Changing Perspectives: Is it Sufficient to Detect Social Bots?

Preprint · April 2018

CITATIONS
0

READS
83

3 authors:

Dennis Assenmacher
University of Münster
7 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Christian Grimme
University of Münster
56 PUBLICATIONS **246** CITATIONS
SEE PROFILE

Lena Adam
University of Münster
5 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Changing Perspectives: Is it Sufficient to Detect Social Bots?

Christian Grimme, Dennis Assenmacher, and Lena Adam
{christian.grimme, dennis.assenmacher, lena.adam}@uni-muenster.de

University of Münster, 48149, Münster, Germany

Abstract. The identification of automated activitiy in social media, specifically the detection of social bots, has become one of the major tasks within the field of social media computation. Recently published classification algorithms and frameworks focus on the identification of single bot accounts. Within different Twitter experiments, we show that these classifiers can be bypassed by hybrid approaches, which on a first glance may motivate further research for more sophisticated techniques. However, we pose the question, whether the detection of single bot accounts is a necessary condition for identifying malicious, strategic attacks on public opinion. Or is it more productive to concentrate on detecting strategies?

Keywords: Social Bots, Online Propaganda, Social Media Analysis, Social Media Computation

1 Introduction

Automation in social media has received enormous attention in scientific and public discussions. Scientific papers [1–5] up to newspapers [6–9] – recently also in Germany [10, 11] – report on the threats posed by automated accounts as well as on the identification of automated profiles during election campaigns like the Brexit vote [12] or the last US Presidential election [13]. Specifically the term Social Bot stands synonym for malicious activities, which aim for manipulation of public opinion or even elections. Consequently and rather straightforward, science focuses on mechanisms to detect these automated profiles based on their individual behavior. Besides descriptive observation techniques, a plethora of automated techniques are available to identify social bots, ranging from machine learning approaches to very simple activity indicators. Basic approaches [14] merely analyze the frequency of an accounts activity (a social bot is postulated, if an activity threshold is passed), sophisticated approaches try to identify behavioral patterns of automated accounts. Probably the most well-known approach of the latter class is the Botometer (formerly known as BotOrNot) service provided by the Indiana University [5, 15].

All approaches, simple up to complex, follow rules that usually describe fully automated behavior of social media accounts. If a human partly or temporarily manages an account, the indicators as well as the trained (machine learning)

models become vague and imprecise in their detection performance. Especially for machine learning approaches, another problem occurs: trained with limited (and manually gathered) ground truth, these methods specialize to detect exposed behavioral and metadata patterns for a given set of accounts within a fixed time interval. Due to high dynamics and changing usage of social media accounts, exposed patterns of these profiles may change also rapidly. This leads to varying accuracy of the trained detection mechanisms and eventually, the (at least temporary) inability to detect before-known social bot accounts.

To empirically support our argument, we first conduct two experiments to highlight the volatility of social bot detection mechanisms under changing usage patterns for social media accounts. Exemplarily, we concentrate on Botometer as the most prominent and rather advanced detection technique. In a first experiment, we construct fully automated social bots, which can be easily detected by simple indicators and Botometer alike, and successively integrate human behavior. During the bots' activity, we analyze the detection performance of Botometer over time. In a second experiment, we implement a set of 30 social bots that actively befriend to Twitter users and expose human like behavior. After a month of constant and fully automated behavior the small bot net starts massive action to promote a topic. Here, we also track the detection performance of Botometer.

Starting from these experimental insights and the discussion of current detection techniques, we pose the principal question, how detection mechanisms for social bots contribute to the prevention of manipulation or propaganda via social media. We propose a shift of perspective from detecting simple account properties towards identifying coordinated strategies, i.e., orchestrated activities of multiple (automated, semi-automated or human-steered) accounts. This shift from the micro-level of social bot detection to the macro-level of strategy detection is a by far greater challenge to research, but certainly of greater importance.

This work is structured as follows: The next section highlights some established and current developments in social bot detection and proposes a taxonomy that identifies two main overall streams of methodology: inferential and descriptive analysis. Thereafter, an experimental study on Botometer as current inferential detection mechanism is presented. Based on this, we pose the principle question, whether detecting automation patterns in single accounts is helpful after all. Based on two case studies on campaigns observed during the German general election in September 2017, we propose a change of perspective towards detecting orchestrated behavior of actors in social media.

2 Detection of Social Bots

With the "Rise of Social Bots" – this wording is also a reference to one of the most recent and influential reviews on the topic [5] – research tackled the detection of automated social media profiles. Early social bot realizations and also many current implementations are simple and merely focused on content amplification. Consequently, detection approaches for this type of bots monitor the activities of suspicious accounts and set (usually rather arbitrary) thresholds

for defining accounts as social bots. Interestingly, a lot of current research is still based on these methods [13, 14].

Current social bot implementations are far more sophisticated. Accounts are created to resemble human accounts and social bots mimic human behavior on the meta data level, i.e., they automatically vary their activity profile, follow a day-night-cycle or befriend and even communicate (in a simple manner) with other accounts. Although there are limits in intelligent interaction [16], with the before mentioned rudimentary techniques, social bots are not detectable anymore. Even human observers may be deluded by these obfuscation techniques. Sophisticated automatic detection mechanisms however, can analyze multiple aspects of the meta data over time and are (sometimes) able to find suspicious patterns in behavior for classifying accounts. Others analyze the behavior of many accounts over time with respect to predefined indicators. Thus, in contrast to Ferrara et al. [5], we divide the current detection techniques in only two classes.

