WILEY

SPECIAL ISSUE ARTICLE

Arming the public with artificial intelligence to counter social bots

Kai-Cheng Yang[1] | Onur Varol[2] | Clayton A. Davis[1] | Emilio Ferrara[3] | Alessandro Flammini[1,4] | Filippo Menczer[1,4] ⬤

[1]Center for Complex Networks and Systems Research, Indiana University, Bloomington, Indiana

[2]Center for Complex Networks Research, Northeastern University, Boston, Massachusetts

[3]Information Sciences Institute, University of Southern California, Los Angeles, California

[4]Indiana University Network Science Institute, Bloomington, Indiana

Correspondence
Filippo Menczer, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408.
Email: fil@indiana.edu

The increased relevance of social media in our daily life has been accompanied by efforts to manipulate online conversations and opinions. Deceptive social bots—automated or semi-automated accounts designed to impersonate humans—have been successfully exploited for these kinds of abuse. Researchers have responded by developing artificial intelligence (AI) tools to arm the public in the fight against social bots. Here we review the literature on different types of bots, their impact, and detection methods. We use the case study of Botometer, a popular bot detection tool developed at Indiana University, to illustrate how people interact with AI countermeasures. A user experience survey suggests that bot detection has become an integral part of the social media experience for many users. However, barriers in interpreting the output of AI tools can lead to fundamental misunderstandings. The arms race between machine learning methods to develop sophisticated bots and effective countermeasures makes it necessary to update the training data and features of detection tools. We again use the Botometer case to illustrate both algorithmic and interpretability improvements of bot scores, designed to meet user expectations. We conclude by discussing how future AI developments may affect the fight between malicious bots and the public.

KEYWORDS
social influence, social media, social networking

1 | INTRODUCTION

During the past two decades, we have progressively turned to the Internet and social media to find news, share opinions, and entertain conversations (Morris & Ogan, 1996; Smith & Brenner, 2012). What we create and consume on the Internet impacts all aspects of our daily lives, including our political, health, financial, and entertainment decisions. This increased influence of social media has been accompanied by an increase in attempts to alter the organic nature of our online discussions and exchanges of ideas. In particular, over the past 10 years we have witnessed an explosion of social bots (Boshmaf, Muslukhov, Beznosov, & Ripeanu, 2013; Lee, Eoff, & Caverlee, 2011), a presence that does not show signs of decline.

Social bots are social media accounts controlled completely or in part by computer algorithms. They can generate content automatically and interact with human users, often posing as, or imitating, humans (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Automated accounts can be harmless and even helpful in scenarios where they save manual labor without polluting human conversations. In this paper, however, we focus on those actors in online social networks that surreptitiously aim to manipulate public discourse and influence human opinions and behavior in an opaque fashion. While more traditional nefarious entities, like malware, attack vulnerabilities of hardware and software, social bots exploit human vulnerabilities, such as our tendencies to pay attention to what appears to be popular and to trust social contacts (Jun, Meng, & Johar, 2017). Unlike other social engineering attacks, such as spear phishing (Jagatic, Johnson, Jakobsson, & Menczer, 2007), bots can achieve scalability through automation. For example, multiple accounts controlled by a single entity can quickly generate posts and make specific content trend or amplify misinformation. They can trick humans and engagement-based ranking algorithms alike, creating the appearance that some person or

opinion is popular. Therefore, defending from social bots raises serious research challenges (Boshmaf, Muslukhov, Beznosov, & Ripeanu, 2012).

Manipulation of public opinion is not new; it has been a common practice since the dawn of humanity. The technological tools of all eras—printed media, radio, television, and the Internet—have been abused to disseminate misinformation and propaganda. The deceptive strategies employed on all these types of channels share striking similarities (Varol & Uluturk, 2018). Nevertheless, social media are particularly vulnerable because they facilitate automatic interactions via software. As a result, social media platforms have to combat a deluge of attacks. Facebook recently announced that 1.5 billion fake accounts were removed over 6 months in 2018.[1] Even a very low miss rate could leave millions of accounts available to be used as bots. In this light, it is not surprising that as many as 9–15% of active Twitter accounts were estimated to be bots in 2017 (Varol, Ferrara, Davis, Menczer, & Flammini, 2017), and that social bots are responsible for generating two thirds of links to popular websites (Wojcik, Messing, Smith, Rainie, & Hitlin, 2018). Public interest in social bots has also dramatically increased during the past few years. A recent Pew survey shows that 66% of Americans are aware of the existence of social bots and a consistent fraction of those believes that they are malicious (Stocking & Sumida, 2018). Despite high awareness, the survey also reveals that many people are not confident in their ability to identify social bots. Part of the difficulty in learning to recognize bots is that they continuously evolve in response to efforts by platforms to eliminate them. Social bots that go undetected by platforms have become increasingly sophisticated and human-like (Varol, Ferrara, Davis, et al., 2017), introducing an arms race between bot creators and the research community. Indeed, industry and academic research about social bots has flourished recently.

A first concern of this paper is to briefly review recent studies about social bots and the development of artificial intelligence (AI) countermeasures. A second goal is to present a study on how people interact with bot detection tools and what insight they gain. This is an important perspective that is often neglected by developers of AI tools. After all, winning this fight requires broad participation by social media users.

2 | LITERATURE REVIEW

There are many types of social bots. In this section, we review recent studies on the diverse characteristics of social bots, their activities, and impact. We then review the literature on bot detection algorithms.

2.1 | Characterization of social bots

Some simple bots only do one thing: They post content automatically. The Twitter account @big_ben_clock is a quintessential representative of this class—it tweets every hour mimicking the real Big Ben. Similar accounts can automatically post or retweet/share to publicize information such as news and academic papers (Haustein et al., 2016; Lokot & Diakopoulos, 2016). These bots are naive and easy to

identify: They share only one type of content and they do not try to misrepresent themselves or their motivations. In 2011, Lee et al. (2011) identified thousands of social bots. They created bots that posted meaningless messages through automated scripts. The underlying expectation was that only other bots and no human would follow these honeypot accounts. The honeypot accounts did end up having many followers. Subsequent analysis confirmed that these followers were indeed bots and revealed the common strategy of randomly following accounts to grow the number of social connections.

More sophisticated bots adopt various strategies to impersonate human users. Social bot developers can populate bot profiles by searching and collecting material from other platforms. A more extreme example of these kinds of bots are identity thieves: they copy usernames, profile information, and pictures of other accounts and use them as their own, making only small changes (Cresci, Di Pietro, Petrocchi, Spognardi, & Tesconi, 2015). State-of-the-art machine learning technologies can be employed as part of the algorithms that automatically generate bot content (Freitas, Benevenuto, Ghosh, & Veloso, 2015). Sophisticated bots can emulate temporal patterns of content posting and consumption by humans. They can even interact with other users by engaging in conversations, commenting on posts, and answering questions (Hwang, Pearce, & Nanis, 2012).

Bots employ a variety of strategies to form an audience. There are bots designed to gather followers and expand social circles, with the goal of exerting some form of influence. Some bots, for example, search the social networks for popular accounts, follow them, and ask to be followed back (Aiello, Deplano, Schifanella, & Ruffo, 2012). This type of infiltration has been proven to be more effective when the target community is topically centered. Bots can identify and generate appropriate content around specific topics to gain trust and attention from people interested in those topics. Such attention often translates in following relationships (Freitas et al., 2015). This finding is even more worrisome when one considers that the targets were programmers, a population expected to be more aware of social bots compared to other users.

Another class of bots, whose activity is less evident, is that of so-called fake followers (Cresci et al., 2015; De Micheli & Stroppa, 2013). These accounts are managed by entities that get paid in exchange for following customers who want to increase their perceived popularity. The fake followers often follow each other, forming a network that lends credibility to each member and allows them to elude being flagged for lack of followers. A few dollars can buy thousands of fake followers. A recent report[2] revealed that reality television stars, professional athletes, comedians, TED speakers, pastors, and fashion models all purchased fake followers on Twitter at least once. The companies that sell fake followers also sell fake engagement (likes, comments, retweets) and even custom-made fake personas.

A group of social bots, sometimes referred to as a botnet, may act in coordination, posting the same content again and again to generate false popularity. This type of behavior is generally harder to detect as the accounts involved, when inspected individually, may appear as genuine. Their behavior becomes conspicuous only when one is able to detect a large number of such accounts acting in a strongly coordinated manner (Chavoshi, Hamooni, & Mueen, 2016). Not surprisingly, hybrid approaches that mix human and automated activity have been

devised to bypass bot and bot-coordination detection systems (Grimme, Assenmacher, & Adam, 2018). A recent study by Cresci, Di Pietro, Petrocchi, Spognardi, and Tesconi (2017) shows that neither humans nor supervised machine learning algorithms can identify this kind of bots successfully.

