Introduction

Hailing as harbingers of democracy during the late 2000s (Howard, et al., 2011), social media have increasingly demonstrated the potential to be abused, which can lead to harmful consequences for their users and society at large (Ferrara, 2015). Many platforms, from fringe to mainstream, have demonstrably been the subject of some form of abuse or manipulation, as reported by practitioners, the research community, or the press, leading to potentially nefarious implications (Ferrara, 2019): For example, mainstream platforms like Facebook, YouTube, or Reddit have been associated with the rampant spread of spam, radical content, misinformation, and conspiracies (Del Vicario, et al., 2016; Tufekci, 2018; de Zeeuw, et al., 2020). Spam was particularly pronounced during the COVID-19 pandemic, and multimedia-based platforms such as Tiktok and Instagram became major drivers of public health misinformation (Shahsavari, et al., 2020; Basch, et al., 2021; Quinn, et al., 2021). Numerous studies attributed the incubation of malicious information to fringe communities: from niche platforms and discussion forums such as Gab, Minds, Parler, Voat, and 4Chan, to the alt-right information ecosystem of podcasts and fake news outlets like Infowars and Breibart, misinformation and conspiracies born out of small radical groups
systematically made it to the mainstream (Zannettou, et al., 2018a; Zannettou, et al., 2018b; Tangherlini, et al., 2020; Papasavva, et al., 2021; Aliapoulios, et al., 2021).

But chief among all platforms, for the extent of reported spam, manipulation, and abuse, might be Twitter: possibly in part due to the ease of access to its data (Williams, et al., 2013), and in part because it has been under public scrutiny via multiple government and federal investigations, Twitter attracted an overwhelming amount of attention by researchers. The platform has been the subject of thousands of academic studies aimed at highlighting a multifaceted typology of problematic behaviors (Zimmer and Proferes, 2014).

It would impossible to summarize this whole literature in one single review: for this reason, in this survey, I will focus on the Twitter-related literature with a special focus on the problem of spam and false accounts, their behavior characterization, detection, use, and effects. This choice is dictated by the ample amount of research that investigates the use of spam, false, and compromised accounts, as well state-sponsored troll farms, bot and automated accounts, in the context of nefarious activities like scams, financial frauds, coordinated political campaigns, orchestrated misinformation, crises response disruption, radical propaganda, etc.

Background

A. Taxonomy of spam and false accounts

Twitter defines four types of behavior as spam in their policy: (a) “commercially-motivated spam, that typically aims to drive traffic or attention from a conversation on Twitter to accounts, websites, products, services, or initiatives;” (b) “inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are;” (c) “coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting;” and, (d) “coordinated harmful activity that encourages or promotes behavior which violates the Twitter Rules.”

Such violations can be enacted by multiple sources, including but not limited to (i) promotional accounts (spammers); (ii) false/fake accounts; (iii) bots, or automated accounts; (iv) accounts with malicious intent (sometimes referred to as trolls); and, (v) compromised and repurposed accounts.

I provide a taxonomy of such accounts next, with the aim to later map back to this taxonomy when discussing the capability of state of the art machine learning models, and the findings of research studies in regard to the extent of different forms of spam on Twitter:

- **Promotional accounts** (or spammers): The category encompasses any account that is created on the platform with the deliberate intention to spam, *i.e.*, promote commercial products or services, with the goal of driving other (legitimate) users’ attention to that. This type of account can be controlled by human users, can be automated, or a hybrid of both. They tend to be disproportionately more active than most typical legitimate Twitter users.

- **False (or fake) accounts**: The category comprises any account that is created on the platform with the intention of boasting the visibility or appearance of influence of other accounts. False accounts are often sold as fake followers, *i.e.*, pay-per-follower services exist that allow Twitter users to purchase followers that will artificially inflate their apparent visibility and popularity. Fake accounts are often inactive or minimally active, they engage predominantly in following other users, and occasionally on rebroadcasting (retweeting) such users’ posts, or (sometimes automatically) liking the content produced by the users they are paid to follow.

- **Bots (or automated) accounts**: This category distinguishes user accounts based on the way their account is controlled. A bot account is an account that is exclusively or predominantly controlled by
software, *i.e.*, more or less complex computer scripts that automate the activity and behavior of the account with the goal to mimic more or less faithfully, the way a human user would behave on the platform. Some bots are disproportionately more active than most human users, since their activity is automated. Simple bots have limited functionalities such as automatically retweeting or liking certain content, whereas bots that rely on more advanced artificial intelligence, are capable of creating content that is often indistinguishable from human-generated content. Bots can also mimic or “clone” the behavior of target human users by simply imitating their temporal and network activities to simulate the patterns of engagement of real human users on Twitter [1].

- **Malicious (or troll) accounts**: This category distinguishes users accounts based on their intent. A common practice in the literature is to consider malicious users to be engaged in inauthentic activities when they are purposely initiating or following an agenda, *i.e.*, political propaganda, conspiratorial campaigns, disinformation, influence operations, toxic attitudes, etc., aimed to artificially steer or distort a Twitter conversation, and/or to deceive legitimate users therein involved. Malicious accounts are often engaged in campaigns that can be organized by state-sponsored foreign governments, influence groups, and occasionally by lone wolves. They are also often engaged in coordinated activity carried out in orchestration with many other such accounts. Malicious accounts can be automated, but are often controlled by human operators, sometimes referred to as trolls (*e.g.*, Russian Internet Research Agency troll factory).

- **Compromised accounts**: This category comprises accounts that were originally created by legitimate users but have since been hacked or compromised, and their control has been passed to somebody else. The original owner might or may not have access to such accounts, hence might be unwillingly and unknowingly complicit in the misuse of their accounts. Often these accounts are dormant and the owner may not notice. Compromised accounts are often sold on black markets, *e.g.*, the Dark Web. They are also often repurposed into botnets.

**B. Creation of spam accounts**

To understand how spam accounts behave, and how to detect and curb them, it is useful to first understand how such accounts are created and operated in the first place.

On the taxonomy above, special attention should be paid to bot accounts, which operate in a manner substantially different from the other four categories of spam accounts, *i.e.*, by means of software automation. The other categories of accounts are usually created and operated manually (*i.e.*, by human users) and their detection warrants a separate discussion, so I will focus on bots.

1. **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) and its evolution or alternatives, to generate human-like content. This framework allows the creation of bots that generate genuine-looking short text on platforms like Twitter, requiring revised strategies 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, Heroku, or behind proxy services, *e.g.*, residential IP proxies (RESIP) such as Brigh Data, SmartProxy, ProxyRack, or OxyLabs (Chiapponi, *et al.*, 2022) in an attempt to disguise their actions.

2. **Open source Twitter bots**
A recent survey discusses readily available Twitter bot-making tools (Daniel and 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 and Millimaggi, 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 bot-making tools and enumerated the most common capabilities. Typical Twitter bots’ automated features include:

- Searching users, trends, and keywords;
- Following users, trends, and keywords;
- Liking content, based on users, trends, and keywords;
- Tweeting and mentioning users and keywords, based on AI-generated content, fixed-templated content, or cloned-content from other users;
- Retweeting users and trending content, and mass tweeting;
- Talking to (replying) other users, based on AI-generated content, templated content, or cloned-content from other users; and finally,
- Pausing activity to mimic human circadian cycles and bursty behaviors, as well as to satisfy API constraints, and to avoid suspension.

According to Daniel and 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.

C. Detection of spam accounts

Most of the work to develop models to automatically detect spam accounts has focused on the detection of bots. This is in part due to the fact that the definition of spam oftentimes entails a normative judgment of the intent behind an account’s activity and behavior (i.e., inferring whether the entity behind the account is attempting to gain from the spamming activity). Conversely, the definition of bot can be operationally defined unequivocally as ‘software-operated’ or ‘automated’ account (as opposed to human-operated account) — irrespective of the account’s intent (hence, bots are not to be considered inherently ‘bad’ or ill-intended, although research shows that the vast majority of bot accounts attempt to disguise or hide their automated nature (Ferrara, et al., 2016).

1. How to detect bots

The problem detection and suspension of bots on social media platforms like Twitter has been well studied for over a decade (Thomas, et al., 2011; Pierri, et al., 2022; Boshmaf, et al., 2013; Freitas, et al., 2015). Some popular bot detection techniques based on machine learning have been pioneered by groups at Indiana University, University of Southern California, and University of Maryland, in the context of the SMISC (Social Media in Strategic Communication) program sponsored by DARPA (U.S. Defense Advanced Research Projects Agency) [2]: In 2015, the DARPA SMISC program organized a DARPA Challenge aimed at detecting bots used within anti-vaccination campaigns on Twitter (Subrahmanian, et al., 2016). More traditional cybersecurity techniques led to the discovery of large bot networks (botnets) on Twitter by various research groups (Boshmaf, et al. 2011; Abokhodair, et al., 2015; Echeverria and Zhou, 2017).

2. Tools for bot detection

The literature on bot detection is extensive. In 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; and, (3) machine learning methods based on the identification of highly-predictive features that discriminate between bots and humans. Undoubtedly, other taxonomies that propose alternative or complementary criteria also exist.
Twitter spam and false accounts prevalence, detection, and characterization: A survey (Gorwa and Guilbeault, 2020; Woolley, 2016).

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

- **Botometer [3]**, a bot-detection tool developed at Indiana University (https://botometer.osome.iu.edu/faq);
- **BotometerLite [4]**, a lightweight bot-detection model developed by the same team of Botometer (https://botometer.osome.iu.edu/botometerlite);
- **Bot Sentinel [5]**, a public dataset of accounts classified as bots or not via machine learning (https://botsentinel.com);
- **Tweet Bot or Not [6]**, an R package for classifying Twitter accounts as bot or not, (https://github.com/mkearney/tweetbotornot);
- **DeBot [7]**, a real-time bot detection system by University of New Mexico’s researchers (https://www.cs.unm.edu/~chavoshi/debot).

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

- Deep learning (Kudugunta and Ferrara, 2018);
- Anomaly detection (Minnich, *et al.*, 2017; Gilani, *et al.*, 2017; Echeverria, *et al.*, 2018);
- Graph-based techniques (Feng, *et al.*, 2022; Feng, *et al.*, 2021);
- Time series analysis (Chavoshi, *et al.*, 2016; Stukal, *et al.*, 2017; Pozzana and Ferrara, 2020).

Finally, numerous recent studies have used mixed approaches that combine some of the ideas above or build upon state-of-the-art advances in artificial intelligence techniques. This includes (i) use of long-short term memory (LSTM) neural networks (Wei and Nguyen, 2019); (ii) generative adversarial networks (GANs) (Wu, *et al.*, 2020); (iii) computational stylometrics (Cardaioli, *et al.*, 2021); (iv) random walk based detection (Fukuda, *et al.*, 2022); (v) embedding based techniques (Kumar, *et al.*, 2021); (vi) transformer based techniques (Martin-Gutierrez, *et al.*, 2021); (vii) learning automata (Rout, *et al.*, 2020); (viii) attention mechanisms (Shen, *et al.*, 2022); and, (ix) ensembles models (Shukla, *et al.*, 2021).

3. **Botometer**

The first tool to become publicly available for computational bot detection was Botometer, originally released as “BotOrNot” in 2014. This tool was used to complement bot analysis in this study, providing some of the state of the art techniques in computational bot detection. Botometer’s version 1 was accompanied by a peer reviewed paper that described its methodology (Davis, *et al.*, 2016).