2.1 Inferential Approaches

The first class of detection approaches is based on the analysis of data from account activities in social media and tries to infer representative patterns for social bot behavior. Sometimes, methods of machine learning are applied to automatically deduce features and rule sets. Those rule sets are then used on not yet classified accounts to get some rating. An early detection mechanism contained in this class is not based on machine learning but manually defines rules for befriending behavior of social bots [17]. Yang and colleagues [18] also use feature extraction techniques from representative behavioral features of human and robotic accounts in the RenRen network to identify meaningful discrepancies of both classes. Based on this, an online sybil detection system for automated accounts is implemented. Another method by Clark et al. [19] tries to identify automated activity on Twitter by focussing on language analysis. The approach identifies natural (human) language patterns to indirectly distill automated produced content. The currently most popular approach for classifying single Twitter accounts is the Botometer (formerly known as BotOrNot) web service[1] provided by the Indiana University [20, 15]. Based on more than 1,000 features used in a random forest classifier, a given Twitter account is analyzed and rated in an interval of 0 (human) and 1 (social bot). This rating can be interpreted as probability for the specific account for being a social bot (or not).

Overall, inference-based methods implicitly assume underlying common characteristics of social bot behavior that need to be explored and described by fixed rule sets. To generate these rules, an annotated data set (ground truth) is needed to extract representative features for human and social bot behavior. All approaches focus on classifying single user accounts in social networks to detect the type of actor (human or machine) behind the curtain.

[1] https://botometer.iuni.iu.edu

2.2 Descriptive Approaches

Different from inferential approaches, the second class of descriptive approaches comprises usually manual observations of specific campaigns in social networks. Examples of such case studies are the detection of a Ukrainian bot net by Hegelich and Janetzko [21]. The authors analyze a large dataset of Twitter posts and metadata by applying frequency indicators and clustering methods. From these insights, they extract evidence for a large bot net that was active during the Ukrainian revolution in 2014. In the same way, using tools from descriptive data analysis, Eccheverria and Zhou [22] identified a large social bot network, which posted Star Wars quotations – probably just to age the Twitter accounts for later use in campaigns. An early clustering approach by Cao et al. [23] for detecting similar behavior in accounts can also be considered as descriptive method. The authors provide a so-called SynchroTrap, which detects loosely synchronized actions of accounts in the context of campaigns. The basic assumption is, that a campaign needs a central, thus synchronized activity of multiple social bot accounts.

A major advantage of the descriptive approaches is their openness towards new and yet unknown strategies. However, they demand an (usually a-posteriori) identification of campaigns. Even in current approaches, it is necessary to integrate human intelligence for the selection of indicators as well as for the interpretation of results. Once a campaign is identified, actors can be investigated and bots can be separated from human accounts.

2.3 A Comment on both classes

The approaches of the defined classes differ in their perspective on social bot detection. The inferential perspective assumes universal patterns to be identified for social bots. The descriptive perspective works case-based and tries to identify social bots from a group of accounts that participate in an observed campaign. Although all approaches have the same goal, the descriptive approaches are inherently context-related. The initial restriction on a topic or campaign indirectly restricts the amount of accounts that has to be considered for detecting social bots. Still, the approaches of the inferential class are predominant in literature and current discussion. They work in a rather context-free manner by identifying bot characteristics for single accounts. On the one hand, this can be of advantage, as these methods are directly applicable to social media accounts. On the other hand, the missing context implies the absence of important indications that could support or falsify the detection result. In the following section, we investigate this ambivalence for the most commonly used indicator Botometer.

3 Experiments

To get first insights into the performance of current bot detection mechanisms, we conducted several Twitter-based experiments. Therefore we used fully automated and hybrid bot approaches to check, whether those mechanisms are

capable to appropriately identify bots. The bot types used in the following experiments are motivated by a taxonomy published by Grimme et al. [16]. They assume three classes of bots ranging from simple automation (for broadcasting and multiplication of content) via human-like acting bots (possibly also containing a hybrid component) to intelligent acting (and content producing) bots. To show that current bot detection mechanisms struggle to appropriately identify social bots, we restrict ourselves to the first two classes. For the third class no productive realization is known yet.

3.1 Experimental Setup

The different experiments are based on a propitiatory social bot framework, that is capable of realizing the before mentioned simple and hybrid bot types of the taxonomy introduced in [16]. Figure 1 visualizes the three core components of the proposed framework: *Account, Bot, and Human*. The account component depends on the underlying social media platform. The remaining components of the framework do not explicitly focus on a specific platform and can be regarded in a more abstract way. The account can be accessed and interacted with, by either the fully automated bot component via an application programming interface (API) or by the human via a web/mobile client. It has to be emphasized that the functionality that can be realized by the automated bot component mainly depends on the provided functionalities of the platform's API. In case of the Twitter platform, the API provides full access to all functionalities that can be used within the web-frontend. Therefore all natural account interactions can be mimicked by the bot component.

