#COVID-19 ON TWITTER: BOTS, CONSPIRACIES, AND SOCIAL MEDIA ACTIVISM

EMILIO FERRARA, UNIVERSITY OF SOUTHERN CALIFORNIA
EMAIL: EMILIOFE@USC.EDU

ABSTRACT

With people moving out of physical public spaces due to containment measures to tackle the novel coronavirus (COVID-19) pandemic, online platforms become even more prominent tools to understand social discussion. Studying social media can be informative to assess how we are collectively coping with this unprecedented global crisis. However, social media platforms are also populated by bots, automated accounts that can amplify certain topics of discussion at the expense of others. In this paper, we study 43.3M English tweets about COVID-19 and provide early evidence of the use of bots to promote political conspiracies in the US, but also as a tool to enable participatory activism to surface information in the English-speaking Twitter that could otherwise be censored in China.

Keywords: social media, bots, coronavirus, COVID-19, conspiracies, participatory activism

INTRODUCTION

At the time of this writing (mid-April 2020) the novel coronavirus (COVID-19) pandemic outbreak has already put tremendous strain on many countries' citizens, resources and economies around the world. Social distancing measures, travel bans, self-quarantines, and business closures are changing the very fabric of societies worldwide. With people forced out of the safety and comfort of their life routines, social media take centerstage, more than ever, as a mirror to global social discussions. Therefore, it is of paramount importance to determine whether online chatter reflects genuine people’s conversations or otherwise may be distorted by the activity of automated accounts, often referred to as bots (a.k.a., social bots, sybil accounts, etc.). The presence of bots has been documented in the context of online political discussion (A. Bessi & Ferrara, 2016; E. Ferrara, 2017; Luceri et al., 2019), public health (Allem et al., 2017; Broniatowski et al., 2018; Hwang et al., 2012; Subrahmanian et al., 2016; Sutton, 2018), civil unrest (Stella et al., 2018), stock market manipulation (Emilio Ferrara, 2015), the spread of false news (Grinberg et al., 2019; Shao et al., 2018; Vosoughi et al., 2018), alongside with other tools such as troll accounts (Badawy et al., 2018, 2019; Bail et al., 2019; Broniatowski et al., 2018; Luceri et al., 2020; Sutton, 2018).

In this paper, we chart the landscape of Twitter chatter within the context of COVID-19 related conversation, in particular to characterize the presence and activity of bots. We leverage a large Twitter dataset that our group has been continuously collecting since January 21, 2020, when the first COVID-19 case was announced in the United States (Holshue et al., 2020). We use combinations of machine learning and manual validation to identify bots, and then use computational tools and statistical analysis to describe their behavior, in contrast with human activity, and their focal topics of discussion.

RELATED WORK

There has already been a wealth of studies that looked into social media dynamics in the context of COVID-19. As of the time of this writing (mid-April 2020), the vast majority of these studies are pre-print papers that provide a timely, yet partial, characterization of online discussion and issues revolving around COVID-19 (Alshaabi et al., 2020; Chen et al., 2020; Cinelli et al., 2020; Gallotti et al., 2020; Gao et al., 2020; Kleinberg et al., 2020; Li et al., 2020; Pennycook et al., 2020; Schild et al., 2020; Singh et al., 2020).

Various studies presented the concept of *social media infodemic,* i.e., the spread of questionable content and sources of information regarding the COVID-19 pandemic, as postulated by (Cinelli et al., 2020). This research illustrates the problem of containing the spread of unverified information about COVID-19, showing that questionable and reliable information spreads according to similar diffusion patterns. Along this line, (Gallotti et al., 2020) suggests that low-quality information anticipates epidemic diffusion in various countries, with the peril of exposing those countries' population to irrational social behaviors and public health risks. Both studies account for large-scale data collection from online platforms like Facebook and Twitter but do not emphasize the importance of information manipulation on such social media. The work by Singh and colleagues also looked at the spatio-temporal dynamics of misinformation spread on Twitter, drawing a picture with similar implications as the two studies above (Singh et al., 2020).

More research is needed, as the information landscape evolves, and more scientific insights are unveiled on the clinical and medical implication of this disease, to understand what qualifies for rumors, misinformation, or disinformation campaigns. For example, information about the possible effectiveness of some treatments could be considered as rumors at a given point in time, in absence of definitive scientific consensus; yet, as more clinical evidence emerges, these may become false claims, hence classified as misinformation: one such example is Hydroxychloroquine, a known anti-malaria drug whose effectiveness in treating SARS-CoV-2 (the coronavirus causing the COVID-19 disease) remains debated at this point in time, and whose potentially lethal side-effects limit large-scale testing.

The work by Pennycook and colleagues epitomizes the seriousness of this problem by showing, with a social experiment including 1,600 participants, that subjects tend to share misinformation and false claims about COVID-19 predominantly because they are unable to determine whether the content is scientifically sound and accurate or not (Pennycook et al., 2020).

Other studies investigated collective attention and engagement dynamics concerning COVID-19 on Twitter. For example, (Alshaabi et al., 2020) analyzed 1,000 unigrams (1-grams) posted on Twitter in 24 languages during early 2020, and compared them with the year prior. The authors emphasize how the global shift in attention to the COVID-19 pandemic is concentrated around January 2020, after the first wave of infections in China started to phase off, and peaked again in early March, when the United States and several other western countries started to get more heavily hit by the pandemic. Their work suggests that social media mirror offline attention dynamics, and hints at the potential implications that diminished collective attention can have on the perception of gravity of this pandemic (or lack thereof).

Various studies investigated emotional and sentiment dynamics on social media conversation pertaining COVID-19 (Gao et al., 2020; Kleinberg et al., 2020; Schild et al., 2020). For example, (Kleinberg et al., 2020) annotated a corpus (N=2,500 tweets and N=2,500 longer texts) producing a ground truth dataset for emotional responses to COVID-19 content. The analysis of this UK-centric corpus suggests that issues pertaining family safety and economic stability are more systematically associated with emotional responses in longer texts, whereas tweets more commonly exhibit positive calls for solidarity.

Contrary to that, recent work based on large-scale multiplatform data collections encompassing Twitter and 4chan illustrates the endemic prevalence of hate speech, especially sinophobia, in both platforms (Schild et al., 2020). Furthermore, the cross-platform diffusion of racial slur and targeted attacks shows how fringe platforms like 4chan are incubators of new hate narratives aimed, in the case of COVID-19, against Asian people; in mainstream social media platforms like Twitter, however, the focus is on putting blame on China for the alleged responsibility in originating the virus and inability to contain it.

On a different note, (Gao et al., 2020) looked at social media sentiment as expression of potential mental health issues associated with social isolation and other side-effects of containment measures enacted to limit the spread of COVID-19 in China. By means of online surveys to complement observational data collected from populare Chinese social media platforms, the authors suggest that social media exposure to outbreak-related content was correlated with increased odds of reporting issues associated with mental health, including depression and anxiety, across different demographics in their population.

Background on COVID-19

The first cases of a novel coronavirus disease (officially named COVID-19 by the World Health Organization on February 11, 2020) were reported in Wuhan, China in late December 2019; the first fatalities were reported in early 2020. The first case in the United States was announced on January 21, 2020 (Holshue et al., 2020): our Twitter data collection aligns with that date (Chen et al., 2020).

The fast-rising infection rates and death toll led the Chinese government to quarantine the city of Wuhan on January 23, 2020.[1] During this period, other countries began reporting their first confirmed cases of the disease, and on January 30, 2020 the World Health Organization (WHO) announced a Public Health Emergency of International Concern. With virtually every country on Earth reporting cases of the disease, and infections rapidly escalating in some regions of the world, including the U.S., Europe and the middle East, WHO has subsequently upgraded COVID-19 to a pandemic.[2] On March 13, 2020 the United States government announced the state of national emergency. Our data collection's end aligns with that date. As of the time of this writing (mid-April 2020), COVID-19 has been reported in every country worldwide, leaving governments all over the globe scrambling for ways to contain the disease and lessen its adverse consequences to their people's health and economy. Infections exceed two million. Fatalities are well over a hundred thousand. There is still no scientific consensus on the effectiveness of any particular treatment; vaccines are not expected to be available to large swaths of the population for at least a year. COVID-19 has been among the trending topics of discussion on Twitter uninterruptedly since early 2020.

[1] https://www.nytimes.com/article/coronavirus-timeline.html
[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

BACKGROUND ON BOTS

What is a bot. A bot (short for robot, a.k.a., social bot, social media bot, or sybil account) is a social media account controlled, predominantly or completely, by a piece of software (a more or less sophisticated artificial intelligence), in contrast with accounts controlled by human users (Emilio Ferrara et al., 2016).

How to create a bot. Early social media bots, in the late 2000s, were created to tackle simple tasks, such as automatically retweeting content posted by a set of sources or finding and posting news from the Web. Today, the capabilities of bots have significantly improved: bots rely on the fast-paced advancements of Artificial Intelligence, especially in the area of natural language generation, and use pre-trained multilingual models like OpenAI's GPT-2 (Radford et al., 2019) to generate human-like content. This framework allows the creation of bots that generate genuine-looking short texts on platforms like Twitter, making it harder to distinguish between human and automated accounts (Alarifi et al., 2016).

The barriers to bot creation and deployment, as well as the required resources to create large bot networks, have also significantly decreased: for example, it is now possible to rely upon bot-as-a-service (BaaS), to create and distribute large-scale bot networks using pre-existing capabilities provided by companies like *ChatBots.io*, and run them in cloud infrastructures like *Amazon Web Services* or *Heroku*, to make their detection more challenging (Emilio Ferrara, 2019).