Botometer continually undergoes improvements and the Indiana University team in charge of this tool periodically releases new and improved versions to the public. Figure 1 illustrates a timeline of Botometer versions published since 2014 to date.
Botometer’s authors define it as “a machine learning tool for bot detection on Twitter” (Yang, et al., 2022). The current version of Botometer is version 4 (v4 ESC, i.e., Ensemble of specialized classifiers). Adopting Botometer can have several advantages, as described in Yang, et al. (2022):

First, it is well maintained and has been serving the community for the past seven years without major outages. It has also been routinely upgraded to stay accurate and relevant. Second, Botometer is easily accessible through both a Web interface and an application programming interface (API). Anyone with a Twitter account can use the Web version for free; researchers with Twitter developer accounts can use the API endpoints to analyze large-scale datasets. The API has a nominal fee for heavy use, which discourages abuse and partially offsets infrastructure and maintenance costs. Third, Botometer is quite popular. It handles around a quarter million daily queries — over half a billion in total since its inception. Finally, Botometer has been extensively validated in the field. Many researchers have applied Botometer in their studies to directly investigate social bots and their impact (Keller and Klinger, 2019; Broniatowski, et al., 2018; Allem, et al., 2020; Fan, et al., 2020), or to distinguish human accounts and bot-like accounts in order to better address their questions of interest (Vosoughi, et al., 2018; Grinberg, et al., 2019; Bovet and Makse, 2019).

4. How Botometer works

The latest version of Botometer (v4 ESC) is described as (Yang, et al., 2022):

- Under the hood, Botometer is a supervised machine learning classifier that distinguishes bot-like and human-like accounts based on their features (i.e., characteristics). Unsupervised methods have also been proposed in the literature (Chavoshi, et al., 2016; Echeverria and Zhou, 2017), but they only allow for the
 detection of specific, predefined behaviors. Therefore they are not suitable to build a general detection tool.

- Botometer considers over 1,000 features that can be categorized into six classes: user profile, friends, network, temporal, content and language, and sentiment (Varol, et al., 2017). For example, the user profile category includes features such as the length of the screen name, whether the account uses the default profile picture and background, the age of the account, etc. The content and language category consists of features such as the number of verbs, nouns, and adjectives in the tweets. For a given account, these features are extracted and encoded as numbers. This way the account can be represented by a vector of feature numbers, enabling machine learning classifiers to process the information.

5. Accuracy and robustness of bot detection tools

The performance and accuracy of different bot detection models has been tested extensively, as well as independently verified by other researchers besides the original developers (Echeverria, et al., 2018; Alipour, et al., 2022).

a) LOBO: A system to evaluate the generalizability of bot detection tools

In 2018, Echeverria, et al. (2018) published the first independent and rigorous evaluation of the accuracy of Botometer. This test was based on version 2 of Botometer (to date, Botometer reached version 4; see Yang, et al., 2022). They used a plethora of datasets and variations of bot detection, and reached the following conclusions:

This can be seen in Table 1, which shows the average botometer scores for each of the bot classes. To evaluate whether this tool would be able to predict the dataset, we provide another metric which is the percentage of the queried accounts that receive a botometer score over 0.5. Because of rate limiting, we only collected botometer scores for up to 1,000 randomly selected accounts belonging to each of the classes. We can see that many of these bot classes are overwhelmingly classified as users, for example, only 2.75% of the Bursty Bots are classified as bots, and less than 10% of DeBot bots (dataset C) are classified as bots. Both of them with average botometer scores less than 0.1, indicating that they are “very likely” to be users. Different bot classes will achieve reasonable and even perfect performance on this bot detection task, but variability between bot classes is very clear.
Table 1: Performance of Botometer (v2) on difference data sets; BTS (%) is the percentage of correctly detected bots. BTS (Avg) is the average bot score. From Echeverria, et al. (2018).

| ID | Name                          | BTS(%) | BTS(Avg) | Size   |
|----|-------------------------------|--------|----------|--------|
| A  | Star Wars Bots                | —      | —        | 357,000|
| B  | Bursty Bots                   | 2.75   | 0.04     | 500,000|
| C  | DeBot                         | 7.67   | 0.09     | 700,000|
| D  | Fake followers                | 96.79  | 0.90     | 721    |
| E  | Social spambots #1            | 92.35  | 0.85     | 551    |
| F  | Social spambots #2            | 99.37  | 0.96     | 3,320  |
| G  | Social spambots #3            | 94.10  | 0.87     | 458    |
| H  | Traditional spambots #1       | 98.28  | 0.93     | 872    |
| I  | Traditional spambots #2       | 100.00 | 0.85     | 1      |
| J  | Traditional spambots #3       | 66.08  | 0.60     | 283    |
| K  | Traditional spambots #4       | 97.81  | 0.90     | 977    |
| L  | ~ 1k followers                | 20.89  | 0.21     | 387    |
| M  | ~ 100K followers              | 10.90  | 0.13     | 534    |
| N  | ~ 1M followers                | 1.32   | 0.02     | 229    |
| O  | ~ 10M followers               | 0.00   | 0.00     | 26     |
| Q  | Fake followers-FSF            | 100.00 | 0.96     | 33     |
| S  | Fake followers-INT            | 100.00 | 0.95     | 64     |
| T  | Fake followers-TWT            | 95.34  | 0.89     | 624    |
| V  | HoneyPot bots (Darpa)         | 27.69  | 0.30     | 2,521  |
| W  | Attack on Ben Nimmo           | 59.09  | 0.54     | 1,558  |
| X  | Attack on Brian Krebs         | 83.05  | 0.78     | 728    |

The general performance of Botometer in this independent analysis is very high, reaching on average over 90 percent accuracy on most bot datasets. Nevertheless, this study highlighted one limitation of the original versions of Botometer, which was a lower performance in generalizing to unseen bot types (e.g., DeBot dataset). This particular problem was addressed by Botometer v3, using a technique illustrated in Yang, et al. (2020).

b) State-of-the-art bot detection benchmark
Alipour, *et al.* (2022) independently tested the accuracy of two state-of-the-art (SOTA) bot detection tools, Botometer and TweetBotOrNot. They used over a dozen datasets, including the DeBot data (Echeverria, *et al.*, 2018), and tested various generalizability and robustness aspects of bot detection. Their analysis reached several conclusions, summarized below.

6. **Botometer results**

Here are two excerpts from Alipour, *et al.* (2022) on Botometer’s performance:

- As shown in Table 2A and Table 2B the Botometer performs well in identifying bot and human accounts on the data on which it is trained, especially for the threshold of 0.5.
- In addition to the datasets that the Botometer was trained on, we used other datasets that Juan Echeverria et al. used in their research in 2018 (Echeverria, *et al.*, 2018). We call these datasets Juan datasets. As illustrated in Table 2, Botometer still performed well except on the DeBot dataset. DeBot dataset was collected by Juan Echeverria, *et al.* by querying the accounts detected as bots by the DeBot Bot Detection Service. It should be noted here that Botometer performed better on this dataset in our replication study than the results reported by Encheverra, *et al.* when they performed the original study (Echeverria, *et al.*, 2018) on the same datasets. This might be because of the improvements of the Botometer’s detection module.

| Dataset                  | >=0.5 | >=0.75 | >=0.8 |
|--------------------------|--------|--------|--------|
| TwiBot-20                | 30%    | 18%    | 16%    |
| Cresci-rtbust-2019       | 80%    | 70%    | 64%    |
| Botometer-feedback-2019 | 82%    | 64%    | 59%    |
| Kaiser                   | 87%    | 79%    | 76%    |
| Vendor-purchased-2019   | 80%    | 41%    | 33%    |
| Pronbots-2019            | 90%    | 84%    | 82%    |
| Astroturf                | 74%    | 63%    | 58%    |
| Political-bots-2019     | 100%   | 46%    | 38%    |
| Botwiki-2019            | 99%    | 97%    | 96%    |

**Table 2:** Percentage of (A) bots and (B) humans detected by Botometer according to training data set and threshold. From Alipour, *et al.* (2022).

7. **TweetBotOrNot results**
Here are two excerpts from Alipour, et al. (2022) on TweetBotOrNot’s performance:

- **Table 3A** and **Table 3B** depict Tweetbotornot’s performance on both Botometer trained and Juan datasets, respectively. These results show that Tweetbotornot does not perform as well as Botometer on these datasets.

- Among the datasets we examined, there is only one, namely Kaiser dataset, on which Botometer and Tweetbotornot detect a relatively similar number of accounts. This similarity in the performance led us to examine whether the bots detected by the Botometer are the same as the bots detected by the TweetBotOrNot. Echeverria, et al. (2018) showed that TweetBotOrNot and Botometer had very little overlap in the bots identified in political data, especially those that contained certain hashtags. However, in our study of different datasets regardless of the content of the tweets, TweetBotOrNot and Botometer overlap a lot in the detected bots. In Table 8 [8], we show the number of records in each dataset, the number of bots correctly detected by Botometer at a threshold of ≥0.5, the number of bots detected by TweetBotOrNot with a bot-probability of >0.5, and the number of common bots (cb) that both systems detected based on the same threshold and probability. According to these results, most of the bots detected by both systems are the same. In general, there is a high overlap in the number of bots detected by TweetBotOrNot with the ones Botometer detected. Most of the time TweetBotOrNot detected bots are a subset of Botometer detect bots, column cb in Table 8.

| Dataset                  | >0.5 | >0.6 | >0.7 | >0.8 |
|--------------------------|------|------|------|------|
| TwiBot-20                | 4%   | 4%   | 3%   | 3%   |
| Cresci-rbust-2019        | 2%   | 2%   | 2%   | 2%   |
| Botometer-feedback-2019 | 30%  | 30%  | 29%  | 29%  |
| Kaiser                   | 69%  | 68%  | 68%  | 67%  |
| Vendor-purchased-2019    | 4%   | 3%   | 3%   | 3%   |
| Pronbots-2019            | 2%   | 2%   | 2%   | 1%   |
| Astro turf               | 1%   | 1%   | 1%   | 1%   |
| Political-bots-2019     | 31%  | 31%  | 31%  | 23%  |
| Botwiki-2019            | 99%  | 99%  | 99%  | 99%  |

| Dataset                  | >0.5 | >0.6 | >0.7 | >0.8 |
|--------------------------|------|------|------|------|
| BurstyUsers              | 2%   | 2%   | 2%   | 2%   |
| Journalist attack bots (Ben Nimmo) | 2%   | 2%   | 2%   | 1%   |
| Journalist attack bots (Brian) | 2%   | 2%   | 1%   | 1%   |
| DeBot                    | 4%   | 4%   | 4%   | 3%   |

**Table 3**: Percentage of bots detected by TweetBotOrNot according to Botometer (A) and Juan (B) training data set and threshold. From Alipour, et al. (2022).
8. Overall conclusions of the 2022 SOTA bot detection benchmark

Alipour, et al. (2022) concluded:

Our results show that the performances of SOTA systems in detecting bot accounts are better than detecting human accounts. Moreover, the performances of SOTA systems do have a substantial overlap in terms of the bots as well as the human accounts they are able to detect. However, the performances of SOTA systems decrease when they are evaluated on totally different datasets than the one used to train them. This decrease is bigger in the performance of Tweetbotornot compared to the performance of Botometer.

9. On the accuracy and robustness of computational bot detection models

Based Echeverria, et al. (2018) and Alipour, et al. (2022), computational bot detection (i.e., bot detection based on machine learning models) is a valid scientific tool; Botometer yields the most accurate classification among the assessed, publicly available tools. As such, Botometer represents the state of the art (SOTA) in computational bot detection to date. The accuracy of the tool is moderately to very high (above 80 percent for almost all data and tasks, with peaks above 95 percent for simpler classification tasks). Botometer is often the go-to tool for tasks that requires scalability (i.e., processing of millions of accounts in limited time), albeit manual double-checking and fine-tuning of models is always advised.

