Digital Weapons in Social Media Manipulation Campaigns

Felipe Cardoso, Luca Luceri, Silvia Giordano
University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Manno, Switzerland
felipe.cardoso@supsi.ch, luca.luceri@supsi.ch, silvia.giordano@supsi.ch

Abstract

In the last few years, social media have been weaponized to spread false information and affect users' opinions. Bots represent one of the main actors associated with manipulation campaigns and interference operations on social media. Uncovering the strategies behind bots' activities is of paramount importance to develop computational tools for their detection. In this paper, we propose to explore bots' behavior by inspecting the digital weapons (i.e., sharing activities) they utilize to spread content and interact with humans in the run-up to the 2018 US Midterm Election. We observe that bots strategically mimicked the human temporal activity (since several months before the election) and balanced their interaction among human and bot population. Human response to bots' deceptive activity is alarming: One in three retweets performed by humans comes from a bot-generated content.

Introduction

False news, fake accounts, low-credibility sources, state-sponsored operators, and so on so forth: The online ecosystem landscape appears loaded of threats and malicious entities disposed to undermine the integrity of social media discussion. Among those, bots (i.e., software-controlled accounts) and trolls (i.e., human operators often state-sponsored) have been recognized as the main responsible actors of manipulation and misinformation operations in diverse contexts (Chen, Lerman, and Ferrara 2020; Yang, Torres-Lugo, and Menczer 2020; Nizzoli et al. 2020; Ferrara 2015). The malicious activity of such actors received enormous resonance in the political domain, where the risk of manipulating public opinion creates concerns for the integrity of voting events (Chen et al. 2020; Stella, Cristoforetti, and De Domenico 2019; Badawy, Lerman, and Ferrara 2019). Hence, developing computational tools to detect orchestrated campaigns is of paramount importance for the health of online platforms and represents a challenging task (Ferrara et al. 2016b). In the last few years, researchers offered several approaches to identify bots (Varol et al. 2017; Yang et al. 2019; Cresci et al. 2019; Mazza et al. 2019), while solutions for unveiling the activity of trolls have been recently proposed (Luceri, Giordano, and Ferrara 2020; Addawood et al. 2019). However, social media abuse persists and the online ecosystem still presents a mix of organic and malicious users (Luceri et al. 2019b; Im et al. 2019).

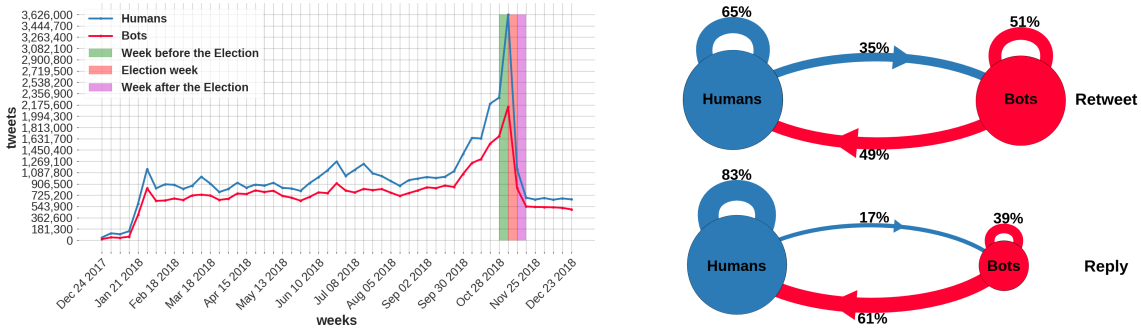
Table 1: Account and tweet statistics

Number (%)	Bots	Humans
accounts	184,545 (19.57%)	758,581 (80.43%)
tweets	40,273,442 (43.2%)	52,939,091 (56.8%)
original tweets	5,677,142 (37.08%)	9,635,364 (62.92%)
retweets	32,746,675 (45.37%)	39,427,132 (54.63%)
replies	1,849,625 (32.30%)	3,876,595 (67.70%)

