/technology, cultural/arts, sports, and medical.

[7] See Abdelal R, Herrera YM, Johnston AI, McDermott R. *Measuring Identity: A Guide for Social Scientists*. Cambridge University Press; 2009.

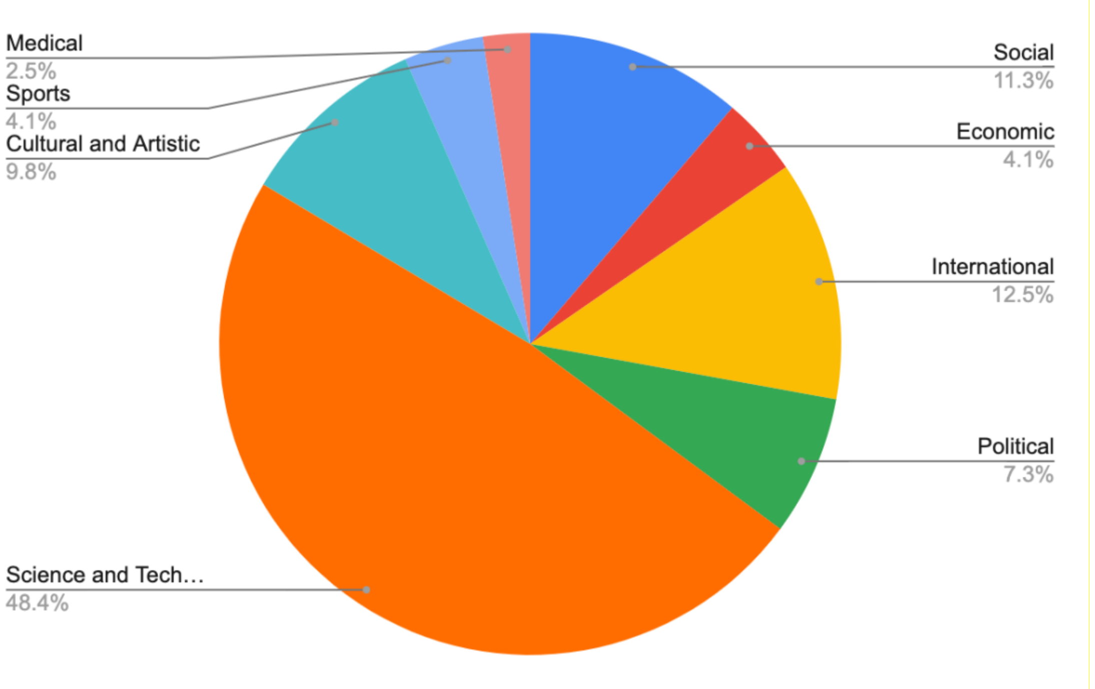

Figure 3. PersianNews Topic Results

We report the proportion of our random sample of tweets fitting into each of the topic classifications in the pie chart in Figure 3. They show a wide spread of discussion across all topics, with the most common topic being science and technology. This reinforces the description of Farsi language Twitter as a site for varied discussion across all spheres of social discussion.

Time series visualization of the proportion of tweets per topic by day is shown in Figure 4. Spikes on certain days point to the face validity of this exercise, with large spikes in cultural topics on major Islamic holidays (green), and spikes in international discussion around major international crises or events (grey).