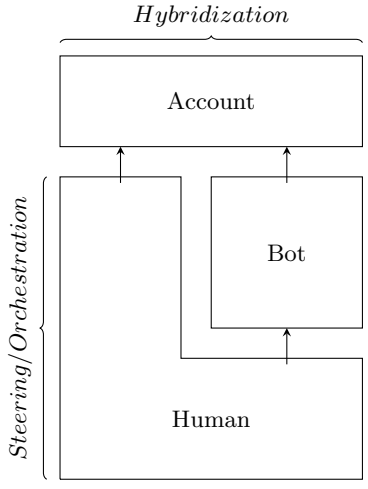


Fig. 1. Conceptual Bot Framework.

The used framework can be adjusted in two different dimensions. The *hybridization* dimension specifies to what extend the bot component, and the human component should interact with the social media platform account. Figure 1 displays an equal share of bot and human interaction.

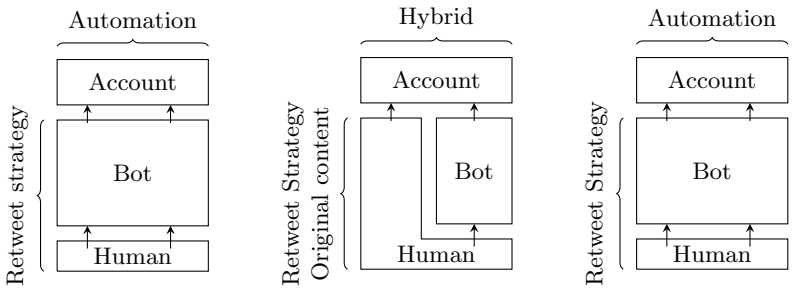
The *steering / orchestration* dimension adjusts the proportion of the individual components. The bot component may consist of different automation mechanisms. A rather simple functionality would be the repetitive multiplication of social media posts (retweeting). Hence this scenario indicates a small steering share for the bot component. For a higher bot steering factor, we could add a day-night-cycle or automated and intelligent following mechanisms. The human component can also be vertically adjusted. Within a simple scenario, human interaction could be reduced to a minimum, such as specifying which kind of content should be promoted or retweeted. In contrast, a prominent human interaction scenario would realize an automated spreading of original but predefined postings. In such a case the social bot needs a variety of tweets as input, which have to be manually created and curated by humans.

Within our experiments, we utilize the Botometer service to analyze the scores for different Twitter accounts [15]. Botometer is a classification system which determines the probability for a given Twitter account being a social bot. Applying supervised learning techniques such as random forests, the system learns a classifier by using 1,000 different account related features. Those features are divided into six different categories: *user*, *friends*, *network*, *content*, *timing*, and *sentiment*. For each category the learning algorithm predicts a bot likelihood. Additionally, an aggregated bot score that considers all available features is provided by the service.

3.2 Pre-experiment: Botometer

The first experiment aims for the analysis of Botometer scores of bot-accounts which expose different behavior over time. Furthermore, we want to examine whether and to which extend human interaction in terms of hybridization is able to bias the assessment of the Botometer scores. Therefore, the experiment is divided into three phases:

Phase 1: At the beginning three different bot accounts are started with new and empty profiles. Each bot account follows a simple retweeting strategy. In this case the bots retweet posts containing the hashtag *#bitcoin* without adding additional texts or comments. Therefore, candidate tweets related to the hashtag *#bitcoin* are picked via the Twitter streaming API. Each bot retweets random posts from the candidate list. Furthermore, the bots follow no day-night-cycle. Their retweet actions are strictly set to specific points in time. Additionally, we set the bot activity to 50 retweets per day. All these regular and simple settings ensure, that a clearly automated basic behavior is exposed by the accounts. We expect Botometer to detect these accounts as social bots with > 0.5 probability. Apart from setting up the bot scripts, there is no human interaction in the first phase of the experiment. Using this



(a) Phase 1 (b) Phase 2 (c) Phase 3

Fig. 2. Conceptual view of the Botometer experiment, divided in three phases of behavior.

fully automated configuration, which is schematically shown in Figure 2 a), the bots ran for two weeks.

Phase 2 : After the initial fully automated phase, two of the three bots are manually curated (starting February 5th). Manual interaction is, for example, tweeting, retweeting, liking of posts related to current incidents, like discussions about soccer games, the weather, or TV series. The manual human intervention follows a typical daily-life structure. An exemplary activity pattern is manual interaction in the morning, at lunch time, and in the evening. With human intervention, the bots do up to ten "human actions" per day, in addition to their basic retweet-strategy. As shown in Figure 2 b) the two accounts are controlled in a hybrid way now. The human intervention is also part of the hybridization-axis, since the human-controlled actions are done directly through the web interface of the account. Using this configuration the bots run two additional weeks. As a baseline, the third bot still follows the simple retweet-strategy, described in Phase 1.

Phase 3 : After two weeks, the human intervention is stopped, and the bot behavior changes back to the configuration of Phase 1, refer to 2 c).

At each phase of the experiment the Botometer score of the bots is calculated on a hourly basis.

Figure 3 shows the development of the Botometer scores of the three bots during the four weeks of the experiment. To display the score per day, the mean of the hourly scores is calculated. Additionally, a regression line for each bot has been computed, in order to analyze the trend of the account classification.