2.2 | Activity and impact of social bots

Bot activity has been reported in several domains, with the potential to affect behaviors, opinions, and choices. Health is one domain of particular concern (Allem & Ferrara, 2018), where we have observed social bots influencing debates about vaccination policies (Broniatowski et al., 2018; Ferrara, Varol, Davis, et al., 2016) and smoking (Allem & Ferrara, 2016; Allem, Ferrara, Uppu, Cruz, & Unger, 2017). Politics is another key domain. During the 2010 U.S. midterm elections, primitive social bots were found to support some candidates and attack their opponents (Metaxas & Mustafaraj, 2012), injecting thousands of tweets pointing to websites with fake news (Ratkiewicz et al., 2011).

During the 2016 U.S. presidential election, social bots were found to generate a large amount of content, possibly distorting online conversations. Our analysis (Bessi & Ferrara, 2016) suggests that retweets were a vulnerable mode of information spreading: There was no significant difference in the amount of retweets that humans generated by resharing content produced by other humans or by bots. In fact, humans and bots retweeted each other substantially at the same rate. This suggests that bots were very effective at getting messages reshared in the human communication channels. We further explored how bots and humans talked about the two presidential candidates. We noted that bots tweeting about Donald Trump generated the most positive tweets (Bessi & Ferrara, 2016). The fact that bots produce systematically more positive content in support of a candidate can bias the perception of the individuals exposed to this content, suggesting that there exists an organic, grassroots support for a given candidate, while in reality it is all artificially generated.

Similar cases of political manipulation were reported in other countries (Stella, Ferrara, & De Domenico, 2018). Our analysis highlighted the presence of Twitter bots during the 2017 French Presidential Election (Ferrara, 2017). We explored tweets related to candidates Marine Le Pen and Emmanuel Macron in the days leading to the election. We identified 18K active bots that pushed the so-called *MacronLeaks* disinformation campaign. Accounts who engaged with these bots were mostly U.S.-based users with preexisting interest in alt-right topics and alternative news media, rather than French users. Hundreds of bots posting about the French elections were also active in the 2016 U.S. Election discussion, suggesting the existence of a black market for reusable political disinformation bots.

The spread of fake news online is another area in which the effect of bots is believed to be relevant (Bessi et al., 2015; Lazer et al., 2018). A study based on 14 million tweets posted during and after the 2016 U.S. presidential election revealed that bots played a key role in the spread of low-credibility content (Shao et al., 2018,b). The study uncovered strategies by which social bots target influential accounts and amplify misinformation in the early stages of spreading, before it becomes viral. Even bots created with good intentions may contribute

to the spread of misinformation. After the Boston marathon bombing, for example, bots started to automatically retweet indiscriminately, without verifying the credibility of the posts being retweeted or that of their sources (Gupta, Lamba, & Kumaraguru, 2013), unintentionally amplifying the diffusion of misinformation.

Social bots have been used to promote terrorist propaganda and proselytize online extremism. By analyzing a sample of 20,000 Islamic State supporting accounts, Berger and Morgan (2015) found that the terrorist organization was actively using social media and bots to spread its ideology. Abokhodair, Yoo, and McDonald (2015) dissected a botnet that tried to misdirect the online discussion during the Syrian civil war in 2012. Ferrara, Wang, Varol, Flammini, and Galstyan (2016) analyzed a dataset of millions of extremist tweets and found that social bots produced some of the content.

2.3 | Bot detection methods

An early attempt by Wang et al. (2012) involved building a crowdsourcing social bot detection platform. This method proved to be effective, but scalability was a prominent issue: while bots can be multiplied at will at essentially no cost, human detectors cannot. Later efforts therefore mostly leveraged machine learning methods.

Approaches based on supervised machine learning algorithms are the most common (Subrahmanian et al., 2016). Supervised approaches depend and often start with the collection of an extensive dataset, with each account labeled as either human or bot. These labels usually come from human annotation (Varol, Ferrara, Davis, et al., 2017), automated methods (e.g., those based on honey pots described earlier; Lee et al., 2011), or botnets that display suspicious behaviors (Echeverria & Zhou, 2017a, 2017b). A critical issue with existing datasets is the lack of ground truth. There is no objective, agreed-upon, operational definition of social bot. One of the factors that explain this is the prevalence of accounts that lie in the gray area between human and bot behavior, where even experienced researchers cannot easily discriminate. Nevertheless, datasets do include many typical bots; using the training labels as proxies for ground truth makes it possible to build practically viable tools (Davis, Varol, Ferrara, Flammini, & Menczer, 2016; Varol, Ferrara, Davis, et al., 2017).

The choice of relevant features used to describe entities to be classified is a critical step of machine learning classifiers. Different choices have been considered, but in general six broad categories of features have been identified as relevant for discriminating between human and bot accounts (Varol, Davis, Menczer, & Flammini, 2018): user metadata, friend metadata, retweet/mention network structure, content and language, sentiment, and temporal features. In the case of supervised learning, after extraction and preprocessing, the features are fed into supervised machine-learning models for training, then the trained models are used to evaluate previously unseen accounts.

Most techniques attempt to detect bots at the *account* level, by processing many social media posts to extract the features listed above. Recently, Kudugunta and Ferrara (2018) proposed a deep neural network approach based on the contextual long short-term memory (LSTM) architecture that exploits content and metadata to detect bots at the *tweet* level. Contextual features are extracted from user

metadata and fed as auxiliary input to LSTM deep nets processing the tweet text. The proposed framework, designed to predict whether a given tweet was posted by a human or a bot, exhibited promising performance on test benchmarks.

While supervised methods have proven to be effective in many cases, they do not perform well at detecting coordinated social bots that post human-generated content (Chen & Subramanian, 2018; Cresci et al., 2017; Grimme et al., 2018). As mentioned earlier, those coordinated bots are not usually suspicious when considered individually. Their detection requires information about their coordination, which becomes available only once the activity of multiple bots is considered. Unsupervised learning methods have been proposed to address this issue. Early attempts combined multiple features via Euclidean distance and then applied clustering or network community algorithms (Ahmed & Abulaish, 2013; Miller, Dickinson, Deitrick, Hu, & Wang, 2014). Cresci, Di Pietro, Petrocchi, Spognardi, and Tesconi (2016) encoded tweet type or content as strings of symbols, then considered the longest common substring to detect coordination. By comparing the time series of accounts sampled from the Twitter streaming application program interface (API), Chavoshi et al. (2016) built an unsupervised tool called DeBot that is able to find accounts tweeting in synchrony, suggesting they are automated. Chen and Subramanian (2018) adopted a similar method to detect bots by finding accounts tweeting similar content. These unsupervised methods have the advantage of focusing on what accounts (or tweets) have in common rather than what is distinct, but also the disadvantage of having to consider quadratic numbers of account pairs in the worst case.

A recent research direction is to test the limits of current bot detection frameworks in an adversarial setting. The idea is to propose methodologies to engineer systems that can go undetected. Cresci, Petrocchi, Spognardi, and Tognazzi (2019) proposed the use of evolutionary algorithms to improve social bot skills. Grimme et al. (2018) employed a hybrid approach involving automatic and manual actions to achieve bots that would be classified as human by a supervised bot detection system. Despite the good intention of pointing to weaknesses in existing systems, this research might also inspire bot creators and give them a competitive advantage.

3 | USER ENGAGEMENT WITH BOT DETECTION TOOLS

Research efforts aimed at bot detection may be valuable to mitigate the undesired consequences of bot activity, but their efficacy is limited by two factors: limited public awareness of the bot problem, and unwillingness to adopt sophisticated tools to combat it. The success of research efforts also critically depends on the capacity to adapt to ever changing and increasingly sophisticated artificial accounts. This, in turn, depends on the ability to engage the public and collect the feedback it provides. It is therefore important to understand how the public adopts, interacts with, and interprets the results of bot detection tools. This perspective is often neglected by research on social bots. In this section, we use Botometer—a bot detection tool developed at Indiana University—as a case study to illustrate efforts aimed at understanding the needs of the users and how their feedback can help improve tools.

