Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Trends in Detection and Characterization of Propaganda Bots

William Williamson and James Scrofani, *Senior Member, IEEE*

***Abstract*—Since the revelations of interference in the 2016 US Presidential elections, the UK's Brexit referendum, the Catalan independence vote in 2017 and numerous other major political discussions by malicious online actors and propaganda bots, there has been increasing interest in understanding how to detect and characterize such threats. We focus on some of the recent research in algorithms for detection of propaganda botnets and metrics by which their impact can be measured.**

***Index Terms*—propoganda bots, botnets, social media,**

I. Introduction

The study of computational propaganda has received increased attention over the past two years due to the role it played in the US 2016 Presidential elections, the UK's Brexit referendum [1], the Catalan independence vote in 2017 [2], and numerous other political discussions since 2010 [3]. It was widely reported in the US media that 'fake news' stories circulated on social media impacted the public discussion during the US 2016 Presidential campaign. Specifically, computational propaganda has been defined as the "assemblage of social-media platforms, autonomous agents, and big data tasked with the manipulation of public opinion [4]." It generally encompasses activities by dedicated, malicious human users (trolls), automated accounts (bots), and cyborgs human curated automated accounts. The primary focus of this paper is on bots. The specific role that was played by propaganda bots in the 2016 US election was explored quantitatively by Bessi and Ferrara [5] who found that almost 19% of the Twitter conversation during final months of the election was attributable to propaganda bots, and that as many as 400,000 propaganda bots may have been active. Twitter has put the number of confirmed bots at 50,258 which highlights a typical discrepancy in bot declaration criteria which we discuss below. The notoriety of these intrusions into political events seems to have spurred greater interest in researching methods for detecting and characterizing social bots, and political propaganda bots in particular.

In the early 2000's bots were primarily employed to conduct DDoS attacks and deliver malware [6]. The Marina, Conficker, and Zeus botnets are particularly well-known examples of this model. These bots had a fairly mechanistic, machine-to-machine mission. In contrast, social bots attempt to fool humans into believing that they (the bots) are other humans. In effect, they must pass a Turing test to accomplish their mission. For example, the Ashley Madison website, which advertised itself as a forum for connecting people interested in adulterous relationships, employed femme-bots to pose as real women in attempts to lure more men to the site. It is estimated that the site had 12,000 actual women participating and 70,000 bots posing as women [7].

Much, indeed most, of the current literature on social bots focuses on one specific social media platform Twitter. This is because the Twitter API provides unrivalled access to details of user interactions, user profiles, and even content [8]. It is reasonable to assume that propaganda bots are being employed to the extent they can be on other social media platforms as well, but there are significant barriers to research on these platforms and that has led to few papers addressing platforms other than Twitter. We address one example in this paper [9].

In this paper we summarize the emerging trends in detection of propaganda bots and techniques applied to characterize their impact and influence. The paper is organized as follows: in section 2, we review some of the most successful recent efforts at detecting propaganda bots. In section 3 we discuss the properties that make propaganda bots effective and means of measuring those properties. Section 4 addresses the need for detection and characterization research to stay abreast of the adversary's continual advances in sophistication of propaganda botnets. In section 5 we make our closing remarks.

II. Detection of Propaganda Bots

Detection of bots has been a subject of study for more than a decade, however the properties of botnets aimed at exerting political influence heretofore referred to as propaganda bots present particular signatures and behaviors that set them apart from other types of botnets. In this paper we focus on methods that have been applied specifically to detection of propaganda bots (or in the more general case, influence bots) operating in the social media ecosystem. As we will see, the goals of widely disseminating a message, creating 'trending' topics, and earning reputation and trust force constraints on propaganda bots that will result in user accounts that look and behave differently from the typical human users on a given social media platform.

While it is our convention to refer to all automata that are aimed at leveraging social media for the purposes of spreading a political ideology as 'propaganda bots,' it is important to note that these bots manifest several different types of approaches to spread their message. Some make no attempt to mimic human behavior and simply post news, fake news, or other information feeds. Others, like the Ashley Madison bots, have

W. Williamson and J. Scrofani are with the Department of Electrical and Computer Engineering, Naval Postgraduate School, Monterey, CA, 93943 USA, e-mail: (see https://my.nps.edu/web/ece/faculty).