Open Source Twitter bots. A recent survey discusses readily-available Twitter bot-making tools (Daniel & Millimaggi, 2020): the authors provide an extensive overview of open-source GitHub repositories and describe how prevalent different automation capabilities, such as tweeting or resharing, are across these tools. According to (Daniel & Millimaggi, 2020), whose survey focused exclusively on repositories for Twitter bots developed in Python, there are hundreds of such open-source tools readily available for deployment. The authors studied 60 such bot-making tools and enumerated the most common capabilities. Typical automated features of such bots include:

(1) searching users, trends, and keywords;
(2) following users, trends, and keywords;
(3) liking content, based on users, trends, and keywords;
(4) tweeting and mentioning users and keywords, based on AI-generated content, fixed-templated content, or cloned-content from other users;
(5) retweeting users and trending content, and mass tweeting;
(6) talking to (replying) other users, based on AI-generated content, templated content, or cloned-content from other users; finally,
(7) pausing activity to mimic human circadian cycles and bursty behaviors, as well as to satisfy API constraints, and to avoid suspension.

According to (Daniel & Millimaggi, 2020), these features can enable bots to carry out various forms of abuse including: denigrate and smear, deceive and make false allegations, spread misinformation and spam, and finally clone users and mimic human interests.

We refer the interested readers to the excellent survey by (Daniel & Millimaggi, 2020) for further details.

How to detect bots. Researchers in cyber-security have highlighted first some potential challenges associated with the detection of bots (Abokhodair et al., 2015; Boshmaf et al., 2013, 2011; Freitas et al., 2015; Thomas et al., 2011). Historically, however, bot detection techniques have been pioneered by groups at Indiana University, University of Southern California, and University of Maryland, in the context of a program sponsored by DARPA (the U.S. *Defense Advanced Research Projects Agency*) aimed at detecting bots used for anti-science misinformation (Subrahmanian et al., 2016).

More recently, large bot networks (botnets) have been discovered on Twitter by various research groups (Abokhodair et al., 2015; Boshmaf et al., 2011; Echeverria & Zhou, 2017; Thomas et al., 2011).

The literature on bot detection has become very extensive (Cresci et al., 2019; Emilio Ferrara, 2018; Emilio Ferrara et al., 2016; Stieglitz et al., 2017; K. Yang et al., 2019).

In (Emilio Ferrara et al., 2016), we proposed a simple taxonomy to divide bot detection approaches into three classes: (1) systems based on social network information; (2) systems based on crowd-sourcing and the leveraging of human intelligence; (3) machine learning methods based on the identification of highly-predictive features that discriminate between bots and humans. Other recent surveys propose complementary or alternative taxonomies that are worth considering as well (Cresci et al., 2019; Stieglitz et al., 2017; K. Yang et al., 2019)

Some openly accessible tools exist to detect bots on platforms like Twitter:

(1) *Botometer*[3], also used here, is a bot-detection tool developed at Indiana University (Davis et al., 2016);
(2) *BotSlayer*[4] is an application to detect and track potential manipulation of information on Twitter;
(3) the *Bot Repository*[5] is a centralized database to share annotated datasets of Twitter bots.

Finally, various models have been proposed to detect bots using sophisticated machine learning techniques, such as:

(1) Deep learning (Kudugunta & Ferrara, 2018),
(2) Anomaly detection (Echeverria et al., 2018; Gilani, Farahbakhsh, et al., 2017; Gilani, Kochmar, et al., 2017; Minnich et al., 2017),
(3) Time series analysis (Chavoshi et al., 2017; Pozzana & Ferrara, 2020; Stukal et al., 2017).

Due to the continuously evolving nature of bots, and the challenges that that poses for detection, in this article we will focus on studying the top and bottom end of the bot score distribution, rather than carrying out a binary classification of accounts into bots and humans. This avoids the conundrum of dealing with borderline cases for which detection can be inaccurate, and conversely to focus on accounts that exhibit clear human or bot traits. Furthermore, the results will be manually validated for accuracy.

[3] Botometer: https://botometer.iuni.iu.edu/
[4] BotSlayer: https://osome.iuni.iu.edu/tools/botslayer/
[5] Bot Repository: https://botometer.iuni.iu.edu/bot-repository/

METHODOLOGY

In this section, we describe the data and methodologies adopted in this paper. We first introduce the unique Twitter dataset at our disposal, covering a significant period of the early COVID-19 outbreak. We then discuss the adopted bot detection strategies, their validation and quality assurance.

DATA

We have been continuously tracking online chatter regarding COVID-19 since mid-January 2020. Our primary focus has been Twitter. The collected full dataset comprises of nearly a hundred million tweets about the outbreak and related events worldwide. We decided to make our dataset public to foster research on this important problem. Information about this dataset (COVID-19 Twitter Dataset[6]) have been recently published (Chen et al., 2020). Some details follow next.

Data collection. We began actively collecting tweets on January 28, 2020, leveraging Twitter's streaming API to follow specific keywords and accounts that were trending at the time. When we started collecting tweets, we also used Twitter's search API on the same keywords to gather related historical tweets. This allows to get all data matching the query that were posted up to one week prior. Thus, the earliest tweets in our collection date back to January 21, 2020. On that day, the first case of coronavirus infection on United States soil was announced to the public (Holshue et al., 2020). Since then, we have incrementally added keywords and accounts to follow based on trending conversations occurring on Twitter at any time. We have collected nearly 100 million tweets from the inception until the time of this writing, constituting approximately a terabyte of compressed raw data.

Tracking content. By continuously monitoring Twitter's trending topics, keywords and sources associated with COVID-19, we did our best to capture conversations related to the coronavirus outbreak. Twitter's streaming API returns any tweet containing the keyword(s) in the text of the tweet, as well as in its metadata (including URLs and embedded images' filenames). We list the keywords that we have been collecting in **Table 1**, along with the date we began tracking them. Due to the evolving nature of the pandemic and online conversations, these trackers expanded as we continued to monitor Twitter for additional keywords and accounts to add to our monitoring list. The version of the dataset used in this paper (v1.1) contains approximately 62M tweets. The data period spans 1/21/2020 to 3/12/2020.

Keywords. By nature of the Twitter API, tweets containing the keyword, or part of it, anywhere in their text or metadata were captured. For example, by tracking "Corona", our data collection would include all tweets containing "Coronavirus", "CoronavirusOutbreak", etc. It's also worth noting that the search is case-insensitive, so it ignores differences in capitalization. Dashes and Unicode characters (e.g., "–") constitute an exception and have to be tracked separately, hence we track "Covid-19" and "COVID–19" in addition to "Covid". We provide some preliminary statistics of the collected tweets. **Table 2** provides some descriptive statistical summary of the data at our disposal. The language breakdown is in **Table 3**. One can immediately appreciate the global nature of the data, with multiple languages represented. In this particular study, however we will focus exclusively on the English subset of 43.3M tweets.

[6] COVID-19-TweetIDs (v1.2): https://github.com/echen102/COVID-19-TweetIDs

Table 1: Examples of tracked keywords (case-insensitive) in our data collection.

Keyword	Tracked Since	Keyword	Tracked Since
Corona	1/21/2020	Outbreak	1/21/2020
CDC	1/21/2020	Covid-19	2/16/2020
Ncov	1/21/2020	Corona virus	3/2/2020
Wuhan	1/21/2020	Covid	3/6/2020
N95	1/21/2020	Sars-cov-2	3/6/2020
Epidemic	1/21/2020	COVID–19	3/8/2020

Table 2: Statistics of our dataset comprising 62M tweets.

Number of tweets	61.98 million	Period of data	1/21/2020 – 3/12/2020
No. original tweets	7.36 million	No. reply tweets	3.52 million
No. retweets	39.9 million	No. all other tweets (quotes, etc.)	11.2 million
Number of users	12.92 million	No. Users with bot scores	1,056,124 (8.2%)
No. Suspended users	486,814	Suspended users with bot scores	15,282 (3.14%)
No. Verified users	88,635	Verified users with bot scores	5,517 (6.22%)

Table 3: Language breakdown (top 10) of our dataset.

Language	No. original	No. replies	No. retweets	No. quoted	Total
English (en)	4,915,281	2,492,012	29,782,240	6,132,604	43,322,137
Spanish (es)	805,504	308,736	3,688,091	1,113,688	5,916,019
Indonesian (in)	193,409	110,970	1,137,288	371,074	1,812,741
French (fr)	186,050	59,426	879,914	637,453	1,755,843
Thai (th)	16,580	3,006	1,009,847	623,184	1,652,617
Portuguese (pt)	188,526	97,060	635,925	310,169	1,231,680
Italian (it)	221,405	54,515	636,092	410,134	1,200,145
Japanese (ja)	101,719	30,822	657,470	178,543	1,090,555
Turkish (tr)	116,562	32,768	300,634	107,322	557,286
Tagalog (tl)	69,998	40,224	141,318	151,973	403,513

Bot Detection and Validation

In this paper, we use three sources to make determinations about the nature of the scrutinized accounts. For bot detection, we rely upon the previously mentioned Botometer API. We also obtained the list of all verified Twitter accounts as of the end of March 2020. Finally, we queried Twitter to determine all accounts in our dataset that have been suspended as of the beginning of April 2020.

Botometer Score. To estimate whether an account is automated by software rather than controlled by a human users, we rely upon Botometer (Davis et al., 2016; K. Yang et al., 2019). Specifically, we use the latest version of Botometer that was available at the time of this research, namely version v3, whose improvements over the previous versions are documented in a recent paper (K. Yang et al., 2019); this includes the ability to leverage newly-found features that are revealing of more sophisticated bot behavior, as well as a broader training set that includes celebrity accounts and other types of accounts (e.g., organization accounts) that were previously wrongly-classified as bots in a rather systematic way. Finally, we use the bot score provided by the API that is specific of English language (as opposed to the so-called “universal score” that is better suited for non-English content). Bot scores have been continuously collected throughout the duration of this study’s data collection, for random sample of the users observed in the discussion. The deliberately decided to collect a random sample, as opposed to focusing on the most active users, to guarantee the collection of bot scores across the entire spectrum of usage activity. At the time of this writing, we collected bot scores for 1,056,124 users, or approximately 8.2% of the total – or about 1 in 12 accounts in our data. To the best of our knowledge, this makes it the largest ever study in terms of number of users annotated with bot scores provided by Botometer. Having such large number of annotations at our disposal allows us to study the top and bottom end of the bot score distribution, rather than doing a binary classification of the accounts, hence focusing solely on accounts that exhibit clear human or bot traits.