D. Prevalence of spam accounts

1. Timeline of recent reports (Reverse chronological ordering)

Various studies have been published reporting on the number of spam accounts, typically focusing on promotional, fake accounts, and bots. Each study employs different methodologies, sample sizes, datasets, etc. In the following, we summarize some of the more transparent reports that have appeared in the last five years.

In June 2022, we analyzed 5.5M users, contributing over 207M retweets and 23M replies, active in the context of broad COVID-19 discussion (Chang and Ferrara, 2022). Here are some highlights of that study:

We randomly sampled 200 accounts (100 from each account) which were validated by three researchers. This yielded an average of 88 percent agreement with the Botometer on average.

The distribution of Botscores within our dataset is visualized in Figure 2, which shows a natural bimodal break between 0.5. Hence, we label users with botscores above 0.5 as bots.

In absolute terms, liberal bots dominate the amount of retweets, taking up to 20 percent of all retweets. In contrast, conservative bots only amount to around 4.5 percent of the total human retweets.
Luceri, et al. (2019) published a study of the prevalence of spam accounts in the context of the 2018 U.S. midterms. In summary, they found:

We classified 21.1 percent of the accounts as bots, which in turn generated 30.6 percent of the tweets in our data set. Overall, Botometer did not return a score for 35,029 users that corresponds to 3.5 percent of the accounts.

99.4 percent of these accounts were suspended by Twitter, whereas the remaining percentage of users protected their tweets turning on the privacy settings of their accounts.

Stella, et al. (2018) measured the number of bots involved in conversations about the 2017 Catalan referendum; in summary:

We monitored the Twitter stream and collected data by using the Twitter Search application programming interface (API), from September 22, 2017, to past the election day, on October 3, 2017 [...] This procedure yielded a large dataset containing 3.6 million unique tweets, posted by 523,000 unique users.

Off-the-shelf learning models were trained on multiple historical ground truth datasets and achieved high detection accuracy (>90 percent) on cross-validation benchmarks. Logistic regression (LR), our reference model for this study,
was selected for its best trade-off between scalability and accuracy: The model is very precise at detecting human accounts precision rate (PR) 98 percent, compared with bot accounts (PR: 92 percent), while detecting nearly all existing bots recall rate (RR) 99 percent, compared with human users retrieval (RR: 88 percent).

We discovered that bots generated specific content with negative connotation that targeted the most influential individuals among the group of Independentists (i.e., Catalan independence supporters).

Bots produced 23.6 percent of the total number of posts during the event.

Wojcik, et al. (2018) examined the prevalence of bots-created tweeted links to news sites. Again, in summary:

Of all tweeted links to popular Web sites, 66 percent are shared by accounts with characteristics common among automated “bots,” rather than human users.

Among popular news and current event Web sites, 66 percent of tweeted links are made by suspected bots — identical to the overall average. The share of bot-created tweeted links is even higher among certain kinds of news sites. For example, an estimated 89 percent of tweeted links to popular aggregation sites that compile stories from around the Web are posted by bots.

A relatively small number of highly active bots are responsible for a significant share of links to prominent news and media sites. This analysis finds that the 500 most-active suspected bot accounts are responsible for 22 percent of the tweeted links to popular news and current events sites over the period in which this study was conducted. By comparison, the 500 most-active human users are responsible for a much smaller share (an estimated 6 percent) of tweeted links to these outlets.

The study does not find evidence that automated accounts currently have a liberal or conservative “political bias” in their overall link-sharing behavior. This emerges from an analysis of the subset of news sites that contain politically oriented material. Suspected bots share roughly 41 percent of links to political sites shared primarily by liberals and 44 percent of links to political sites shared primarily by conservatives — a difference that is not statistically significant. By contrast, suspected bots share 57 percent to 66 percent of links from news and current events sites shared primarily by an ideologically mixed or centrist human audience.

Figure 3 illustrates the overwhelming prevalence of bot-produced links to popular news.
Figure 3: Analysis of 379,841 tweeted links to 925 news and current events Web sites over the time period 27 July–11 September 2017. From Wojcik, et al. (2018).

I published an analysis on disinformation and bot operations ahead of the 2017 French presidential election (Ferrara, 2017). That study noted that:

By using a Logistic Regression model trained on the 10 metadata and activity features described above, we obtain very accurate user classification on the cross-validated tests (92 percent accuracy, 89 percent AUC-ROC). We adopted this model to detect all bots and separate them from human users in our dataset.

Out of 99,378 users involved in MacronLeaks, our model classified 18,324 of them as social bots, and the remainder of 81,054 as human users. The fraction of social bots amounts for about 18 percent of the total users involved in the campaign.

Remarkably, among the top 15 social bots detected by our framework, four accounts have been so far deleted, seven have been suspended, and two have been quarantined by Twitter. Two accounts are still active, and they may be the result of misclassification. Overall, this example of manual verification suggests that we obtained 13 correct bots out of 15 detected, an accuracy of nearly 87 percent.

Varol, et al. (2017) reported on the prevalence of bots on Twitter using a combination of machine learning bot detection and human validation. They remarked, in part, that:

We also estimated the fraction of bots in the active English-
speaking population on Twitter. We classified nearly 14M accounts using our system and inferred the optimal threshold scores that separate human and bot accounts for several models with different mixes of simple and sophisticated bots. Training data have an important effect on classifier sensitivity. Our estimates for the bot population range between 9 percent and 15 percent.

Using mixtures with different ratios of accounts from the manually annotated and honeypot datasets, we obtain an accuracy ranging between 0.90 and 0.94 AUC.

Using an optimal threshold, we measured false positive and false negative rates at 0.15 and 0.11 respectively in our extended dataset. In these experiments, human annotation is considered as ground truth.

Bessi and Ferrara (2016) published a study of social bot interference in political discussion revolving around the 2016 U.S. presidential election. In this study, we highlight that the fraction of bots engaged in the 2016 election was approximately 14.4 percent; they were responsible for approximately 18.45 percent of content.

In Table 4, I summarize the findings of these studies to draw a comprehensive conclusion about the estimated prevalence of spam or bot accounts on Twitter:

| Study                        | Spam account prevalence | Spam content prevalence | Accuracy/Confidence | Focus             |
|------------------------------|-------------------------|-------------------------|---------------------|-------------------|
| Chang and Ferrara, 2022      | Unspecified             | 24.5% of retweets       | 76% annotator agreement | Bots              |
| Luceri, et al., 2019         | 21.1%                   | 30.6%                   | 99.4%               | Bots              |
| Stella, et al., 2018         | Unspecified             | 23.6%                   | 92–98% accuracy     | Bots              |
| Wojcik, et al. (2018)        | Unspecified             | 66% of popular content  | 86% accuracy        | Bots              |
| Ferrara, 2017                | 18.3%                   | Unspecified             | 87–92%              | Bots/trolls       |
| Varol, et al., 2017          | 9–14%                   | Unspecified             | 90–94% accuracy; 88% annotator agreement | Bots/spammers    |
The prevalence of accounts that engage in some type of spam activities, as defined by Twitter, is consistently reported between approximately 9 and 21 percent, with an average estimate at approximately 12.5 percent plus or minus 2 percent of the total Twitter accounts.

The prevalence of content that is considered spam has been measured on a broader range spanning from approximately 18 percent to 30 percent so it can be assumed to average at, or in excess of, 20 percent of the overall content appearing on Twitter.

The studies relied on a plethora of methodologies including but not limited to machine learning, human annotations, and combinations thereof, and when available, report accuracy and/or confidence typically above 80 percent and often above 90 percent. The reported margin of error (both in favor and against spam classification, \textit{i.e.}, false positives and false negatives) renders the estimates reliable and robust.

2. Honeypots and black markets for spam and false accounts

There have been recent studies that attempted to identify spam accounts using strategies not based on machine learning or crowdsourcing. The two most common alternative strategies have included the use of honeypots to attract spammers, and the scrutiny of black markets (such as the dark Web) or services that provide spam and false accounts as a service for purchase. Here are some notable examples:

Mazza, \textit{et al.}'s (2022) ReadyToAbuse study tracked tens of thousands of Twitter accounts (including compromised accounts) that were sold on the \textit{buyaccs} Web site, a marketplace for purchasing fake and compromised accounts on various social media platforms, most prominently Twitter. According to their study:

- We passively monitored a popular Twitter account merchant \textit{buyaccs.com} from June 2019 to the end of July 2021 and detected 63,358 fake accounts for sale as a result.

- We focused our analysis on the fake accounts that produced at least one tweet throughout 2020, identifying 5,457,758 tweets from 23,579 accounts.

- We uncovered four coordinated campaigns conducted by fake accounts and showed more details on their inauthentic behavior on Twitter.

In 2017, I published a study that unveiled the MacronLeaks disinformation campaign in the run-up to the 2017 French presidential election, and characterized the role that spam bots played in it (Ferrara, 2017). Besides unveiling the presence of over 18,000 bot accounts, I also documented the repurposing of at least 800 such accounts, which I found active as well in the discourse surrounding the 2016 U.S. presidential election. Investigative reports demonstrated that such accounts had been sold and purchased on black markets, specifically on the dark Web.

Lee, \textit{et al.} (2011) built a honeypot system designed to attract spammers on Twitter. They found:

- The system ran from December 30, 2009 to August 2, 2010.
During that time the social honeypots tempted 36,043 Twitter users, 5,773 (24 percent) of which followed more than one honeypot.

We found that Twitter eventually suspended the accounts of 5,562 (or 23 percent of the total polluters identified by the social honeypots). We observe that of the 5,562, the average time between the honeypot tempting the polluter and the account being suspended was 18 days.

E. Effects of spam accounts

The presence of spam and false accounts has negative effects on the quality of user experience and the general health of a given platform. A plethora of negative effects of spam accounts have been documented through various rigorous peer-reviewed research. Some studies documented illicit activities perpetrated by spam accounts via Twitter, such as scams and fraud, while others focused on deception and influence campaigns. I will document some of these studies next.

1. Scams and frauds

These forms of abuse are rampant on Twitter. Spam accounts are often implicated in various forms of scams and frauds, most prominently (a) by means of cryptocurrency Ponzi schemes and/or pump-and-dump campaigns; (b) by promoting counterfeit and/or knock-off products; (c) by linking to Web sites selling illicit or unapproved drugs; (d) by advertising advance fees scams; (e) by promoting phishing Web sites; and, (f) by linking to malware downloads. Its worth noting that Twitter sanctions most of these behaviors but appears to fail to suspend accounts that are involved in such behavior (Nizzoli, et al., 2020; Phillips and Wilder, 2020; Nghiem, et al., 2021; Eigelshoven, et al., 2021).

Mazza, et al. (2022) were able to quantify the effects of at least one of the spam campaigns that they discovered, a loss of over US$90,000 from unwitting Twitter users:

Among our results is the identification of four campaigns based on coordinated inauthentic behavior (CIB), exposing attempts to influence the political debate and aggressive spamming aimed at deceiving users or advertising counterfeit products. We also uncovered that one of these campaigns managed to steal at least $90K from deceived users. Notably, the CIB instances adopted different strategies: amplify content through retweets; impersonate reputable accounts to infiltrate ongoing conversations; leverage the hype surrounding specific events to gain visibility; use third-party content to build a network and then spam own contents.

Nizzoli, et al. (2020) charted the landscape of cryptocurrency manipulations via social media. They investigated 5.7M Twitter users responsible for 16.8M crypto-related posts. From their study:

We collected a large dataset, composed of more than 50M messages published by almost 7M users on Twitter, Telegram and Discord, over three months. We performed bot detection on Twitter accounts sharing invite links to Telegram and Discord channels, and we discovered that more than 56 percent of them were bots or suspended accounts.