Botometer is based on a supervised machine learning approach (Davis et al., 2016; Varol, Ferrara, Davis, et al., 2017). Given a Twitter account, Botometer extracts over 1,000 features relative to the account from data easily provided by the Twitter API, and produces a classification score called *bot score*: the higher the score, the greater the likelihood that the account is controlled completely or in part by software, according to the algorithm. Because some of the features are based on English-language content, the bot score is intended to be used with English-language accounts. To evaluate non-English accounts, Botometer also provides a language-independent score that is produced by a classification model trained excluding linguistic features. Botometer additionally reports six subscores, each produced by a model based on a distinct subset of features. The score names refer to the feature classes: user meta-data, friends, content, sentiment, network, and timing. The subscores are provided to help users identify which features contribute to the overall score.

The earliest version of Botometer (known as "BotOrNot") became available to the public in May 2014. Free access is offered through both a web interface (botometer.org) and an API. In the past 4 years, Botometer has constantly increased its basin of adoption, and has provided data for a number of influential studies, including by Vosoughi, Roy, and Aral (2018) and the Pew Research Center (Wojcik et al., 2018). Even more importantly, Botometer is used by many regular Twitter users. Currently, it handles over a quarter million requests every day (Figure 1), while the website receives over 500 daily visits. Botometer also supports several third-party tools, including browser extensions, bot-hunting bots, and localized versions in non-English-speaking countries.

Between August and October 2018, we conducted a user experience survey with participants recruited among the visitors of the Botometer website. The survey contained two questions with required answers and a few optional ones. It also allowed respondents to enter some free-text comments. We collected usable answers from 731 participants; they are listed in Figure 2 together with the questions. A few interesting facts emerge. First, more than one third of participants
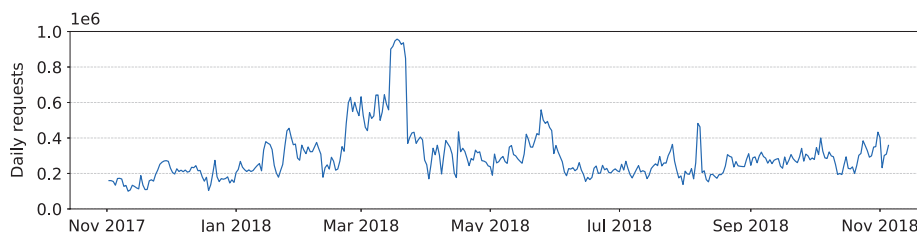

FIGURE 1 Number of daily requests handled by Botometer
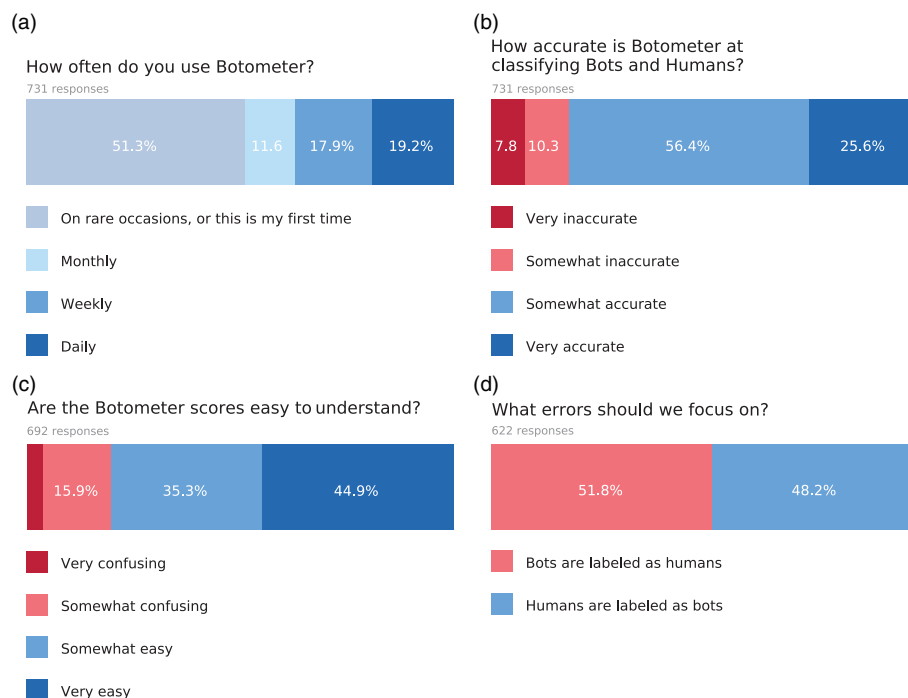
FIGURE 2 Botometer user experience survey questions and responses. Questions (a) and (b) are required; others are optional

use Botometer at least weekly, implying that detecting bots is becoming a recurring need for at least some users (Figure 2(a)). Second, over 80% of the users believe Botometer is accurate in classifying bots and humans (Figure 2(b)). Third, over 80% of the users find the bot scores presented by Botometer easy to understand (Figure 2(c)). Although these numbers are encouraging, we are aware of self-selection bias, as respondents of the survey tend to be active users of Botometer. Finally, users seem to be equally worried by false positives (humans misclassified as bots) and false negatives (bots misclassified as humans) as shown in Figure 2(d). A more nuanced picture of this last observation emerges when these results are disaggregated into frequency-of-usage categories: Figure 3 shows that occasional users care more about false positives. A typical example of usage that might elicit this concern is when an individual checks whether their own account looks like a bot. On the other hand, frequent users care more about false negatives. For example, an advertiser interested in the

number of fake followers of a celebrity paid to endorse a product may be more worried about missing bots.

Botometer is based on Random Forest classifiers (Breiman, 2001). The score produced by such a model is the fraction of trees that classify the account under examination as a bot. This interpretation of the bot score is obscure to the typical user, who is unlikely to have background knowledge in machine learning. Because a bot score is defined in the unit interval, it is tempting to interpret it as the *probability* that the account is a bot—if the score is 0.3, say, then 30% of accounts with similar feature values are bots. Another interpretation is that such an account is 30% automated. Judging from early user feedback, such incorrect interpretations were prevalent. Despite an explanation offered on the website, the confusion persisted. To address this concern and bridge the gap between classifier output and user expectations, in the next section we describe three changes in how scores are presented: a non-linear re-calibration of the model, a linear rescaling of scores to the [0,5] range, and the introduction of a new statistics called complete automation probability (CAP). These changes were incorporated into the latest version of Botometer (v3), launched in May 2018, together with other improvements described in Section 5. Users seem to appreciate these changes, as shown in Figure 4.

4 | BOT SCORE INTERPRETABILITY

4.1 | Model calibration

As previously described, the raw output of the random forest classifier is not a probability. Consider the classifier C and two Twitter accounts A and B such that $C(A) = x$ and $C(B) = y$ with $x < y$. Here we can say that account A is less likely to be a bot than account B, but we cannot say that account A is x% likely to be a bot. While this is not a problem
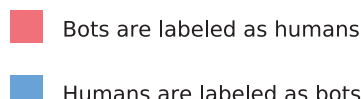

FIGURE 3 Responses to the Botometer user experience survey question about error type concerns (Figure 2(d)), grouped by frequency of usage. Respondents who use Botometer daily or weekly are considered frequent users

If you used Botometer before and after May 2018, how do the two versions compare?

297 responses

| 9.8 | 46.1% | 39.7% |

- ■ New version is much worse
- ■ New version is a little worse
- ■ New version is a little better
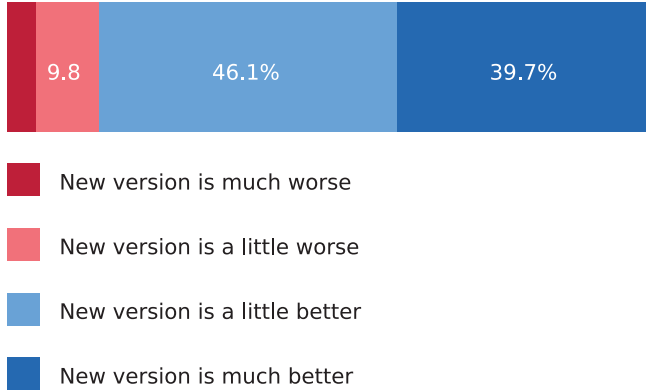- ■ New version is much better

FIGURE 4 Botometer user experience survey result: Comparison between Botometer v2 and v3

for measures like area under receiver operating characteristic curve (AUC), which quantifies the probability that a classifier will rank a randomly chosen bot higher than a randomly chosen human, it does present an obstacle when interpreting bot scores.