Verified Twitter accounts. On March 27, 2020 we obtained the full list of Verified Twitter accounts from PushShift.io – a service that tracks the entire Twitter firehose stream and provides data and analytics to researchers. As of that date, there were 351,660 Verified accounts on Twitter as a whole. Out of that number, we observed 88,635 verified accounts that are present in the dataset at hand.

Suspended Twitter accounts. On April 2, 2020 we queried the Twitter API, specifically the *users/lookup* endpoint, to determine out of all the nearly 13M users, how many were suspended. As of that date, out of the total there were 486,814 suspended accounts in our dataset, accounting for roughly 3.7%, or about 1 in 27, of the total user population in our data. These accounts have been suspended in the time span between last observed activity in our data and time of our query (i.e., April 2, 2020).

STATISTICAL ANALYSIS

BOT SCORE RANK DISTRIBUTION

Bot detection is hard. Bot-making tools continuously evolve, and the capabilities of bots improve while available bot detection techniques catch up. For this reason, it becomes increasingly harder to classify users “in the wild” preserving high degrees of accuracy across the whole spectrum of human-to-bot likeness. For such a reason, in this study we focus on the top and bottom end of the bot score rank distribution, and isolate accounts in the top decile (i.e., top 10^{th} percentile of the bot score distribution) and flag them as *high bot score accounts*; conversely, we isolate users in the bottom decile (i.e., bottom 10^{th} percentile of the distribution), and refer to them as *low bot score accounts*. We will only draw distinctions at the aggregate level between these two groups, without making any further inference, either binary or probabilistic classification, of the nature of any given account.

In **Table 4**, we show the percentile rank distribution of bot scores and average values of a subset of selected user activity features, namely (i) the total number of tweets posted by each users, (ii) the proportion of COVID-19 related tweets observed in our data, (iii) and the account age, measured as the number of days elapsed between the creation of the account and their first COVID-19 tweet in our data. The distribution portrays such aggregate statistics every fifth percentile of the bot scores.

Some striking patterns emerge. As the bot scores increases, the number of total tweets posted by users on average decreases. For example, users in the bottom 5th percentile (0.05) have posted on average over 15 thousand tweets. Conversely, accounts in the top 5th percentile, have posted on average only about 1,600 total tweets. A similar pattern emerges with account age. Accounts in the bottom end of the bot score distribution have been active on average for almost three thousand days (or 8 years!) as opposed to accounts with the higher bot scores, whose average age is less than three years.

However, the trend is reversed when looking at the fraction of COVID-19 related tweets: accounts with higher bot scores post significantly more COVID-19 tweets than those in the lower end of the distribution. In fact, for accounts in the top 5th percentile of the bot score distribution, the ratio of COVID-19 tweets to their total is 0.81%, whereas for the bottom 5th percentile this ratio is 0.03%. In other words, accounts with the highest bot scores post about 27 times more about COVID-19 than those with the lowest bot scores.

Suspensions across the bot score spectrum vary from about 2% to approximately 3.5%, with accounts having bot scores in the 60-80 percentile being more likely to get suspended. Concluding, only 81 accounts (0.1%) with bot scores in the top decile are verified. Accounts in the bottom decile are on average twenty times more likely to be verified (avg. ~2%) than the top decile (avg. 0.1%).

Overall, the insights drawn from the bot score rank distribution analysis suggest that an investigation to characterize the behavior of suspicious accounts with high bot scores is warranted.

Table 4: Rank distribution of bot scores and account activity average metrics, along with suspended and verified statistics.

Percentile	Bot Score	Total Tweets	COVID-19 Ratio %	Account Age	Suspended %	Verified %
0.05	0.03	**15312**	0.03	2909	2.34	1.9
0.1	0.04	13287	0.04	**2919**	2.12	**1.97**
0.15	0.04	10986.5	0.06	2864	2.11	1.73
0.2	0.05	9223	0.07	2742	2.05	1.48
0.25	0.05	8009	0.09	2585	2.11	1.34
0.3	0.06	7151	0.1	2418	2.29	1.11
0.35	0.07	6592.5	0.12	2232	2.42	0.94
0.4	0.08	6036.5	0.13	2054	2.56	0.8
0.45	0.09	5566	0.15	1827	2.61	0.74
0.5	0.1	5122	0.17	1551	2.9	0.59
0.55	0.12	4621	0.19	1399	3.09	0.54
0.6	0.14	4375	0.21	1237	3.3	0.44
0.65	0.16	4016	0.24	1176	3.42	0.37
0.7	0.19	3752	0.27	1134	3.42	0.34
0.75	0.22	3601	0.3	1136	**3.47**	0.41
0.8	0.27	3416	0.33	1187	3.21	0.29
0.85	0.34	3510	0.35	1314	2.74	0.24
0.9	0.44	4178.5	0.31	1503	2.48	0.17
0.95	0.59	5112	0.27	1609	2.07	0.15
1	**0.79**	1626.5	**0.81**	1184.5	2.24	0.06

BOT SCORE DISTRIBUTION VALIDATION

To provide additional insights in the bot score distribution, we leverage the annotations of verified and suspended users. In **Figure 1**, we illustrate the histogram of bot scores for verified and suspended accounts in our dataset. The two distributions are statistically very significantly different (Mann-Whitney rank test, p-value<0.001).

Whereas approximately 90% of the verified accounts have bot scores lesser than 0.1, the bot score of suspended accounts is much more broadly distributed, with approximately half of the suspended accounts exhibiting scores higher than 0.1. Whereas in the lower end of the distribution there are both suspended and verified accounts – which is to be expected, since accounts can be suspended for various reasons, not just for being automated – the upper end of the distribution does not contain almost any verified user, but it exhibits hundreds of suspended accounts. This suggests that there is a correlation between account suspension and increased bot likeness.

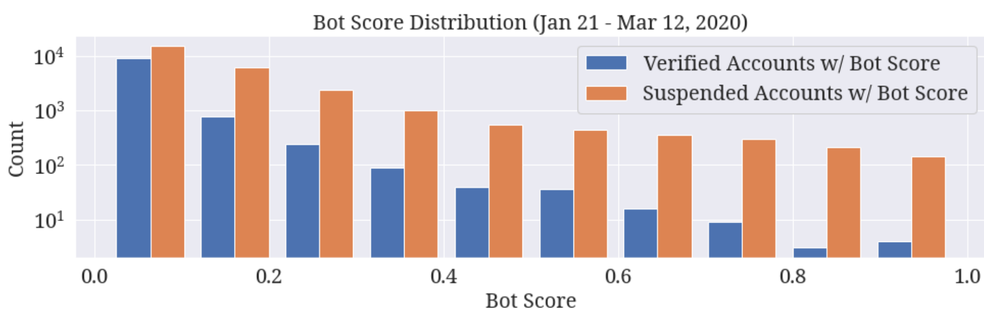

Figure 1: Distribution of bot scores for verified and suspended accounts in our dataset.

HIGH AND LOW BOT SCORE USER ANALYSIS

STATISTICAL ANALYSIS OF PREDICTIVE FEATURES

In line with recent analysis by (K.-C. Yang et al., 2019), as well as our prior work (E. Ferrara, 2017), we report six basic account meta-data features that are known to carry predictive power in the differentiation between bot and human users, namely (i) topical tweets (in this case, COVID-19 related tweet count), (ii) total number of tweets, (iii) number of followers, (iv) number of friends, (v) number of favorited tweets, and (vi) number of times the account was added to a list by other users.

In **Figure 2**, we show the strip plots of the distributions for all features, for accounts in the bottom decile of the bot score distribution (left plot) as well as for accounts in the top decile (right plot). The strip plots convey all observations alongside samples of the underlying distribution data, displayed as jitter over the bot plot that is randomly sampled from the underlying distribution. Visual comparison of the two plots illustrates immediately a striking difference in the distributions, which is confirmed by statistical analysis: the feature-pairwise Mann-Whitney rank tests are all strongly significant, all p-values<0.001. In line with **Table 4**, low bot score users appear to have posted and favorited significantly more over longer periods, have more friends and followers, and finally have been added to more lists by other users than accounts with highest bot scores; the latter, however, have posted significantly more COVID-19 related tweets over a significantly shorter life span.

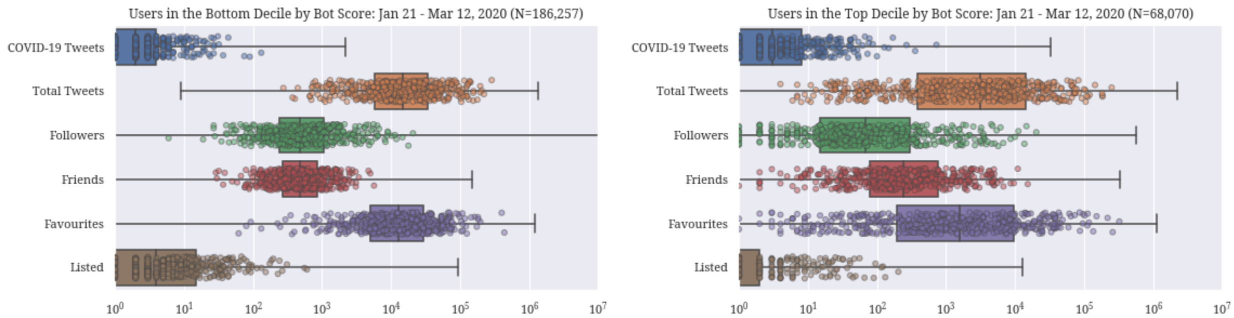

Figure 2: Distributions of user activity features for low bot score users (left) versus high bot score accounts (right).

The specifics of the underlying distributions portrayed in **Figure 2** can ben observed in **Figure 3**. The six plots illustrate the probability density function pdf(x) of users in the top (blu lines) and bottom (orange lines) deciles by bot scores. One can further appreciate how users that exhibit more human-line traits, in the bottom 10th percentile of bot scores, have significantly different activity and engagement metrics than accounts exhibiting bot-like scores in the top 10th percentile. Combined, **Figures 2** and **3** corroborate the hypothesis that the two sets of accounts behave in significantly different ways, and subsequently they are perceived differently by other users, and this also affects the prevalence of COVID-19 related activity.