We uncover the pivotal role of Twitter bots in broadcasting
invite links to deceptive Telegram and Discord channels [...] we observed that 93 percent of the invite links shared by Twitter bots point to Telegram pump-and-dump channels, shedding light on a little-known social bot activity.

We reported on 15 Twitter botnets that are responsible for the 75.4 percent of invite links to pump-and-dump channels. 

**Figure 4** shows a botnet of spam accounts all promoting the same Telegram crypto-scam account: it is evident that these accounts were part of a coordinated network; only a fraction of the accounts in this botnet were suspended by Twitter, whereas the majority of them (56.3 percent) remained active and engaged with the crypto spam campaign.

**Figure 4**: Invite link network highlighting deceptive Twitter accounts. A large portion of Twitter have a deceptive nature (56.3 percent). The typical star structures frequently correspond to botnets promoting a single channel. We found 69 botnets with a size of at least 10 elements. From Nizzoli, *et al.* (2020).
2. Stock market manipulation

Reports of suspected influence of spam accounts on stock market transactions have been circulating for over a decade.

In March 2011, Dan Mirvish, a Huffington Post contributor and media expert, linked the Internet popularity of actor Anne Hathaway to spikes in the stock value of Warren Buffett’s Berkshire-Hathaway, stating “When Anne Hathaway makes headlines, the stock for Warren Buffett’s Berkshire-Hathaway goes up [...] It looks like all the automated, robotic stock trading programs are picking up the same chatter on the Internet about “Hathaway” as the IMDb’s StarMeter, and they’re applying it to the stock market.” [9] This was later named the Hathaway effect.

Hwang, et al. (2012) studied the Hathaway effect more broadly, postulated that spam bots on social media could exacerbate the effects of stock market manipulation.

The issue became such a concern that in November 2014, the U.S. Securities and Exchange Commission (SEC) posted an alert to encourage awareness about social media stock manipulation and fraud schemes [10].

Perhaps the most egregious example of the effect of Twitter spam on the stock market, and malicious use of spam accounts to pump and dump stocks, involves a little-known tech company called Cynk.

Seeking Alpha was first to report suspicious spam activity around the SCYNK stock (see Figure 5) [11]. The New York Post later stated that “Seeking Alpha found out early on about Cynk’s otherwise inexplicable run that ‘no fewer than a dozen stock promoters’ were hyping the stock in their tweets” [12].
Forbes [13] summarized this stock market manipulation event in this way:

Last summer, a small, unknown social media company called Cynk Technology (CYNK) experienced the kind of runup that is the stuff of Wall Street legend. In less than a month, its stock price rose from six cents to nearly twenty-two dollars a share, an increase of more than 36,000 percent.

The combination of bots and HFT algorithms in a market can amount to a conspiracy of virtual dunces, with bot programs pumping out fake “news” that will prompt an HFT algorithmic program to buy or sell, which sends a trading signal to other algorithmic programs, the whole thing snowballing until either until a completely worthless company is worth $6 billion or the

Figure 5: Spam accounts pumping $CYNK on Twitter. From Seeking Alpha.
entire market spirals into a crash, wiping out billions of dollars.

The SEC eventually suspended trading on CYNK, figuring it for a pump-and-dump, but by then the damage was done and the scammers had won. CYNK is now back to trading at a penny.

An independent analysis carried out by the analyst firm Product Reviews used Botometer version 1 (BotOrNot) stating that “On a website called Bot or Not, we searched some of the Twitter accounts promoting this stock and it seems like most of these are robot accounts rather than humans.” [14]

The stock market manipulation and related events were later investigated by the SEC and the perpetrator was charged with microcap fraud [15].

Tardelli, et al. (2020) focused on characterizing financial disinformation campaigns on Twitter:

Aim to shed light on this issue by investigating the activities of large social botnets in Twitter, involved in discussions about stocks traded in the main US financial markets. We show that the largest discussion spikes are in fact caused by mass-retweeting bots.

We focus on characterizing the activity of these financial bots, finding that they are involved in speculative campaigns aimed at promoting low-value stocks by exploiting the popularity of high-value ones.

These accounts appear as untrustworthy and quite simplistic bots, likely aiming to fool automatic trading algorithms rather than human investors.

Our findings pave the way for the development of accurate detection and filtering techniques for financial spam.

3. Political disinformation

Efforts to distort political discourse via Twitter has been reported since the early 2010s (Ratkiewicz, et al., 2011; Metaxas and Mustafaraj, 2012). But on the onset of the 2016 U.S. presidential election, a rather new phenomenon was observed on Twitter, in concert with bots (Bessi and Ferrara, 2016), and hyper-partisan campaigns: the spread of false news and the coordination of disinformation campaigns (Allcott and Gentzkow, 2017; Marwick and Lewis, 2017; Mele, et al., 2017). Disinformation is the deliberate spread of false or inaccurate information — to be distinguished from misinformation, where there is no planned agenda or intent to spread inaccuracies or falsehood. The adoption of automated accounts such as bots in the context of disinformation campaigns is particularly concerning because there is the potential to reach a critical mass large enough to dominate public discourse and alter public opinion (Ferrara, et al., 2016; Woolley and Howard, 2016; Marwick and Lewis, 2017); this could steer the public’s attention away from facts and redirect it toward manufactured, planted information.

In August 2017, I published an analysis on disinformation and bot operations ahead of the 2017 French presidential election (Ferrara, 2017). The study highlighted:

We monitored the Twitter stream between 27 April and 7 May 2017 (Election Day), and collected a very large dataset containing nearly 17 million tweets related to the 2017 French
presidential election. Within it, we identified the subset related to the MacronLeaks disinformation campaign.

Out of 99,378 users involved in MacronLeaks, our model classified 18,324 of them as social bots, and the remainder of 81,054 as human users. The fraction of social bots amounts for about 18 percent of the total users involved in the campaign.

We uncovered that accounts used to support then-presidential candidate Trump before the 2016 U.S. election have been brought back from a limbo of inactivity (since November 2016) to join the MacronLeaks disinformation campaign. Such anomalous usage patterns point to the possible existence of a black market for reusable political-disinformation bots.

Here are some statistics regarding the performed activity and influence that the bot accounts accrued over the short 10 days before the election: (a) Bots posted on average 2.86 MacronLeaks-related tweets; (b) Bots obtained on average 1,382 followers (\(\sigma = 22,282\)); (c) Bots friended on average 1,058 users (\(\sigma = 12,190\)); (d) Bots favorited on average 228 tweets (\(\sigma = 924\)); and, (e) Finally, bots have been listed on average on 7.42 lists (\(\sigma = 90.3\)).

Im, et al. (2020) published the StillOutThere study, aimed at assessing whether Russian trolls were still present and active on Twitter and the extent of their continued influence operations. The authors remarked:

- develop machine learning models that predict whether a Twitter account is a Russian troll within a set of 170K control accounts; and,
- demonstrate that it is possible to use this model to find active accounts on Twitter still likely acting on behalf of the Russian state.

4. Political deception

Another form of influence exerted by spam accounts is political deception. According to Woolley (2016), Howard, et al., (2018), and Woolley and Howard (2017), the behavior of political spam accounts can be categorized relative to their intentions:

- In political discussions, manufacturing consensus, to enhance the perception of popularity or influence of an entity (political actor, party, organization, etc.);
- Bolstering opinions and voices, to amplify the platform and audience that an entity will receive;
- Cementing polarization, by increasing the inflammatory or divisive nature of an issue or agenda;
- Increasing chaos and confusion, by posting inaccurate information, disinformation, and rumors;
- Algorithmic manipulation, to trick recommendation and ranking systems used by social media platforms, and give higher visibility to certain actors, viewpoints, or campaigns.

5. Conspiracy amplification

The issue of conspiracy amplification has been studied extensively recently (Wang, et al., 2022; Sharma, et al., 2022). In Ferrara, et al. (2020), we illustrated the role that bots played in amplifying content generated by alt-right or hyper-partisan news outlets during the 2020 U.S. presidential campaigns.

Figure 6 shows that there is an unmistakable correlation between the fraction of bots engaged in the spread of a given link to a news site, and the likelihood of that the news site being hyper-partisan or conspiratorial.
The more conspiratorial or hyperpartisan a news site is, the more likely it is to see a substantial amount of bots spreading its content.

Hyper-partisan news outlets like One America News Network (OANN) and Infowars see the greatest proportion of their user base tweeting QAnon material. These two outlets also have the highest average bot scores. Left-leaning news outlets such as the New York Times and Washington Post have low bot scores and proportion of QAnon users. The proportion of users using QAnon keywords is highly correlated with the average Botometer score (correlation coefficient: 0.947) across the spectrum of left, right, and neutral outlets.

![Figure 6: Proportion of users utilizing QAnon hashtags and mean botscore for each media outlet, dot size indicating relative number of tweets. From Ferrara, et al. (2020).](image)

6. Coordinated influence campaigns

Spam and false accounts have been implicated in coordinated influence campaigns, as disclosed by Twitter as well as numerous research studies published recently. Here are some examples:

Pacheco, et al. (2020) investigated a coordinated influence campaign, later attributed to Russia, against the Syria Civil Defense (also known as the White Helmets). They reported that:

We unveil coordinated groups using automatic retweets and content duplication to promote narratives and/or accounts. The results also reveal distinct promoting strategies, ranging from the small groups sharing the exact same text repeatedly, to complex “news website factories” where dozens of accounts
Twitter spam and false accounts prevalence, detection, and characterization: A survey

synchronously spread the same news from multiple sites.

Similar results were independently published by Wilson and Starbird (2021):

Our findings reveal a network of social media platforms from which content is produced, stored, and integrated into the Twitter conversation. We highlight specific activities that sustain the strategic narratives and attempt to influence the media agenda.

Nizzoli, et al. (2021) demonstrated the presence of coordinated influence campaigns in the context of the 2019 U.K. general election:

We empirically demonstrate that coordination and automation are orthogonal concepts. Thus, our framework can complement long-studied techniques for detecting automation, manipulation, and inauthenticity.

This study demonstrated that coordinated behavior and automation (i.e., bot accounts) are two sides of the same coin (see Figure 7), and therefore different means of detection needed to be developed and employed to guarantee the detection of orchestrated influence campaigns. Solely relying on bot detection techniques won’t suffice.

---

**Figure 7:** “Correlation between coordination and use of automation, in terms of bot scores estimated by Botometer and accounts suspended by Twitter. Both indicators of possible automation appeared to be two orthogonal and largely uncorrelated with respect to coordination.” From Nizzoli, et al. (2021).
In Sharma, et al. (2021), we demonstrated that the detection of coordinated accounts implicated in orchestrated influence campaigns is achievable by means of machine learning, obtaining accurate, scalable, and reliable detection systems.

We verified the effectiveness of the proposed method and training algorithm on real-world social network data collected from Twitter related to coordinated campaigns from Russia’s Internet Research Agency targeting the 2016 U.S. presidential elections, and to identify coordinated campaigns related to the COVID-19 pandemic.

Leveraging the learned model, we find that the average influence between coordinated account pairs is the highest. On COVID-19, we found coordinated groups spreading anti-vaccination, anti-masks conspiracies that suggest the pandemic is a hoax and political scam.

7. Public health misinformation

Researchers have shown that spam and false accounts played a major role in the spread of public health misinformation. Numerous studies, both prior to the COVID-19 pandemic and in the aftermath, demonstrated that rumors, unscientific or unsubstantiated claims, unsupported or unverified medical information, and public health misinformation thrived on Twitter in part, or largely due to the activity of spam accounts. For example:

In the midst of the Ebola crisis, I portrayed the landscape of Twitter discussion on it (Ferrara, 2015). We postulated that:

Figure 8 shows the spread of concern and fear-rich content during the same interval of time (some example tweets are reported in Table 5). Positioning, size, and colors again represent the prominence of the accounts involved in the discussion and different Twitter topical communities. We can observe how panic and fear spread virally, reaching large audiences at great diffusion speed: the exposure to content that leverage human fears obfuscates our judgment and therefore our ability to discriminate between factual truths and exaggerated news. In turn, this fosters the spreading of misinformation, rumors, and unverified news, potentially creating panic in the population, and the generation of further, more negative content.

