Behavioral changes associated with interacting with bots on Twitter

Zakaria Babutsidze
SKEMA BUsiness School
Université Côte d'Azur (GREDEG)
zakaria.babutsidze@skema.edu

Dorian Vincileoni
SKEMA Business School
Université Côte d'Azur (GREDEG)
dorian.vincileoni@skema.edu

Abstract

We study changes in Twitter users' behavior associated with interacting with automated bots on the platform. Based on the list of malicious bots identified and shut down by Twitter in the wake of 2016 US presidential election, we are able to identify about 54 thousand human Twitter users who have interacted with automated accounts. We first establish the baseline pattern of user behavior in the period before interaction and then measure behavioral changes observed around the period of interaction. Using a quasi-experimental research design, we document economically and statistically significant quantitative and qualitative changes in users' behavior and discuss their implications.

1. Introduction

The proliferation of online social networks (e.g. Facebook, Twitter, etc.) has created significant challenges for modern societies. One of them is the easiness of implementing automated accounts (hereafter bots) who could pedal certain type of content through the network without regard for the veracity of this content. Bots on Twitter, Facebook and YouTube have been explicitly linked to disinformation campaigns during the run-up to 2016 US presidential elections [1, 2], as well as in other digitally-intensive political processes outside the United States [3, 4]. Even though a part of research community questions whether bot activities have changed the outcome of the 2016 US elections [5], the fact that they contribute to a larger problem of spread of fake news, low-credibility and inflammatory content is without a doubt [6, 7, 8]. The scale of inflicted damage of such activities could reach devastating proportions even for established democracies [9, 10].

Yet, not much is know about who interacts with bots and how (if at all) this interaction alters their behavior on social networks themselves. Do these people increase their engagement on the platform (i.e. social network) as a result of interaction? Is this increase long-lasting? Are they able to generate more engaging content? Current paper attempts to answer these questions by using an extensive data set of Twitter users who have interacted with a recognized bot. A list of 2,752 malicious bots that were shut down by the platform was forwarded to the United States congress and made public in November 2017. We have identified the 50 most influential bots from this list. We have further identified a large portion of Twitter users who have interacted with these bots. In this paper we study the subset of Twitter users who have interacted at least once with one of these 50 most influential bots. We show that such interaction is associated with quantitative, as well as qualitative changes in twitter user behavior. Interestingly, our results indicate that these changes start significantly before the first official interaction with the identified bot.

2. Decisions before and after interacting with the bot

2016 US presidential election has put the spotlight on electronic social networks and the role they play in modern society. Electronic social networks facilitate maintenance of social relationships, but they also assume the role of an information broker [11]. The attachment of the latter function to electronic social networks might, however, be problematic as they are designed to prioritize engaging rather than trustworthy content [8]. Significant effort toward counteracting consequences of such design have come from major social networks over the past five years. However, these are in a direct clash with business models of electronic social networks and, thus, can go only so far.

With the rise of artificial intelligence, the automation of human-computer interaction has also become popular. Specific "conversational bots" have been developed and successfully applied in multiple settings ranging from medical [12] to human resource management [13] applications. Such has also penetrated

URI: https://hdl.handle.net/10125/79705
978-0-9981331-5-7

social networks with the development of "social bots". However, their purpose and aim, as well as consequences stand currently questioned [14]. This is particularly problematic as social networks are infested with bots that are masked under the guise of humans [15, 7] and can employ sophisticated strategies to sway public opinion [16]. On top of this, given adaptation capabilities of such automated agents, the coincidence between the bot design purpose and ultimate outcome cannot be guaranteed [17].

Given an important potential for impacting information circulation and human decision process, the actual impact of social bots is currently questioned. While one part of research community reports on high efficiency of automated agents on social networks [6, 7, 18], the other set of authors questions the size of such impact. For example, [19] demonstrates that bots are significantly less central than verified accounts during the discussions around contentious political events, and [20] reports that only a small proportion of vaccine-critical information that reaches active US Twitter users comes from bots. Overall findings indicate that public's trust in social bots, and therefore the potential for the impact on their decision-making process, depends on the area of inquiry [21].

Under such circumstances, from the individual decision-making perspective, it is important to ask two questions: Why do people interact with bots? And does this interaction change their behavior?