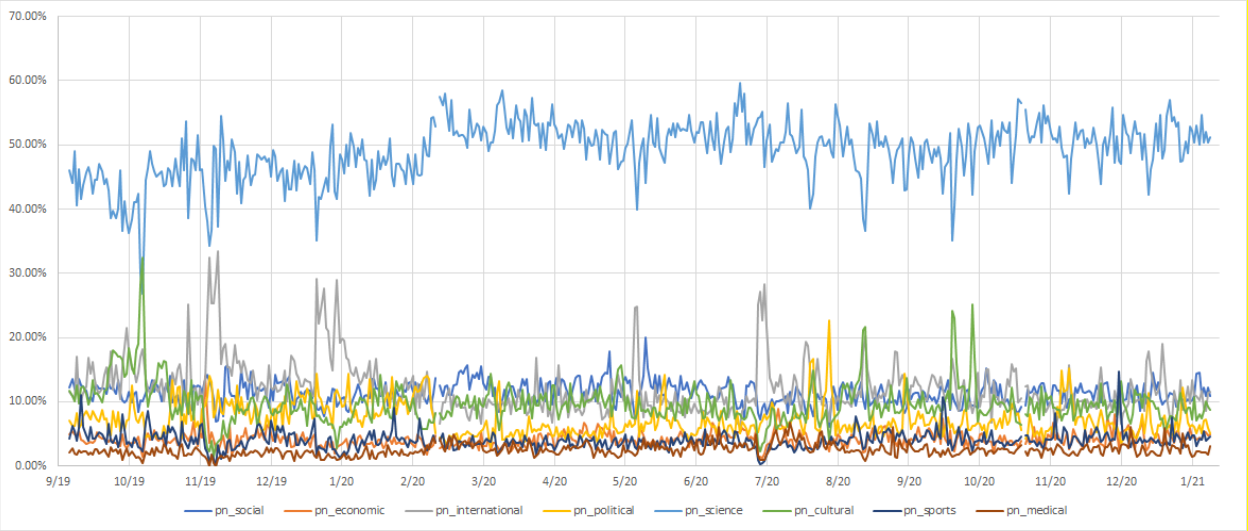

Figure 4: Time Series Trends of Tweets by News Topic

Second, we deployed a neural network trained specifically on named entities from Farsi language media in order to identify specific locations, events, people, and organizations mentioned in our dataset (Poostchi et al., 2016; Poostchi et al., 2018). The top 15 entities identified by the NER classifier in each category are presented in Table 2, translated into English.

The most frequent locations discussed were other countries and Iranian cities with a focus on the surrounding countries of the Mideast, pointing to widespread discussion of the Iranian near abroad. The top events were largely either Islamic religious ceremonies or elections (both domestic and specific foreign elections). Also of note though is that the 15th most referenced event was the national student university exams.

Both domestic and international organizations are well represented among organizations, with references just outside the top fifteen to at least two terrorist organizations (ISIS #17 and Taliban #23). In addition, two soccer teams (Persepolis and Esteghlal) are the most discussed non-governmental organizations.

The most discussed individual person is former President Trump (at nearly double the second place individual), while President Biden comes in at #14, indicating wide discussion of American presidents. The rest of the identified people are mostly religious or political figures within the country (mullah, imam, Rouhani, Khamenei, Qassem Soleimani) but also include Reza Pahlavi, the son of the former monarch living in exile.

Table 2. Named Entity Recognition (NER) Classifier – Top 15 Entities by Category

#	Location	Event	Person	Organization
1	Iran	Aban	Trump	Iranian
2	America	Ashura	Hossein	Iranian
3	Tehran	Arbaeen	Al-Hossein	America
4	Iraq	Eid Bayat	Rouhani	Islamic Republic
5	China	Aban uprising	Ali	Sepah/IRGC
6	Karbala	Arbaeen (40th) of Hussein	Qassem Soleimani	American
7	Europe	week of holy defense	Khamenei	Iranians
8	Syria	night of Yalda	Khamenei	Persepolis
9	Lebanon	Quds day	Khomeni	Esteghlal
10	Afghanistan	world day (day of the world)	Soleimani	Israel
11	Turkey	American elections	Qassem	United Nations
12	Israel	Aban 98 (November 2019)	Reza Pahlavi	Iraqi
13	Saudi Arabia	Presidential election	Navid Afkari	Ukrainian
14	Middle East	Student Entrance exams (konkur)	Biden	Iranian regime
15	Mashad	elections	Mohamed	English

Overall, each category of named entity recognition points to broad patterns of discussion in addition to specifically interesting outliers that we can explore in more detail in the ethnographic sections.

Finally, we tokenized the sample tweets and applied a naive topic modelling approach to the texts (using Latent Dirichlet Allocation, or LDA). We iteratively modeled the tweets on an assumed integer number of independent topics from three to fifty and identified the model with six topics as the one that maximized coherence (0.329) and perplexity (-9.95). In Table 3, we summarize the most statistically salient Farsi words from each of the topics translated into English, along with the percentage of tweets that were classified as that topic.

Table 3. Summary of LDA Topics and Salient Words

#	Representative Salient Words	Summary of Topic	%
1	prison, trial, freedom, peace	Legal & Political	11.1%
2	K-pop, follow, live, young	Entertainment	13.6%
3	imam, God, Islam, guardianship	Religion & Islam	14.3%
4	parliament, republic, Rouhani, home country	Domestic Politics	14.0%
5	this, that, really, very	Misc (other)	27.8%
6	vaccine, execution, exams, weather	Current Events	19.2%

Topic 6, current events is the second most frequently discussed topic with representative salient words such as vaccine, execution, and exams. Note that although it is the top topic in terms of percentage, topic 5 is largely comprised of miscellaneous words with very little semantic meaning such as “this”, “that” and “other”, pointing to this being the residual category into which tweets that don’t match the other topics that have been sorted. Thus, current events (Topic 6) discussed above is the top substantive topic at 19.2%. Topic 2, entertainment, is almost exclusively terms related to K-Pop such as “@weareoneEXO” and “EXO”, referring to the popular K-Pop band EXO. Topics 3 (14.3%) and 4 (14.0%) demonstrate how Iranians use Twitter to discuss religion and domestic politics, using word like “parliament”, “republic”, “Islam” and “guardianship”. These two topics combined with topic 1 on legal and political issues total nearly 40% of all the total summarized topics, showing how Iranians often discuss politics, current events, and domestic issues in their online discourse.