Proper contextualization of the raw classifier output requires comparison to the rest of the scores in the training set. For example, maximum-likelihood estimation is employed by Varol, Ferrara, Davis, et al. (2017) to find the appropriate cutoff for binary classification: the best score above which an account is classified as a bot and below which it is classified as human. This threshold approach is useful for scientists counting bots in a collection (Gorwa, 2017; Haustein et al., 2016; Shao, Ciampaglia, et al., 2018; Stella et al., 2018; Suárez-Serrato, Roberts, Davis, & Menczer, 2016), but it is not intuitive for humans evaluating a single account.

A *calibrated* classifier C' is one whose output can be interpreted as a probability, that is, one in which $C'(A) = x$ means that the classifier estimates as x% the probability that account A is a bot. As a corollary, the binary classification threshold of a well-calibrated model is 50%. Calibrated models are ideal from the point of view of interpreting their results. To generate such a calibrated model C' from our classifier C, we want a calibration function F such that $F \circ C = C'$.

To calibrate Botometer, we employed Platt's scaling (Niculescu-Mizil & Caruana, 2005), a logistic regression model trained on classifier outputs. Figure 5 presents the initial model outcomes and the calibrated scores. Note that this mapping shifts scores within the unit interval but preserves order, therefore leaving the AUC unchanged. The figure also shows reliability diagrams for raw and calibrated scores (DeGroot & Fienberg, 1983). We split the unit interval into 20 bins; an instance in the training dataset is assigned to a bin based on its predicted (raw) score. For each bin, the mean predicted score is computed and compared against the fraction of true positive cases. In a well-calibrated model, the points should align with the diagonal. Observe that the blue line on the right side of Figure 5 is steepest in the middle of the range; this is because most uncalibrated bot scores fall near the middle of the unit interval. Since users, when presented with a single uncalibrated bot score, do not know that most scores fall

into this relatively narrow range, they are misled into perceiving uncertainty about classification of most accounts. The flatter red line in the plot shows that each bin has approximately the same number of scores in the calibrated model.

As mentioned earlier, the percentage of bot accounts is estimated around 9–15% (Varol, Ferrara, Davis, et al., 2017), which means the probability of a randomly selected account being automated is low. The calibrated score cannot reflect this fact, since it is a likelihood estimation based on our roughly balanced (half humans, half bots) training datasets. In the next subsection we offer a Bayesian posterior probability that overcomes this problem. The bot score presented on the current Botometer website (v3) is the calibrated likelihood score, linearly transformed to a [0, 5] scale in order to differentiate it from the probability that an account with that score is a bot. The API, whose users tend to be more advanced, also provides uncalibrated scores.

4.2 | Complete automation probability

Users of Botometer are asking the question: "Is this account a bot, or not?" While model calibration makes the bot scores easier to interpret, a calibrated model still lacks one additional piece of information necessary to answer this "bot or not" question: the background level of bots in the population. To see why, consider that the less likely a claim, the more evidence is required to accept it. In our case, the rarer the bots, the higher the score needed to have confidence that an account is a bot. This connection between likelihood and background probability is formalized by Bayes' theorem. The "bot or not" question can be expressed mathematically as $P(\text{Bot}|S)$, the conditional probability that an account is a bot given its bot score S. Applying Bayes's rule allows us to rewrite the conditional probability as:

$$P(\text{Bot}|S) = P(\text{Bot})\frac{P(S|\text{Bot})}{P(S)}. \tag{1}$$

We call $P(\text{Bot}|S)$, the quantity we are after, the *posterior* probability. The *prior* probability $P(\text{Bot})$ is the background probability that any randomly chosen account is a bot. The $\frac{P(S|\text{Bot})}{P(S)}$ term is called the *evidence*. It compares the *likelihood* that a bot has score S, $P(S|\text{Bot})$, with the probability that any account has that score. Since this is a ratio of functions of the score S, it is not affected by transformations of S as done in model calibration. Therefore we use raw, uncalibrated scores in our calculations.

To calculate the posterior probability, we can expand the denominator in the evidence term of Equation 1:

$$P(S) = P(S|\text{Bot})P(\text{Bot}) + P(S|\text{Human})P(\text{Human})$$
$$= P(S|\text{Bot})P(\text{Bot}) + P(S|\text{Human})(1 - P(\text{Bot})). \tag{2}$$

The task is then to obtain distributions for the likelihoods $P(S|\text{Bot})$ and $P(S|\text{Human})$. The model's training data provides empirical distributions of scores for both humans and bots (Figure 6); density estimation can be used to find a probability density function likely to produce a particular empirical distribution. Binning is the simplest approach to density estimation, sometimes employing a sliding window. However, this approach proved unsuitable because quantization
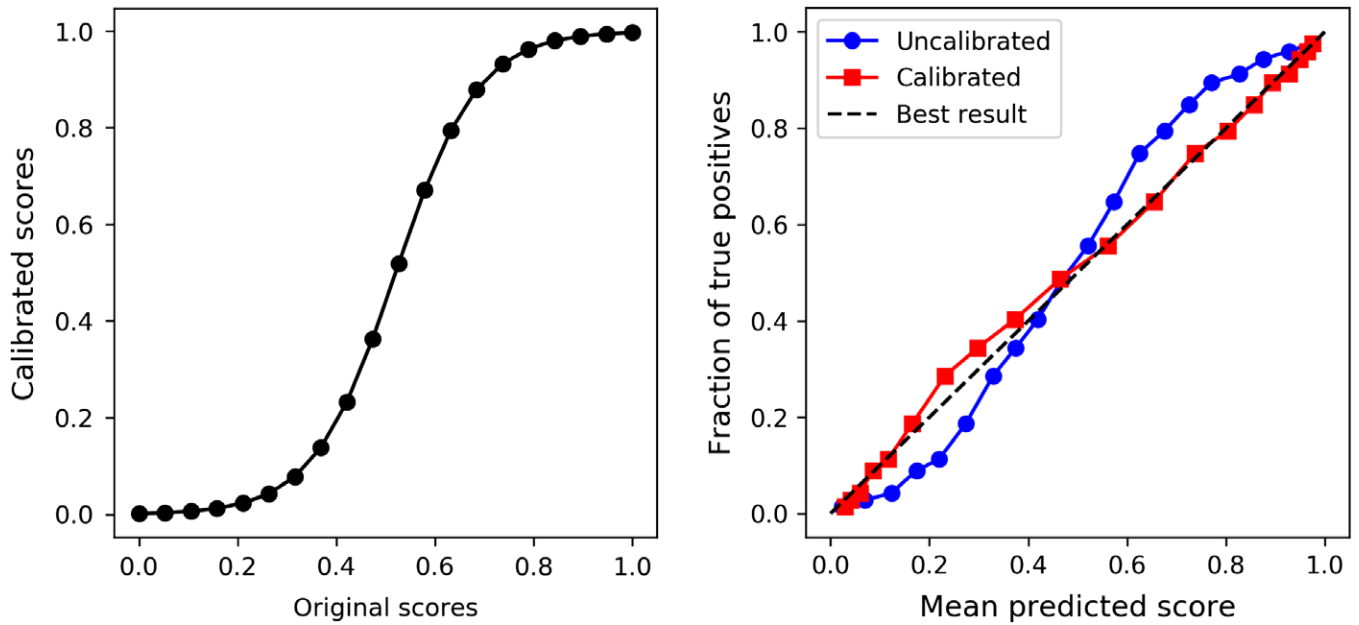
FIGURE 5 Calibration of the bot scores. The mapping function projects raw classifier outputs to calibrated scores (left). Reliability curves plot true positive rates against mean predicted scores (right). The calibrated curve indicates higher reliability because it is closer to the unbiased diagonal line
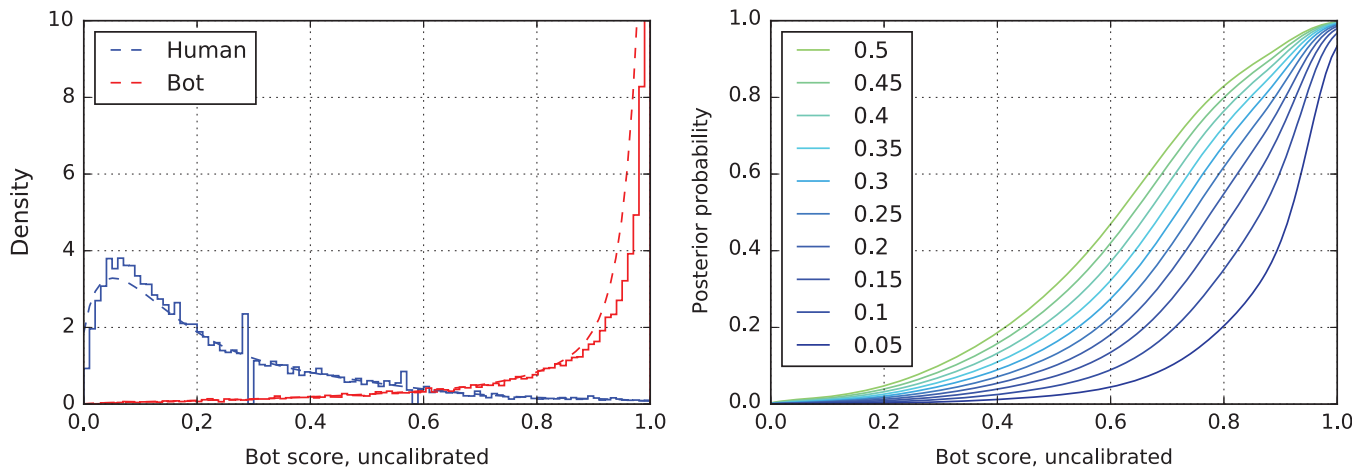