Even though the empirical part of the paper concentrates on the latter question, it is worth mentioning that when contemplating the former question the fact whether you know that the interaction counter-part is a bot or not is an important variable. Besides potential concealment of bot identity, social network users might get exposed to bots due to their friends' choices. Given that trust levels are high among close social contacts [22], information streaming through this medium might also enjoy high trust [23]. Through this mechanism, bots may be able to affect structure and functioning of social systems.

Even if the account is identified as a bot, the user might decide to engage with such automated entity. As postulated by the theory of embodied cognition, the presence of (ro)bots in a social system affects the way humans perceive social norms [24]. Research in Computers as Social Actors initiated by [25] has demonstrated particular ease with which humans anthropomorphise computer artifacts. Thus, bots could readily be accepted as parts of the social system. Known presence of bots in the system could make the interaction (even if confrontational) with bots more acceptable [26]. Research studying differences in interaction patterns across human-to-human and human-to-bot interactions reveals striking similarities across the two setups [27] indicating particular ease in treating automated agents similar to human users.

Somewhat more importantly for evaluating the impact of automated agents, human decision-making might be altered as a result of interacting with bots. Emotional contagion theory argues that human to human interaction can be a medium for emotional transfer [28]. Through a longitudinal study, [29] have shown that this process can result in a large scale diffusion emotion diffusion process. Further, [9] have demonstrated that such emotional contagion can happen on electronic social networks too (on Facebook in this case). A different stream of research has demonstrated that talking to chatbots can change human motivations [30]. Given this potential, and theoretical studies indicating high potential impact of such changes [31], behavioral changes associated with interaction with social bots is worthy of our attention.

3. Data

Our data collection has started by distilling the list of 2,752 bots down to a more manageable set. The task consisted of identifying the most influential automated accounts. Given that all these accounts are now defunct, we could not collect detailed data on their behavior. However, traces of all accounts remain in posts of currently functional accounts. As a first step we have extracted all data on all 2,752 bots available on the platform as of September 2020. In order to measure the respectfulness of the bot we counted the number of verified accounts that have interacted with each of the bots. 50 bots that command the highest number of verified accounts mentioning them in the posts were retained for the analysis. Table 1 gives presents the list of the bots retained for the analysis. This list contains some famous bots impersonating organizations and individuals from both ends of the political spectrum, like @ten_gop (and its variants) which is supposed to be the account of the Tennessee Republican Party and was an integral part of the Muller investigation, and @crystal1johnson who was supposed to be a "black lives matter activist" who became famous by the fact that Twitter CEO Jack Dorsey retweeted some of its content.

The next step consisted of identifying all Twitter users who were recorded as having interacted with one of the 50 most influential bots. In this work interaction counts as a voluntary communication (i.e., writing a post, or commenting somebody else's post) in which the user includes the bot's Twitter handle. In

@10_gop
@andyhashtagger
@anzgri
@berkhoff85
@bizgod
@blackeyeblog
@blackmattersus
@blacknewsoutlet
@blacktolive
@bleepthepolice
@blk_voice
@bydrbre
@chicagodailynew
@chistpost
@chrisuport_port
@coldwar20_ru
@comradzampolit
@crystal1johnson
@dailylosangeles
@dailysanfran
@danageezus
@exquote
@giselleevns
@gloed_up
@jenn_abrams
@kadirovrussia
@lavrovmuesli
@lukas_rosler
@margosavazh
@maxdementiev
@neworleanson
@novostidamask
@novostimsk
@novostispb
@onlinecleveland
@pamela_moore13
@pigeontoday
@politweecs
@realten_gop
@southlonestar
@ten_gop
@thefoundingson
@thisiskate
@todaybostonma
@todayinsyria
@todaymiami
@todaypittsburgh
@trayneshacole
@tribunaonline24
@usa_gunslinger

Table 1. Automated accounts (bots) included in the analysis.

order not to confound the effects of interacting with multiple bots, we have excluded all users who have interacted with multiple bots (out of the original list of 2,752). As a result the data contains accounts of users who have been recorded to interact with only one (of 50) bots (potentially multiple times). The exact time of the first interaction is an essential part of the data structure and will be used as an analysis tool in what follows. Acknowledging that verified accounts are somewhat different from an average member of Twitter community, we have excluded all verified users from the analysis.