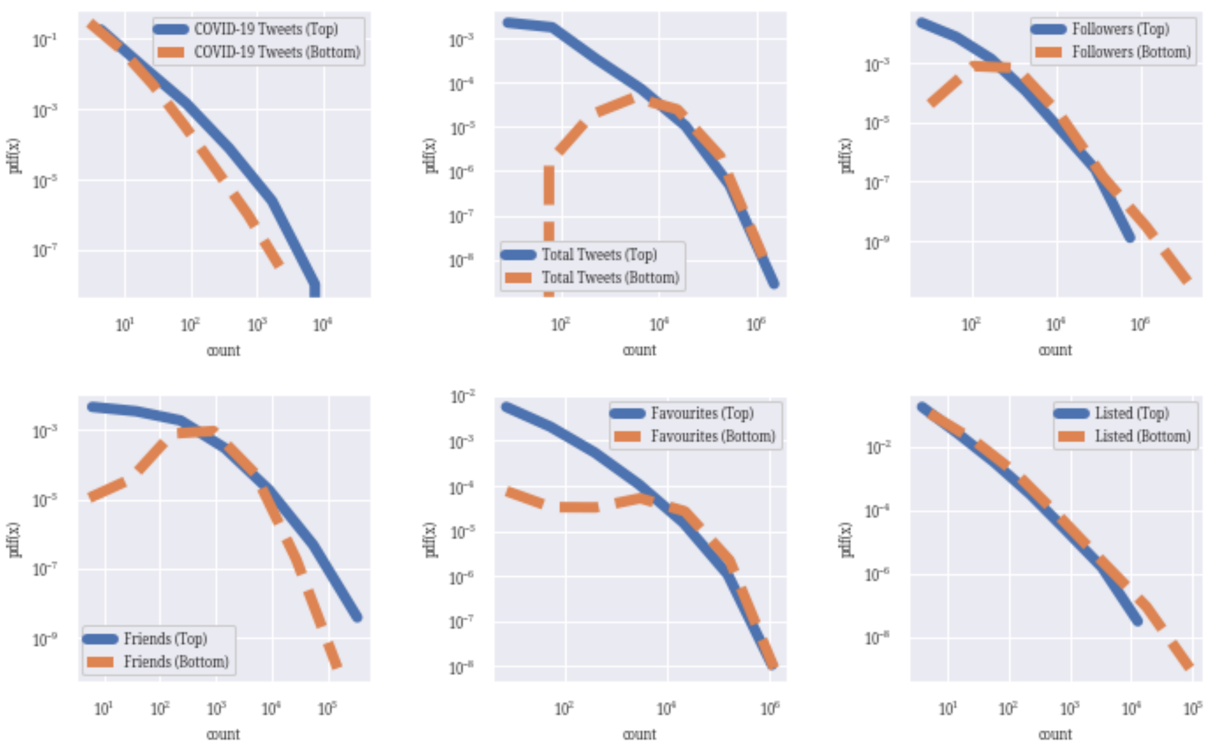

Figure 3: Distributions of six user activity features for users in the top and bottom deciles by bot score distribution.

AGE AND PROVENANCE ANALYSIS

Our final investigation in the characteristics of high and low bot score accounts centers around age and provenance of the users. The analysis above suggested than on average users with higher bot scores in our COVID-19 dataset also exhibit shorter account age. Account age and prevalence of activity related to COVID-19 appear to be very strongly correlated features.

To further investigate this relation, in **Figure 4** we show the distributions of account age of users at the time when they joined the COVID-19 discussion. The histogram portrays two very distinct stories for accounts in the top and bottom deciles of the bot score distribution: the former appear to be joining the COVID-19 in the early days since their creation: in fact, the average amount of time elapsed between account creation and first COVID-19 post for high bot score users is less than 100 days. In other words, the vast majority of high bot score accounts have been created relatively in proximity to the emergence of the COVID-19 outbreak and jump on this discussion with high intensity shortly after their creation. For example, over 80,000 high bot score accounts have been created between 50 and 100 days prior to their first COVID-19 tweet. Conversely, it is apparent that accounts in the bottom decile of the bot score distribution have been created significantly prior to the events.

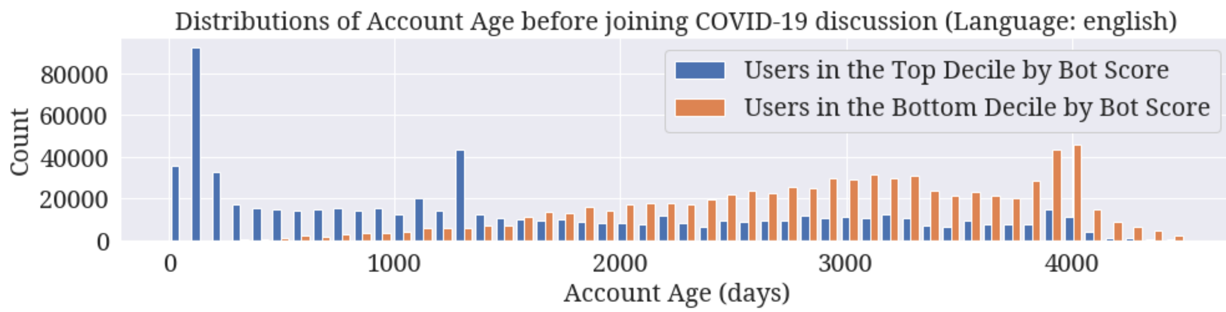

Figure 4: Age distribution of the accounts in the top and bottom decile of the bot score distribution at the time of joining the COVID-19 Twitter discussion.

The second aspect that we evaluate is the provenance of these accounts. The attribution of account provenance is a well-known challenging task, because the Twitter API does not provide crucial forensic information that would be required for exact provenance attribution, such as the IP address of the machine or VPN server an account connects from. However, in lieu of such information, the best proxy at our disposal is the ability to reconstruct the server and data center that dispatched each tweet. There are two data centers, namely DC10 and DC11, that dispatch tweets, and 30 servers associated with these data centers. A simple language analysis of the tweets originating from the two data centers clearly suggests that DC10 is used to dispatch tweets originating in Asia, South-East Asia, Russia, and the Middle East. Conversely, DC11 dispatches tweets originating predominantly from Europe and the Americas.

Figure 5 illustrates remarkable differences in the data center connectivity patterns between accounts in the top and bottom deciles by bot scores. In particular, it appears evident that DC10 (the data center that serves the Eastern world) dispatches almost 50% more tweets originating from high bot score accounts (blue bars) than from low bot score ones (orange bars). The opposite patterns appear to be true for DC11 (the data center serving the Western world): DC11 dispatches two-third more tweets from low bot score users than from top score ones. We can only speculate about the origin of such difference, in absence of the investigative tools necessary to get a definitive answer: prior investigations carried out by Twitter unveiled the systematic presence of information operations based in countries such as Russia, Iran, North Korea, and China. We speculate that a fraction of these may be carried out using bots, which in turn leave a digital trail associated with the data center used to dispatch tweets.

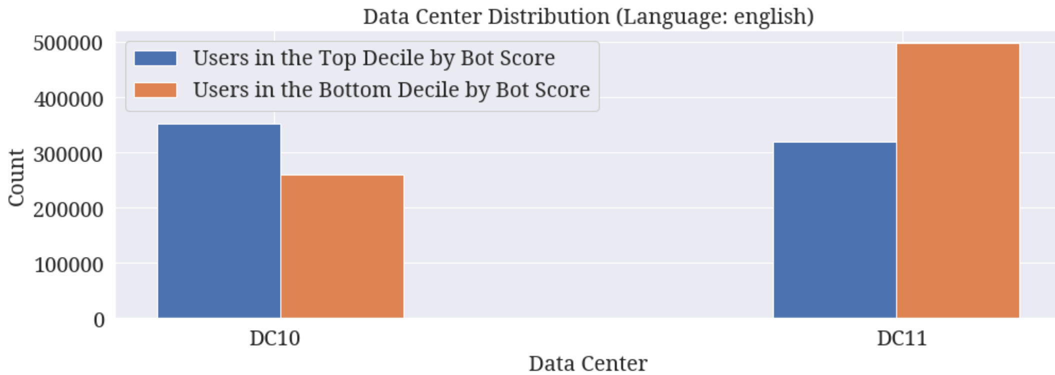

Figure 5: Distributions of the data center provenance of accounts with high and low bot scores.

Furthermore, **Figure 6** provides additional insights on the discrepancies observed at the level of server connectivity. If no correlation existed between bot likeness and internet connectivity patterns, one would expect the same number of tweets originating from each server as a function of bot scores. However, Figure 6 clearly shows that certain servers serve many more high bot score users, and conversely others serve significantly more low bot score users.

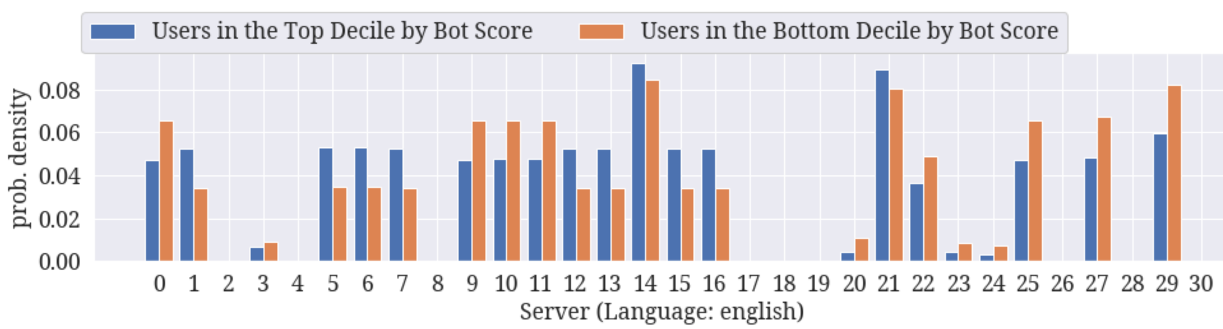

Figure 6: Distributions of the server used by accounts with high and low bot scores.