We also proposed a strategy to tackle this issue: “This self-reinforced mechanism can be interrupted by the implementation of ad-hoc intervention campaigns, designed to be effective on specific targets, based on their characteristics and susceptibility.”

| Time | Content |
|------|---------|
|      |         |
| Date/Time     | Message                                                                 |
|--------------|--------------------------------------------------------------------------|
| 2014-10-17   | @FoxNews WHY WON'T YOU REPORT THE CRUISE SHIP IS BEING DENIED ENTRY INTO BELIZE? CRUCIAL FOR CLOSING OUR BORDERS! |
| 2014-10-17   | @FoxNews Those hc workers are totally wacky. Y today do they decide to tell some1 they handled specimens from Ebola pt, esp on ship&not sick? |
| 2014-10-17   | @NBCNews Many countries are refusing entry from flights from Ebola nations. Why aren’t you asking the WH about their crazy policy? |
| 2014-10-17   | @CNN Why two doctors treated for EBOLA & INFECTED no one else including CARE GIVERS but it’s not the same for TX? Who screwed up? |
| 2014-10-17   | @FoxNews Are they paying these potential Ebola victims money to get on ships and planes? I am beginning to really wonder |
| 2014-10-17   | OMG! He cld b infected “@NBCNews: Who is man wearing plain clothes during an #Ebola patient’s transfer?” via @NBCDFW |
Figure 8: The spreading of fear-rich content across different Twitter communities (retweets and mentions). From Ferrara (2015).

Broniatowski, et al. (2018) examined how Twitter bots and trolls affected online information on health.

Broniatowski, et al. compared bots’ to average users’ rates of vaccine-relevant messages, which we collected online from July 2014 through September 2017. We estimated the likelihood that users were bots, comparing proportions of polarized and antivaccine tweets across user types. We conducted a content
They discovered statistically significant evidence that “Compared with average users, Russian trolls, sophisticated bots, and ‘content polluters’ tweeted about vaccination at higher rates. Whereas content polluters posted more antivaccine content, Russian trolls amplified both sides. Unidentifiable accounts were more polarized and antivaccine. Analysis of the Russian troll hashtag showed that its messages were more political and divisive.”

Additionally “Malicious online behavior varies by account type. Russian trolls and sophisticated bots promote both pro- and antivaccination narratives. This behavior is consistent with a strategy of promoting political discord. Bots and trolls frequently retweet or modify content from human users.”

This study demonstrated that public health misinformation campaigns (later attributed to Russia) that leveraged spam accounts, bot accounts, sophisticated bots, and trolls, disproportionately engaged with antivaccine discourse on Twitter (Figure 9). Second, such bots and trolls aimed to sow discord by amplifying both pro- and anti-vax arguments causing confusion and disagreement among other users.

**Figure 9:** Bots’ likelihood of tweeting about vaccines compared with average Twitter users: 14 July 2014–26 September 2017. From Broniatowski, *et al.* (2018).

Since early 2020, with the COVID-19 pandemic, a plethora of studies investigated spam, false, and bot accounts, focusing on their role and effects contributing to the spread of public health misinformation related to COVID-19. The evidence has been overwhelmingly in support of the notion that spam and false accounts have had nefarious effects on the spread of misinformation about COVID-19. A recent survey by
the World Health Organization (WHO) attempted to summarize some of these findings (Gabarron, et al., 2021). Numerous other studies have been published that provide further evidence of COVID-related public health misinformation on Twitter (Chen, et al., 2021; Muric, et al., 2021; Rao, et al., 2021; Jiang, et al., 2021; Chen, et al., 2022).

8. Crisis response disruption

Researchers have shown that Twitter has acted as a responsive social sensor in times of crises, reflecting the information-seeking efforts of those involved in different kinds of emergencies or natural disasters, such as earthquakes, tsunamis, fires, hurricanes, and floods (Sasaki, et al., 2010; Yates and Paquette, 2011; Yin, et al., 2015).

However, what happens if the conversation in the context of an emergency is polluted by spam? Various studies have investigated the effect of spam and false accounts in the context of crises.

Gupta, et al. (2013) were among the first to study the effect of false rumors spreading after the Boston Marathon bombing of 2012. They reported:

We analyzed one such media i.e., Twitter, for content generated during the event of Boston Marathon Blasts, that occurred on April, 15th, 2013. A lot of fake content and malicious profiles originated on Twitter network during this event.

Our results showed that 29 percent of the most viral content on Twitter, during the Boston crisis were rumors and fake content; while 51 percent was generic opinions and comments; and rest was true information.

Figure 10 illustrates the coordinated strategies suspected accounts employed to spread rumors or spam in the context of crises.

Figure 10: “Network of suspended accounts (retweets/replies/mentions) created during the Boston blasts. We see four different forms of interactions amongst the suspected profiles (left to right): Single Links, Closed Community, Star Topology, and Self Loops.” From Gupta, et al. (2013).
Zhao, *et al.* (2015) published a study of rumors on Twitter. They developed a technique to detect clusters of Twitter posts whose topic is a disputed factual claim. They found that:

> On a typical day of Twitter, about a third of the top 50 clusters were judged to be rumors, a high enough precision that human analysts might be willing to sift through them.

One week after the Boston bombing, the official Twitter account of the Associated Press (AP) was hacked. The hacked account sent out a tweet about two explosions in the White House and the President being injured and within 60 seconds after the hacked AP account sent out the rumor about explosions in the White House, there were already multiple users enquiring about the truth of the rumor (Figure 11).

**Remarks:** this study is important because first, it quantifies in ⅓ the amount of rumor clusters among Twitter discussions revolving around a crisis, an alarmingly large amount of spam or false content. Second, it shows evidence that *post hoc* mitigation strategies based on suspension are ineffective since rumors spread virally within minutes.

**Figure 11:** False tweets about an alleged explosion in the White House, a hack later attributed to the Syrian Electronic Army. Within seconds, the rumor went viral. From Zhao, *et al.* (2015).

9. *Extremism, radical propaganda, and recruitment*

Berger and Morgan (2015) documented the myriads of efforts and associated strategies that ISIS and other extremist groups adopted on Twitter. The authors observed that:

> “ISIS uses several practices designed to amplify its apparent support on Twitter, including ‘bots’ (computer software that creates activity on a social media account in the absence of a human user) and spam (purchased tweets promoting ISIS content). We were able to compensate for this, although not perfectly.”

> “In the overall Census Dataset, around 400 nonclient apps were
Conclusions and recommendations

Why is Twitter the emblematic case to study social media manipulation and abuse? In this survey, I aimed to provide a non-comprehensive summary of the literature that appeared over the last decade that tackles the problem of abuse and manipulation of social media, with a focus on spam and false accounts on Twitter. I argued that dealing with this problem is of paramount importance and that Twitter represents the most emblematic case as it suffered from many forms of abuse that have had major societal implications.

In reviewing the research, I highlighted how Twitter has suffered from the most prototypical forms of abuse. Whether it was the study of the spread of misinformation and conspiracies, the detection of financial scams and frauds aimed to depauperate its users, the unveiling of state-sponsored influence campaigns, or toxic and hateful speech, Twitter has been the go-to choice for researchers and by far the most studied platform in the current social media landscape. Although naturally, Twitter is not the sole platform to suffer from these problems, I argued that it can serve as a testbed to understand and combat these problems, which can lead to developing tools and techniques that can then be applied to other mainstream platforms.

In the following, I provide a set of recommendations that will hopefully pave the way for an agenda of future research, cognizant policy- and decision-making, and an informed public.

Technological solutions

Fostering data access. Arguably, at least in part due to the ease of accessing its data, Twitter has been the subject of much academic (and beyond) scrutiny (Williams, et al., 2013). Albeit this might on the surface appear to be detrimental to the public-facing image of the platform, I argue that instead Twitter has been at the forefront in terms of practices fostering data access to the research community at large. The ability to access in a systematic way, both in real-time and via search, massive droves of data via the Twitter APIs has enabled thousands of studies, which in turn allowed us to garner a better understanding not only of the problems that Twitter has been suffering from, but also of their implications, and some possible solutions. In this regard, open data practices similar to what Twitter has been doing for years should be cherished, and I for one would welcome more platforms embracing the same open data policies.

The need for transparent algorithmic practices. Data access is a necessary but not sufficient condition to enable the study and understanding of social media manipulation and abuse. Platforms employ increasingly complex algorithms to curate content, rank news feeds, recommend interactions, offer personalized information, serve targeted ads, etc. The need for transparency grows with the complexity of solutions, in particular for transparent practices enabling algorithmic auditing (Venkatadri, et al., 2019; Raji, et al., 2020). Twitter, much like many other mainstream platforms (Ali, et al., 2019; Ali, et al., 2021), suffers from various forms of algorithmic biases (Bartley, et al., 2021). These issues can possibly be exacerbated by the fact that platforms continually operate as massive-scale experiments, where A/B testing of different functionalities leads to different subsets of users being served slightly different versions of the platform, which can lead to new, unforeseen problems only affecting certain population subgroups (Hunter and Evans, 2016). These issues are incredibly hard to study from an outsider’s perspective, without a direct pathway to transparently access and audit the algorithms and the experiments carried out by the platforms.
A place for artificial intelligence and machine learning. The artificial intelligence/machine learning (AI/ML) research community has made giant leaps in the development of accurate and scalable techniques to identify different forms of platform abuse, including false and spam accounts. Hundreds, maybe thousands of peer-reviewed papers, are published each year that propose a variety of approaches to leverage advances in AI/ML to tackle complex problems such as detecting spam, compromised accounts, state-sponsored troll farms, bot and automated accounts, influence campaigns, extremist propaganda, cryptocurrency manipulation, and so on. Many of these techniques are also tested and evaluated by third-party independent scrutiny, and their limitations and strengths are well investigated. Some of these approaches are provided as services (e.g., as publicly accessible APIs), and some are continuously maintained and even improved over time. These are valid rationales for considering AI/ML solutions to tackle complex problems at the scale that some of these massive platforms like Twitter operate.

An online “Real ID” and other identity verification systems. Although AI/ML might provide significant help in mitigating some issues, arguably some challenges in this space are harder than others. For example, for platforms that do not impose a strict enforcement (such as Facebook) on an account’s identity matching an individual’s identity, guaranteeing the integrity and authenticity of user accounts in other ways is paramount. This is in part because research has systematically found that, hiding behind the anonymity veil, users tend to misbehave or act in anti-social manners significantly more than when their identity is known. Anonymity comes with a high price, to quote Davenport (2002) who stated that “by allowing anonymous communication we actually risk an incremental breakdown of the fabric of our society”. I recently postulated that to certify user identity and information veracity we might rely on blockchain technologies (Ferrara, 2019). Since then, some approaches inspired by the same principles, such as humanID [16], have been proposed to verify and certify users’ identity while keeping them distinctly separated from publicly visible account information on a platform.