All identified accounts were passed through "Botometer" [32] in order to evaluate whether they themselves exhibited bot-like behavior. In order to make sure we analyzed the behavior of human twitter users, all accounts where automation suspicion was above 10% were disqualified.[1] Another important feature of sample selection was user location. As large portion of the bot activity revolved around 2016 US presidential elections we have decided to concentrate on Twitter users who resided in the US only. We have used self-reported user locations for this purpose. This is the best proxy available to us to make sure we are studying a fairly homogeneous sample with respect to the subject matter. Given that election campaigns have a strongly state-specific flavor we only retained accounts where the state of the user could be identified. This narrowed down our sample to 54,106 twitter users. These users are spread across all US states (and Washington DC). Representation of states ranges from California, Texas and New York accounting for 12.66%, 8.78% and 8.22% of users respectively to South Dakota, Wyoming and Nevada with 0.12%, 0.11% and 0.07%. Distribution of accounts across the bots is even more asymmetric. 58% of the users in our dataset have interacted with one particularly powerful bot (@ten_gop). The second most popular bot (@crystal1johnson) accounts for only 7.43% of users. 37 or 50 bots each accounts for less then one percent of users in our dataset.

Users interact with bots at varying intensity. About 69% of the users interacts with the corresponding bot only once. On the extreme of the distribution, the dataset also contains a user who is recorded interacting with a bot 526 times. The average number of interactions across all accounts is 2.21 (S.d.=5.72). Bot interaction duration also varies across the users. This ranges from 0 (for users with one interaction) to interaction stretched over the five year period. The mean interaction duration is 20 days (S.d.=60).

Once the Twitter user has been retained for the analysis, we have collected their Twitter timeline. This includes all of their original tweets from the inception of the account until December, 31 2020. Unfortunately we are not able to collect information about content that these users have been re-tweeting. Absence of re-tweets constraints the analysis in this paper to the original content generated by the users included in the study. A typical twitter user in our dataset has has about 8.3 years of experience on the platform, and writes about 4.3 original tweets per day. In the timeline collection process, we are able to recover each tweet's content, as well as its timestamp, number of likes, retweets and comments it has received (as of December 31, 2020). Using tweet content we are able to characterize these tweets in multiple ways. We measure tweet's length (in terms of number of words), number of @s, number of hashtags, number of images, number of links, etc. Data contains over 702 million tweets.

In order to study the change in tweeting behavior we need to choose the time unit. Given the size of the dataset, we have chosen to study users' monthly behavior. We measure characteristics of each tweet, and aggregate data on monthly level in order to describe account's behavior. Resulted panel data set includes

[1] Botometer automation score distribution is bi-modal separating highly likely humans from highly likely bots. Expanding automation cut-off to 20% adds only about 3,000 additional users to the data set.

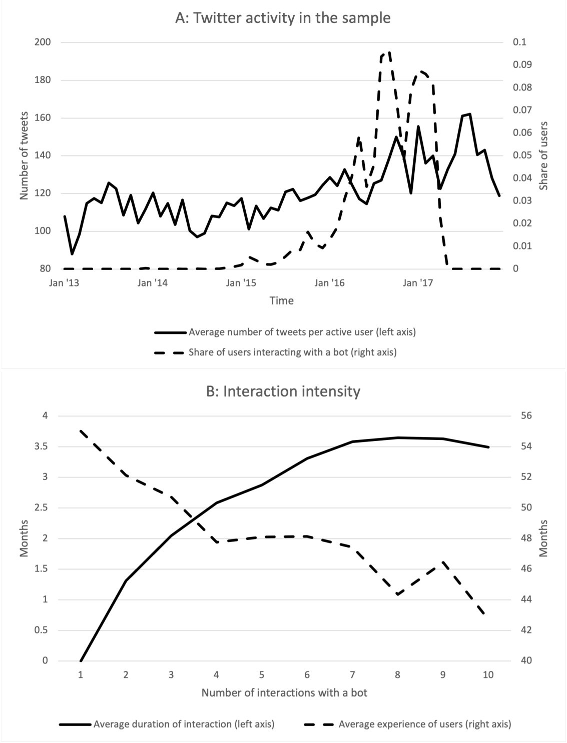

Figure 1. Descriptive characteristics of the sample.

nearly over 5.4 million user-month observations.