For all social bots, the Botometer score of the simple retweeting phase 1 converges to a score of 0.5. A score of 1.0 in this case means that the account is most certainly controlled by a bot, where a score of 0.0 means, that the account apparently only contains human-steered interactions. The authors of Botometer state, that a score around 0.5 enables to no precise statement, whether the account is steered by a human, or a bot [15]. Hence the behavior of the Botometer measurement for our simple bots is astonishing. Obviously, already the very

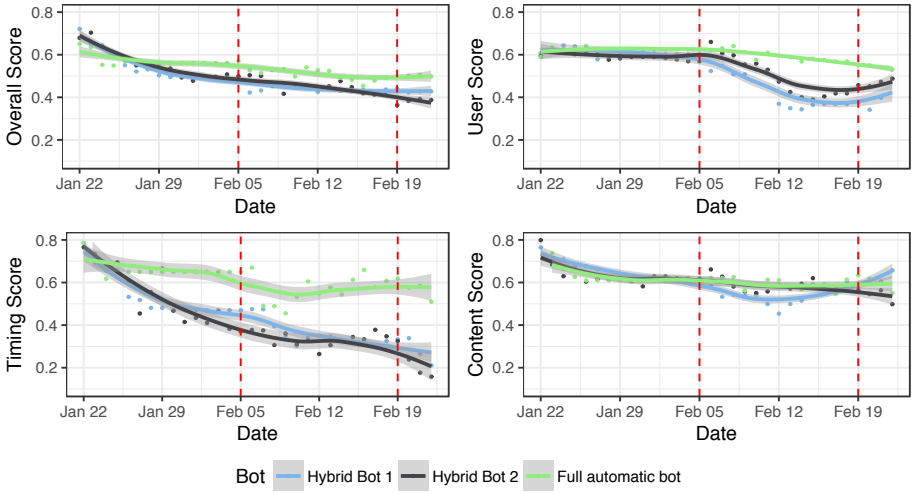


Fig. 3. Botometer scores over time, including trends. Top left: overall score; top right: user score; bottom left: timing score; bottom right: content-related score.

simple and regular implementation of activity leads to the inability to classify the accounts. At the same time, we find that the start of phase 2 shows no change of the score development. With some inter-bot variance, the *overall* Botometer score converges to a a range of 0.3 to 0.5 at the end of the experiment. To get more information on the effects on our hybrid interaction, we take a deeper look at the sub-scores of Botometer. Exemplary, the development for three of the five sub-scores is shown in Figure 3.

Considering the *user* sub-score, it is obvious, that the start of human intervention on February 5th leads to a strong decrease of the scores for the hybrid bots. The score of the fully automated account never drops below 0.5. After February 19th – the end of human intervention – the user scores of the two hybrid bots increase again. Amongst other features the *user* sub-score takes into account the features "number of tweets/retweets/mentions/replies (per hour and total)" [15]. Certainly, these features change significantly during the human intervention in Phase 2. Another feature, which may lead to a decrease of all the accounts, is the continuously changing "age of the account".

An even more obvious change of the score range in phase 2, is noticeable in the *Timing* sub-score. This score is based on calculation of time ranges between two consecutive tweets/retweets/mentions. The human intervention in phase 2 massively improves the scores of the hybrid bots. The timing sub-score dropped under 0.4, whereas the value of the fully automated account ranges about 0.5. Within the third phase of the experiment, the scores of the hybrid bots decrease even further and reach a score range of around 0.3 to 0.2. This might be caused by the change in tweeting activity at the transition from phase 2 to phase 3.

Analyzing the sub-score *Content* indicates that the human intervention seems to have almost no impact on the features of this score. Within all phases of the experiment, the mean of the three values ranges between 0.7 and 0.4. Since the content is changed from merely retweeting bitcoin tweets to original text post, pictures, etc., this behavior is surprising. An explanation of this behavior could be the fact, that Botometer is trained on English profiles and content. The bots tweeted mainly in German, so the available detection patterns are possibly not able to properly classify the content.

The sub-scores *Friends* and *Networking* (not shown here) have no impact on the overall-score as well. This might be due to the fact that the human intervention was limited on posting activities. No network activities have been done, neither by the automated nor by the human influenced account. The sub-score *Sentiment*, is – like the sub-score content – composed of different text-based features. Furthermore, there is no observable difference between the scores of the automated and the hybrid accounts. This might again be, due to the fact that the algorithm is trained on English data.

3.3 A Social-bot-driven campaign

The second experiment has been conducted between January 5 and February 5 in 2018. Within this study we investigate the impact of a coordinated strategy to push a predefined hashtag or topic, respectively. The main goal is to check, if bot accounts that are part of the attack, can be detected by the Botometer service and whether our attack is able to actually trigger a new trend on the twitter platform. In order to conduct the experiment we constructed a hashtag that should encourage users to actively join the the twitter conversation. To ensure user's participation, we tried to gamify the whole setting: using the hashtag #songmoji, Twitter users are asked to post titles of different songs, only by relying on emoticons. Figure 5 shows an exemplary songmoji which was prepared in advance of our study. The complete experiment was conducted in two different phases, namely

1. building a follower network and
2. pushing the predefined hashtag by spreading tweets through the network.