URI: https://hdl.handle.net/10125/60148
ISBN: 978-0-9981331-2-6

arguably passed a Turing test in that they successfully interact with human users, engaging them in conversation.

A. Detection Metrics

When evaluating the performance of various bot detection classifiers or detection algorithms, one might think that defining a success metric should be straight forward, but this is not the case. The most commonly accepted metrics in the literature are precision and recall, and the F_1 score, which attempts to balance precision and recall. Others report in terms of true positive rate (TPR) and false positive rate (FPR), or more efficiently the area under the curve (AUC) which refers to the receiver operating characteristic (ROC) curve, which plots TPR vs. FPR. These terms are defined as follows [9]:

$$\begin{aligned} \text{TP} &= \text{number of true positive declarations,} \qquad (1)\\ \text{FP} &= \text{number of true negative declarations,}\\ \text{TN} &= \text{number of false positive declarations,}\\ \text{FN} &= \text{number of false negative declarations,}\end{aligned}$$

$$\text{TPR} = \frac{TP}{(TP+FN)},$$

i.e., percentage of bots correctly identified as bots,

$$\text{FPR} = \frac{FP}{(FP+TN)},$$

i.e., percentage of legitimate users misclassified as bots, AUC: normalized to range [0,1], where 1 indicates perfect performance,

$$\text{Precision} = \frac{TP}{(TP+FP)},$$

$$\text{Recall} = \frac{TP}{(TP+TN)},$$

and

$$F_1 = 2\,\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

While many researchers report precision as their primary metric, Morstatter, et al. [10] argue that the F_1 score is the most relevant since merely reporting precision does not properly account for false negatives, and thus many bots will go undetected. However, the operational implication of a declaration that a specific user is or is not a bot carries real world consequences. Apart from a strictly research perspective, merely declaring bot or human has no value. The value proposition is associated with the ensuing action implied by that declaration. For example, Twitter has a policy of suspending user accounts that are deemed to be bots and in recent years they have inadvertently suspended the accounts of legitimate users who were mistaken as bots [11]. This obviously has negative business consequences for a social media provider, so Twitter tends to err on the side of caution and minimizes the change of false positives at the expense of false negatives. Some authors mention this as a cause for discrepancy in assessment of level of bot activity when comparing their detection algorithm to Twitter's [10], [12].

B. Data sets

As with most supervised learning approaches, a major challenge is finding sufficient labeled data in a training set that is representative of the data that will be encountered 'in the wild.' There are several publicly available data sets for detection and/or characterization of propaganda bots. We briefly describe some of the more widely used ones here.

DARPA conducted a Twitter Bot Challenge [12] in 2015 and created a synthetic data set in which the 39 bots were of course known with 100% certainty. While this data is no longer available to the public, for the duration of the challenge it provided an excellent ground truth, making possible a head-to-head comparison of the techniques used by the six competing teams. We describe those results in the next section.

Lee, et al. [13] set up 60 honeypot Twitter accounts to lure bots to interact, and over the course of seven months in 2011, collected 23,869 bot followers. The rationale in structuring the honeypot accounts is that by tweeting frequently, tweeting randomly chosen content, frequently using @ reply messages to each other and tweeting often on trending topics, the honeypot accounts would be exhibiting behavior attractive to bots, but not of interest to humans. Twitter eventually suspended 23% of the harvested accounts, with the suspensions lagging the honeypot detections by an average of 18 days. The remaining 77% of the accounts were also classified as bots by the researchers based on cluster analysis which revealed 4 distinct classes of behavior. This data set of bot accounts is publicly available.

The University of Indiana [14] combined 15,000 of the bot accounts from Lee's honeypot data with 16,000 human accounts, which they verified by manually annotating 3,000 of the accounts. The human accounts were obtained via Twitter's API over a three-month period starting in October 2015. The annotated accounts were drawn 300 from each decile of the team's initial bot classification score. This merged data set is publicly available and researchers can compare their classification results to those published by the University of Indiana on this data set in the above reference.