Figure 1 plots a number of descriptive characteristics of the data measured on monthly level. Panel A plots the average number of tweets per user in our dataset in the period of January 2013 through December 2017. This shows stability in overall twitter behavior in the sample with a slight (but sustained) increase in activity around 2016 election period. The same panel also plots the share of users who have had their first recorded interaction happening in a given month. This curve clearly shows the peak of bot activity in the US. Further analysis of data shows that there is some seasonality in twitter behavior. Namely, there are significant differences in twitting activity across various months of the year. This indicates that we need to take into account this feature in order to precisely identify the effect of interaction with a bot.

Panel B in Figure 1 visualizes how the number of interactions with the bot are related to the duration of the interaction. Not surprisingly we see that while two interactions are usually spread across one month period, 8 interactions span over 3.5 months. This panel also demonstrates that there is a significant difference across the profiles of people who interact with bots at different intensities. Namely people who interact with bots once have (on average) had 55 months experience on twitter, while people who interact with bots 10 times have only 43 months of experience (a year less) at the time when they first interacted with the bot. This could mean that younger people are more likely to interact with bots over extended periods of time. It could, however, also indicate that there is a significant amount of accounts created on twitter with the purpose of interacting with (known) bots (which would drive down the average experience of the user who heavily interacts with the bot). Nevertheless, this indicates that we need to control for the user's experience in our analysis.

4. Estimation method

The main statistical challenge in the process of correctly estimating behavioral changes associated with interaction with bots is identifying an appropriate counterfactual to which target user behavior should be compared. Matching (either exact, coarse exact, or propensity score) across Twitter users is not feasible as it requires collecting data on extensive part of Twitter users who have not interacted with bots. Appropriate methodology for identifying such users is not obvious. An alternative way for estimating relevant effects is to use only the data of bot-interacting (treated) individuals and take advantage of the fact that they interact with bots at different points in time. We use the latter approach.

Using the data set described in the previous section, and given the fact that all important characteristics of behavior are count variables, we estimate following two (account-level fixed effects, panel) Poisson models. First we fit the model

$$\begin{aligned} numTweets_{i,t} = {} & \mathcal{B}^{50}int \\ & + \beta_i numInt + \beta_e experience \\ & + \beta_m month + \epsilon_i + \epsilon_{i,t}, \end{aligned} \tag{1}$$

where $numTweets_{i,t}$ corresponds to the number of Tweets by user i in month t and $numInt$ measures the number of interactions the user had with the bot (throughout the whole Twitter activity period). It is important to control for this variable as the difference in interaction intensity might be an indication of inherent differences across users (for example explicit intention to interact with the automated agent). $experience$ measures the number of months since the account joined Twitter. It is important to control for this feature as data shows clear and steady increase in frequency of tweeting with the increase in experience. $month$ is an identifier of the current month of the year and is intended to control for yearly cycles in user behavior on Twitter (for example, our data shows that users produce much more tweets in January than in any other month of the year).

Most importantly, $\mathcal{B}^{50}$ collects 50 dummy variables identifying months in the focus period. Namely, these collect the identifier of the month when i first interacted with the bot, but also identifiers of each of 24 months prior to first interaction interaction, 24 months following the first interaction, and one common dummy for all months after the period (i.e. beyond two years after interaction). This allows us to estimate the profile of behavior of a typical account.[2]

Once we have established the tweeting intensity behavior of an account, we estimate a set of characteristics per tweet within each month using

$$\begin{aligned} charTweet_{i,t} = \mathcal{B}^{50}int &+ numTweets_{i,t} \\ &+ \beta_i numInt + \beta_e experience \\ &+ \beta_m month + \epsilon_i + \epsilon_{i,t}. \end{aligned} \tag{2}$$

Here, $charTweet_{i,t}$ would measure the total number of words (in all tweets of user i in month t) in order to measure the change in twitting style of the user, total number of @s in order to measure user's willingness to engage with other users. This variable could also measure the impact of user's content. This could be accomplished by measuring total number of engagements with user's content in a given month (number of retweets, number of likes and number of comments).

Given that all dependent variables are count measures, and that we have a panel structure, we estimate equations (1) and (2) using panel Poisson estimator. Results presented in the following section are based on these estimations and are reported along with robust standard errors clustered on user level.