Using social network analysis (SNA), a directed network graph was constructed of the most frequently mentioned 126 notes and 1,268 edges and categorized into eight communities detected based on modularity. As shown in Figure 5, the green cluster represents a conservative faction, mentioning users such as Supreme leader Khamenei's Farsi account and @mb_ghalibaf, conservative speaker of the Iranian parliament and former mayor of Tehran. Expatriate and reformist entities such as Reza Pahlavi (descendant of Iran's former monarch) comprise the blue cluster. Other groups and individuals living outside Iran are also included in the orange cluster on the right, including @pmoIran (People's Mojahedin of Iran) and @maryam_rajavi_p, leader of the People's Mojahedin of Iran (PMOI/MEK). Similar to the LDA analysis above, a distinct and separate community in red represents entertainment and K-pop with mentions such as @exohatch, and @xLittleActive, mostly consisting of fan accounts for Iranian followers of Korean pop band EXO (@weareEXO).

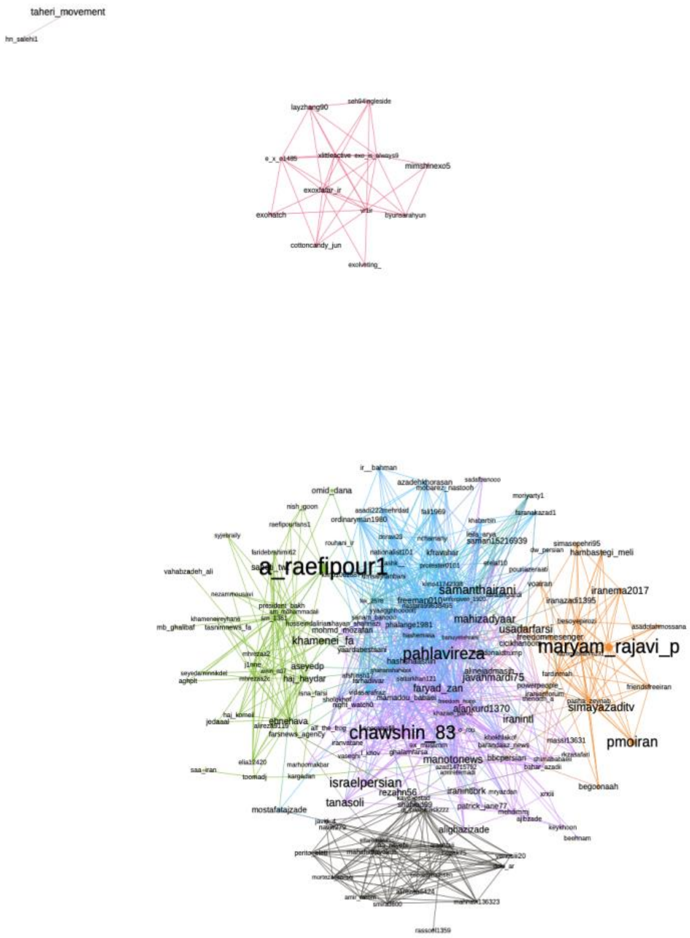

Figure 5: Mentions Network Graph

At face value, these tools show a cacophony of discussion, across the entire gamut of human expression. However, we argue that the identity contestation framework is a fruitful way to approach this data through an ethnographic lens, because it assumes that the *noise* is in fact the latent process of contestation over the content of Iranian national identity. What we can pull out of the above data are patterns of contestation: Iranians discussing what they are, what they should be, how they relate to surrounding nations, and how their identity integrates with the global community.

3.2 Informational Contestation: What Iran and Iranians Are

Iranians use Twitter to discuss national identity and share information regarding domestic and international politics. In some cases this is complementary to the regime's official line. For example, after the US assassination of Qassem Soleimani in January 2020, many Iranians expressed outrage at this breach of international norms and praised/mourned the IRGC-Quds force commander as a martyr (Hashemi & Wilson, 2020).

However, social media's role in the expression of national identity can be a double-edged sword. Just a few days later, this discourse quickly flipped to critiquing the government explanation of the airliner that was shot down, demonstrating informational correction by the population asserting a counter narrative. "Ukrainian airliner/aircraft" (هواپیمای_اوکراینی) was the sixth most popular hashtag used in January 2020. Table 4 lists the top 20 hashtags on Farsi Twitter in January 2020, clearly demonstrating how in the wake of the national crisis, social media is being used to express a range of national identity.