A data set was collected by a University of Arizona and Carnegie-Mellon University team during the Arab Spring movement in Libya from Feb 2011 to Feb 2013. [10] Tweets were collected via the Twitter API and selected by a set of keyword hashtags provided by subject matter experts. Initial labeled data was based on Twitter account suspensions and deletions, which accounted for 7.5% of the harvested accounts. The authors assert that this percentage is low due to Twitter's tendency to err on the side of caution when suspending accounts. The same team employed a honeypot approach which drew 3,602 bots and combined this with a set of 3,107 human accounts. The human accounts were obtained by beginning with a set of 10 manually verified human users and collecting accounts that they were following, assuming humans generally don't follow malicious bots. Both of these data sets are available.

C. Detection and classification techniques

Most of the literature discusses bot detection as a binary classification task. [10], [14], [12], [15], [16], [17] Indeed, the University of Indiana has made their detection algorithms available through a public web interface to an application titled BotOrNot [18]. This tool (now known as Botometer and available at https://botometer.iuni.iu.edu/) returns a bot score when a query is submitted for a given username.

Most of the literature to date has documented research on feature-based binary classification. The set of features varies between approaches as does the number of features used. At one extreme, Varol, et al. [14] incorporate 1,150 features in their classification algorithm that forms the basis for the BotOrNot engine. At the other end of the spectrum, Duh, Rupnik, and Korosak [1] claim comparable performance with a single feature – the "tweetstorm" – derived from the temporal properties of Twitter behavior. Ironically the latter study used the BotOrNot tool to benchmark 'ground truth.'

Most researchers seem to have focused on four categories of features: user profiles, network connections, activity/behavior, and semantic content. These features were used to some extent by each of the top 3 teams in the DARPA Twitter Bot Challenge [12]. Within these broad categories there are some features or combinations of features which consistently prove to be predictive. User profile data has proven useful in identifying simple bots. Cases where there are obvious mismatches between name and gender, or where a profile picture or location are missing have proven to be strong indicators of a bot [19]. Other user information includes type of platform used to access the site. Bots rarely come from mobile devices, while humans often do, so platform is also a good indicator [9].

The social ties formed by bots usually are a very strong feature differentiating human from bot accounts. Several of the teams we studied had good success with variations of features based on the ratio of followers to followees [20], [21], [22]. Bots tend to follow other bots, and humans tend to follow other humans [14]. The impact of propaganda bots is realized when humans follow or retweet bots.

One of the strongest subcategories of the behavior and activity features is temporal behavior. Because propaganda botnets are attempting to draw attention to their topic, they frequently act in temporally correlated ways that create a detectable signature. [1], [14], [23], [24]. Other behavioral features include following circadian cycles for appropriate time zones, diversity in rate and topic of tweets, etc.

Some researchers also rely strongly on semantic content and sentiment for classification [17]. Reliance on content has two weaknesses it is not language agnostic, so the applicability of the tool is limited to the languages known by the research team [17], [20]. Further, the keywords and hashtags must be updated for each emerging topic of study. Still, it is common practice to filter based on topical keywords when assembling an initial data set associated with a particular political event or movement [23].

The metrics reported for most studies usually indicate good precision, however, the data sets often vary from experiment to experiment. This may be indicative of having tuned classification parameters to a specific topic domain. (some researchers claim they outperform other techniques on the selected data set, when the competitor's technique claims higher precision on their respective data set). At present there is no widely accepted set of labeled data that has the community wide acceptance that the ImageNet data set [25] does for the machine vision community.

The general process learned from the DARPA Twitter Bot Challenge is followed by all of the above-mentioned classification approaches. That process can be summarized as [12]:

1) Manually verify a set of label data. This can be done by subject matter heuristics, or by using blocked or banned accounts (implying trust in the social media provider's algorithms and experts.)
2) Clustering and outlier detection. This helps determine most salient features for classification. Standard clustering algorithms are usually used: k-NN and k-means being most common.
3) Classification and outlier analysis. Standard machine learning algorithms are normally used.