CONTENT ANALYSIS

The goal of the following analysis is to identify patterns of information production and consumption that are distinctive of the high (resp., low) bot score populations. To this aim, we will isolate all tweets produced by these two groups and carry out contrastive analysis in the adoption patterns of keywords and hashtags, trying to surface the differences, if any, that likely automated accounts exhibit with respect to likely human users. Next, we detail the preprocessing steps taken to curate the textual content (tweets) produced and consumed by the two groups.

Preprocessing. The first step was to isolate tweets in English language. The Twitter API provides the estimation of the language, which we leveraged to select a subset of approximately 43.3M tweets. Out of this set, we isolated the tweets produced by the high (resp., low) bot score users (i.e., the users in the top and bottom deciles of the bot score distributions). This produced 671,774 tweets total tweets for the *high* *bot score accounts*, and 756,940 tweets for the *low bot score users.* Starting from these tweets, we will extract the distinctive hashtags and n-grams preferentially adopted by the two groups of accounts.

Tweet type disaggregation. On twitter, there are four modalities of posting: (i) original tweets; (ii) reply tweets; (iii) quoted tweets; (iv) retweets. Each of these mechanisms is used for different purposes. An original tweet is posted any time a user composes a new tweet from scratch. A reply is an answer to another tweet, typically posted by another user, albeit it is possible to reply to one's own tweets. A quote embeds another tweet and adds original text typically as a commentary; it's once again possible to quote one's own tweets, however typically a quote embeds tweets posted by other users. Finally, a retweet is a one-click operation that allows to reshare on one's timeline another tweet that will appear without any modifications or commentary (again, typically posted by another user, despite it's possible to retweet one's own tweets). In the following analysis, we will disaggregate according to these four communication mechanisms, since they have different aims, and can also be abused in distinctive ways. By means of this disaggregation, we obtain the following subsets of tweets:

(i) 50,483 and 83,342 original tweets posted by high and low bot score accounts, respectively;
(ii) 10,852 and 50,756 reply tweets posted by high and low bot score accounts, respectively;
(iii) 70,432 and 153,304 quote tweets posted by high and low bot score accounts, respectively;
(iv) 540,007 and 468,539 retweets posted by high and low bot score accounts, respectively.

It's worth observing how *high bot score accounts* appear to disproportionally predilect the adoption of retweets (which is a one-click, or if you wish a one-line-of-code operation) whereas *low bot score users* tend to produce significantly more original, reply, and quoted tweets. For example, *low bot score users* produce nearly five times more reply tweets, and more than twice quoted tweets, than *high bot score accounts*.

N-gram extraction. We will carry out a systematic n-gram analysis to surface common sub-sentences that tend to occur frequently in the tweets. An n-gram is simply a sequence of *n* words. We carried out n-gram extraction for n=1, 2, and 3. For n=1 we obtain 1-grams, a.k.a. unigrams. For n=2 we obtain 2-grams, a.k.a. bigrams. Finally, for n=3 we obtain 3-grams, a.k.a. trigrams.

Each tweet in the corpus is curated according to the following cleaning protocol. First, the tweet text is lower-cased. Then, link (URLs) are removed, alongside user mentions (username of other Twitter accounts which are preceded by the "@" symbol), and hashtags (terms preceded by the "#" symbol) – a hashtag analysis is indeed carried out separately. Special characters are also removed, to clean the tweet text from non-linguistic symbols such as ampersands, etc. Finally, stop-words, common English-language terms that include short function words, as well as non-lexical words, are also removed using the "nltk" Python library. Finally, for each n-gram we check that at least one word is longer than 4 characters, to remove common n-grams typical of online slang, such as "lol" or "ah ah", etc.

The n-grams extracted by this process will be analyzed in the next sections.

Hashtag extraction. The Twitter API provides the list of hashtags included in each tweet, which we leverage to extract all hashtags from all 671,774 tweets in the *high bot score accounts* group (resp., 756,940 tweets in the *low bot score users* group). For each tweet, we remove hashtags that contain the keywords that we used to seed the data collection, which are listed in **Table 1**. For example, if a tweet contains two hashtags, e.g., #coronavirus and #ncov2019, the former would be removed because it contains the keyword "corona" that we used to seed the data collection. The removal of hashtags containing seed keywords make surfacing other interesting hashtags easier. We decided to avoid any other preprocessing step, such as removing hashtags shorter than some number of characters, or hashtags that do not appear above a certain threshold, in order to avoid skewing the results in any way. This means that at times we will observe hashtags that exhibit low volume, for example hashtags that contain a misspell, e.g., "coronoavirus" can be observed.

FINDINGS FROM CONTENT ANALYSIS

In the following, we discuss two core findings that we our analysis highlighted. First, we discuss how *high bot score accounts* appear to be engaged in the spread of conspiracies and political propaganda, in stark contrast with the comparison group of *low bot score users* that is instead engaged in the discussion of public health concerns. Second, we highlight how *high bot score accounts* are especially concerned about pushing civic engagement narratives related to concerns of China's limiting freedom of speech about the outbreak, in opposition to the control group discussing calls for solidarity and help.

CONSPIRACIES & POLITICAL PROPAGANDA

The spread of conspiracies on online social media is a well-established issue. Numerous studies have been devoted to understand, for example, how unscientific claims circulate online (Alessandro Bessi et al., 2015; Kahan et al., 2012; Scheufele & Krause, 2019; Vicario et al., 2016), or how conspiratorial narratives are constructed online (Andrews et al., 2016; Arif et al., 2016; Starbird, 2017), especially in the context of political ideology (E. Ferrara, 2017; Marwick & Lewis, 2017). Our analysis highlights the emergence of similar issues in the context of the conversation related to COVID-19.

Distinctive n-grams. In **Figure 7**, we show the time-series of the top 10 distinctive trigrams produced by the two populations, the *high* and *low bot score users*, in original tweets. We also display an associated word cloud for each group. An n-gram is considered distinctive in this analysis if it appears in the top 10 of a group (e.g., the *low bot score users*) but does not appear in the top 10 of the other (e.g., the *high bot score accounts*). This allows us to surface the most popular characteristic linguistic trends in each group. Each time-series in Figure 7 shows the daily volume of tweets containing a given trigram in that population exclusively. It is important to underscore, again, that this is the prevalence of the n-grams in each group, and not in the whole Twitter population – this is in order to give a perspective on the relative volume of content produced by each group relative to each other, to enable a contrastive analysis.

Findings. The top panel of **Figure 7** provides ample evidence that *high bot score accounts* are predominantly posting conspiratorial content of political nature alongside with COVID-19 information. Various alt-right popular narratives can be isolated which are aimed at pushing divisive political ideology. The relative volume of activity associated with these narratives driven by *high bot score accounts* is in all comparable with that of *low bot score users* concerned with public health risks (bottom panel of **Figure 7**).

Figure 7: Time-series of the top 10 distinctive trigrams in original tweets for low (bottom) and high (top) bot score accounts. Right: word clouds of top distinctive keywords in the tweets associated with the narratives displayed on the left.

The top distinctive trigrams in this analysis contain keywords such as *QAnon*, which according to Wikipedia,[7] is "is a far-right conspiracy theory detailing a supposed secret plot by an alleged "deep state" against U.S. President Donald Trump and his supporters". *QAnon* has been extensively adopted by alt-right activists to foster participatory advocacy on social media (Zuckerman, 2019), but it has also been abused by the Russian Internet Research Agency (IRA) to push conspiratorial and divisive narratives (Cosentino & Cosentino, 2020a, 2020b). *QAnon* appears alongside with other known alt-right terms, e.g., *GreatAwakening*, *Inf0wars*, WWGWGA (Where We Go One We Go All), *WeThePeople*, and *PatriotsFights* (de Saint Laurent et al., 2020). The results are consistent when considering unigrams and bigrams.

[7] https://en.wikipedia.org/wiki/QAnon

Distinctive hashtags. In **Figure 8**, we illustrate the time-series of top 10 distinctive hashtags characterizing the *low* and *high bot score accounts,* in original tweets*.* As for the n-gram analysis, hashtags are deemed distinctive to a group if they appear in the top 10 of that group and do not appear in the top 10 of the other. We note how the top 10 of distinctive hashtags for the *low bot score users* contains, again, predominantly hashtags that are associated with the public health aspects of the COVID-19 pandemic. The most common distinctive hashtags include news-related terms like #breaking, mentions of influential actors (e.g., #trump), organizations (e.g., #who), and countries (e.g., #iran), alongside with the COVID-19 hashtags used to characterize the disease-related topic, including #2019ncov, #ncov, and #ncov2019.

On the other hand, for *high bot score accounts*, we observe once again a picture compatible with the n-gram analysis discussed above. Alongside with news-related hashtags such as #news and #smartnews, we observe some alt-right hashtags such as #qanon and #greatawakening.

Validation. Taken together, the n-gram and hashtag analyses paint a picture suggestive of the fact that *high bot score accounts* are injecting content with conspiratorial narratives charged with alt-right ideology. These hypotheses are further validated with a process of manual verification and coding.

The first step in our validation was to determine how many verified accounts in the *high bot score* population posted about conspiracies and political propaganda. For sake of illustration, we discuss the results for the trigram analysis discussed above. This validation step determined that only 9 verified accounts posted content (out of 1803 total, or 0.05%). Conversely, for *low bot score users*, the number of verified accounts was 215 (out of 1037 total, or 20.7%). In other words, in the trigram analysis above, *low bot score users* were 414 times more likely to be verified than *high bot score accounts.* The former population consisted predominantly of established social media accounts that reported on public health concerns related to the pandemic. On the other hand, the *high bot score accounts*, of which only 0.05% is verified, may have used the COVID-19 conversation as a vector to promote conspiratorial narratives.