Table 4. Hashtags Used During January 2020

Rank #	Hashtag	English	Count
1	انتقام_سخت	hard revenge	394264
2	قاسم_سلیمانی	qassem soleimani	208144
3	قاسم_سليماني	qassem soleimani	189503
4	فوری	urgent	158291
5	ایران	Iran	153662
6	هواپیمای_اوکراینی	Ukrainian airliner/aircraft	115548
7	EXO	EXO	102287
8	IraniansDetestSoleimani	IraniansDetestSoleimani	92197
9	PahlaviRepresentsIranians	PahlaviRepresentsIranians	79268
10	HardRevenge	HardRevenge	68654
11	ترامپ	Trump	68149
12	سردار_سلیمانی	General Soleimani	67542
13	آمریکا	America	51547
14	سیستان_و_بلوچستان	Sistan and Baluchistan	50680
15	IranProtests	IranProtests	47886
16	براندازم	I will overthrow (the regime)	45009
17	FreeDetainedIranianProtesters	FreeDetainedIranianProtesters	44640
18	کاسبان_خون	businessman's blood?	44178
19	حاج_قاسم_سلیمانی	Haj Qasem Soleimani	41857
20	iHeartAwards	iHeartAwards	41817

In other instances, the alternative channels provided by social media platforms and private messaging applications are used to actively spread information being suppressed officially by the government. For example, Iranians also used Twitter to counter government disinformation during the COVID-19 pandemic. The government strategically delayed acknowledging or releasing public announcements of the outbreak in order to prevent a decrease in voter turnout for the February 21, 2021 national elections. However, as shown in Figure 5, a time series plot of the frequency of keyword usage, Farsi Twitter saw discussion of coronavirus immediately after the first cases in Iran, overwhelming discussion of the election even on the day of the election.

While Iranians used Twitter to fill the gap left by the government's lack of public service announcements following the initial outbreak in Iran's religious city of Qom, there were also several user-generated disinformation campaigns that spread dangerous falsehoods about cures and remedies for the virus. In March 2020, 44 Iranians died after drinking toxic alcohol or methanol, following a rumor that doing so would cure the virus (Bote, 2020). Alcohol is illegal in the Islamic Republic, so consumption of this bootlegged toxic alcohol led to what is now believed to be over 200 deaths. Other rumored alternative cures and preventative measures perpetuated online were lemon in hot water, ginger and turmeric and other spices as well as gargling with vinegar.

Examination of Persian language tweets from the week of the outbreak and election presented in Table 5 reveal that 15% of the top hashtags used were related to coronavirus while over 50% mention elections or voting. This demonstrates that the government's disinformation campaign during the rigged election was countered by the population on Twitter, a platform that is technically banned in the country.

Coronavirus comprises three of the top 20 (~ 15%) hashtags used in Persian Twitter during the week of the elections. كرونا (Corona) was the top hashtag, used over 18,000 times. Qom, where the virus was first reported, is the 17th most used hashtag. Hashtags related to elections or parliament comprise over 50% of the top hashtags used, including phrases like تحريم_انتخابات (I vote because) (#6) and تحريم_انتخابات (boycott elections) (#12).

Table 5. Top 20 Hashtags Used During February 2020 National Elections

#	Hashtag	Count	English
1	کرونا	18072	corona
2	مجلس_قوی	18016	strong parliament
3	انتخابات	16907	elections
4	می_آیم_چون	15693	We/I come because
5	NoVote4Terrorists	13576	NoVote4Terrorists
6	رای_میدهم_چون	5727	I vote because
7	ایران	5273	Iran
8	فوری	3922	instantaneous/urgent
9	شجریان	3681	Sharjarian
10	FATF	3581	FATF
11	ایران_قوی	3565	strong Iran
12	تحریم_انتخابات	3534	boycott elections
13	رای_بی_رای	3317	vote without vote/vote no vote
14	ویروس_کرونا	2758	coronavirus
15	انتخابات_مجلس	2744	parliamentary/Majlis elections
16	تهران	2575	Tehran
17	قم	2566	Qom
18	رای_دادم	2491	I voted
19	کروناویروس	2104	coronavirus
20	رای_من_سرنگونی	2048	my vote (was) overthrown/destroyed

As shown in Figure 5, Farsi tweets began mentioning corona at the end of January 2020 with a sharp rise in the use of the term during the first week of February. Yet, the first three weeks of February, discussions on Twitter revolved mainly around the election. On February 18th, the first case of COVID-19 was reported in the country.[8] We observe a clear spike in mentions of corona/COVID-19 in the data (on February 19 and again on February 26-27). From that moment on, corona dominated the discussions online. The only exception is the day of the election (February 21) when mentions of coronavirus were briefly (and barely) overtaken by the discussion of elections.