So far we have discussed feature-based classifiers that, for the most part, rely on traditional machine learning classification algorithms: regression, decision trees, support vector machines, and random forest (RF). Most of the teams whose research we have discussed above evaluate all of these methods and report which technique worked best with their feature set. In most of the cases we studied the random forest approach provides the best results for example, the Indiana University Team [14] tried each of the standard classifier routines in the scikit-learn library and found that RF was the most accurate algorithm for their features, yielding a 0.95 AUC for the initial dataset and an 0.85 for the expanded data set with FPR of 0.15 and FNR of 0.11.

In addition to these more traditional approaches, several novel approaches have been reported recently, and we will mention a few here. In their study of BREXIT related propaganda, Duh, Rupnik, and Korosak [1] employ an Ising spin-glass model borrowed from physics to measure correlation of temporal activity. Cai, Li, and Zhang [26] report the first application of deep learning to social bot detection using the Arab Spring honeypot data [10] and report competitive results with an F1 score of 87%. Two unsupervised learning approaches have also been reported. The first is the DeBOT algorithm [24], which employs dynamic time warping as a robust means of correlating temporal behavior. In addition, DeBOT is capable of running in near real time, as opposed to forensically. Another unsupervised approach is found in the Associative Affinity Factor Analysis (AFAA) algorithm developed by Sadiq, et al. [27] which used patterns from tweets by celebrities who were self-declared supporters of Clinton or Trump in the 2016 election as data to create behavior models for their own 3,000 bot army. Their bots injected data into the discussion. An unsupervised learning algorithm based on Multi-Factor Analysis was then applied for bot detection, and hierarchical clustering was then used to separate the bots into factions (Democrat or Republican) based on content. Finally, we mention a very brief concept paper by Cresci, et al. [28] which suggests the idea of using genetic algorithms to generate likely variations in botnet design. His premise is that the diversity of potential botnets would provide researchers with opportunities to anticipate the development of more sophisticated botnets, reducing the risk of technical surprise. The author notes that such approaches have been helpful in spam filtering.

Lastly we mention the only paper which we considered that looked at a social media platform other than Twitter, namely Sina-Weibo, the Chinese microblogging tool that is functionally similar to Twitter. Dan and Jieqi [9] obtained a dataset of about 3,000 humans and 3,000 bots directly from Sina-Weibo. Several features were selected, and one feature at a time was removed to measure each feature's importance to classification. The most salient being: fan/follower ratio, retransmission rate (like retweet), and platform type (mobile or PC). Both random forest and C4.5 classifier algorithms were tested and random forest slightly outperformed with an F1 score of 0.94. It is easy to see that the methodology and results are similar to those obtained for Twitter.

III. Characterizing Influence

In order to fully appreciate the impact of computational propaganda, it is not enough to simply quantify the proportion of bots participating in political discourse, or even to estimate the volume of traffic they are generating. These measures are certainly useful and are usually a necessary first step in characterization. However, they do not provide a complete picture of actual influence.

Influence could manifest itself in a variety of ways affecting the character of the debate in a social media forum by moving the debate towards or away from a given position or exacerbating polarization as documented in the case of the Catalan election [2], or obfuscating real issues by hijacking the discussion and leading it off track (smokescreening) as seen in the Syrian uprising [29]. Of even more concern is the prospect that influence of online bots will manifest itself in real world actions as opposed to merely affecting the character of the online discussion. Arguably, this is already happening. ISIS has already run online propaganda campaigns aimed at encouraging youth to become radicalized [13] in support of their recruitment efforts.