It is, therefore, of fundamental importance to carry on research to uncover both the strategies and the driving forces behind the activity of malicious actors on social media with the objective of advancing their detection. Along this research direction, in this paper, we inspect the activity of bot accounts on Twitter during the year approaching the 2018 US Midterm Election. In particular, we aim to explore the strategies developed by automated accounts for spreading information and interacting with social media users, while resembling human appearance and avoiding detection. Specifically, we focus on the sharing activities that Twitter users can employ to create content (i.e., *original tweets*), reshare others' tweets (i.e., *retweets*), and respond to others' tweets (i.e., *reply*). We investigate how and to what extent bots use such *digital weapons* over time to identify cues that might enable an early detection of orchestrated campaigns.

Twitter Data and Accounts

To perform our analysis, we leverage the dataset published in (Wrubel, Littman, and Kerchner 2019), which gathers election-related messages shared on Twitter during the year (from January to December) of the Midterm Election (November 6, 2018). From the set of released tweet IDs, we collect 126M tweets by querying the Twitter API. Given the massive amount of accounts to be classified (bot vs. human), we consider to narrow the analysis to the tweets published by the set of $\sim$1M accounts classified in (Luceri et al. 2019a; Deb et al. 2019) by means of Botometer (https://botometer.iuni.iu.edu/). In our dataset, we recog-



(a) Number of shared tweets over time

(b) Interaction between humans and bots

Figure 1: Sharing activities of bot and human accounts

nized 943k of these accounts, which were responsible for about 93M tweets, whose details are shown in Table 1.

Results

In this Section, we evaluate how bots strategically perform their activity over time by examining their strategies to spread information and interact with other users. In Table 1, we report the number (and percentage) of shared tweets by bot and human accounts. Humans represent the largest portion (80%) of the accounts and shared more content (about 57% of the tweets) with respect to the bot population (20% of the accounts and about 43% of the tweets). However, bots were more active than humans: On average, a bot shared three times the number of tweets shared by a human user (218 vs. 70 tweets/account), which is consistent with bots' purpose of flooding online platforms with a high volume of tweet (Bessi and Ferrara 2016). In particular, retweeting is the most used sharing activity for both humans (74% of the time) and bots (81% of the time). As expected, bots heavily rely on the retweet action as it represents a simpler operation for automated accounts with respect to the creation of an original tweet or a reply, which requires the usage of more sophisticated techniques based on natural language processing and understanding (Ferrara 2019). This is also in line with previous findings (Luceri et al. 2019b), further confirming that retweets have been employed as bots' favorite digital weapon to resonate messages and create an illusion of consensus within online communities (Ferrara et al. 2016a).

In Fig. 1a, we show the number of tweets shared weekly by humans and bots over all the period under analysis. Interestingly, bots followed a similar temporal pattern of human users over the whole year, which might suggest that bots attempted to resemble human behavior by strategically mimicking their temporal activity (Luceri et al. 2019b). This is also true for the disentangled sharing activities, whose analogous plots are not shown in the interest of space. Indeed, the number of tweets shared by human and bot over time positively correlate ($\rho > 0.97$, p-value < 0.001) for each kind of activity (i.e., original tweets, retweets, replies). Although original tweets and replies require advanced AI techniques, the volume (and temporal patterns) of such activities performed by bots approaches human activities volume (see Table 1), especially if we consider that humans outnumber bots. This suggests that also more sophisticated bots operated along with less advanced *spammers*. Also, as expected, the content shared by both classes of account increases as the election approaches, and a peak of activity is noticeable during the election week. However, it is worth noting how bots started emulating human activity since the beginning of the year, further confirming how early detecting coordinated campaigns is a challenging task. Finally, in Fig. 1b, we show how bots and humans interact with each other in terms of retweets and replies. Node size is proportional to the percentage of accounts in each group for a given sharing activity, while edge size is proportional to the percentage of interactions between each group. We consider each group separately, thus, the edge size gives a measure of the group propensity to interact with the other groups. Interestingly, bots balanced their interactions with humans (49% of the time) and within the bot population (51% of the time), confirming bots' strategy to amplify messages to elicit an illusion of public consensus. A similar but less balanced pattern can be appreciated for replies, which might suggest that bots mainly use reply to engage with humans. By considering humans interaction, it can be noticed that about one in three retweets performed by humans comes from a bot-generated content, which represents an alarming signal for the spread of misinformation and its impact on people's beliefs.