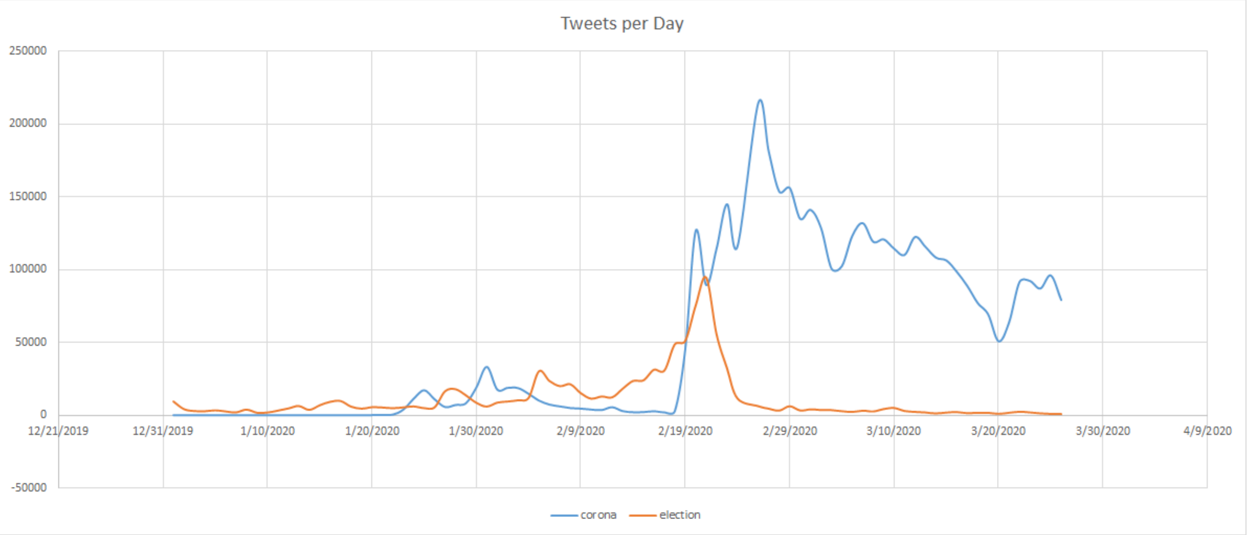

Figure 5. Number of Tweets Over Time: Election vs. Coronavirus

Iranian discussion about COVID-19 on Twitter before and during the national election demonstrates how the government's disinformation attempt to push forward with the rigged elections was countered by the population's detailed accounts of the outbreak in the country. In their tweets, Iranians shared information about the number of cases, precautionary health measures and documented (health care facilities and other) responses to the outbreak. Furthermore, these conversations took place on Twitter which is officially banned by the government, showing both the high level of Iranian circumvention skills as well as the power of social media platforms to challenge state narratives and misinformation campaigns. Social media is important because it breaks the dictator's monopoly on information.

[8] Source: https://www.worldometers.info/coronavirus/

3.3 Policy Conflict: What Iran and Iranians Should Be

Twitter serves as a site to directly and indirectly discuss what Iran's policies should be. Topics on Farsi Twitter include both explicit and more subtle political discussion on issues ranging from the death penalty to the use of art and comedy to critique the country's politics and politicians. For example, Iranians used Twitter to raise awareness of and discuss their discontents regarding the widespread corruption and electoral mismanagement surrounding the 2020 Majlis (Parliamentary) national elections. Using hashtags such as "No Vote" or "Boycott Elections" many Iranian users pointed to the extensive electoral fraud and corruption. In a historically unprecedented vetting, the Guardian Council disqualified more than one third of potential legislative candidates, including 90 incumbents (Bizaer and Rasheed, 2020).

Iranians are known for their use of satire and art for political commentary across different forms of communication (Rahimi 2015). Still today, Iranians use platforms like Twitter, Telegram, WhatsApp and SMS to share jokes that poke fun at political figures and criticize government policies, expressing their political opinions while avoiding detection by authorities. For example, the program Parazit, often referred to as 'The Daily Show of Iran' first aired prior to the disputed 2009 presidential elections. Broadcast from outside of the country through Voice of America, Parazit was known for its subversive comedy and the use of satire to critique the country's politics and politicians. Iranians even use satire and jokes to discuss serious topics such as the COVID-19 outbreak. A study analyzing 45,000 Farsi tweets related to coronavirus found that the top category of content was satire (28%), followed by news (24%) and opinion (18%) (Hosseini et al., 2020).

Anti-Execution Movement

The trending hashtag #DoNotExecute was used by the anti-execution movement to protest the country's use of the death penalty, especially executions of political prisoners or those involved in anti-government protests. While the use of the death penalty has recently declined in the country, Iran still has one of the highest numbers of executions, second only to China. In 2019, Iran had over 250 confirmed executions (Amnesty International, 2020).

In July 2020, Iran's supreme court upheld the death sentences of three young men (Saeed Tamjidi, Mohammad Rajabi, and Amir Hossein Moradi) accused of involvement in the November 2019 protests in response to a sharp rise in gas prices and led to a nearly total internet shutdown. Later that summer, 27-year-old wrestler Navid Afkari was sentenced to death for killing a security officer during protests in 2018 (that were also in response to rising petrol prices). Hashtags calling on authorities not to execute quickly began trending on Iranian Twitter in both English and Farsi language (ex. #SaveNavidAfkari and #Don’tKillOurNavid (#نویدمان_را_اعدام_نکنید) (Stryer 2020). Despite these anti-execution hashtag campaigns gaining significant attention from the international community and human rights groups, on September 12, 2020 27 year old Iranian wrestler and athlete activist Navid Afkari was executed on charges of killing an IRGC security guard during the November 2018 anti-government protests.