Several researchers have tested the efficacy of influence tactics by creating their own influence bot nets and measuring their effects [27], [30]. To date the results have varied widely. Again, there is no canonical data set on which tools can be compared. In one study, Murthy, et al. [31] tasked 12 human users to send Tweets during the UK Prime Minster's online Q&A session in an attempt to influence the discussion. Half of the human users had bots attached to follow their accounts and retweet. The experiment did not result in a significant change in topic. The failure to influence was assessed to be due to the fact that the users and bots were all new accounts created just before the experiment and therefore lacked established connectivity. At the opposite end of the spectrum, Aiello [30] was able to create a bot on a book enthusiast social media forum which was designed to have no distinctive human features, and no preexisting trust relationships, yet this bot quickly became one of the most popular users on the forum. The success was attributed to a quirk in the design of the site. The site left a record in every user's 'guestbook' of the other users who had visited their page. Since the bot was programmed to crawl the site and visit all pages, every user on the site saw the bot as a guest on their page. Curiosity drove many users to look at the bot's page, and at this point the popularity algorithm (similar to Page Rank) determined that the bot must be extraordinarily popular and began recommending it to human users.

A. Metrics of Influence

Social Media companies such as Twitter typically compute their own influence scores. Twitter uses 'Klout.' There is a commercial advantage to being able to offer product incentives, such as test drives of new cars, to influential users. Google's Page Rank algorithm is another widely used means of calculating centrality in a network.

Bots have a long tradition of exploiting these internal metrics to their advantage. In the online gaming world, some users are willing to pay real-world money to purchase in-game advantages [15], so there is financial incentive to develop bots who will play the game to score in-game items or advantages to sell to human users.

An intuitively satisfying measure of influence, used by Abokhodair, et al. [29] is to compute a cumulative rank sum of the retweets among the top 100 retweets by human users that were originally tweeted by bots. This measure gives a direct indication of the extent to which humans have read bot-generated content and thought highly enough of their messages and their credibility that they shared them with other humans. Similarly, a study of the effectiveness of ISIS online propaganda [21] looked for human users who were not tweeting or retweeting ISIS content at the beginning of their 18-month study, but were at the end, and found that 25,538 ISIS supporters created 54,358 infected users' during the period of the study. If measured as a contagious disease, the contagion rate is similar to SARS, Ebola, or HIV/AIDS.

IV. Countermeasures to Detection

With any technology which becomes weaponized, there is an inevitable race between development of game changing technology, developing countermeasures to that technology, and developing counters to the countermeasures. Computational propaganda is no different. There are at least three distinct generations of social bots in existence at this time [23]:

1) Simple bots – few behaviors and unconvincing profiles.

2) Convincing ‘sock puppets’ – hundreds or thousands of accounts following each other, to appear as a community of like-minded citizens
3) Advanced ‘sock puppets’ – like gen 2, but precisely targeted to infiltrate existing communities of interest and inject their own messages in order to reframe the discussion in their terms. Persona’s may exist across multiple social media platforms.

We have described methods of detection that involve identifying characteristic features which differentiate the behavior of propaganda bots from the behavior of human engaging in political dialogue on social media It would stand to reason that if malicious actors want to better hide their bots, they would adopt more human-like behavior. In fact, there is evidence that this type of countermeasure has been employed. [29]. The Indiana University team saw a 10% drop in AUC simply by the addition of their manually annotated data, suggesting that feature weights need frequent adjustment to adapt to even normal behavior pattern changes [14].

However, there is a limit to how far an influence bot can disguise itself and still have the desired impact [23]. In order to promote a topic vociferously enough to ensure that the ranking algorithms elevate that to a trending topic, a propaganda botnet must create a sufficiently high number of messages in a short amount of time. To do so would require either a large number of bots acting in an abnormally coherent fashion, or a smaller number of bots tweeting at abnormally high rates. As we have seen, [1], [23], [24] both of these tactics produce temporal signatures that are detectable to some degree by state of the art classifiers.

Similarly, a botnet must have sufficient connectivity with other users in the social network to promulgate its message. One study characterized this type of information flow as a ‘viral cascade’ [32] which generates a fairly distinct signature.