Figure 4 visualizes both phases within the conceptual view of our proposed bot framework.

For the first phase, we created 30 distinct twitter accounts, each of them consisting of different meta-data such as profile image, hobbies and user location. Within a period of 28 days, all bot accounts automatically increased their reach by following twitter accounts which tweeted about different predefined topics. We focused on trending German hashtags, since the experiment was aimed to a German audience. It should be emphasized that during the first phase, the accounts only retweeted content. None of the accounts actively tweeted any original content.

In preparation to the second phase, a set of 120 unique #songmoji tweets was manually created. This pool of original tweets was used by the social bots

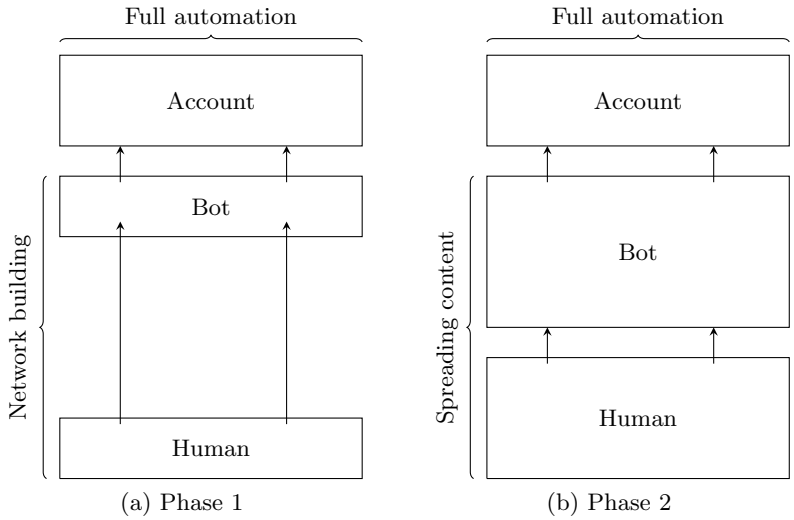


Fig. 4. Conceptual view of the campaign experiment.

to massively spread the hashtag through their follower network. Additionally, all bots automatically liked tweets published by users which adapted the #songmoji hashtag. Furthermore our bots retweeted #songmoji tweets, which were posted by other users. In order to avoid that our bots would be banned by Twitter, because of content spamming, we restricted the actual tweet and retweet frequency to a high but human achievable number of 75 posts per day.

Within Figure 6 the average Botometer scores of all 30 bots over the experimental duration (until the accounts were suspended by Twitter) are visualized. For almost all scores, there is a significant drop, starting at the beginning of the second phase. Especially the average user score drops to a minimum of 0.25. This drop can be explained by the fact that within the second phase, the bots initially started to spread the original tweets that were manually created beforehand. Due to the fact that Botometer's user feature measures, amongst other,

Fig. 5. Example of a predefined "Songmoji".

the number of tweets and retweets of an account, it is not a surprising result that this score drops most. We also observed that at the beginning of the second phase, many users, which showed the willingness to participate at our emoticon game, followed our bot account. Hence, we can also explain the drop of Botometer's Friend and Network score. All in all, we see that an automated, coordinated strategy, executed by more or less simple but orchestrated bot programs cannot be detected by the Botometer service at an individual account level.

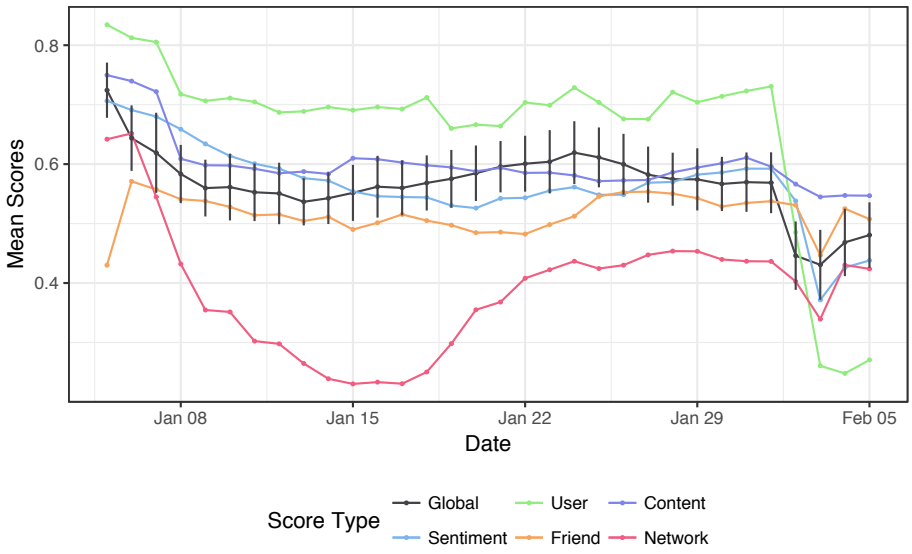


Fig. 6. Average bot scores over time including all average sub-scores.