5. The effects of interacting with bots

Figure 2 presents results of estimations described in the previous section. Here we focus on a four-year period around the first recorded voluntary interaction with the bot and describe the behavioral profile of users. We plot results in terms of percentage change in our variables of interest compared to the account's behavior in the period from the account's inception to two years prior the first interaction. The period prior to two years before first interaction with the bot constitutes the baseline and is used to set a typical behavioral profile for a given user. Table 2 presents selected coefficients from these estimations. In the table, estimates are presented in the form of incidence rate ratios.

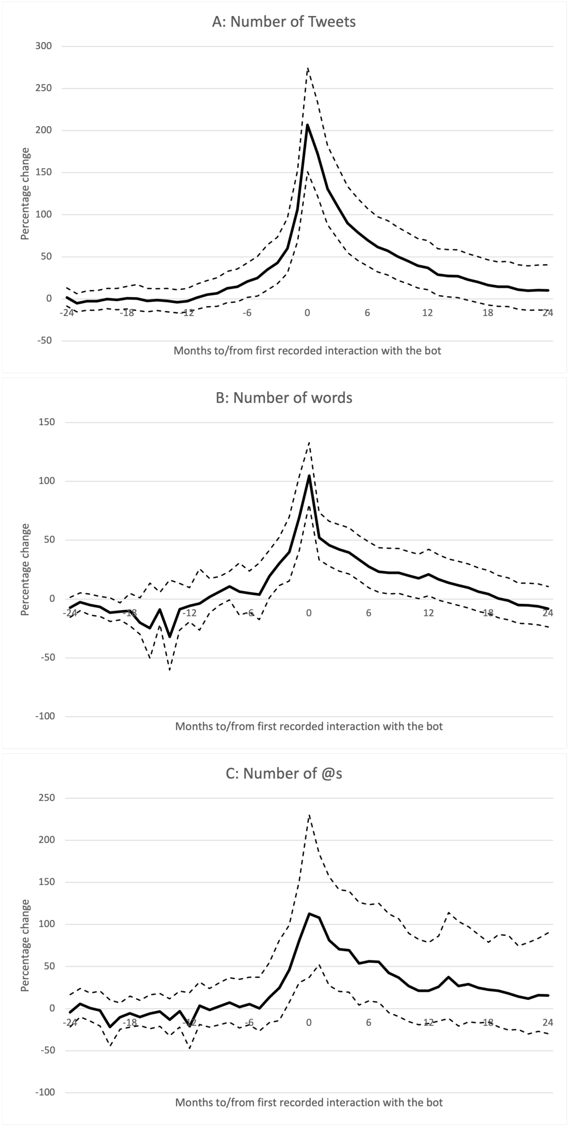

Figure 2. Tweeting behavior in relation to the interaction with the bot.

Notes: All values are calculated in relation to tweeting behavior observed prior to two years before the first interaction with a bot. Dashed lines delimit 95% confidence interval.

Three panels in figure 2 describe the temporal behavior of number of tweets per month, as well as number of words and number of @'s generated by the typical user. Notice, that because $numTweets_{i,t}$ is included in equation (2), number of words and number of references (i.e., @'s) measure averages per generated tweet in a given month.

The results show significant alteration of tweeting behavior by users during the time period around the interaction with the confirmed social bot. This behavior change is most prominent in terms of the

[2] We could also include bot-level, as well as state-level fixed effect identifiers. However, given that we only have data on people who are located within one state, and they interact with only one bot, account level fixed effects estimator takes both of these into account.

	Tweets	Words	@s	Engagements
t-24	1.016	0.926	0.956	1.544
	(0.55, 0.76)	(0.04, 0.10)	(0.10, 0.66)	(0.25, 0.04)
t-12	0.972	0.941	0.791	1.409
	(0.07, 0.70)	(0.07, 0.44)	(0.16, 0.26)	(0.42, 0.25)
t-6	1.206	1.049	1.054	1.635
	(0.10, 0.03)	(0.09, 0.57)	(0.14, 0.70)	(0.06, 0.17)
t	3.067	2.048	2.128	3.875
	(0.31, 0.00)	(0.13, 0.00)	(0.49, 0.00)	(1.54, 0.00)
t+1	2.722	1.522	2.081	3.623
	(0.28, 0.00)	(0.10, 0.00)	(0.33, 0.00)	(1.59, 0.00)
t+2	2.304	1.457	1.812	3.546
	(0.24, 0.00)	(0.10, 0.00)	(0.32, 0.00)	(1.51, 0.01)
t+3	2.098	1.421	1.704	4.004
	(0.22, 0.00)	(0.10, 0.00)	(0.30, 0.00)	(1.92, 0.01)
t+6	1.697	1.275	1.563	4.324
	(0.17, 0.00)	(0.09, 0.00)	(0.28, 0.00)	(2.30, 0.01)
t+12	1.369	1.210	1.213	5.331
	(0.15, 0.00)	(0.10, 0.02)	(0.24, 0.32)	(3.72, 0.02)
t+18	1.164	1.042	1.225	5.023
	(0.14, 0.19)	(0.09, 0.42)	(0.24, 0.29)	(3.44, 0.02)
t+24	1.102	0.918	1.156	4.049
	(0.14, 0.43)	(0.08, 0.37)	(0.29, 0.57)	(3.11, 0.07)