FIGURE 6 Likelihood distributions (left) and posterior probabilities (right) for the complete automation probability calculation. The left plot shows the binned, empirical bot score distribution for accounts labeled human and bot, along with dashed lines displaying the density estimate for each. On the right, posterior probability curves are calculated for several choices of the prior, $P(Bot)$

artifacts in the Botometer classifier output lead to discontinuities in the density functions.

The existence of these artifacts suggests curve-fitting or kernel density estimation (KDE) methods. KDE is widely used for generating distribution functions from empirical data, but the commonly used implementations assume unbounded support, that is, a theoretically infinite range of possible bot score values. Since the support of our distribution is necessarily the unit interval, the use of KDE in this case leads to biased estimates near the interval boundaries (Leblanc, 2012).

Seeking to avoid this systematic bias, Botometer uses curve fitting, choosing among curves that naturally admit a bounded support. One well-known candidate meeting this criterion is the beta distribution. However, three different fitting algorithms (maximum-likelihood, method of moments, and negative log-likelihood) all failed to produce

adequate fits. This is likely due to the shape of the $P(S|$ Human$)$ curve, with its mode near, but not at, zero. Ultimately we chose a lesser-known family of curves for our fit: Bernstein polynomials (Babu, Canty, & Chaubey, 2002). Implementation is straightforward as computing values of a polynomial curve is simple, even one with high degree. Experiments showed that a 40th-degree polynomial fit produced satisfactory results (Figure 6).

With likelihood curves generated from our classifier testing data, it is straightforward to calculate the evidence term in Bayes' rule. The other term is the prior $P(Bot)$, the background probability of a given account being a bot. This is necessary because the training data does not include information on how common bots are in the wild. Failure to take the prior into account leads one to overestimate the number of bots in a sample due to humans being more common than bots,

leading to false positives outnumbering false negatives. The best we can do is use an informed estimate for the prior. In general, one could allow users to estimate the prior probability; a user might have information about the proportion of bots in a particular sample. For example, a sample of suspicious accounts would require a high prior probability. Botometer uses $P(\text{Bot}) = 0.15$, corresponding to a previously published estimate of the proportion of bots on Twitter (Varol, Ferrara, Davis, et al., 2017).

Taken together, the prior and the evidence allow us to calculate the posterior probability that an account with a particular bot score S is a bot. In Botometer, the posterior is called CAP. As shown in Figure 6, this probability is generally more conservative than the calibrated bot score, reflecting the relative rarity of bots. The reported CAP estimates the probability that the account is indeed a bot, and gives end users the information they need to act on the data returned. For example, there are Twitter bots that use Botometer to call out other bot accounts. Whereas these bot-hunting bots previously had to pick a somewhat-arbitrary threshold, now they can use the CAP to make an informed decision based on their tolerance for the risk of falsely labeling accounts.

5 | UPDATING DETECTION MODELS

A supervised machine learning tool is only as good as the data used for its training. Social bots evolve rapidly, and even the most advanced algorithms will fail with outdated training datasets. Therefore it is necessary to update classification models, using newly available data as well as feedback collected from users. At the same time, one must continuously evolve the set of features that may discriminate between human behaviors and increasingly complex bot behaviors. Let us again use the Botometer case to illustrate the need for retraining bot detection models via new data and feature engineering.

In the past few years, several bot datasets collected by colleagues in academia have been included in our training data. Such data are now shared with the public at large via a platform called *Bot Repository* (botometer.iuni.iu.edu/bot-repository). Table 1 lists the datasets currently included and used for training, while Table 2 summarizes the different versions of Botometer and the growing training dataset it employed.

Retraining does not only mean adopting new datasets as they become available, but also engineering new features—or redesigning features that are already in use—to best describe the behaviors of new classes of bots. In light of journalistic evidence and other academic research, the list of features employed by Botometer was recently enriched to incorporate new ones that are designed to capture bots employed in information operations. Examples include:

- Time zones: an account would be suspicious if its profile indicates the US Eastern Standard Time zone while most of its followers appear to be in the Moscow Standard Time zone.
- Language metadata: anomalous patterns in language use and audience language preferences can be revealed by a low fraction of neighbors (friends, followers) having the same language as a target account.

TABLE 1 Training data used by Botometer. CNetS refers to the Center for Complex Networks and Systems Research at Indiana University

Dataset name	#Bots	#Human	Notes
caverlee	22,179	19,276	Honeypot-lured bots and sample human accounts (Lee et al., 2011)
varol-icwsm	826	1,747	Manually labeled bots and humans sampled by Botometer score deciles (Varol, Ferrara, Davis, et al., 2017)
cresci-17	10,894	3,474	Spam bots and normal humans (Cresci et al., 2017)
pornbots	21,963	0	Pornbots shared by Andy Patel (github.com/r0zetta/pronbot2)
celebrity	0	5,970	Celebrity accounts collected by CNetS team
vendor-purchased	1,088	0	Fake followers purchased by CNetS team
botometer-feedback	143	386	Botometer feedback accounts manually labeled by author K.-C.Y.
political-bots	62	0	Automated political accounts run by @rzazula (now suspended), shared by @josh_emerson on Twitter
Total	57,155	30,853	All datasets available at botometer.iuni.iu.edu/bot-repository

- Device metadata: we capture the types of devices and platforms used for posting tweets, as well as the entropy across different platforms.
- Content deletion patterns: highly active accounts frequently create and delete content to hijack user attention without revealing too much information on their excessive posting behavior. We incorporated features to estimate content production rates based on the account creation date, metadata on total tweets, recent

TABLE 2 Datasets and numbers of features used for training subsequent versions of Botometer models

Version (activity period)	#Features	Training datasets	Notes
v1 (1 May 2014–3 May 2016)	1,150	caverlee	Initial version of Botometer (known as BotOrNot at the time)
v2 (3 May 3 2016–11 May 2018)	1,150	caverlee, varol-icwsm	Version used in results presented by Varol, Ferrara, Davis, et al. (2017)
v3 (11 May 2018–Present)	1,209	caverlee, varol-icwsm, cresci-17, pornbots, vendor-purchased, botometer-feedback, celebrity	New features introduced to capture more sophisticated behaviors and to comply with Twitter changes.

content volume, and inter-event times between recent posts—the mismatch between these statistics can be used as a proxy of content deletion. An account that appears to have tweeted once per day on average over the past 5 years, but has generated 500 tweets per day during the last 2 days, should raise flags.

When updating a supervised learning framework for bot detection, a central question is whether, or to what extent, the retrained model can recognize new classes of bots on which the model has not been trained. To explore this question about generalization, we trained models on increasingly large and complex datasets, and measured how such models perform both on the training dataset (via cross-validation) and on datasets explicitly excluded from the training dataset. Accuracy results are shown in Figure 7. We note that the addition of datasets about unseen classes of bots in the training data does not seem to degrade model performance: the accuracy remains high across models trained on increasingly richer datasets. In other words, the present framework seems flexible enough to accommodate the learning of new examples without forgetting the old ones. Novel datasets used for testing, in general, lead to lower performance compared to when models can learn from those instances. But models tend to generalize better when classifying human accounts. Bot instances with similar data used for training the model, such as spam bots, also yield better generalization. The most significant gain is observed by adding training data about porn bots, which have very few tweets and realistic profile details. All models in this experiment are trained using the full set of 1,209 features used in the latest version of Botometer. Overall, the current model trained with the new features and on the new datasets improved the cross-validation accuracy from 0.95 to 0.97 AUC.