National Entrance Exams

The Iranian University Entrance Exam (Konkour) is a notoriously difficult national entrance exam that has created a higher education crisis in the country. From July through August 2020, one of the top trending hashtags called for the government to postpone or cancel national entrance exams (تعویق_تمام_کنکورهای_سراسری - postpone all national entrance exams; لغو_امتحانات_نهایی_حضوری - cancel the exams; سلامت_دانشجو - student health). While there was already sharp criticism of the exams in the country (Kamyab 2015), these discontents were compounded by the COVID-19 pandemic and the frustrations felt by Iranian youth due to high unemployment and cultural restrictions. Despite these public health concerns, the 2020 Konkour exams continued as planned in August 2020 with nearly 1.4 million candidates, an over 270,000 increase from the previous year (Tehran Times, 2020).

While the two examples above represent the use of Twitter to discuss socio-political issues, this dimension should not only be understood as active political contestation and organization, but also in terms of expression of other more mundane policy conflict and public opinion.

3.4 Transnational Collaboration

Twitter is also used to connect with “near” international elements such as the diaspora and culturally and linguistically similar groups that are able to then collaborate because of these commonalities. Farsi Twitter allows Iranians to form transnational connections and collaborate with external allies in the region. The linguistic similarities and overlap between Farsi and other languages in the region also facilitates cross-national collaboration in different areas and information sharing on topics such as political violence and human rights abuses and violations.

A particularly prominent example of this cross-national connection is the transnational feminist communications and collaboration between Iran and Afghanistan as well as other countries calling for gender equality in the region. Shortly after the 1979 Islamic Revolution, female spectators were banned from football and other stadiums, though many women continued to attend matches despite this restriction by disguising themselves as men. When 29-year-old female soccer fan Sahar Khodayari died from setting herself on fire after being arrested on charges of trying to enter a soccer stadium in September 2019, Afghans showed their support for the right of Iranian women to enter soccer stadiums on Twitter using the hashtag Blue Girl, a nickname given to Khodayari based on her favorite soccer team. The hashtag was used by both Iranian and Afghani users to call for equal access to soccer stadiums. Aghan users expressed that they felt sympathy with Iranian fans as Afghani women were also banned from stadiums under Tabliban rule (Noori 2019).

In another case of transnational solidarity, Iranians used Twitter to express support and sympathy for Afghans after the November 2020 terrorist attacks on Kabul University by using the popular hashtag Jane Pedar Kojasti, a phrase that roughly translates to “Where Are You My Dear Daughter”, quoting a father who was frantically trying to contact his daughter who was killed in the attacks. On November 6, 2020 Iran lit Azadi Tower in the Pakistani flag and projected images and the hashtag #jane-Pedar-kojasti in solidarity with the Afghan people and the families of victims (TehranPicture, 2020; Honaronline, 2020). On Twitter and other social media platforms, users created commemorative content such as videos and songs while using the phrase Jane Pedar Kojasti (Taranehaye Mandegar, 2020).

While many users might self-censor their online content or refrain from expressing their opinions for fear of punishment or retaliation by the state, transnational communications made possible by Twitter, Instagram and other social media platforms allows for transnational connection between Iranians living in the country and those living outside of Iran. Internet communication technologies have provided Iranians with cost effective communication methods. Previously, colleagues and family members would have to spend significant time, money and resources to communicate across borders, using phone cards and paying high international calling rates. Social media drops the financial and danger threshold for keeping the population connected to both its global diaspora, and the larger international community.

Even if these transnational campaigns might not bring about concrete change in policy ‘on the ground’, digital communication technologies create opportunities for movements to disseminate information widely and rapidly, connect with existing members and recruit new supporters and collect monetary and other resources required to promote the groups’ cause and achieve its aims. Twitter permits users to connect with Iranians within the country and in the diaspora, as city name hashtags are often used to provide nearly real time updates on events on the ground such as national crisis or natural disaster. For example, immediately after the initial outbreak of COVID-19 in the country in February 2019, Iranian users disseminated information on the virus on Twitter and other social media platforms, using city names as hashtags to provide local updates.

3.5 Connecting with Global Culture

As seen through the trending topics such as education, capital punishment and public health above, much of Iranian discourse on Twitter is political in nature. However, Iranians also use the social media platform for entertainment purposes and to discuss non-political topics (especially Korean pop music (K-pop) and Korean television dramas) and to connect with foreign global culture on artistic and humanistic dimensions. Due to Iran’s lax regulations on intellectual property, pirated media such as music, films, and television series are widespread. Cultural products are readily exchanged through personal networks and black markets with impunity, allowing Iranians access to content banned by the Islamic government.

From the Revolution of 1979 to the turn of the millennium, the most coveted content was from western media outlets, but beginning in the early 2000s, Iranian state-sponsored television began airing Korean TV dramas, leading to a spread of the already popular Korean Wave into Iran. In our data, during months when major political events are not dominating social media in Iran, the top hashtags and mentions on Farsi Twitter are dominated by K-Pop, many of which are in English or Korean, despite the tweet text itself being in Farsi. Some attribute the popularity of Korean television and films in Iran to the similarities in the cultures of the two countries. Both countries place high value on tradition, modesty, and family (Koo 2020).