There is reason to believe that propaganda bots will continue to develop more sophistication, however. A recent in-depth study of a specific botnet which operated for 35 weeks during the Syrian uprising sheds light on the construction and operation of an actual 3rd generation propaganda botnet [29]. There are two very interesting features of this botnet. First is that the botnet grew over time, rather than the entire network bursting onto the scene simultaneously, as in Murthy’s experiment [31]. The network started with two bots, grew to about 80 by the end of the 35 weeks and had a total of 130 bots participate during the lifespan of the network before Twitter shut it down in week 35. The other interesting property is that there were several discernible types of bots in the network. Some were only run for two weeks and were thought to be prototypes. The others can be divided into three classes.

1) Core bots – the network began with two in week one and grew to 64 by week 28. These had a tweet rate of rate of 1 tweet/1.8 min, or 1800/week. About 50% are retweets of other core bots and the generator bot.
2) Peripheral bots – There were 15 of these. Their behavior made it hard to classify them as bot or human. Their tweet rate was less than 70/week.
3) Generator bot – there was only one of these. It created original tweets at a rate of 2,100/week.

The diversity of this botnet’s ecosystem is likely a harbinger of things to come as more experience is gained by malicious actors.

V. Conclusions

The phenomena of employing bots to spread propaganda is quite new, and the tactics are still developing. Likewise, the techniques used to detect and mitigate these bots are still in the formative stages. The successes today have largely been in forensically determining the impact of computational propaganda.

The field is beginning to converge on certain best practices and best features for detection and characteristics of propaganda botnets, but results are still inconsistent, and as bots continue to become more sophisticated, the task of mitigating them, and doing so in real time will become significantly more challenging.

All but one of the studies discussed in this paper have used Twitter as their source of data. The Twitter API makes it exceptionally easy for researchers to gather all the features they need. Looking only at Twitter obviously introduces selection bias when trying to infer influence trends in the general population. This consideration should provide incentive to examine other social media platforms. However, other platforms are more challenging for instance Facebook does not generally make content available, and WhatsApp now provides end-to-end encryption for its users. With the recent changes in EU privacy laws, and Facebook CEO Mark Zuckerburg called in front of Congress over privacy abuses in the 2016 election, it is reasonable to expect a more restrictive environment in the future. Indeed in early 2018 a number of initiatives to fund further research into combating fake news a.k.a. computational propaganda have emerged [33]. It is uncertain what the outcome will be as academic researchers find open access to social media platforms becoming more restrictive, while major social media companies find more incentive to fund research for their own proprietary solutions.

References

[1] A. Duh, M. Slak Rupnik, and D. Korošak, “Collective behavior of social bots is encoded in their temporal twitter activity,” *Big Data*, vol. 6, no. 2, pp. 113–123, 2018.

[2] M. Stella, E. Ferrara, and M. De Domenico, “Bots sustain and inflate striking opposition in online social systems,” *arXiv preprint arXiv:1802.07292*, 2018.

[3] N. Dutta and A. K. Bhat, “Use of social media for political engagement: A literature review,” in *Fourteenth AIMS International Conference on Management, MICA, Ahmedabad*, 2016.

[4] S. C. Woolley and P. N. Howard, “Automation, algorithms, and politics— political communication, computational propaganda, and autonomous agentsintroduction,” *International Journal of Communication*, vol. 10, p. 9, 2016.

[5] A. Bessi and E. Ferrara, “Social bots distort the 2016 US presidential election online discussion,” *SSRN*, vol. 21, no. 11, 2016.

[6] M. Fossi, D. Turner, E. Johnson, T. Mack, T. Adams, J. Blackbird, S. Entwisle, B. Graveland, D. McKinney, J. Mulcahy *et al.*, “Symantec global internet security threat report,” *White paper, Symantec enterprise security*, vol. 1, 2009.

[7] A. Newitz, “Ashley madison code shows more women, and more bots,” 2015.

[8] Twitter, Inc., “Tweet data dictionaries,” https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects, accessed 25 May 2018.

[9] J. Dan and T. Jieqi, “Study of bot detection on sina-weibo based on machine learning,” in *Service Systems and Service Management (ICSSSM), 2017 International Conference on*. IEEE, 2017, pp. 1–5.