Social and policy solutions

The urge for responsible policy-making. Technology-based solutions might help make a dent in the problems above but are bound to be insufficient, and history shows they are likely to fail if not backed up by sensible policy-making. Twitter has been under public scrutiny and subject to multiple kinds of investigations, as did other mainstream platforms such as YouTube and Facebook. But the expectation that social media companies might be able to implement magic bullet solutions to complex problems like abuse and manipulation is not bound in reality. Policy-makers must enact proposals that are sensible and rooted in scientific consensus, and ascertain what is feasible to expect in terms of technological solutions, and what instead requires new legislature and new policies. This is where corporations, researchers, and law- and policy-makers need to cooperate for the greater good. One successful example of such convergence revolves around the problem of regulating social media bots: the U.S. Senate investigation into the 2016 U.S. presidential election helped expose the now well-known foreign-sponsored efforts led by Russia, which revealed a need for legislating bots. The California B.O.T. Act (“Bolstering Online Transparency”; see Stricke [2020]) and the U.S. Senate bill “Bot Disclosure and Accountability Act of 2019” [17], have been since formulated, with the former becoming operational in July 2019, to regulate what activities automated accounts can and cannot engage with in the context of online political discussion.

Academic and corporate social responsibility. Governments, however, do not have the technical and scientific capabilities and knowhow to face these big challenges with top down approaches. A need arises for socially responsible industry standards to tackle these problems. Without close collaborations between researchers in academia, industry, and third-party practitioners, the hope is faint to come full circle and propose socio-technical solutions that can work in combating the problem of social media spam and abuse. A major obstacle to success is the fact that some abuse practices are not necessarily technologically detectable (e.g., requests to retweet, trading followers, directed-message persuasion, etc.): an equitable algorithmic and socially responsible information ecosystem cannot come to fruition without a set of shared rules aimed at normalizing what behaviors are tolerated and what represent abuse, rather than ad hoc Terms of Service rules that change from platform to platform.
Twitter spam and false accounts prevalence, detection, and characterization: A survey

**The role of research sponsors and funders.** Much responsibility has been put onto the academic and industrial research endeavor to come up with socio-technical solutions to these problems. Nevertheless, I argue that we deal with pressing societal issues and social media spam and abuse can lead to tangible, nefarious impacts for democracy and society at large. Hence, our research must be at the top of public funding agendas, in stark contrast with the limited number of projects and programs that have been devoted to sponsoring research into these problems. It is in the public interest that social media abuse and manipulation should be mitigated, and not solely the responsibility of platforms to ensure so. Therefore, supporting research into socio-technical solutions with public funds is warranted now more than ever.

**Our future might also depend upon an educated public.** Whether major problems like misinformation can ever be solved with the sole use of socio-technical solutions has been subject of much academic debate. Hence, there is a role for literacy toward an educated Internet-based hyper-connected society. But it would be wishful thinking to expect that literacy would suffice to defeat problems like social media manipulation. For example, thanks to the leaps in artificial intelligence, it is nearly impossible for a non-trained human to distinguish between organic and machine-generated content (e.g., a short, tweet-like text, or a profile picture): AI techniques based on large generative models for text (e.g., GPT3 and its successors) and multimedia (DALL-E 2, Midjourney, Stable Diffusion, etc.) can trick human judgment, so even if a literate or tech-savvy user is exposed to such manipulated or AI-generated content, there is a limit to our cognitive and judgemental abilities to ascertain whether what we see is true or machine-generated, or whether what we read is factual or fabricated.

Literacy will prepare the future generation to question everything online, but the full trifecta of socio-technical solutions, sensible policy-making, and an informed public will be necessary to truly uproot the problems of social media spam and abuse, and their possibly devastating effects.

**About the author**

Emilio Ferrara is Professor at the University of Southern California, Research Team Leader at the USC Information Sciences Institute, and Principal Investigator at the USC/ISI Machine Intelligence and Data Science (MINDS). Ferrara uses AI for modeling and predicting human behavior in techno-social systems and has published 200+ papers on these topics. Ferrara served as the trial expert on the issue of spam and false accounts for the Delaware Court of Chancery in the *Twitter v. Musk* litigation.

E-mail: emiliofer [at] usc [dot] edu

**Acknowledgements**

The author is indebted to Professor Tim Weninger (University of Notre Dame), and Professors Patrick Warren and Darren Linvill (Clemson University) for their invaluable feedback.

**Notes**

1. The definition of what a social bot is has seen some debate in the literature. A bot (short for robot) generally refers to an entity operating in a digital space that is controlled by software rather than human. Bots have been categorized according to various taxonomies (Gorwa and Guilbeault, 2020; Woolley, 2016). In this taxonomy, I use the term bot as a shorthand to social bot, a concept that refers to a social media account controlled, predominantly or completely, by computer software (a more or less sophisticated artificial intelligence), in contrast with accounts controlled by human users. I introduced this definition in (Ferrara, *et al.*, 2016), which to date remains by far the most cited article on the subject of Twitter bots, with over 2,000 citations. This definition is in line with the recommendations of (Gorwa and Guilbeault,
2020] who provided the most comprehensive survey on the typologies of bots (cf., p. 9: “we suggest that automated social media accounts be called social bots”).

2. DARPA Social Media in Strategic Communication (SMISC) program (archived) Web site: https://www.darpa.mil/program/social-media-in-strategic-communication.

3. A description of Botometer taken from the official Web site: “Botometer is a machine learning algorithm trained to calculate a score where low scores indicate likely human accounts and high scores indicate likely bot accounts. To calculate the score, Botometer compares an account to tens of thousands of labeled examples. When you check an account, your browser fetches its public profile and hundreds of its public tweets and mentions using the Twitter API. This data is passed to the Botometer API, which extracts over a thousand features to characterize the account’s profile, friends, social network structure, temporal activity patterns, language, and sentiment. Finally, the features are used by various machine learning models to compute the bot scores. We do not retain any data other than the account’s ID, scores, and any feedback optionally provided by the user.”

4. Description of BotometerLite taken from the Web site: “BotometerLite is a lightweight and scalable bot detector for Twitter. Unlike Botometer, BotometerLite only requires one tweet to perform bot detection, which gives it the following advantages: 1. It can efficiently analyze large amounts of accounts; 2. It allows bot detection on historical data, whereas Botometer only returns the most recent status.”

5. Description of Botsentinel taken from the official Web site: “Bot Sentinel is a free non-partisan platform developed to classify and track inauthentic accounts and toxic trolls. The platform uses machine learning and artificial intelligence to classify Twitter accounts, and then adds the accounts to a publicly available database that anyone can browse.”

6. Description of TweetBotOrNot taken from the official Web site: “Uses machine learning to classify Twitter accounts as bots or not bots. The default model is 93.53% accurate when classifying bots and 95.32% accurate when classifying non-bots. The fast model is 91.78% accurate when classifying bots and 92.61% accurate when classifying non-bots.”

7. Description of DeBot taken from official Web site: “DeBot is real-time bot detection system. The project started on Feb 2015 and it has been collecting data since Aug 2015. High correlation in activities among users in social media is unusual and can be used as an indicator of bot behavior. DeBot identifies such bots in Twitter network. Our system reports and archives thousands of bot accounts every day. DeBot is an unsupervised method capable of detecting bots in a parameter-free fashion. In March 2017, DeBot has collected over 730K unique bots. Since we are detecting and archiving Twitter bots on a daily basis, we can offer two different service: bot archive API and on-demand bot detection platform.”

8. See Table 8 in Alipour, et al., (2022), omitted here.

9. D. Mirvish, 2011. “The Hathaway Effect: How Anne gives Warren Buffett a rise,” Huffington Post (2 March), at https://www.huffpost.com/entry/the-hathaway-effect-how-a_b_830041, accessed 23 November 2022.

10. U.S. Securities and Exchange Commission (SEC), 2014. “Updated investor alert: Social media and investing — Avoiding fraud” (12 November), at https://www.sec.gov/enforce/investor-alerts-and-bulletins/ia_socialmediafraud, accessed 23 November 2022.

11. “Promoters push market cap to $655 million despite $39 in assets and no revenue,” Seeking Alpha (18 June 2014), at https://seekingalpha.com/article/2274553-cynk-technology-promoters-push-market-cap-to-655-million-despite-39-in-assets-and-no-revenue-100-percent-downside, accessed 23 November 2022.

12. R. Morgan, 2014. “SEC sees price and reality not in ‘Cynk’,” New York Post (11 July), at https://nypost.com/2014/07/11/sec-sees-price-and-reality-not-in-cynk/, accessed 23 November 2022.
References

N. Abokhodair, D. Yoo, and D.W. McDonald, 2015. “Dissecting a social botnet: Growth, content and influence in Twitter,” CSCW ’15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 839–851. doi: https://doi.org/10.1145/2675133.2675208, accessed 23 November 2022.

A. Alarifi, M. Alsaleh, and A. Al-Salman, 2016. “Twitter Turing test: Identifying social machines,” Information Sciences, volume 372, pp. 332–346. doi: https://doi.org/10.1016/j.ins.2016.08.036, accessed 23 November 2022.

M. Ali, P. Sapiezynski, A. Korolova, A. Mislove, and A. Rieke, 2021. “Ad delivery algorithms: The hidden arbiters of political messaging,” WSDM ’21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 13–21. doi: https://doi.org/10.1145/3437963.3441801, accessed 23 November 2022.

M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, 2019. “Discrimination through optimization: How Facebook’s ad delivery can lead to biased outcomes,” Proceedings of the ACM on Human-Computer Interaction, volume 3, issue CSCW, article number 199, pp. 1–30. doi: https://doi.org/10.1145/3359301, accessed 23 November 2022.

M. Aliapoulios, E. Bevensee, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, and S. Zannettou, 2021. “A large open dataset from the Parler social network,” Proceedings of the International AAAI Conference on Web and Social Media, volume 15, number 1, pp. 943–951. doi: https://doi.org/10.1609/icwsm.v15i1.18117, accessed 23 November 2022.

S.A. Alipour, R. Orji, and N. Zincir-Heywood, 2022. “Security of social networks: Lessons learned on Twitter bot analysis in the literature,” ARES ’22: Proceedings of the 17th International Conference on Availability, Reliability and Security, article number 124, pp 1–9. doi: https://doi.org/10.1145/3538969.3544450, accessed 23 November 2022.

H. Allcott and M. Gentzkow, 2017. “Social media and fake news in the 2016 election,” Journal of Economic Perspectives, volume 31, number 2, pp. 211–236. doi: https://doi.org/10.1257/jep.31.2.211, accessed 23 November 2022.

J.-P. Allem, P. Escobedo, and L. Dharmapuri, 2020. “Cannabis surveillance with Twitter data: Emerging topics and social bots,” American Journal of Public Health, volume 110, number 3, pp. 357–362. doi: https://doi.org/10.2105/AJPH.2019.305461, accessed 23 November 2022.
C.H. Basch, Z. Meleo-Erwin, J. Fera, C. Jaime, and C.E. Basch, 2021. “A global pandemic in the time of viral memes: COVID-19 vaccine misinformation and disinformation on TikTok,” *Human Vaccines & Immunotherapeutics*, volume 17, number 8, pp. 2,373–2,377.
doi: https://doi.org/10.1080/21645515.2021.1894896, accessed 23 November 2022.

J.M. Berger and J. Morgan, 2015. “The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter,” *Brookings Project on U.S. Relations with the Islamic World, Analysis Paper*, number 20, at https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf, accessed 23 November 2022.

A. Bessi and E. Ferrara, 2016. “Social bots distort the 2016 U.S. presidential election online discussion,” *First Monday*, volume 21, number 11, at https://firstmonday.org/article/view/7090/5653, accessed 23 November 2022.
doi: https://doi.org/10.5210/fm.v21i11.7090, accessed 23 November 2022.

Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, 2013. “Design and analysis of a social botnet,” *Computer Networks*, volume 57, number 2, pp. 556–578.
doi: https://doi.org/10.1016/j.comnet.2012.06.006, accessed 23 November 2022.

Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, 2011. “The socialbot network: When bots socialize for fame and money,” *ACSAC ’11: Proceedings of the 27th Annual Computer Security Applications Conference*, pp. 93–102.
doi: https://doi.org/10.1145/2076732.2076746, accessed 23 November 2022.

A. Bovet and H.A. Makse, 2019. “Influence of fake news in Twitter during the 2016 US presidential election,” *Nature Communications*, volume 10, article number 7 (2 January).
doi: https://doi.org/10.1038/s41467-018-07761-2, accessed 23 November 2022.

D.A. Broniatowski, A.M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S.C. Quinn, and M. Dredze, 2018. “Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate,” *American Journal of Public Health*, volume 108, number 10, pp. 1,378–1,384.
doi: https://doi.org/10.2105/AJPH.2018.304567, accessed 23 November 2022.

M. Cardaioli, M. Conti, A. Di Sorbo, E. Fabrizio, S. Laudanna, and C.A. Visaggio, 2021. “It’s a matter of style: Detecting social bots through writing style consistency,” *2021 International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–9.
doi: https://doi.org/10.1109/ICCCN52240.2021.9522339, accessed 23 November 2022.

H.-C.H. Chang and E. Ferrara, 2022. “Comparative analysis of social bots and humans during the COVID-19 pandemic,” *Journal of Computational Social Science*, volume 5, pp. 1,409–1,425.
doi: https://doi.org/10.1007/s42001-022-00173-9, accessed 23 November 2022.

N. Chavoshi, H. Hamooni, and A. Mueen, 2016. Identifying correlated bots in Twitter,” In: E. Spiro and Y.Y. Ahn (editors). *Social informatics. SocInfo 2016. Lecture Notes in Computer Science*, volume 10047. Cham, Switzerland: Springer, pp. 14–21.
doi: https://doi.org/10.1007/978-3-319-47874-6_2, accessed 23 November 2022.

E. Chen, J. Jiang, H.-C.H. Chang, G. Muric, and E. Ferrara, 2022. “Charting the information and misinformation landscape to characterize misinfodemics on social media: COVID-19 infodemiology study at a planetary scale,” *JMIR Infodemiology*, volume 2, number 1, e32378.
doi: https://doi.org/10.2196/32378, accessed 23 November 2022.

E. Chen, H. Chang, A. Rao, K. Lerman, G. Cowan, and E. Ferrara, 2021. “COVID-19 misinformation and the 2020 US presidential election,” *Harvard Kennedy School Misinformation Review* (3 March), at
S. Feng, Z. Tan, H. Wan, N. Wang, Z. Chen, B. Zhang, Q. Zheng, W. Zhang, Z. Lei, S. Yang, X. Feng, Q. Zhang, H. Wang, Y. Liu, Y. Bai, H. Wang, Z. Cai, Y. Wang, L. Zheng, Z. Ma, J. Li, and M. Luo, 2022. “TwiBot-22: Towards graph-based Twitter bot detection,” *arXiv*:2206.04564 (9 June). doi: https://doi.org/10.48550/arXiv.2206.04564, accessed 23 November 2022.

E. Ferrara, 2019. “The history of digital spam,” *Communications of the ACM*, volume 62, number 8, pp. 82–91. doi: https://doi.org/10.1145/3299768, accessed 23 November 2022.

E. Ferrara, 2017. “Disinformation and social bot operations in the run up to the 2017 French presidential election,” *First Monday*, volume 22, number 8, at https://firstmonday.org/article/view/8005/6516, accessed 23 November 2022. doi: https://doi.org/10.5210/fm.v22i8.8005, accessed 23 November 2022.

E. Ferrara, 2015. “Manipulation and abuse on social media,” *ACM SIGWEB Newsletter* (Spring), article number 4, pp. 1–9. doi: https://doi.org/10.1145/2749279.2749283, accessed 23 November 2022.

E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, 2020. “Characterizing social media manipulation in the 2020 U.S. presidential election,” *First Monday*, volume 25, number 11, at https://firstmonday.org/article/view/11431/9993, accessed 23 November 2022. doi: https://doi.org/10.5210/fm.v25i11.11431, accessed 23 November 2022.

E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, 2016. “The rise of social bots,” *Communications of the ACM*, volume 59, number 7, pp. 96–104. doi: https://doi.org/10.1145/2818717, accessed 23 November 2022.

C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, 2015. “Reverse engineering socialbot infiltration strategies in Twitter,” *ASONAM '15: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 25–32. doi: https://doi.org/10.1145/2808797.2809292, accessed 23 November 2022.

M. Fukuda, K. Nakajima, and K. Shudo, 2022. “Estimating the bot population on Twitter via random walk based sampling,” *IEEE Access*, volume 10, pp. 17,201–17,211. doi: https://doi.org/10.1109/ACCESS.2022.3149887, accessed 23 November 2022.

E. Gabarron, S.O. Oyeyemi, and R. Wynn, 2021. “COVID-19-related misinformation on social media: a systematic review,” *Bulletin of the World Health Organization*, volume 99, number 6, pp. 455–463A. doi: https://doi.org/10.2471/BLT.20.276782, accessed 23 November 2022.

Z. Gilani, E. Kochmar, and J. Crowcroft, 2017. “Classification of Twitter accounts into automated agents and human users,” *ASONAM '17: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 489–496. doi: https://doi.org/10.1145/3110025.3110091, accessed 23 November 2022.

R. Gorwa and D. Guilbeault, 2020. “Unpacking the social media bot: A typology to guide research and policy,” *Policy & Internet*, volume 12, number 2, pp. 225–248. doi: https://doi.org/10.1002/poi.3.184, accessed 23 November 2022.

N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, 2019. “Fake news on Twitter during the 2016 U.S. presidential election,” *Science*, volume 363, number 6425 (25 January), pp. 374–378. doi: https://doi.org/10.1126/science.aau2706, accessed 23 November 2022.

A. Gupta, H. Lamba, and P. Kumaraguru, 2013. “$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter,” 2013 *APWG eCrime Researchers Summit*, pp. 1–12.
P.N. Howard, S. Woolley, and R. Calo, 2018. “Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration,” *Journal of Information Technology & Politics*, volume 15, number 2, pp. 81–93. doi: https://doi.org/10.1080/19331681.2018.1448735, accessed 23 November 2022.

P.N. Howard, A. Duffy, D. Freelon, M.M. Hussain, W. Mari, and M. Maziad, 2011. “Opening closed regimes: What was the role of social media during the Arab Spring?” *Project on Information Technology & Political Islam, Working Paper*, 2011.1, at https://hdl.handle.net/2027.42/117568, accessed 23 November 2022.

D. Hunter and N. Evans, 2016. “Facebook emotional contagion experiment controversy,” *Research Ethics*, volume 12, number 1, pp. 2–3. doi: https://doi.org/10.1177/1747016115626341, accessed 23 November 2022.

T. Hwang, I. Pearce, and M. Nanis, 2012. “Socialbots: Voices from the fronts,” *Interactions*, volume 19, number 2, pp. 38–45. doi: https://doi.org/10.1145/2090150.2090161, accessed 23 November 2022.

J. Jiang, X. Ren, and E. Ferrara, 2021. “Social media polarization and echo chambers in the context of COVID-19: Case study,” *JMIRx Med*, volume 2, number 3, e29570. doi: https://doi.org/10.2196/29570, accessed 23 November 2022.

T.R. Keller and U. Klinger, 2019. “Social bots in election campaigns: Theoretical, empirical, and methodological implications,” *Political Communication*, volume 36, number 1, pp. 171–189. doi: https://doi.org/10.1080/10584609.2018.1526238, accessed 23 November 2022.

S. Kudugunta and E. Ferrara, 2018. “Deep neural networks for bot detection,” *Information Sciences*, volume 467, pp. 312–322. doi: https://doi.org/10.1016/j.ins.2018.08.019, accessed 23 November 2022.

S. Kumar, S. Garg, Y. Vats, and A. Singh Parihar, 2021. "Content based bot detection using bot language model and BERT embeddings," *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pp. 285–289. doi: https://doi.org/10.1109/ICCCSP52374.2021.9465506, accessed 23 November 2022.

K. Lee, B. Eoff, and J. Caverlee, 2011. “Seven months with the devils: A long-term study of content polluters on Twitter,” *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, number 1, pp. 185–192. doi: https://doi.org/10.1609/icsm.v5i1.14106, accessed 23 November 2022.

L. Luceri, A. Deb, A. Badawy, and E. Ferrara, 2019. “Red bots do it better: Comparative analysis of social bot partisan behavior,” *WWW ’19: Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1,007–1,012. doi: https://doi.org/10.1145/3308560.3316735, accessed 23 November 2022.

D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. Belmonte Hernández, A. Lozano-Diez, and F. Álvarez, 2021. “A deep learning approach for robust detection of bots in Twitter using transformers,” *IEEE Access*, volume 9, pp. 54,591–54,601. doi: https://doi.org/10.1109/ACCESS.2021.3068659, accessed 23 November 2022.

A. Marwick and R. Lewis, 2017. “Media manipulation and disinformation online,” *Data & Society* (15 May), at https://datasociety.net/library/media-manipulation-and-disinfo-online/, accessed 23 November 2022.
M. Mazza, G. Cola, and M. Tesconi, 2022. “Ready-to-(ab) use: From fake account trafficking to coordinated inauthentic behavior on Twitter,” Online Social Networks and Media, volume 31, 100224. doi: https://doi.org/10.1016/j.osnem.2022.100224, accessed 23 November 2022.

N. Mele, D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, and C. Mattsson, 2017. “Combating fake news: An agenda for research and action” (2 May), at https://shorensteincenter.org/combating-fake-news-agenda-for-research/, accessed 23 November 2022.

P.T. Metaxas and E. Mustafaraj, 2012. “Social media and the elections,” Science, volume 338, number 6106 (26 October), pp. 472–473. doi: https://doi.org/10.1126/science.1230456, accessed 23 November 2022.

A. Minnich, N. Chavoshi, D. Koutra, and A. Mueen, 2017. “BotWalk: Efficient adaptive exploration of Twitter bot networks,” ASONAM ’17: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 467–474. doi: https://doi.org/10.1145/3110025.3110163, accessed 23 November 2022.

G. Muric, Y. Wu, and E. Ferrara, 2021. “COVID-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies,” JMIR Public Health and Surveillance, volume 7, number 11, e30642. doi: https://doi.org/10.2196/30642, accessed 23 November 2022.

H. Nghiem, G. Muric, F. Morstatter, and E. Ferrara, 2021. “Detecting cryptocurrency pump-and-dump frauds using market and social signals,” Expert Systems with Applications, volume 182, 115284. doi: https://doi.org/10.1016/j.eswa.2021.115284, accessed 23 November 2022.

L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, 2021. “Coordinated behavior on social media in 2019 UK general election,” Proceedings of the International AAAI Conference on Web and Social Media, volume 15, pp. 443–454. doi: https://doi.org/10.1609/icwsm.v15i1.18074, accessed 23 November 2022.

L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, and E. Ferrara, 2020. “Charting the landscape of online cryptocurrency manipulation,” IEEE Access, volume 8, pp. 113,230–113,245. doi: https://doi.org/10.1109/ACCESS.2020.3003370, accessed 23 November 2022.

D. Pacheco, A. Flammini, and F. Menczer, 2020. “Unveiling coordinated groups behind White Helmets disinformation,” WWW ’20: Companion Proceedings of the Web Conference 2020, pp. 611–616. doi: https://doi.org/10.1145/3366424.3385775, accessed 23 November 2022.

A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. De Cristofaro, 2021. “‘Is it a coincidence?’: An exploratory study of QAnon on Voat,” WWW ’21: Proceedings of the Web Conference 2021, pp. 460–471. doi: https://doi.org/10.1145/3442381.3450036, accessed 23 November 2022.