Lastly, we manually investigated the content of tweets in both groups discussed in the trigram analysis of original tweets. We document some findings next.[8] As for the *high bot score accounts*, the most popular tweets were posted by accounts that exhibit clear automation patterns; they have high friends/followers ratio and posted thousands of tweets about COVID-19 following templated formats. For example, we found hundreds tweets that contain references to news about COVID-19, followed by sequences of hashtags such as those displayed in **Figure 8** (top panel) in combination with a link, typically to Youtube videos (many of which had already been taken down by Youtube as of the time of this paper). Beside Youtube, the most referenced sources included various hyperpartizan news site, such as InfoWars, ZeroHedge, etc. Typical sensationalistic headlines suggest that: (i) the virus was made in Wuhan labs; (ii) the virus is a "globalist biological weapon"; (iii) the virus was imported into China by the US military; (iv) products imported from China may be infected with the virus. As for *low bot score users,* they reference to traditional news sources, and the most referenced accounts are the US President, the CDC, the WHO and various established news organizations including both left and right leaning newspapers.

[8] Due to privacy requirements and in line with Twitter's Terms of Service, we here only refer to anonymized examples of tweets, removing any information that could be used for the reidentification of the authors.

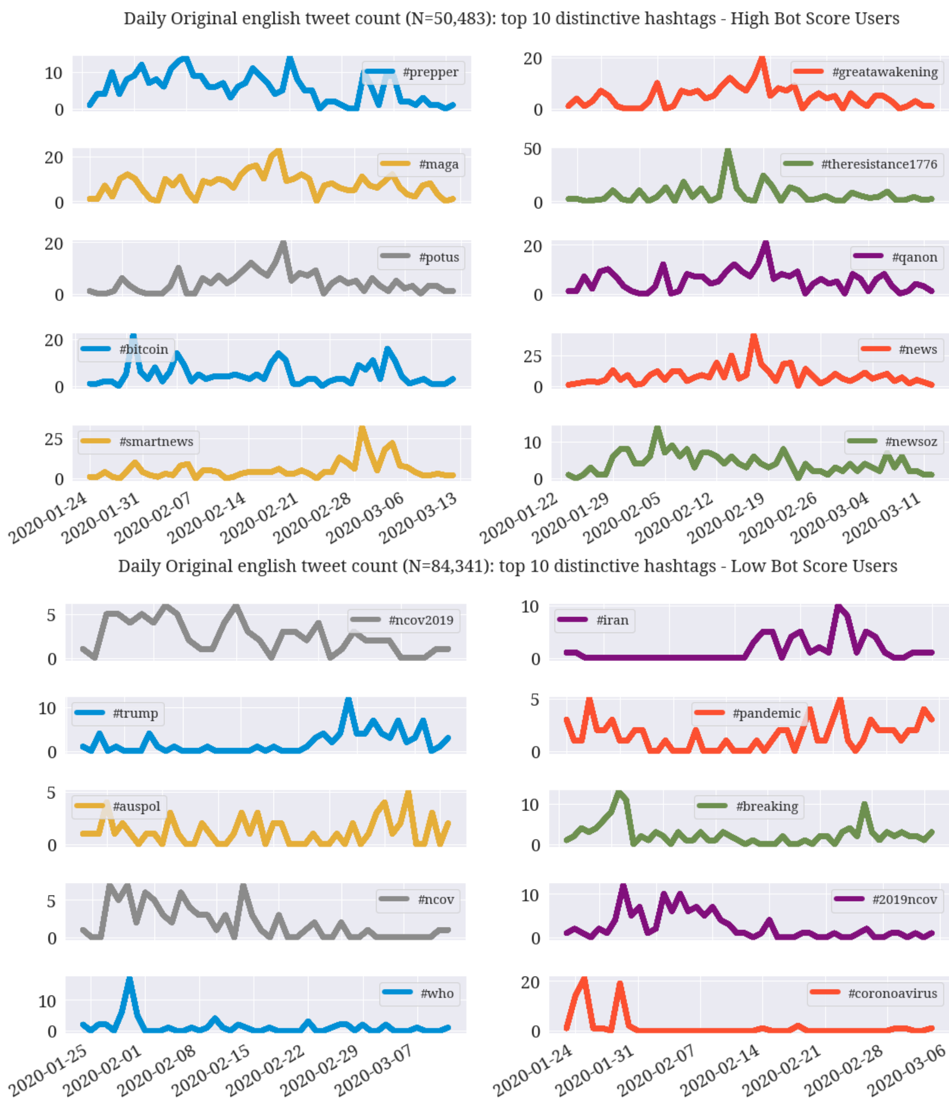

Figure 8: Time-series of the top 10 distinctive hashtags in original tweets for low (bottom) and high (top) bot score accounts.

Participatory Activism

A wealth of research has been published about the democratizing nature of social media. Researchers emphasized how social media foster participatory activism, at times enabling the coordination of protests or the diffusion of information that would be otherwise censored by some governments (Comunello & Anzera, 2012; DeLuca et al., 2012; Howard et al., 2015).

Whereas the adoption of bots for political interference has been documented extensively and in numerous countries and election events (Woolley, 2016), so far to the best of our knowledge the possible use of bots for participatory political activism has gone unreported. Our analysis, discussed next, suggests that automation might have been used to spread more efficiently information on the English-speaking Twitter about events occurring in China that may have otherwise been censored by the government.

Distinctive n-grams. In **Figure 9**, we show the time-series of the top 10 distinctive trigrams produced by the two populations, the *high* and *low bot score users*, in quoted tweets. We also display an associated word cloud for each group. Analogously to the analysis carried out for original tweets, in the case of quoted tweets an n-gram is considered distinctive if it appears in the top 10 of a group (e.g., the *low bot score users*) but does not appear in the top 10 of the other (e.g., the *high bot score accounts*). The time-series in **Figure 9** show the daily volume of quoted tweets containing a given trigram in that population.

Distinctive hashtags. Figure 10 illustrates the time-series of top 10 distinctive hashtags characterizing the *low* and *high bot score accounts,* in quoted tweets. In a fashion similar to above, hashtags are deemed distinctive to a group if they appear in the top 10 of that group and do not appear in the top 10 of the other.

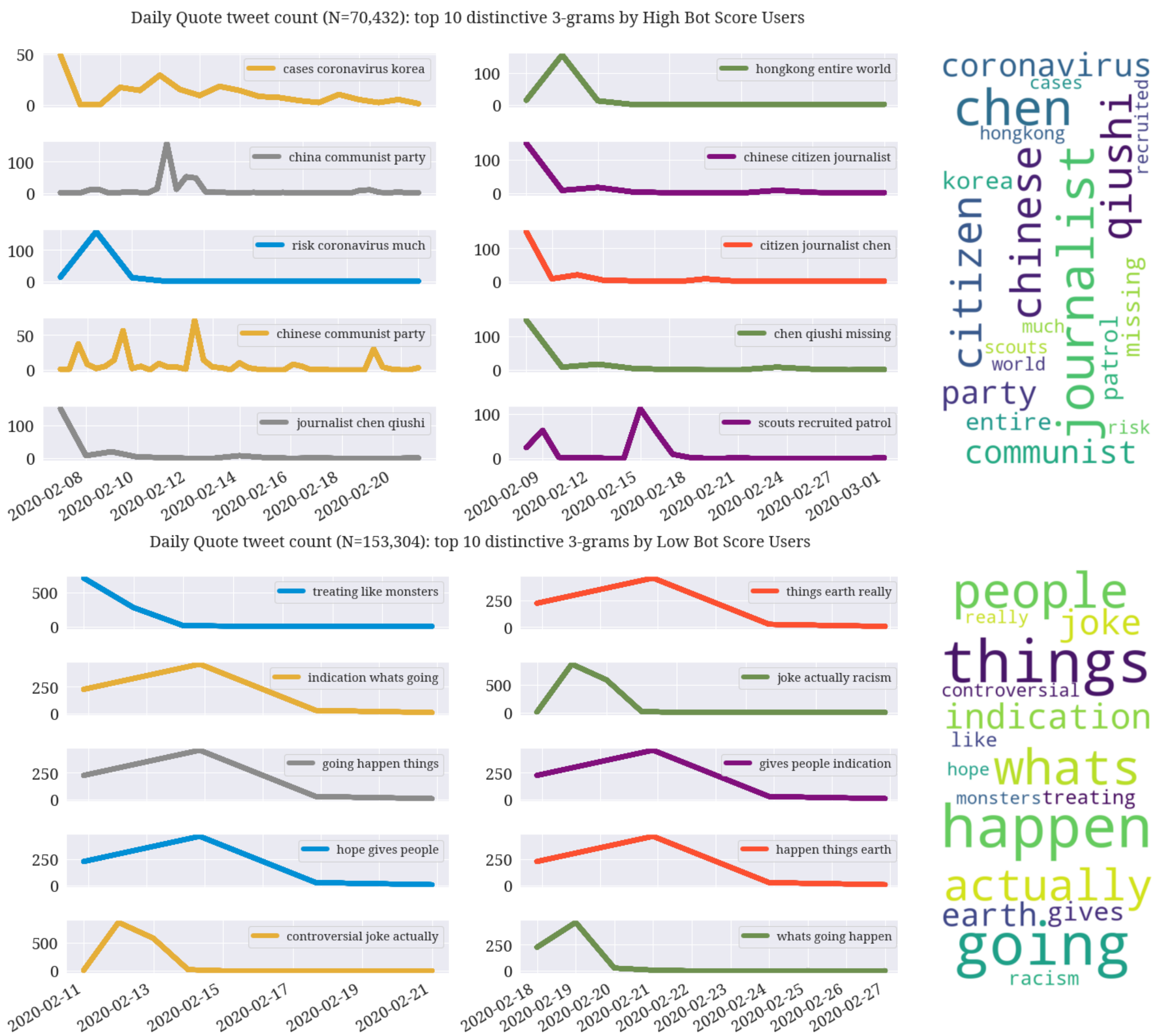

Figure 9: Daily distributions of top 10 distinctive bigrams in retweets for low (bottom) and high (top) bot score accounts