Although the results indicate that the Botometer service was not able to individually classify our bot accounts correctly, all of them were suspended by Twitter after two days of spreading the hashtag. In our case it was not the Twitter platform itself that detected the bots, but other Twitter users. In contrast to individually analyzing each bot and its actions, the participating users noticed the aggressive behavior of the bot net, e.g, that all of their #songmoji tweets were instantly liked by several bot accounts. Some users reported the accounts to Twitter, which resulted in a ban of the accounts to temporarily prevent them from tweeting. An exemplary user reaction leading to the ban can be seen in Figure 7.

Fig. 7. Detection of our bot army by a user (translated from German, anonymized).

4 On the Importance of Strategy Detection

The previously presented experiments on detection approaches for social bots suggest two main conclusions:

1. Although there are tools available, which base on state-of-the-art pattern recognition, their detection quality is depending on previously learned patterns. Obviously, it is easy to create social bots that bypass these patterns in a largely automated fashion. When human interaction is combined with automatic behavior, profiles cannot reliably be classified anymore by these approaches.
2. Human analytic capabilities are in principal able to detect social bot behavior, as our second experiment demonstrated. The humans, however, do not only focus on specific patterns in single account behavior (micro level). They observe macro effects of multiple automated agents as unusual behavior and sort out the actors participating in a campaign.

While the first conclusion may motivate further research to find even more sophisticated approaches for social bot detection, the second conclusion certainly challenges the current way of social bot detection. Current social bot detection is merely the identification of possible vehicles for information or disinformation in social media. Manipulation or propaganda, however, is the result of applying complex strategies or campaigns in and between social media channels as well as in the "real world". Therein multiple types of content may be used by multiple types of users and groups over long or short periods of time. Often, social media campaigns are accompanied by information and campaigns outside social media.

Considering all this, we wonder: Is it necessary to know a single social bot account, and how do we identify specific threats or strategic attacks to public opinion? And even more pointed: Does it really matter, what kind of actor – human or social bot – is part of a malicious campaign?

Here, we demonstrate our argument with two identified orchestrated campaigns during the German governmental election in September 2017. With the

help of multiple indicators, their combination, and the integration of human intelligence, we identified and verified two coordinated (luckily unsuccessful) manipulative attacks. We find that it is of minor importance, whether the participating accounts are automated or not; the challenging task is to identify the orchestrated behavior of accounts.

4.1 Case 1: A Troll Attack to the TV debate of candidates

In this case study, we present a short summary of an analysis of Twitter usage by troll accounts during the TV debate between the German chancellor Angela Merkel and her contender Martin Schulz (social democrats), with an emphasis on detecting organized communication.

As data source we use German language tweets from the Twitter Gardenhose stream (1 % sample) and from the Decahose stream (a fair 10% sample of all tweets), which contain topic-related hash tags (for details refer to [24]). For this case study, we gathered data between 6:00 pm and 11:59 am on September 3, 2017, resulting in 111,317 tweets.

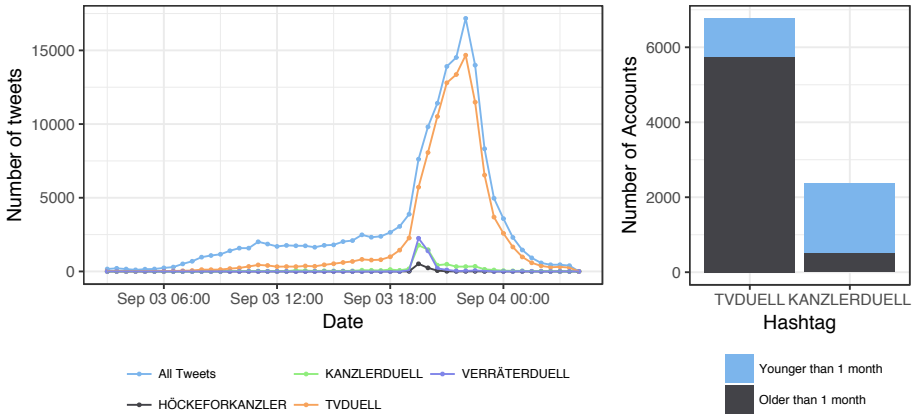


Fig. 8. Important indicators for the first case-study. The figure on the left hand side shows a time series of the activities during the TV debate. The figure on the right hand side shows the proportion of new and old accounts active for two hashtags.

In contrast to existing studies, we employ multiple indicators, some of which are the tweet/retweet relation, the age of twitter accounts, trending hashtag frequency and time series for a descriptive analysis. As a first result, we find that a very high number of new accounts simultaneously tried to push the new hashtag #verräterduell (traitor duel) by combining it with the already existing (and during the TV debate trending) hashtags #kanzlerduell (chancellor duel). The accounts are younger than one month and have mostly been used for retweeting

existing content (without commenting it), to a fraction of 79%. Figure 8 (right) shows the disproportionately high amount of young accounts for the hashtag #kanzlerduell compared to the major hashtag #tvduell for the considered observation. Additionally, Figure 8 (left) gives an impression of the development of several hashtags over time. The campaign is visible as a small activity peak at the beginning of the overall activity peak on Twitter just before the TV debate started.

We presume that what we have documented, was an attempt of an orchestrated attack by human-steered accounts on Twitter that tried to establish a pejorative hashtag hooked onto a neutral one by means of about 380 Twitter accounts, many of which have been established just for being used for this or similar purposes during the election phase. Interestingly, our findings are confirmed by an investigative BuzzFeed publication that refers to an inside report of chat groups that planned to push the mentioned hashtags [25].