Iranian engagement with foreign media is not limited to Korean culture. In 2014, several young Iranians were arrested and charged for indecent behavior after posting a video of themselves singing and dancing to the hit song “Happy” by Pharrel Williams. Within six months the video had been viewed by over one million Iranian viewers on YouTube. The incident led to a campaign using the hashtag #freehappyIranians to protest the arrests (BBC, 2020).

Twitter allows users to connect globally despite severe media censorship and restrictions as seen by the Korean Wave in the IRI. This is especially reflected in our URL analysis, in which YouTube and Spotify links are among the top shared domains. Through the sharing of user-generated content, Iranians (women in particular) actively lead the fandom culture through their daily activities, creating a sense of solidarity with both domestic and global users (Koo, 2020).

4. Why it Matters: Broader Implications of Social Media as Data

With any collection of observational data of private citizens, it is important to consider the ethical dimensions of data collection and usage. The Belmont Report (1976) created a framework for ethics in human research that emphasized informed consent by research subjects. However, social media research is not performed at the human scale at which the Belmont principles were formulated and does not scale to research on millions of individuals in aggregate (Bailey, Dittrich, and Kenneally, 2013).

In addition, the uniquely public nature of social media is important. While someone who posts a tweet has not signed a consent form authorizing the use of that tweet for a specific research project, there is an implicit notion that its public posting is an offer so that others might read it, interpret it, or indeed, research it. The Menlo Report (2012) and subsequent clarifications to the IRB Common Rule (HCIRB/BRP/DCCPS, 2019) emphasized clearly that observational social media data does not meet the definition of human subjects research unless the information is obtained through intervention or includes information that is both identifiable and private. This dataset is explicitly observational and includes unequivocally only publicly available posts.

Researchers should remain aware of the risks created by investigating social media posts originating from within autocracies in which citizens are persecuted and have limited access to justice. However, an often-overlooked core Belmont principle is the right of individuals to *participate* in research. That is, a core ethical mandate of scientists is to include others as broadly as possible in our learning about the world. While we should not be cavalier about the safety of those living in autocracies, neither should we exclude their voices from our data collection, lest we do the work of dictators for them. Collecting, reading, and understanding the social media posts of the oppressed is a force for inclusion and democracy (Wilson, 2022).

Analysis of Farsi Twitter also represents an exciting new frontier in computational social science by presenting the opportunity to discuss the possibilities and limitations of using social media as data. With increased travel restrictions surrounding the COVID-19 pandemic, this approach will become increasingly valuable as scholars experience obstacles to conducting field research in person. This will be particularly true of scholars studying the developing world as the global South is disproportionately impacted by the pandemic and will likely experience far longer recovery trajectories as a result.

While there are limitations, analyzing the Farsi Twittersphere is still important because it allows for analysis under circumstances that prevent scholars from traveling to a country to conduct field work and research in a safe or ethical manner. We believe that in the new normal that will follow the COVID-19 pandemic, digital ethnography will increase in popularity, signaling an exciting shift where more researchers are using social media as data than ever before.

More broadly, this project demonstrates two ways in which social media data is a “weapon of the weak” by enabling access to large scale data for digital ethnography. First, it enables individuals with constrained resources who are otherwise blocked from field work, whether abroad or in their home country or region. This lowers the threshold for original data collection, putting it within reach of students and scholars with minimal budgets.

Secondly, it is a data source that explicitly gives access to the thoughts and culture of non-elites on a grand scale. While the digital divide is a very real issue - those with the resources and technical know-how to circumvent government blocks are certainly not representative of the population - this is still an important step in expanding the choir of voices from whom we can hear in our research.

Social media data democratizes research by expanding the pools of who can listen and who can be heard. That is particularly true with this research project, which we argue should be a boon for researchers of Iranian culture, politics, and society both within Iran and around the world.

References

Abdelal, R. (Ed.). (2009). *Measuring identity: A guide for social scientists*. Cambridge University Press.

Amnesty International. (2019, December 16). *Iran's authorities carrying out vicious post-protest crackdown*. Amnesty International. https://www.amnesty.org/en/latest/news/2019/12/iran-thousands-arbitrarily-detained-and-at-risk-of-torture-in-chilling-post-protest-crackdown/

Amnesty International. (2020, April 21). *Death penalty in 2019: Facts and figures*. Amnesty International. https://www.amnesty.org/en/latest/news/2020/04/death-penalty-in-2019-facts-and-figures/

Bailey, M., Dittrich, D., Kenneally, E., & Maughan, D. (2012). The Menlo Report. *IEEE Security & Privacy Magazine*, *10*(2), 71–75. https://doi.org/10.1109/MSP.2012.52

BBC News. (2014, September 19). Iran: Happy video dancers sentenced to 91 lashes and jail. *BBC News*. https://www.bbc.com/news/world-middle-east-29272732