Table 2. Selected coefficients from fixed effects panel Poisson regressions.

Notes: Number of tweets is estimated using equation (1). Number of words, @s and engagements is estimated using equation (2). Number of engagements combines number of likes, retweets and replies. Coefficients (*t*s) correspond to months, with *t* indicating the treatment month. Coefficients are reported in terms of incidence rate ratios. In brackets are presented (standard error, p-value)s.

tweeting frequency. Panel A of figure 1 indicates that during the month when users first interact with the bot they write 200% more tweets (tripling the frequency) than in a baseline period. This increase in frequency is statistically significant. In following months tweeting frequency seems to drop gradually and becomes statistically indistinguishable from the behavior in the baseline period after 14 months from the first interaction with the bot.

Panels B and C of the same graph demonstrate that behavioral change does not concern only the frequency of tweeting, but also the characteristics of the produced original content. We see that during the month when the account first interacted with the bot, length of a typical tweet, in terms of the number of words, doubles (a 100% increase over the baseline level). Users also seem to use twice more @'s per tweet than usual (during the baseline period), which is an indication that accounts engage with more discussions with other members of the platform. Similar to tweeting frequency, these behavioral changes also dissipates over the following 12 months (in fact, the increase in the usage of @'s becomes statistically indistinguishable from the pre-interaction/baseline period during the 8th months after first interaction). Besides their statistical significance, all these effects are economically very large. Similar, but albeit less distinct patterns can be noticed in the usage of hashtags, sharing of photos, videos and URL's by users who interact with identified social bots.

One important characteristic of these results, however, is that they do not seem to be "caused" by the interaction with the account to a large extent. In fact, we see the uptick in all variable profiles starting from about six months before the recorded interaction. There could be two reasons for this. Firstly, users start being exposed to content generated/diffused by bots before they engage in the discussion with them and this prompts them to change their behavior. Alternatively, it might be that (for some reason) users start participating in discussion on social networks and within about six months they end up interacting with the automated agent. Unfortunately our data does not allow to distinguish between these two explanations. However, we can exclude the possibility that these patterns are driven by the latter mechanism, as in this case observing pronounced peaks during months of first interaction in all three panels of Figure 2 would be extremely unlikely.[3]

All three characteristics reported in Figure 2 describe the behavior of the user. However, one can also ask if engaging with a social bot, and thus, with its ecosystem would also alter performance of the original content generated by the users. To answer this question we model engagement with user's content using equation (2). Engagement measure adds likes, comments and retweets of all original posts by the user in a given month. Estimates of the change in this characteristic are presented on Figure 3. These results point to the fact that tweets by users during the month of interaction with bots enjoy four times higher engagement than usual. Similarly, the engagement seems to be more persistent than behavioral changes in Figure 2 – increase in engagement remains significant over 18 months following the first interaction with the bot. Unlike behavioral measures, the peak in engagement arrives a year after initial interaction with the bot. The uptick also starts later – only three months before the first recorded interaction. At the same time, estimates for engagement measures seem to be less precise than those for behavioral measures as indicated by a wide margin

[3] Recall that first bot interaction months of great majority of our users are spread around at least a 14-months period from February 2016 to March 2017.