6 | DISCUSSION AND PERSPECTIVES

Research on online manipulation, especially social bot detection, is a rapidly growing field; its contributions are yielding numerous academic publications, tools, and datasets. As long as the arms race between bot creators and bot detection systems continue, this research will remain relevant. Beyond the recent research developments, let us make a few considerations on how the online ecosystem may change in the next several years. We focus on the perspectives of social bots and AI countermeasures.

6.1 | Future of social bots

In all likelihood, bots will become regular and acknowledged parts of the social media ecosystems of the future. Despite all the nefarious uses of social bots, there are great examples of bots providing valuable services to social media users (Mønsted, Sapieżyński, Ferrara, & Lehmann, 2017). Optimistic viewpoints expect an increase in service bots and regulations that will mitigate the impact of bots with malicious intentions.

Recent developments of deep-learning models create opportunities to employ sequence-to-sequence models for automatic conversations (Li et al., 2016; Li, Galley, Brockett, Gao, & Dolan, 2015), deep

reinforcement learning for controlling account behaviors (He et al., 2016; Mnih et al., 2015; Serban et al., 2017), and generative adversarial networks for creating fake profile images and content (Goodfellow et al., 2014; Zhao et al., 2017). Some of these methodologies have already demonstrated a capability to surpass human performance on various tasks, such as face recognition and game playing (Mnih et al., 2015; Taigman, Yang, Ranzato, & Wolf, 2014). Deep-learning approaches can also be used to detect sentiment online (Severyn & Moschitti, 2015; Socher et al., 2013; Tang et al., 2014), allowing the next generation of social bots to be more perceptive about human emotions. Successful implementation and further advances of these technologies can pose serious threats, since discriminating between organic human behavior and automation is likely to become a more and more challenging task, even for experts.

The future of social bots may be shaped by legal, ethical, and political considerations as much as technological ones. Recent events, including election manipulations discussed in Section 2.2, have led to loud public calls for platforms to apply stricter safeguards against abuse and for policymakers to enact regulations of social media platform in several countries. While platforms are in theory free to police their systems according to their own terms of service and community standards, they are under increasing pressure from politicians. On the one hand, this may lead to needed actions, such as more robust measures to counter impersonation and deception by social bots. On the other hand, political interference could also be counterproductive. For example, aggressive suspension of abusive bots that spread false or misleading political content has been criticized by some politicians as censorship (Kang & Frenkel, 2018; Tynan, 2018); even research on bot detection has been targeted by partisan attacks and labeled as suppression of free speech (Mervis, 2014).

Another set of legal questions relate to whether bots have rights (Schroeder, 2018). Are bots entitled to journalistic protections such as those that shield reporters from compelled disclosure of sources? Do they have free-speech rights? The latter question is attracting significant attention following a recent California law requiring online bots to identify themselves as such when they interact with humans, for example promoting a political candidate during elections (Sacharoff, 2018). While this law seems to strike a reasonable balance between free speech and regulation, a combination of U.S. Supreme Court decisions about the First Amendment—the free speech rights of corporations, the right against compelled speech, the right to anonymous speech, and the right to lie—may lead to the conclusion that states cannot compel entities to disclose the automated nature of online content. These types of questions are yet to be addressed by the courts, and may have huge repercussions on social media platform regulations designed to protect the information ecosystem from manipulation by bots.

6.2 | Future of AI countermeasures

So far, the technical response to increasingly sophisticated bots has largely focused on increased sophistication in detection approaches. As reviewed in Section 2.3, recently proposed tools for social bot detection employ deep learning, evolutionary algorithms, and unsupervised learning. These techniques can provide solutions for specific
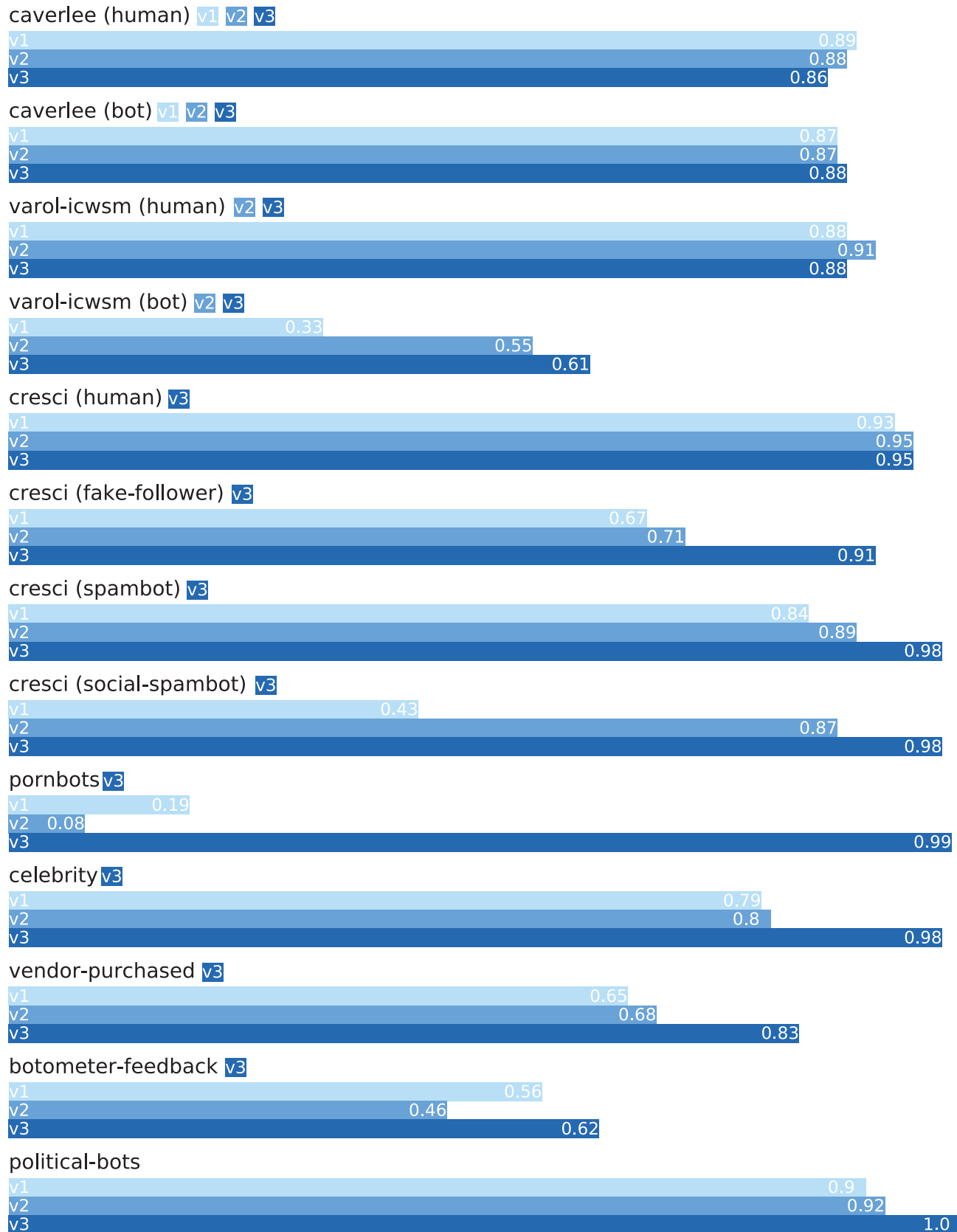
caverlee (human) v1 v2 v3
| | |
v1 0.89
v2 0.88
v3 0.86

caverlee (bot) v1 v2 v3
v1 0.87
v2 0.87
v3 0.88

varol-icwsm (human) v2 v3
v1 0.88
v2 0.91
v3 0.88

varol-icwsm (bot) v2 v3
v1 0.33
v2 0.55
v3 0.61

cresci (human) v3
v1 0.93
v2 0.95
v3 0.95

cresci (fake-follower) v3
v1 0.67
v2 0.71
v3 0.91

cresci (spambot) v3
v1 0.84
v2 0.89
v3 0.98

cresci (social-spambot) v3
v1 0.43
v2 0.87
v3 0.98

pornbots v3
v1 0.19
v2 0.08
v3 0.99

celebrity v3
v1 0.79
v2 0.8
v3 0.98

vendor-purchased v3
v1 0.65
v2 0.68
v3 0.83

botometer-feedback v3
v1 0.56
v2 0.46
v3 0.62

political-bots
v1 0.9
v2 0.92
v3 1.0

FIGURE 7   Model accuracy (percentage of instances correctly classified) based on in-sample evaluation via cross-validation and testing on out-sample datasets. In-sample evaluation is indicated by the version tags, which list the versions of Botometer in which each dataset was used for training. See Tables 1 and 2 for further information on datasets and Botometer versions, respectively

problems, such as addressing the shortage of human-labeled training data on bot activity and analyzing online sentiment. For example, we plan to employ deep-learning models to replace lexicon-based sentiment analysis in the Botometer feature extraction pipeline.