4.2 Case 2: A Social Bot Campaign during the German general election

The second analysis was also performed in the context of the German general election and focuses on the activity of social bots, which distribute advertisement for programmatic details of a (small) German party (Freie Wähler). Although the distribution of political advertisement is ethically unproblematic in principal, the respective party proclaimed not to use social bots for campaigns and demanded the flagging of automated profiles in social networks.

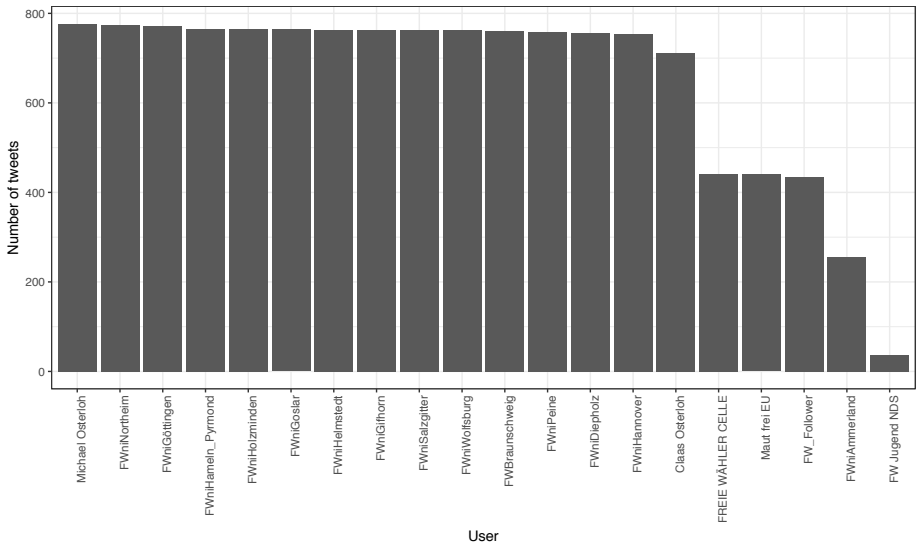


Fig. 9. Most active user accounts for #freiewaehler

German language tweets containing general-election-related hashtags were taken from the Twitter Gardenhose (1 % sample) and Decahose (fair 10 % sample) streams starting at September 10, 2017 until September 25, 2017 (one day after the election), resulting in about 5.5 million tweets.

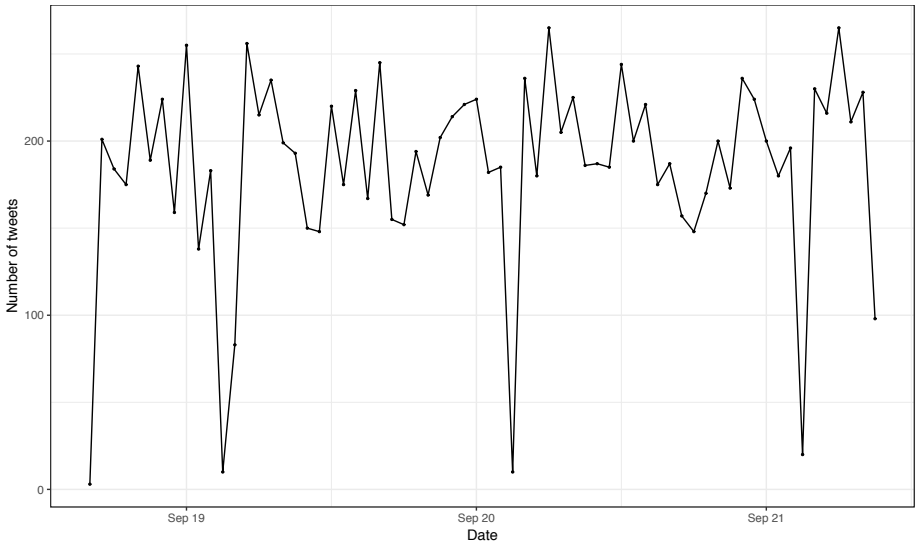


Fig. 10. Number of tweets for #freiewaehler over time.

The indicators in Figures 9 and 10 expose clear patterns of automated behavior. The first indicator simply measures the overall activities for the 20 most active accounts. Interestingly, at least the eighteen most active accounts expose very similar activity behavior. Additional proof of automated actions is provided by the activity time series. We notice a regular drop of activity to almost zero activity at 3:00 am every night. This is caused by a standard network reset procedure at this time. Note, that the use of social bots in this context was later confirmed by the responsible candidate of the respective party – after he was confronted with our findings.

5 Discussion and future directions

The cases shown above highlight campaigns, which were conducted in two extreme ways. One used almost certainly only human actors (trolls). The other one applied social bots to spread content. Both campaigns, however, were centrally coordinated and followed a specific goal, namely spreading ideological content on Twitter to reach a larger audience. In that process, the vehicles for content distribution – humans or bots – are only of secondary interest. The foremost

challenge is to identify the strategy as such. This would have not been possible by analyzing arbitrary user accounts using current detection techniques. With a lot of luck, we would have found some of the very simple bots applied in case 2. The first campaign – promoted by humans – would have been undiscovered. After discovering the campaigns however, we were able to perform a detailed and forensic analysis of the contributing accounts, classifying them as troll or automated accounts, and even finding the responsible actors behind the campaigns.