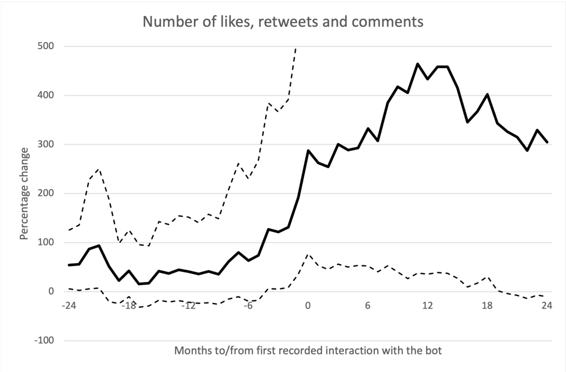

Figure 3. Engagement with user's content.

Notes: Values are calculated in relation to tweeting behavior observed prior to two years before the first interaction with a bot. Dashed lines delimit 95% confidence interval. Given that upper limit of the confidence interval is irrelevant for the positive effect, the plot was scaled to highlight the profile of the estimate itself.

around estimates.

6. Conclusion

In this paper we have analyzed behavioral changes associated with the interaction with automated bot on Twitter. Our results indicate significant changes in human Twitter users' behavior. Around the time period when user interacts with a bot we see a strong take off in frequency of posting. In addition, we see statistically and economically significant qualitative changes that accompany this quantitative break. We see users writing longer posts, we also see them engaging closer with the community by a heavier use of @s, hashtags, images and URLs.

As an important caveat, it is worthy to note that our study has no sure way to measure exact share of this change that is attributable to the interaction with the exposure to information coming from bots. The reason for this is that even though we have information on when a given user has explicitly engaged in a discussion with a bot, there is no way of identifying the exact point in time when they first got exposed to the information coming from the bot. It is possible, in fact plausible, that information peddled by the bot has been present on user's Twitter wall quite some time before they decided to react to it by Tweeting at, or replying to, the bot. This, in fact could be an explanation for the pattern of all measured effects starting to break from baseline predictions about six month prior to official interaction with the bot.

Increase in tweeting frequency and engagement with the community as a result of interacting with an entity diffusing controversial information is somewhat surprising. It goes against the well-known spiral of silence theory [33] which argues that such interaction would result in decrease in the incentive to express one's opinion [34, 35]. However, it is important to notice that in this respect our data might suffer from selection bias. We identify study subjects by the very fact that they engage in the discussion with the bot. People who might suffer from spiral of silence, however, will most likely not interact with the bot (choosing silence). As a result a large portion of users who are exposed to bot's influence but choose to abstain from interacting with them are missing from our study. One way to closely examine this possibility would be to extend the study to include the tweet sentiment analysis. Given that we have access to complete tweet content and can establish the baseline for each user's tweet sentiment, it would be interesting to examine potential breaks in this feature around the bot interaction period. This would also allow us to closely evaluate the emotional contagion effects of bots [7]. This very task of sentiment analysis of our 702 million tweets, indeed, constitutes the first step of our future research.

Another worthy step is the additional data collection effort in order to establish a comparable set of twitter users who have never interacted with any of the identified bots. Such complementary dataset will allow us to set up a more credible baseline for precisely estimating the extent of behavioral change associated with the interaction with the bot.

References

[1] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265, 2018.

[2] Y. Golovchenko, C. Buntain, G. Eady, M. A. Brown, and J. A. Tucker, "Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 u.s. presidential election," *The International Journal of Press/Politics*, vol. 25, no. 3, pp. 357–389, 2020.

[3] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *CoRR*, vol. abs/1707.00086, 2017.

[4] P. Howard and B. Kollanyi, "Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum," *SSRN Electronic Journal*, 06 2016.

[5] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, pp. 211–36, May 2017.

[6] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts,

and J. L. Zittrain, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[7] M. Stella, E. Ferrara, and M. De Domenico, “Bots increase exposure to negative and inflammatory content in online social systems,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12435–12440, 2018.

[8] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, no. 1, p. 4787, 2018.

[9] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, “Experimental evidence of massive-scale emotional contagion through social networks,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[10] A. Badawy, K. Lerman, and E. Ferrara, “Who falls for online political manipulation?,” in *Companion Proceedings of The 2019 World Wide Web Conference*, WWW ’19, (New York, NY, USA), p. 162–168, Association for Computing Machinery, 2019.

[11] C. Chike-Obuekwe, J. Choudrie, A. Nwanekezie, D. Sudaram, and G. Peko, “Investigating the use, adoption and diffusion of online social network adoption (facebook vs twitter), within the older adult population (50+) in hertfordshire uk,” in *Proceedings of 53rd Hawaii International Conference on Systems Sciences*, pp. 4464–4473, HICSS, Hawaii, USA, 2020.