The detection of coordinated bots will continue to be a challenge. Monitoring conversations and coordinated activities among groups of accounts may reveal anomalous behaviors. While the unsupervised methods reviewed in Section 2.3 have proven effective at identifying some coordinated bots with high precision, they can only check a single feature. As a result, only narrow groups of bots that coordinate in specific ways can be captured. To keep up with all kinds of coordinated manipulation, new algorithms are needed to leverage the many possible dimensions of similarity between suspicious accounts. Another research challenge that could be relevant for the dismantling of coordinated bot attacks is the detection of promoted campaigns at their early stages (Ferrara, Varol, Menczer, & Flammini, 2016; Varol, Ferrara, Menczer, & Flammini, 2017). Future efforts should aim for a proactive approach to mitigate the effects of social bots by engaging with them early (Cresci, Petrocchi, Spognardi, & Tognazzi, 2018).

Our recent analysis points to the existence of many distinct types of social bots, such as spam bots, mention bots, cyborgs, and many others (Mitter, Wagner, & Strohmaier, 2014). Botometer employs an ensemble model to learn complex rules able to recognize multiple types of bot behavioral patterns. Detection models need to be able to rapidly adapt to changes in both human and bot behaviors. The iterative process of detecting novel accounts by expanding training datasets, presented in Section 5, is an effective approach but requires extensive and continual manual efforts. We are currently exploring ways to automatically detect novel bot behavior classes and build specialized classifiers. Using multiple models may lead to better discrimination between human behavior and specific bot strategies.

Machine learning is playing a growing role in decision making on issues of broad societal importance (Barocas & Selbst, 2016; Munoz, Smith, & Patil, 2016). As AI systems employ more sophisticated learning methodologies, model complexity increases and systems become vulnerable to overlooked biases (Kirkpatrick, 2016). Social bot detection systems are not immune to the problem of algorithmic bias. Models trained on manually annotated data might inherit the biases of annotators. Algorithmic interpretability would make it easier to spot such biases, and in fact a recent European Union regulation grants users the right to ask for an explanation of an algorithmic decision that was made about them (Goodman & Flaxman, 2016). Unfortunately, many machine learning algorithms, including those used by Botometer, do not make it easy to interpret their classifications. Deep-learning methods are even more challenging in this respect. As bot detection tools are used to make important decisions, such as whether an account should be suspended, the issue of algorithmic bias is an important direction for future research.

As discussed in Section 3, technical solutions should take into account how humans interact with them. Ultimately, the fight against online manipulation must be carried out by our entire society and not just AI researchers. Media literacy efforts may be critical in this respect (Lazer et al., 2018).

Much public debate revolves around how humans will interact with AI. Because of the significant financial and political incentives for developing benign and malicious social bots, this domain provides us with a concrete and realistic scenario to imagine the future of both collaborative and adversarial human-bot interactions.

ENDNOTES

[1] A Blueprint for Content Governance and Enforcement. www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[2] The Follower Factory. www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html

ORCID

Filippo Menczer https://orcid.org/0000-0003-4384-2876

REFERENCES

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). *Dissecting a social botnet: Growth, content and influence in Twitter*. Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW), Vancouver, Canada, 839–851.

Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. *Computer Communications*, *36*(10–11), 1120–1129.

Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2012). People are strange when you're a stranger: Impact and influence of bots on social networks. *Links*, *697*(483,151), 1–566.

Allem, J.-P., & Ferrara, E. (2016). The importance of debiasing social media data to better understand e-cigarette-related attitudes and behaviors. *Journal of Medical Internet Research*, *18*(8), e219.

Allem, J.-P., & Ferrara, E. (2018). Could social bots pose a threat to public health? *American Journal of Public Health*, *108*(8), 1005–1006.

Allem, J.-P., Ferrara, E., Uppu, S. P., Cruz, T. B., & Unger, J. B. (2017). E-cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR Public Health and Surveillance*, *3*(4), e98.

Babu, G. J., Canty, A. J., & Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, *105*(2), 377–392.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Berger, J. M., & Morgan, J. (2015). *The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter* (The Brookings Project on US Relations with the Islamic World, Analysis Paper No. 20). Washington, DC: The Brookings Institution.

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS One*, *10*(2), e0118093.

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

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2012). *Key challenges in defending against malicious social bots*. Paper presented at the Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats (LEET), San Jose, CA.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, *57*(2), 556–578.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., ... Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Chavoshi, N., Hamooni, H., & Mueen, A. (2016). *Debot: Twitter bot detection via warped correlation*. Paper presented at ICDM, Barcelona, Spain, 817–822.

Chen, Z., & Subramanian, D. (2018). An unsupervised approach to detect spam campaigns that use botnets on Twitter. arXiv preprint arXiv:1804.05232.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, *80*, 56–71.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2016). DNA-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, *31*(5), 58–64.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). *The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race*. Paper presented at the Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 963–972.

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2018). *From reaction to proaction: Unexplored ways to the detection of evolving spambots*. Paper presented at the Companion of The Web Conference, Lyon, France, 1469–1470.

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2019). On the capability of evolved spambots to evade detection via genetic engineering. *Online Social Networks and Media*, *9*, 1–16.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). *BotOrNot: A system to evaluate social bots*. Paper presented at the Proceedings of the 25th International Conference on Companion on World Wide Web, Montreal, Canada, 273–274.

De Micheli, C., & Stroppa, A. (2013). *Twitter and the underground market*. Paper presented at the 11th Nexa Lunch Seminar, Turin, Italy.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The Statistician*, *32*, 12–22.

Echeverria, J., & Zhou, S. (2017a). Discovery of the Twitter bursty botnet. arXiv preprint arXiv:1709.06740.

Echeverria, J., & Zhou, S. (2017b). *Discovery, retrieval, and analysis of the 'star wars' botnet in Twitter*. Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Sydney, Australia, 1–8.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, *22*(8). https://doi.org/10.5210/fm.v22i8.8005

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2016). *Detection of promoted social media campaigns*. Paper presented at the Tenth International AAAI Conference on Web and Social Media, Cologne, Germany, 563–566.

Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., & Galstyan, A. (2016). *Predicting online extremism, content adopters, and interaction reciprocity*. Paper presented at the Proceedings of the International Conference on Social Informatics, Bellevue, WA, 22–39.

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. (2015). *Reverse engineering socialbot infiltration strategies in Twitter*. Paper presented at the Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–32.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative adversarial nets*. Paper presented at the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, Canada, 2672–2680.

Goodman, B., & Flaxman, S. (2016). European union regulations on algorithmic decision-making and a "right to explanation". arXiv preprint arXiv:1606.08813.

Gorwa, R. (2017). In S. Woolley & P. N. Howard (Eds.), *Computational propaganda in Poland: False amplifiers and the digital public sphere* (Working Paper 2017.2). Oxford, England: Project on Computational Propaganda.

Grimme, C., Assenmacher, D., & Adam, L. (2018). *Changing perspectives: Is it sufficient to detect social bots?* Paper presented at the International Conference on Social Computing and Social Media, Las Vegas, NV, 445–461.

Gupta, A., Lamba, H., & Kumaraguru, P. (2013). *$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on Twitter*. Paper presented at the eCrime Researchers Summit (eCRS), San Francisco, CA, 1–12.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter. *Journal of the Association for Information Science and Technology*, *67*(1), 232–238.

He, J., Ostendorf, M., He, X., Chen, J., Gao, J., Li, L., & Deng, L. (2016). Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. arXiv preprint arXiv:1606.03667.