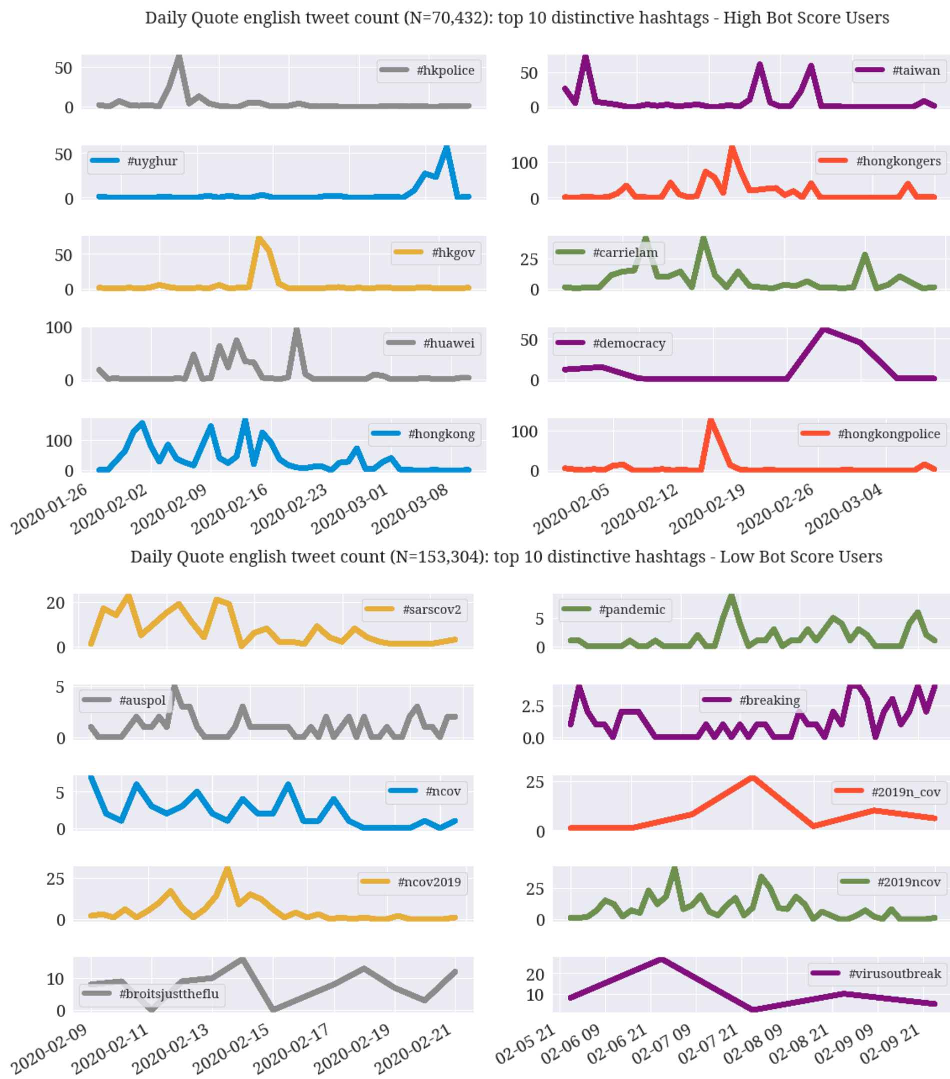

Figure 10: Daily distributions of top 10 distinctive hashtags in quotes for low (left) and high (right) bot score accounts.

Findings. Several time-series in **Figures 9 and 10** show spikes occurring in association with major offline events; most spikes happen to be concentrated in the earlier phase of the observation period, approximately around late January and the beginning of February, when the collective attention was still on China and the progression of the COVID-19 epidemic in Wuhan, the capital of Hubei province, and later Guangdong. The baseline *low bot score users* again discuss COVID-19 topics of general interest.

Numerous trigrams as well as top distinctive hashtags for the *high bot score accounts* appear to promote different narratives, all revolving around human rights and freedom of speech issues, that are compounded alongside with COVID-19. References to the ongoing Hong Kong protests against China's posture are often presented in relation to COVID-19. Other recurring narratives proposed by likely bot accounts include the relation between China and Taiwan, and human right violations against the Uyghur population. Finally, reports of censorship include references to missing journalist Chen Qiushi, one of the first to report on COVID-19 in China, who as of this writing, is reportedly held in custody by the government.

Conclusions

COVID-19 is a global crisis and with people being pushed out of physical spaces due to containment measures, online conversation on social media becomes one of the primary tools to track social discussion. In fact, topics of conversation related to COVID-19 have been trending, uninterruptedly or so, ever since the beginning of the outbreak in early 2020. We leveraged a large-scale data collection tracking in real-time COVID-19 tweets since January 21, 2020, the day the first COVID-19 case was reported on US soil. The dataset we adopt here goes through March 12, 2020, the day before the United States government announced the state of national emergency due to the COVID-19 pandemic.

In this paper, we provided an early characterization of the prevalence of accounts that are likely automated and that post content in relation to the ongoing COVID-19 pandemic. To the best of our knowledge, this is the first study to provide some evidence that *high bot score accounts* are used to:

(i) Promote political conspiracies and divisive hashtags alongside with COVID-19 content;
(ii) Enable participatory activism to shed light on issues that may otherwise be censored in China.

Our work paints a mixed picture where accounts that are likely automated have been used for social good as well as in malicious manners. On the one hand, we observed how *high bot score accounts* use COVID-19 as a vector to promote visibility of ideological hashtags that are typically associated with the alt-right.

But we have also observed how likely bots are used to foster democratic discussion, uncover issues that could otherwise get ignored or censored, and ultimately attract attention on freedom of speech and human right violations that have been reported in China.

The evidence we provide suggests that bots and automation are simply tools: they can be used for good, e.g., to bring to light issues that would otherwise get censored or ignored, or conversely can be abused to distort online narratives in order to promote political ideologies. Therefore, blanket-solutions to limit or prohibit the adoption of automation tools to produce content online, while having the positive benefits of shielding from abuse, may have the side effect of disempowering groups who may be using these tools to promote information diversity and openness to the benefit of society. A more nuanced discussion about the regulation of automation on social media platforms is therefore warranted.

Limitations of this study

Our study has several limitations. First and foremost, despite the sheer size of the dataset at hand, we are only observing a small fraction, approximately 1%, of the overall Twitter conversation, through the lens of the Twitter API. This has been shown to introduce some biases toward over-represented topics (Morstatter et al., 2013), and COVID-19 has been the most spoken-about topic of discussion ever since the beginning of the pandemic. Second, another form of bias is automatically introduced when selecting keywords to follow. For example, despite our dataset exhibits dozens of languages, English content represents over two-third of the overall tweets. To mitigate this bias, we concentrated only on English content for this analysis, despite the fact that several interesting phenomena related to the scope of this work may be observed in other languages as well.

The third and most important limitation is related to the challenge of bot detection. Detecting bots is quite hard. Even the most sophisticated machine learning tools have varying levels of accuracy, especially when applied “in the wild” to identify bots in live conversations. To mitigate this issue, in this work we carried out three forms of additional validation, collecting data for suspended and verified accounts, and manually inspecting data of particular relevance or interest. However, these solutions also have inherent limitations: for example, the list of suspended accounts only reflects Twitter’s policies to intervene and ban an account, but does not provide the rationale for the decision; furthermore, the manual assessment can be viable for the scrutiny of few case studies, like those tackled in this paper, but is not a scalable strategy to carry out large scale studies.

These limitations open up for further studies and delineate a path for forthcoming research.

Future Work

Fake news attracted a large share of attention in the research community, partly as a consequence of the 2016 US election: various studies investigated their prevalence (Allcott & Gentzkow, 2017; Allen et al., 2020; Bovet & Makse, 2019; Grinberg et al., 2019; Vosoughi et al., 2018), the reason of their origin (Pennycook et al., 2018; Pennycook & Rand, 2019), and their impact (Barberá et al., 2018; Guess et al., 2020). Computational work has been devoted to the detection of fake news and for automatic fact checking (Ciampaglia et al., 2015; Rashkin et al., 2017; Shu et al., 2017; Volkova et al., 2017).

The COVID-19 pandemic is already proving challenging in terms of containment of misinformation and rumors, as the information landscape evolves, especially when concerned with the effectiveness of treatments, the feasibility of vaccines, or the risks associated with interrupting containment measures.

In the future, we will focus on understanding the (mis)information landscape, as well as the broader implication that (mis)information has on the public, how it informs choices and policies, and influences public opinion. Our emphasis will shift on the behavior of human users, but we will also investigate to understand whether interference operations may be present, much similarly to what occurred with troll- and bot-based operations by the Russian IRA in 2016 (Badawy et al., 2019, 2018; Bail et al., 2019; A. Bessi & Ferrara, 2016; Broniatowski et al., 2018; Sutton, 2018).

Acknowledgements

The author is grateful to the members of his lab for their invaluable help and support, in particular Emily Chen for initializing the tweet collection, Shen Yan for collecting the list of suspended accounts, and Ashok Deb for collecting the Botometer scores; the author is also grateful to Jason Baumgartner (PushShift.io) for sharing the list of verified Twitter accounts.

About the Author

Emilio Ferrara is Research Assistant Professor of Computer Science at the University of Southern California and Research Team Leader at the USC Information Sciences Institute. His research focuses on characterizing information diffusion and manipulation of social media. He is a recipient of the 2016 Complex System Society Junior Scientific Award, and he received the 2016 DARPA Young Faculty Award.