[10] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, “A new approach to bot detection: striking the balance between precision and recall,” in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2016, pp. 533–540.

[11] A. H. Wang, “Detecting spam bots in online social networking sites: a machine learning approach,” in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2010, pp. 335–342.

[12] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer *et al.*, “The darpa twitter bot challenge,” *arXiv preprint arXiv:1601.05140*, 2016.

[13] K. Lee, B. D. Eoff, and J. Caverlee, “Seven months with the devils: A long-term study of content polluters on twitter.” in *ICWSM*, 2011, pp. 185–192.

[14] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” *arXiv preprint arXiv:1703.03107*, 2017.

[15] J. Oh, Z. H. Borbora, D. Sharma, and J. Srivastava, “Bot detection based on social interactions in mmorpgs,” in *2013 International Conference on Social Computing*. IEEE, 2013, pp. 536–543.

[16] B. Sengar and B. Padmavathi, “P2p bot detection system based on map reduce,” in *Computing Methodologies and Communication (ICCMC), 2017 International Conference on*. IEEE, 2017, pp. 627–634.

[17] M. Kantepe and M. C. Ganiz, “Preprocessing framework for twitter bot detection,” in *Computer Science and Engineering (UBMK), 2017 International Conference on*. IEEE, 2017, pp. 630–634.

[18] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, “Botornot: A system to evaluate social bots,” in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 273–274.

[19] E. Van Der Walt and J. Eloff, “Using machine learning to detect fake identities: Bots vs humans,” *IEEE Access*, vol. 6, pp. 6540–6549, 2018.

[20] M. Kaya, S. Conley, and A. Varol, “Visualization of the social bot’s fingerprints,” in *Digital Forensic and Security (ISDFS), 2016 4th International Symposium on*. IEEE, 2016, pp. 161–166.

[21] E. Ferrara, “Contagion dynamics of extremist propaganda in social networks,” *Information Sciences*, vol. 418, pp. 1–12, 2017.

[22] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier, “When social bots attack: Modeling susceptibility of users in online social networks,” *Making Sense of Microposts (# MSM2012)*, vol. 2, no. 4, pp. 1951–1959, 2012.

[23] C. Francois, V. Barash, and J. Kelly, “Measuring coordinated vs. spontaneous activity in online social movements,” 2018.

[24] N. Chavoshi, H. Hamooni, and A. Mueen, “Debot: Twitter bot detection via warped correlation.” in *ICDM*, 2016, pp. 817–822.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[26] C. Cai, L. Li, and D. Zengi, “Behavior enhanced deep bot detection in social media,” in *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on*. IEEE, 2017, pp. 128–130.

[27] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, “Aafa: Associative affinity factor analysis for bot detection and stance classification in twitter,” in *Information Reuse and Integration (IRI), 2017 IEEE International Conference on*. IEEE, 2017, pp. 356–365.

[28] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, “From reaction to proaction: Unexplored ways to the detection of evolving spambots,” in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 1469–1470.

[29] N. Abokhodair, D. Yoo, and D. W. McDonald, “Dissecting a social botnet: Growth, content and influence in twitter,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 839–851.

[30] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, “People are strange when youre a stranger: Impact and influence of bots on social networks,” *Links*, vol. 697, no. 483,151, pp. 1–566, 2012.

[31] D. Murthy, A. B. Powell, R. Tinati, N. Anstead, L. Carr, S. J. Halford, and M. Weal, “Automation, algorithms, and politics— bots and political influence: A sociotechnical investigation of social network capital,” *International Journal of Communication*, vol. 10, p. 20, 2016.

[32] E. Shaabani, R. Guo, and P. Shakarian, “Detecting pathogenic social media accounts without content or network structure,” in *Data Intelligence and Security (ICDIS), 2018 1st International Conference on*. IEEE, 2018, pp. 57–64.

[33] G. Haciyakupoglu, J. Y. Hui, V. Suguna, D. Leong, and M. F. B. A. Rahman, “Countering fake news: A survey of recent global initiatives,” 2018.