Therefore we strongly suggest a shift of perspective in current bot detection. As inherently included (but not strictly pursued) by the descriptive approaches and partly addressed by a very recent work of Varol et al. [26], we believe that automated strategy and campaign detection is of major importance for defending against malicious attacks of social bots and human actors alike.

The scientific challenges are to identify patterns in campaigns and attacks rather than in behavior of single actors. This certainly requires – apart from longitudinal observations (time dimension) – to consider data from multiple social media / online platforms (spatial dimension). In the end, this can provide methods, which are able to deal with human-driven, fully automated as well as hybrid campaigns and attacks in cyberspace.

Acknowledgement

This work is part of the PropStop project, which is funded by the German Federal Ministry of Education and Research (FKZ 16KIS0495K). The authors are also supported members of the ERCIS network.

References

1. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: The Socialbot Network: When Bots Socialize for Fame and Money. In: Proceedings of the 27th Annual Computer Security Applications Conference. ACSAC '11, New York, NY, USA, ACM (2011) 93–102
2. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Key Challenges in Defending Against Malicious Socialbots. In: Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats. LEET'12, Berkeley, CA, USA, USENIX Association (2012) 1–4
3. Messias, J., Schmidt, L., Oliveira, R., Benevenuto, F.: You followed my bot! Transforming robots into influential users in Twitter. First Monday (2013)
4. Maréchal, N.: Automation, algorithms, and politics— when bots tweet: Toward a normative framework for bots on social networking sites (feature). International Journal of Communication **10**(0) (2016)
5. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The Rise of Social Bots. (2016) 59,96–104
6. Tynan, D.: Social Spam is taking over the Internet (Apr 2012) http://www.itworld.com/article/2832566/it-management/social-spam-is-taking-over-the-internet.html.

7. Fredheim, R.: Putin's bot army part one: a bit about bots. online (2013) http://quantifyingmemory.blogspot.co.uk/2013/06/putins-bots-part-one-bit-about-bots.html.
8. Elliott, C.: The readers' editor on pro-Russia trolling below the line on Ukraine stories. online (may 2014) http://www.theguardian.com/commentisfree/2014/may/04/pro-russia-trolls-ukraine-guardian-online.
9. Ohlheiser, A.: Trolls turned tay, microsofts fun millennial ai bot, into a genocidal maniac (Mar 2016)
10. Rosenbach, Marcel, S.: Internet-Kommentare von Automaten: AfD will im Wahlkampf Meinungsroboter einsetzen (Oct 2016)
11. Pfaffenzeller, M.: Bundestagswahlkampf: CDU erwägt Einsatz von Chatbots (Mar 2017)
12. Howard, P.N., Kollanyi, B.: Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. (2016)
13. Kollanyi, B., Howard, P.N., Woolley, S.C.: Bots and Automation over Twitter during the US Election. Technical Report Data Memo 2016.4, Project on Computational Propaganda. www.politicalbots.org, Oxford, UK
14. Neudert, L.M.N.: Computational propaganda in Germany: A cautionary Tale. Technical report, Project on Computational Propaganda. www.politicalbots.org (2017)
15. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online Human-Bot Interactions: Detection, Estimation, and Characterization. (2017)
16. Grimme, C., Preuss, M., Adam, L., Trautmann, H.: Social Bots: Human-Like by Means of Human Control? Big Data **5**(4) (2017) 279–293
17. Paradise, A., Puzis, R., Shabtai, A.: Anti-Reconnaissance Tools: Detecting Targeted Socialbots. IEEE Internet Computing **18**(5) (2014) 11–19
18. Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B.Y., Dai, Y.: Uncovering Social Network Sybils in the Wild. ACM Trans. Knowl. Discov. Data **8**(1) (February 2014) 2:1–2:29
19. Clark, E.M., Williams, J.R., Galbraith, R.A., Jones, C.A., Danforth, C.M., Dodds, P.S.: Sifting robotic from organic text: A natural language approach for detecting automation on Twitter. Journal of Computational Science **16** (2016) 1–7
20. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: A system to evaluate social bots. CoRR **abs/1602.00975** (2016)
21. Hegelich, S., Janetzko, D.: Are social bots on twitter political actors? Empirical evidence from a Ukrainian social botnet. In: International AAAI Conference on Web and Social Media. (2016) 579–582
22. Echeverra, J., Zhou, S.: The 'Star Wars' botnet with >350k Twitter bots. CoRR **abs/1701.02405** (2017)
23. Cao, Q., Yang, X., Yu, J., Palow, C.: Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. CCS '14, New York, NY, USA, ACM (2014) 477–488
24. Grimme, C., Assenmacher, D., Adam, L., Preuss, M., Stockdiek, J.F.H.L.: Bundestagswahl 2017: Social-Media-Angriff auf das #kanzlerduell? Technical Report 2017.1, Project PropStop (www.propstop.de), Münster, Germany
25. Schmehl, K.: Diese geheimen Chats zeigen, wer hinter dem Meme-Angriff #Verräterduell aufs TV-Duell steckt
26. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. EPJ Data Science **6**(1) (2017) 13