References

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a social Botnet: Growth, content and influence in twitter. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 839–851. https://doi.org/10.1145/2675133.2675208

Alarifi, A., Alsaleh, M., & Al-Salman, A. M. (2016). Twitter turing test: Identifying social machines. *Information Sciences*, *372*, 332–346. https://doi.org/10.1016/j.ins.2016.08.036

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. In *Journal of Economic Perspectives* (Vol. 31, Issue 2, pp. 211–236). American Economic Association. https://doi.org/10.1257/jep.31.2.211

Allem, J.-P., Ferrara, E., Uppu, S. P., Cruz, T. B., & Unger, J. B. (2017). E-Cigarette Surveillance With Social Media Data: Social Bots, Emerging Topics, and Trends. *JMIR Public Health and Surveillance*, *3*(4), e98. https://doi.org/10.2196/publichealth.8641

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Alshaabi, T., Minot, J. R., Arnold, M. V., Adams, J. L., Dewhurst, D. R., Reagan, A. J., Muhamad, R., Danforth, C. M., & Dodds, P. S. (2020). *How the world’s collective attention is being paid to a pandemic: COVID-19 related 1-gram time series for 24 languages on Twitter*. http://arxiv.org/abs/2003.12614

Andrews, C., Fichet, E., Ding, Y., Spiro, E. S., & Starbird, K. (2016). Keeping up with the tweet-dashians: The impact of “official” accounts on online rumoring. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *27*, 452–465. https://doi.org/10.1145/2818048.2819986

Arif, A., Shanahan, K., Chou, F. J., Dosouto, Y., Starbird, K., & Spiro, E. S. (2016). How information snowballs: Exploring the role of exposure in online rumor propagation. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *27*, 466–477. https://doi.org/10.1145/2818048.2819964

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 258–265. https://doi.org/10.1109/ASONAM.2018.8508646

Badawy, A., Lerman, K., & Ferrara, E. (2019). Who falls for online political manipulation? *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 162–168. https://doi.org/10.1145/3308560.3316494

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2019). Assessing the Russian Internet Research Agency’s impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/PNAS.1906420116

Barberá, P., Tucker, J. A., Guess, A., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018).

Social media, political polarization, and political disinformation: a review of the scientific literature.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11). https://doi.org/10.5210/fm.v21i11.7090

Bessi, Alessandro, Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, *57*(2), 556–578. https://doi.org/10.1016/j.comnet.2012.06.006

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The Socialbot Network: When bots socialize for fame and money. *ACM International Conference Proceeding Series*, 93–102. https://doi.org/10.1145/2076732.2076746

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7. https://doi.org/10.1038/s41467-018-07761-2

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Chavoshi, N., Hamooni, H., & Mueen, A. (2017). DeBot: Twitter Bot Detection via Warped Correlation. *Big Data*, *5*(4), 817–822. https://doi.org/10.1109/icdm.2016.0096

Chen, E., Lerman, K., & Ferrara, E. (2020). *#COVID-19: The First Public Coronavirus Twitter Dataset*. http://arxiv.org/abs/2003.07372

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0128193

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). *The COVID-19 Social Media Infodemic*. http://arxiv.org/abs/2003.05004

Comunello, F., & Anzera, G. (2012). Will the revolution be tweeted? A conceptual framework for understanding the social media and the Arab Spring. *Islam and Christian-Muslim Relations*, *23*(4), 453–470. https://doi.org/10.1080/09596410.2012.712435

Cosentino, G., & Cosentino, G. (2020a). From Pizzagate to the Great Replacement: The Globalization of Conspiracy Theories. In *Social Media and the Post-Truth World Order* (pp. 59–86). Springer International Publishing. https://doi.org/10.1007/978-3-030-43005-4_3

Cosentino, G., & Cosentino, G. (2020b). Polarize and Conquer: Russian Influence Operations in the United States. In *Social Media and the Post-Truth World Order* (pp. 33–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-43005-4_2

Cresci, S., Spognardi, A., Petrocchi, M., Tesconi, M., & Pietro, R. Di. (2019). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. *26th International World Wide Web*

Conference 2017, WWW 2017 Companion, 963–972. https://doi.org/10.1145/3041021.3055135

Daniel, F., & Millimaggi, A. (2020). On Twitter Bots Behaving Badly: A Manual and Automated Analysis of Python Code Patterns on GitHub. *Journal of Web Engineering*, *18*(8), 801–836. https://doi.org/10.13052/jwe1540-9589.1883

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to Evaluate Social Bots. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 273–274. https://doi.org/10.1145/2872518.2889302

de Saint Laurent, C., Glaveanu, V., & Chaudet, C. (2020). Malevolent Creativity and Social Media: Creating Anti-immigration Communities on Twitter. *Creativity Research Journal*, *32*(1), 66–80. https://doi.org/10.1080/10400419.2020.1712164

DeLuca, K. M., Lawson, S., & Sun, Y. (2012). Occupy Wall Street on the Public Screens of Social Media: The Many Framings of the Birth of a Protest Movement. *Communication, Culture & Critique*, *5*(4), 483–509. https://doi.org/10.1111/j.1753-9137.2012.01141.x

Echeverria, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Stringhini, G., & Zhou, S. (2018). LOBO: Evaluation of Generalization Deficiencies in Twitter Bot Classifiers. *Proceedings of the 34th Annual Computer Security Applications Conference on - ACSAC '18*, 137–146. https://doi.org/10.1145/3274694.3274738

Echeverria, J., & Zhou, S. (2017). Discovery, retrieval, and analysis of the'Star wars' botnet in twitter. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 1–8. https://doi.org/10.1145/3110025.3110074

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, *22*(8).

Ferrara, Emilio. (2015). Manipulation and abuse on social media. *ACM SIGWEB Newsletter*, *Spring*, 1–9. https://doi.org/10.1145/2749279.2749283

Ferrara, Emilio. (2018). *Measuring Social Spam and the Effect of Bots on Information Diffusion in Social Media* (pp. 229–255). Springer, Cham. https://doi.org/10.1007/978-3-319-77332-2_13

Ferrara, Emilio. (2019). The history of digital spam. *Communications of the ACM*, *62*(8), 82–91. https://doi.org/10.1145/3299768

Ferrara, Emilio, Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The Rise of Social Bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 25–32. https://doi.org/10.1145/2808797.2809292

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). *Assessing the risks of "infodemics" in response to COVID-19 epidemics*. http://arxiv.org/abs/2004.03997

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., Wang, Y., Fu, H., & Dai, J. (2020). Mental Health

Problems and Social Media Exposure During COVID-19 Outbreak. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3541120

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). *An in-depth characterisation of Bots and Humans on Twitter*. http://arxiv.org/abs/1704.01508

Gilani, Z., Kochmar, E., & Crowcroft, J. (2017). Classification of twitter accounts into automated agents and human users. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 489–496. https://doi.org/10.1145/3110025.3110091

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/SCIENCE.AAU2706

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 1–9. https://doi.org/10.1038/s41562-020-0833-x

Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S. I., Kim, L., Tong, S., Lu, X., Lindstrom, S., … Pillai, S. K. (2020). First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine*, *382*(10), 929–936. https://doi.org/10.1056/NEJMoa2001191

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Mazaid, M. (2015). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2595096

Hwang, T., Pearce, I., & Nanis, M. (2012). Socialbots: Voices from the fronts. *Interactions*, *19*(2), 38–45. https://doi.org/10.1145/2090150.2090161

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735. https://doi.org/10.1038/nclimate1547

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). *Measuring Emotions in the COVID-19 Real World Worry Dataset*. http://arxiv.org/abs/2004.04225

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*, 312–322. https://doi.org/10.1016/j.ins.2018.08.019

Li, J., Xu, Q., Cuomo, R., Purushothaman, V., & Mackey, T. (2020). Data Mining and Content Analysis of Chinese Social Media Platform Weibo During Early COVID-19 Outbreak: A Retrospective Observational Infoveillance Study. *JMIR Public Health and Surveillance*. https://doi.org/10.2196/18700

Luceri, L., Deb, A., Giordano, S., & Ferrara, E. (2019). Evolution of bot and human behavior during elections. *First Monday*. https://doi.org/10.5210/fm.v24i9.10213

Luceri, L., Giordano, S., & Ferrara, E. (2020). *Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election*. http://arxiv.org/abs/2001.10570

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society Research

Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Minnich, A., Chavoshi, N., Koutra, D., & Mueen, A. (2017). BotWalk: Efficient adaptive exploration of twitter bot networks. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 467–474. https://doi.org/10.1145/3110025.3110163

Morstatter, F., Urgen Pfeffer, J. ¨, Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Seventh International AAAI Conference on Weblogs and Social Media*. www.aaai.org

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., McPhetres, J., Zhang, Y., & Rand, D. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *PsyArXiv [Working Paper]*, 1–24. https://doi.org/10.31234/OSF.IO/UHBK9

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pozzana, I., & Ferrara, E. (2020). Measuring bot and human behavioral dynamics. *Frontiers in Physics*, *8*, 125. https://doi.org/10.3389/FPHY.2020.00125

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI Blog. https://openai.com/blog/better-language-models/

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. https://www.aclweb.org/anthology/D17-1317/

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 201805871. https://doi.org/10.1073/PNAS.1805871115

Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., & Zannettou, S. (2020). *"Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19*. http://arxiv.org/abs/2004.04046

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). *A first look at COVID-19 information and misinformation sharing on Twitter*. http://arxiv.org/abs/2003.13907

Starbird, K. (2017). Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *Eleventh International AAAI Conference on Web and Social Media*. www.aaai.org

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 201803470. https://doi.org/10.1073/pnas.1803470115

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). *Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts*. http://arxiv.org/abs/1710.04044

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting Bots on Russian Political Twitter. *Big Data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016). The DARPA Twitter Bot Challenge. *Computer*, *49*(6). https://doi.org/10.1109/MC.2016.183

Sutton, J. (2018). Health communication trolls and bots versus public health agencies' trusted voices. In *American Journal of Public Health* (Vol. 108, Issue 10, pp. 1281–1282). American Public Health Association Inc. https://doi.org/10.2105/AJPH.2018.304661

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: An analysis of twitter spam. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 243–258. https://doi.org/10.1145/2068816.2068840

Vicario, M. Del, Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647–653. https://www.aclweb.org/anthology/P17-2102/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Woolley, S. (2016). Automating power: Social bot interference in global politics | First Monday. *First Monday*, *21*(4). https://firstmonday.org/article/view/6161/5300

Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2019). *Scalable and Generalizable Social Bot Detection through Data Selection*. http://arxiv.org/abs/1911.09179

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

Zuckerman, E. (2019). QAnon and the Emergence of the Unreal. *Journal of Design and Science*, *6*. https://doi.org/10.21428/7808da6b.6b8a82b9