Conclusion

Bots represent one of the most recognized threats for the integrity of social media. In this paper, we examined how bots strategically perform their online activity over time during the year approaching the 2018 US Midterm Election. We observed how bots reproduce humans' temporal patterns for every sharing activity since the year before the election and we recognized their propensity of flooding online platforms with retweets. Also, we noticed the usage of increasingly advanced bot accounts capable of creating tweets and engaging with humans in conversations. Finally, our analysis presents an alarming finding: One in three posts re-shared by humans is an original content created by bots. In our future endeavors, we aim to extend this analysis to further characterize bots' strategy over time.

References

Addawood, A.; Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *ICWSM*, 15–25.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who falls for online political manipulation? In *WWW*, 162–168.

Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 us presidential election online discussion. *First Monday* 21(11).

Chen, W.; Pacheco, D.; Yang, K.-C.; and Menczer, F. 2020. Neutral bots reveal political bias on social media. *arXiv preprint arXiv:2005.08141*.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. #covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

Cresci, S.; Petrocchi, M.; Spognardi, A.; and Tognazzi, S. 2019. Better safe than sorry: an adversarial approach to improve social bot detection. In *Proceedings of the 10th ACM Conference on Web Science*, 47–56.

Deb, A.; Luceri, L.; Badawy, A.; and Ferrara, E. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. *arXiv:1902.00043*.

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

Ferrara, E.; Varol, O.; Menczer, F.; and Flammini, A. 2016b. Detection of promoted social media campaigns. In *Tenth International AAAI Conference on Web and Social Media*, 563–566.

Ferrara, E. 2015. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* (Spring):4.

Ferrara, E. 2019. The history of digital spam. *Communications of the ACM* 62(8):82–91.

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2019. Still out there: Modeling and identifying russian troll accounts on twitter. *arXiv:1901.11162*.

Luceri, L.; Deb, A.; Badawy, A.; and Ferrara, E. 2019a. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of the 2019 World Wide Web Conference*.

Luceri, L.; Deb, A.; Giordano, S.; and Ferrara, E. 2019b. Evolution of bot and human behavior during elections. *First Monday* 24(9).

Luceri, L.; Giordano, S.; and Ferrara, E. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. *Proceedings of the 2020 International Conference of Web and Social Media (ICWSM)*.

Mazza, M.; Cresci, S.; Avvenuti, M.; Quattrociocchi, W.; and Tesconi, M. 2019. Rtbust: exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, 183–192.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; Tesconi, M.; and Ferrara, E. 2020. Charting the landscape of online cryptocurrency manipulation. *arXiv preprint arXiv:2001.10289*.

Stella, M.; Cristoforetti, M.; and De Domenico, M. 2019. Influence of augmented humans in online interactions during voting events. *PloS one* 14(5).

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Int. AAAI Conference on Web and Social Media*, 280–289.

Wrubel, L.; Littman, J.; and Kerchner, D. 2019. 2018 U.S. Congressional Election Tweet Ids.

Yang, K.-C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies* e115.

Yang, K.-C.; Torres-Lugo, C.; and Menczer, F. 2020. Prevalence of low-credibility information on twitter during the covid-19 outbreak. *arXiv preprint arXiv:2004.14484*.