Hwang, T., Pearce, I., & Nanis, M. (2012). Social bots: Voices from the fronts. *Interactions*, *19*(2), 38–45.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100.

Jun, Y., Meng, R., & Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, *114*(23), 5976–5981.

Kang, C., & Frenkel, S. (2018). Republicans accuse Twitter of bias against conservatives. *The New York Times*.

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM*, *59*(10), 16–17.

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*(October), 312–322.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Leblanc, A. (2012). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics*, *64*(5), 919–943.

Lee, K., Eoff, B. D., & Caverlee, J. (2011). *Seven months with the devils: A long-term study of content polluters on Twitter*. Proceedings of the AAAI International Conference on Web and Social Media (ICWSM), Barcelona, Spain.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. arXiv preprint arXiv:1603.06155.

Lokot, T., & Diakopoulos, N. (2016). News bots: Automating news and information dissemination on Twitter. *Digital Journalism*, *4*(6), 682–699.

Mervis, J. (2014). An internet research project draws conservative ire. *Science*, *346*(6210), 686–687.

Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. *Science*, *338*(6106), 472–473.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, *260*, 64–73.

Mitter, S., Wagner, C., & Strohmaier, M. (2014). A categorization scheme for socialbot attacks in online social networks. arXiv preprint arXiv:1402.6288.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS One*, *12*(9), e0184148.

Morris, M., & Ogan, C. (1996). The internet as mass medium. *Journal of Communication*, *46*(1), 39–50.

Munoz, C., Smith, M., & Patil, D. (2016, May). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President.

Niculescu-Mizil, A., & Caruana, R. (2005). *Predicting good probabilities with supervised learning*. Paper presented at the Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 625–632.

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *ICWSM*, Barcelona, Spain, 297–304.

Sacharoff, L. (2018). Do bots have first amendment rights? *Politico*.

Schroeder, J. (2018). Are bots entitled to free speech? *Columbia Journalism Review*.

Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., ... Bengio, Y. (2017). A deep reinforcement learning chatbot. arXiv preprint arXiv:1709.02349.

Severyn, A., & Moschitti, A. (2015). *Twitter sentiment analysis with deep convolutional neural networks*. Paper presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 959–962.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS One*, *13*(4), e0196087.

Smith, A., & Brenner, J. (2012). *Twitter use 2012* (Pew Internet & American Life Project, 4). Washington, DC: Pew Research Center.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. Paper presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, 1631–1642.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440.

Stocking, G., & Sumida, N. (2018). *Social media bots draw public's attention and concern*. Washington, DC: Pew Research Center.

Suárez-Serrato, P., Roberts, M. E., Davis, C., & Menczer, F. (2016). *On the influence of social bots in online protests*. Paper presented at the International Conference on Social Informatics, Bellevue, WA, 269–278.

Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., et al. (2016). The DARPA Twitter bot challenge. *Computer*, *49*(6), 38–46.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). *Deepface: Closing the gap to human-level performance in face verification*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 1701–1708.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). *Learning sentiment-specific word embedding for Twitter sentiment classification*. Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Baltimore, MD, 1555–1565.

Tynan, D. (2018). Facebook accused of censorship after hundreds of us political pages purged. *The Guardian*.

Varol, O., Davis, C. A., Menczer, F., & Flammini, A. (2018). Feature engineering for social bot detection. In G. Dong & H. Liu (Eds.), *Feature engineering for machine learning and data analytics* (pp. 311–334). Boca Raton, FL: Chapman and Hall/CRC Press.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). *Online human-bot interactions: Detection, estimation, and characterization*. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), Montreal, Canada.

Varol, O., Ferrara, E., Menczer, F., & Flammini, A. (2017). Early detection of promoted campaigns on social media. *EPJ Data Science*, *6*(1), 13.

Varol, O., & Uluturk, I. (2018). Deception strategies and threats for online discussions. *First Monday*, *22*(5). https://doi.org/10.5210/fm.v22i5.7883

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2012). Social turing tests: Crowdsourcing Sybil detection. arXiv preprint arXiv:1205.3856.

Wojcik, S., Messing, S., Smith, A., Rainie, L., & Hitlin, P. (2018). *Bots in the Twittersphere*. Washington, DC: Pew Research Center.

Zhao, J., Xiong, L., Jayashree, P. K., Li, J., Zhao, F., Wang, Z., ... Feng, J. (2017). *Dual-agent GANs for photorealistic and identity preserving profile face synthesis*. Paper presented at the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, 66–76.

AUTHOR BIOGRAPHIES

Kai-Cheng Yang is a Ph.D. Student in Informatics at Indiana University, Bloomington. He received B.S. and M.S. degrees in Theoretical Physics from Lanzhou University, China in 2014 and 2017. He is interested in social network analysis using computational modeling methods, with a focus on bot detection.

Onur Varol is a postdoctoral research associate at Northeastern University in the Center for Complex Network Research. He completed his PhD in Informatics at Indiana University, Bloomington (USA). His thesis focused on analysis of manipulation and threats on social media and was awarded the 2018 University Distinguished Ph.D. Dissertation Award. He has developed the Botometer system to detect social bots on Twitter and his team ranked top 3 worldwide at the 2015 Twitter DARPA Challenge. Efforts on studying social bots have yielded publications in prestigious venues such as the International Conference of Web and Social Media (ICWSM), Nature Communications, the Web Conference, and Communications of the ACM. He is currently working on quantifying success of online personas and impact of their actions. He is also modeling user interactions, leveraging online data across multiple platforms to understand conscious and unconscious behaviors.
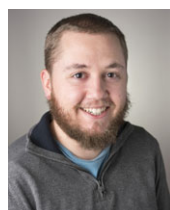
Clayton A. Davis is a PhD candidate in Informatics at the IU School of Informatics, Computing, and Engineering. He holds BS and MA degrees in Mathematics. His research interests and contributions span the development of big data platforms for social media analytics, machine learning algorithms for the detection of social bots, design of crowdsourcing platforms for collection of sensitive data, the role of online social media in large-scale social movements, and the application of network science to online public health interventions. He believes strongly in engaging the public with his research; between Kinsey Reporter, the Observatory on Social Media, and Botometer, web interfaces for Davis' projects have served millions of queries to thousands of internet users.

Dr. Emilio Ferrara is Research Assistant Professor and Associate Director of Informatics & Data Science at the USC Department of Computer Science, Research Team Leader at the USC Information Sciences Institute, and Principal Investigator at the USC/ISI Machine Intelligence and Data Science (MINDS) group. Ferrara's research interests include using AI for modeling and predicting human behavior in techno-social systems. Ferrara has published over a hundred articles on social networks, machine learning, and network science, appeared in venues like Proceeding of the National Academy of Sciences, Communications of the ACM, Physical Review Letters, and his research has been featured on all major news outlets. He was named 2015 IBM Watson Big Data Influencer, he received the 2016 DARPA Young Faculty Award, the 2016 Complex Systems Society Junior Scientific Award, and the 2018 DARPA Director's Fellowship. His research is supported by DARPA, IARPA, Air Force, and Office of Naval Research.

Dr. Flammini (PhD in Physics from the Institute of International Studies in Trieste, Italy, 1996) is Professor in the School of Informatics, Computing and Engineering at Indiana University in Bloomington. Dr. Flammini's background is in Statistical Mechanics, and, before joining Indiana University, has held research positions in different institutions, including the Cavendish Lab at the University of Cambridge and the Ralph Parson Lab at the Massachusetts Institute of Technology. His main research interests are in the area Computational Social Science. In particular, Dr. Flammini's work focuses on the empirical analysis and modeling of information diffusion on complex networks, such as the Web and online social media. Dr. Flammini's research has been sponsored by several agencies, including DARPA, NSF, the McDonnell Foundation, and Democracy Fund.

Filippo Menczer is a Professor of Informatics and Computer Science at Indiana University, Bloomington. He holds a Laurea in Physics from the Sapienza University of Rome and a PhD in Computer Science and Cognitive Science from the University of California, San Diego. Dr Menczer is an ACM Distinguished Scientist, a Fellow of the Center for Computer-Mediated Communication, a Senior Research Fellow of The Kinsey Institute, and a Board Member of the IU Network Science Institute. His research focuses on Web and data science, network science, computational social science, and science of science. His work on the spread of information and misinformation in social media has been covered in many U.S. and international news sources. Menczer received multiple service awards and currently serves as Associate Editor of the *Network Science* journal and on the editorial boards of *EPJ Data Science* and *PeerJ Computer Science*.