Bizaer, M., & Rasheed, Z. (2020, January 27). *Mass disqualification of candidates add to discontent in Iran*. https://www.aljazeera.com/news/2020/1/27/mass-disqualification-of-candidates-add-to-discontent-in-iran

Bote, J. (2020, March 10). *44 dead in Iran from drinking toxic alcohol fake coronavirus cure*. USA Today. https://www.usatoday.com/story/news/world/2020/03/10/44-dead-iran-drinking-toxic-alcohol-fake-coronavirus-cure/5009761002/

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*, *53*(6), 3831–3847. https://doi.org/10.1007/s11063-021-10528-4

Hashemi, L. M., & Wilson, S. L. (2020). *Analysis | If any Iranians supported Soleimani's killing, it would've been dissidents on Twitter. The opposite happened.* Washington Post. Retrieved January 27, 2020, from https://www.washingtonpost.com/politics/2020/01/08/twitter-is-where-iranian-dissidents-might-support-soleimanis-killing-opposite-happened/

HCIRB/BRP/DCCPS (2019). Human Subjects Considerations for Social Media Research.

Honaronline. (2020, November 7). *Iran sympathized with victims of Kabul terrorist attack*. Honaronline. http://www.honaronline.ir/Section-news-2/154856-iran-sympathize-with-victims-of-kabul-terrorist-attack

Hosseini, P., Hosseini, P., & Broniatowski, D. (2020). Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Proceedings of the 1st Workshop on NLP for

COVID-19 (Part 2) at EMNLP 2020, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.26

Kamyab, S. (2015). The University Entrance Exam Crisis in Iran. *International Higher Education*, *51*. https://doi.org/10.6017/ihe.2008.51.8010

Koo, G. Y. (2020). Riding the Korean Wave in Iran. *Journal of Middle East Women's Studies*, *16*(2), 144–164. https://doi.org/10.1215/15525864-8238160

Marchant, J., Sabeti, A., Bowen, K., Kelly, J., & Heacock Jones, R. (2016). #Iranvotes: Political Discourse on Iranian Twitter During the 2016 Parliamentary Elections. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2799271

Mechkova, Valeriya, Daniel Pemstein, Brigitte Seim, Steven Wilson. 2020. *Digital Society Project Dataset v2.*

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, Behavior, and Social Networking*, *16*(9), 708–713. https://doi.org/10.1089/cyber.2012.0334

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. [Bethesda, Md.]: The Commission.

Noori, H. (2019, September 22). *Afghan women stand in solidarity with Iranians after "Blue Girl" Sahar Khodayari's death*. The National. https://www.thenationalnews.com/world/asia/afghan-women-stand-in-solidarity-with-iranians-after-blue-girl-sahar-khodayari-s-death-1.913627

Official Persian Twitter. (2021). توییتر فارسی. Telegram. /s/OfficialPersianTwitter?before=200570

Poostchi, H., Zare Borzeshi, E., Abdous, M., & Piccardi, M. (2016). PersoNER: Persian Named-Entity Recognition. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3381–3389. https://aclanthology.org/C16-1319

Poostchi, H., Zare Borzeshi, E., & Piccardi, M. (2018, May). BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: The First Entity-Annotated Persian Dataset. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1701

Rahimi, B. (2015). Satirical cultures of media publics in Iran. *International Communication Gazette*, *77*(3), 267–281. https://doi.org/10.1177/1748048514568761

Sanchez, R. (2019, December 13). Bullets and blackout: Inside four days of killing in Iran. *The Telegraph*. https://www.telegraph.co.uk/news/2019/12/13/bullets-blackout-inside-four-days-killing-iran/

Sapra, B. (2021, February 22). *A visualization of Iran's internet shutdown*. WIRED Middle East. http://wired.me/technology/iran-2019-internet-shutdown-map-opte-project/

Sreberny, A., & Khiabany, G. (2010). *Blogistan: The internet and politics in Iran*. I. B. Tauris, Distributed in the United States and Canada exclusively by Palgrave Macmillan.

Stryer, R. (2020, September 16). #Don'tExecute: A semi-successful campaign against capital punishment in Iran. *Atlantic Council*. https://www.atlanticcouncil.org/blogs/iransource/dontexecute-a-semi-successful-campaign-against-capital-punishment-in-iran/

Taranehaye Mandegar .ترانه های ماندگار (2020) *Majid Kharatha—Jane pedar kojasti—* جان پدر کجاستی مجید خراطها *(OfficialAudio)*. https://www.youtube.com/watch?v=l3z7ZRgSUlg

TehranPicture. (2020, November 6). *Video Mapping Sympathy with the people of Afghanistan*. TehranPicture; TehranPicture. %2f%2fwww.tehranpicture.ir%2fen%2falbum%2f6712%2fVideo-Mapping-Sympathy-with-the-people-of-Afghanistan

Wilson, Steven L. (Forthcoming 2022). *Social Media as Social Science Data*. Cambridge University Press

Yang, K., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115