R. Phillips and H. Wilder, 2020. “Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites,” 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), pp. 1–8. doi: https://doi.org/10.1109/ICBC48266.2020.9169433, accessed 23 November 2022.

F. Pierri, L. Luceri, and E. Ferrara, 2022. “How does Twitter account moderation work? Dynamics of account creation and suspension during major geopolitical events,” arXiv:2209.07614v2 (19 September). doi: https://doi.org/10.48550/arXiv.2209.07614, accessed 23 November 2022.

I. Pozzana and E. Ferrara, 2020. “Measuring bot and human behavioral dynamics,” Frontiers in Physics (22
E.K. Quinn, S.S. Fazel, and C.E. Peters, 2021. “The Instagram infodemic: Cobranding of conspiracy theories, coronavirus disease 2019 and authority-questioning beliefs,” Cyberpsychology, Behavior, and Social Networking, volume 24, number 8, pp. 573–577.
doi: https://doi.org/10.1089/cyber.2020.0663, accessed 23 November 2022.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, 2019. “Language models are unsupervised multitask learners,” at https://d4mucfpwyv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, accessed 23 November 2022.

I.D. Raji, A. Smart, R.N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, 2020. “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing,” FAT* ‘20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44.
doi: https://doi.org/10.1145/3351095.3372873, accessed 23 November 2022.

A. Rao, F. Morstatter, M. Hu, E. Chen, K. Burghardt, E. Ferrara, and K. Lerman, 2021. “Political partisanship and antiscience attitudes in online discussions about COVID-19: Twitter content analysis,” Journal of Medical Internet Research, volume 23, number 6, e26692.
doi: https://doi.org/10.2196/26692, accessed 23 November 2022.

J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, 2011. “Detecting and tracking political abuse in social media,” Proceedings of the International AAAI Conference on Web and Social Media, volume 5, number 1, pp. 297–304.
doi: https://doi.org/10.1609/icwsm.v5i1.14127, accessed 23 November 2022.

R.R. Rout, G. Lingam, and D.V.L.N. Somayajulu, 2020. “Detection of malicious social bots using learning automata with URL features in Twitter network,” IEEE Transactions on Computational Social Systems, volume 7, number 4, pp. 1,004–1,018.
doi: https://doi.org/10.1109/TCSS.2020.2992223, accessed 23 November 2022.

T. Sakaki, M. Okazaki, and Y. Matsuo, 2010. “Earthquake shakes Twitter users: Real time event detection by social sensors,” WWW ’10: Proceedings of the 19th international conference on World Wide Web, pp. 851–860.
doi: https://doi.org/10.1145/1772690.1772777, accessed 23 November 2022.

S. Shahsavari, P. Holur, T. Wang, T.R. Tangherlini, and V. Roychowdhury, 2020. “Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news,” Journal of Computational Social Science, volume 3, number 2, pp. 279–317.
doi: https://doi.org/10.1007/s42001-020-00086-5, accessed 23 November 2022.

K. Sharma, E. Ferrara, and Y. Liu, 2022. “Characterizing online engagement with disinformation and conspiracies in the 2020 US presidential election,” Proceedings of the International AAAI Conference on Web and Social Media, volume 16, pp. 908–919, and at https://ojs.aaai.org/index.php/ICWSM/article/view/19345/19117, accessed 23 November 2022.

H. Shen, X. Liu, and X. Zhang, 2022. “Boosting social spam detection via attention mechanisms on Twitter,” Electronics, volume 11, 1129.
doi: https://doi.org/10.3390/electronics11071129, accessed 23 November 2022.

H. Shukla, N. Jajtapat, and B. Patil, 2021. “Enhanced Twitter bot detection using ensemble machine learning,” 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 930–936.
M. Stella, E. Ferrara, and M. De Domenico, 2018. “Bots increase exposure to negative and inflammatory content in online social systems,” *Proceedings of the National Academy of Sciences*, volume 115, number 49 (4 December), pp. 12,435–12,440. doi: https://doi.org/10.1073/pnas.1803470115, accessed 23 November 2022.

B. Stricke, 2020. “People v. robots: A roadmap for enforcing California’s new online bot disclosure act,” *Vanderbilt Journal of Entertainment & Technology Law*, volume 22, number 4, article number 6, pp. 839–894, and at https://scholarship.law.vanderbilt.edu/jetlaw/vol22/iss4/6, accessed 23 November 2022.

D. Stukal, S. Sanovich, R. Bonneau, and J.A. Tucker, 2017. “Detecting bots on Russian political Twitter,” *Big Data*, volume 5, number 4, pp. 310–324. doi: https://doi.org/10.1089/big.2017.0038, accessed 23 November 2022.

V.S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, 2016. “The DARPA Twitter bot challenge,” *Computer*, volume 49, number 6, pp. 38–46. doi: https://doi.org/10.1109/MC.2016.183, accessed 23 November 2022.

T.R. Tangherlini, S. Shahsavari, B. Shahbazi, E. Ebrahimpizadeh, and V. Roychowdhury, 2020. “An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the Web,” *PLoS ONE*, volume 15, number 6 (16 June), e0233879. doi: https://doi.org/10.1371/journal.pone.0233879, accessed 23 November 2022.

S. Tardelli, M. Avvenuti, M. Tesconi, and S. Cresci, 2020. “Characterizing social bots spreading financial disinformation,” In: G. Meiselwitz (editor). *Social computing and social media. Design, ethics, user behavior, and social network analysis. Lecture Notes in Computer Science*, volume 12194. Cham, Switzerland: Springer, pp. 376–392. doi: https://doi.org/10.1007/978-3-030-49570-1_26, accessed 23 November 2022.

K. Thomas, C. Grier, D. Song, and V. Paxson, 2011. “Suspended accounts in retrospect: An analysis of Twitter spam,” *PIMC ’11: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 243–258. doi: https://doi.org/10.1145/2068816.2068840, accessed 23 November 2022.

Z. Tufekci, 2018. “YouTube, the great radicalizer,” *New York Times* (10 March), at https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html, accessed 23 November 2022.

O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, 2017. “Online human-bot interactions: Detection, estimation, and characterization,” *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, number 1, pp. 280–289. doi: https://doi.org/10.1609/icewsm.v11i1.14871, accessed 23 November 2022.

G. Venkatadri, P. Sapiezynski, E.M. Redmiles, A. Mislove, O. Goga, M. Mazurek, and K.P. Gummadi, 2019. “Auditing offline data brokers via Facebook’s advertising platform,” *WWW ’19: The World Wide Web Conference*, pp. 1,920–1,930. doi: https://doi.org/10.1145/3308558.3313666, accessed 23 November 2022.

S. Vosoughi, D. Roy, and S. Aral, 2018. “The spread of true and false news online,” *Science*, volume 359, number 6380 (9 March), pp. 1,146–1,151. doi: https://doi.org/10.1126/science.aap9559, accessed 23 November 2022.

E.L. Wang, L. Luceri, F. Pierri, and E. Ferrara, 2022. “Identifying and characterizing behavioral classes of
radicalization within the QAnon conspiracy on Twitter,” arXiv:2209.09339 (19 September). doi: https://doi.org/10.48550/arXiv.2209.09339, accessed 23 November 2022.

F. Wei and U.T. Nguyen, 2019. “Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings,” 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 101–109. doi: https://doi.org/10.1145/2675133.2675208, accessed 23 November 2022.

S.A. Williams, M.M. Terras, and C. Warwick, 2013. “What do people study when they study Twitter? Classifying Twitter related academic papers,” Journal of Documentation, volume 69, number 3, pp. 384–410. doi: https://doi.org/10.1108/JD-03-2012-0027, accessed 23 November 2022.

S. Wojcik, S. Messing, A. Smith, L. Rainie, and P. Hitlin, 2018. “Bots in the Twittersphere,” Pew Research Center (9 April), at https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/, accessed 23 November 2022.

S.C. Woolley, 2016. “Automating power: Social bot interference in global politics,” First Monday, volume 21, number 4, at https://firstmonday.org/article/view/6161/5300, accessed 23 November 2022. doi: https://doi.org/10.5210/fm.v21i4.6161, accessed 23 November 2022.

S.C. Woolley and P. Howard, 2017. “Computational propaganda worldwide: Executive summary,” Oxford Internet Institute, at https://demtech.oii.ox.ac.uk/research/posts/computational-propaganda-worldwide-executive-summary/, accessed 23 November 2022.

S.C. Woolley and P.N. Howard, 2016. “Political communication, computational propaganda, and autonomous agents — Introduction,” International Journal of Communication, volume 10, at https://ijoc.org/index.php/ijoc/article/view/6298, accessed 23 November 2022.

B. Wu, L. Liu, Y. Yang, K. Zheng, and X. Wang, 2020. “Using improved conditional generative adversarial networks to detect social bots on Twitter,” IEEE Access, volume 8, pp. 36,664–36,680. doi: https://doi.org/10.1109/ACCESS.2020.2975630, accessed 23 November 2022.

K.-C. Yang, E. Ferrara, and F. Menczer, 2022. “Botometer 101: Social bot practicum for computational social scientists,” Journal of Computational Social Science, volume 5, pp. 1,511–1,528. doi: https://doi.org/10.1007/s42001-022-00177-5, accessed 23 November 2022.

K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, 2020. “Scalable and generalizable social bot detection through data selection,” Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, number 1, pp. 1,096–1,103. doi: https://doi.org/10.1609/aaai.v34i01.5460, accessed 23 November 2022.

D. Yates and S. Paquette, 2011. “Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake,” International Journal of Information Management, volume 31, number 1, pp. 6–13. doi: https://doi.org/10.1016/j.ijinfomgt.2010.10.001, accessed 23 November 2022.

J. Yin, S. Karimi, A. Lampert, M. Cameron, B. Robinson, and R. Power, 2015. “Using social media to enhance emergency situation awareness,” Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 4,234–4,238, and at https://www.ijcai.org/Proceedings/15/Papers/602.pdf, accessed 23 November 2022.

S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M Sirivianos, G. Stringhini, and G. Suarez-Tangil, 2018a. “On the origins of memes by means of fringe Web communities,” PIMC ’18: Proceedings of the Internet Measurement Conference 2018, pp. 188–202.
Twitter spam and false accounts prevalence, detection, and characterization: A survey

doi: https://doi.org/10.1145/3278532.3278550, accessed 23 November 2022.

S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn, 2018b. “What is Gab? A bastion of free speech or an alt-right echo chamber?” *WWW ’18 Companion: The 2018 Web Conference Companion*, pp. 1,007–1,014. doi: https://doi.org/10.1145/3184558.3191531, accessed 23 November 2022.

Z. Zhao, P. Resnick, and Q. Mei, 2015. “Enquiring minds: Early detection of rumors in social media from enquiry posts,” *WWW ’15: Proceedings of the 24th International Conference on World Wide Web*, pp. 1,395–1,405. doi: https://doi.org/10.1145/2736277.2741637, accessed 23 November 2022.

M. Zimmer and N.J. Proferes, 2014. “A topology of Twitter research: Disciplines, methods, and ethics,” *Aslib Journal of Information Management*, volume 66, number 3, pp. 250–261. doi: https://doi.org/10.1108/AJIM-09-2013-0083, accessed 23 November 2022.

---

**Editorial history**

Received 10 November 2022; revised 15 November 2022; revised 17 November 2022; accepted 22 November 2022.

---

This paper by Emilio Ferrara is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Twitter spam and false accounts prevalence, detection, and characterization: A survey by Emilio Ferrara. *First Monday*, volume 27, number 12 (December 2022). doi: https://dx.doi.org/10.5210/fm.v27i12.12872