[12] X. Fan, D. Chao, Z. Zhang, D. Wang, X. Li, and F. Tian, “Utilization of self-diagnosis health chatbots in real-world settings: Case study,” *Journal of Medical Internet Research*, vol. 23, p. e19928, Jan 2021.

[13] S. Majumder and A. Mondal, “Are chatbots really useful for human resource management?,” *International Journal of Speech Technology*, 2021.

[14] G. Murtarelli, A. Gregory, and S. Romenti, “A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots,” *Journal of Business Research*, vol. 129, pp. 927–935, 2021.

[15] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pp. 280–289, AAAI Press, 2017.

[16] V. L. Kitzie, E. Mohammadi, and A. Karami, ““life never matters in the democrats mind”: Examining strategies of retweeted social bots during a mass shooting event,” *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 254–263, 2018.

[17] P. Sweetser and M. Aitchison, “Do game bots dream of electric rewards? the universality of intrinsic motivation,” in *Proceedings of FDG’20 International Conference on the Foundations of Digital Games*, ACM, New York, NY, USA, 2020.

[18] X. Liu, “A big data approach to examining social bots on twitter,” *Journal of Services Marketing*, vol. 33, no. 4, pp. 369–379, 2019.

[19] S. González-Bailón and M. De Domenico, “Bots are less central than verified accounts during contentious political events,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, 2021.

[20] A. G. Dunn, D. Surian, J. Dalmazzo, D. Rezazadegan, M. Steffens, A. Dyda, J. Leask, E. Coiera, A. Dey, and K. D. Mandl, “Limited role of bots in spreading vaccine-critical information among active twitter users in the united states: 2017–2019,” *American Journal of Public Health*, vol. 110, pp. S319–S325, 2021/05/05 2020.

[21] N. Aoki, “An experimental study of public trust in ai chatbots in the public sector,” *Government Information Quarterly*, vol. 37, no. 4, p. 101490, 2020.

[22] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, “Social phishing,” *Commun. ACM*, vol. 50, p. 94–100, Oct. 2007.

[23] A. Bessi and E. Ferrara, “Social bots distort the 2016 u.s. presidential election online discussion,” *First Monday*, vol. 21, Nov. 2016.

[24] M. L. Anderson, “Embodied cognition: A field guide,” *Artificial Intelligence*, vol. 149, no. 1, pp. 91–130, 2003.

[25] C. Nass and Y. Moon, “Machines and mindlessness: Social responses to computers,” *Journal of Social Issues*, vol. 56, no. 1, p. 81–103, 2000.

[26] M. Wischnewski, R. Bernemann, T. Ngo, and N. Krämer, “Disagree? you must be a bot! how beliefs shape twitter profile perceptions,” in *Proceedings of CHI’21 Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, 2021.

[27] J. Dev and L. J. Camp, “User engagement with chatbots: A discursive psychology approach,” in *Proceedings of CUI’20 Conference on Conversational User Interfaces*, ACM, New York, NY, USA, 2020.

[28] R. A. Easterlin, “Explaining happiness,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 19, pp. 11176–11183, 2003.

[29] J. H. Fowler and N. A. Christakis, “Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study,” *BMJ*, vol. 337, 2008.

[30] P. B. Brandtzaeg and A. Følstad, “Chatbots: changing user needs and motivations,” *Interactions*, vol. 25, no. 5, pp. 38–43, 2018.

[31] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, “Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks,” *European Journal of Information Systems*, vol. 28, no. 4, pp. 394–412, 2019.

[32] A. Rauchfleisch and J. Kaiser, “The false positive problem of automatic bot detection in social science research,” *PLOS ONE*, vol. 15, pp. 1–20, 10 2020.

[33] E. Noelle-Neumann, “The spiral of silence a theory of public opinion,” *Journal of Communication*, vol. 24, no. 2, pp. 43–51, 1974.

[34] G. Neubaum and N. C. Krämer, “Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media,” *Media Psychology*, vol. 20, no. 3, pp. 502–531, 2017.

[35] G. Neubaum and N. C. Krämer, “What do we fear? expected sanctions for expressing minority opinions in offline and online communication,” *Communication Research*, vol. 45, no. 2, pp. 139–164, 2018.