GOTCHA BOT DETECTION: CONTEXT, TIME AND PLACE MATTERS

Rose Marie Santini, Débora Salles, Fernando Ferreira, Felipe Grael

https://doi.org/10.1590/SciELOPreprints.5974

Submitted on: 2023-04-28
Posted on: 2023-05-05 (version 1)
(YYYY-MM-DD)

GOTCHA BOT DETECTION: CONTEXT, TIME AND PLACE MATTERS

Rose Marie Santini
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
ORCID: https://orcid.org/0000-0003-0657-7217

Débora Salles
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
ORCID: https://orcid.org/0000-0002-3436-6698

Fernando Ferreira
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
ORCID: https://orcid.org/0000-0003-3455-2316

Felipe Grael
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
ORCID: https://orcid.org/0000-0002-9261-379X

ABSTRACT

Bot detection is increasingly relevant considering that automated accounts play a disproportionate role in spreading disinformation, controlling social interactions, influencing social media algorithms and manufacturing public opinion online for different purposes. Definition, description and detection of automated manipulation techniques have proved a challenge as technology quickly advances in reach and sophistication. Considering the high contextual character of social science research, the employment of off-the-shelf detection tools raises questions regarding the applicability of machine learning systems in different cases, times and places. Thus, our purpose is to discuss the role of computational methods focusing on understanding the limitations and potential of machine learning systems to identify bots on social media platforms. To address it, we analyze the performance of Botometer, a widely adopted detection tool, in a specific domain (Amazon Forest Fires) and language (Portuguese) and propose a supervised machine learning classifier, called Gotcha, based on Botometer's framework and trained for this specific dataset. We also question how our classifier behaves and evolves over time and perform tests to evaluate the generalization capabilities of the retrained model. Our results demonstrated that supervised methods do not perform well with datasets that present features on which the system was not directly trained, such as language and topic. Hence, our study shows that a successful computational model does not always guarantee reliable results, applicable to a specific real case. Our findings indicate the need for social scientists to confirm the reliability of different tools created and tested only through the prism of computational studies before applying them to empirical social science research.

Keywords: Bot detection, machine learning algorithm, Brazil, computational propaganda

INTRODUCTION

The use of social bots to strategically change election results has been broadly discussed and documented worldwide since 2016, for controlling the narratives during political campaigns and to skew the political debate (Bastos & Mercea, 2019; Mønsted et al., 2017; Santini et al., 2021). Indeed, research has shown that bots have extrapolated electoral campaigns and have been used to control social interactions and manufacture public opinion online for different purposes - i.e., to disrupt mainstream media coverage (Zhdanova & Orlova, 2017), to undermine civic protests (Arif et al., 2018), to make up audience and influence website ratings (Santini et al., 2020) and to spread low credibility content on different issues (Shao et al., 2018).

The relevance of bot detection is even greater considering that scientific evidence consistently shows that social bots play a disproportionate role in spreading disinformation (Schuchard et al., 2019) and influencing social media algorithm curation (Giansiracusa, 2021). Despite the importance of effective detection of automated malicious campaigns in strengthening a platform's security (Rauchfleisch & Kaiser, 2020), responses by social media companies have been inadequate to contrast computational manipulation (Beatson et al., 2021).

Measuring the social implications of fake and automated accounts in orchestrated campaigns on social media is an open challenge for the research community (Chen & Subramanian, 2018; Cresci et al., 2017; Luceri et al., 2019; Varol et al., 2017). Definition, description and detection of automated manipulation techniques have proved a challenge as technology quickly advances in reach and sophistication, setting off an “algorithmic race” (Santini et al., 2018) for commercial and state players. The boundary between humanlike and bot-like behavior is becoming blurred and trivial detection strategies, such as focusing on high volume of content generation, are less and less successful in identifying sophisticated bots (Ferrara et al., 2016).

Moreover, Briant (2021) points out that, despite the boom in media coverage on bot influence on social media discussions, academic communication research has relied heavily upon investigative journalism on manipulation strategies instead of using primary evidence and rigorous methodological approaches. The lack of computational analysis suggests that these studies underestimate the scale of the bot phenomena. Although public awareness generated by such reporting is capable of leading to important changes in platform behavior (Barrett & Kreiss, 2019; Napoli, 2021), a growing body of knowledge on bot detection techniques have evolved to encompass more complex and elusive manipulation strategies employed by botmasters (Cresci et al., 2017, 2018; Yang et al., 2013).

The most emblematic methods for automation identification combine machine learning with manual qualitative analysis curation to detect messaging patterns (Howard & Kollanyi, 2016). Researchers frequently employ established tools like Botometer in their method approaches, as its large body of peer-reviewed publications lend authority and legitimacy to their analysis (Rauchfleisch & Kaiser, 2020). Thus, there is a pressing need to understand how social science researchers can use the available machine learning tools to be applicable in different cases, times and places, considering social science findings are highly contextually dependent (van Atteveldt & Peng, 2018).

We have addressed this literature gap by analyzing the performance of an existing off-the-shelf tool with published validity results, in order to understand how well it performs in a specific domain (Amazon Forest Fires) and in a specific language (Portuguese). Given the novelty and complexity of the bot detection task, we aim to develop an analytical model to understand what limitations Botometer might have when dealing with different datasets, topics and language. We have specifically tested the classifiers performance in the Brazilian Twittersphere domain concerning the climate debate using longitudinal analysis.

BACKGROUND AND FRAMING

The concern about robots and their impact on society is not new, it has been around since the early days of computational science. However, the research on social bots has gained attention over the last 10 years with the increasing use of automated accounts to mimic human behavior on social media platforms and their ubiquity in political conversation online. Researchers such as Ratkiewicz et al. (2011) and Ferrara et al. (2016) were fundamental in building the preliminary understandings of what a social bot is, how to detect these automated accounts and what their possible effects are.

Social bots are defined as false online identities that try to emulate and possibly alter human behavior using computer scripts to automatically produce content and interact on social media (Ferrara et al., 2016; Varol et al., 2017). However, there is a growing body of knowledge showing that social bots are not restricted to software programs acting on online platforms. Instead, researchers have been considering social bots as any type of bot-like automated behavior, such as humans manually copying or retweeting content repeatedly in a robotlike way for the purposes of manipulating online conversation (Fernquist et al., 2018).

This widened definition of bots goes hand in hand with a certain unreliability when it comes to detecting bots (Guimarães et al., 2021). There are numerous cases of Twitter accounts that are cyborgs and hybrid accounts (a combination of automation and human curation) and others that look like bots but are actually human and vice versa. For example, malignant bots are often automated for some percentage of time, whether that is intra-day or over their life cycle, particularly if the “bot master” wants to age the account so that it doesn't look like it was created for one purpose (DiResta, 2018). Hence, the identification of bots is an important but complicated task. Due to the mutable nature of bots, coupled with their continuous online presence and inevitable interplay with human users, defining a bot is anything but an exact science (Bastos & Mercea, 2018), as researchers lack accuracy in identifying their characteristics, activity patterns, automation degree and profile types. Even identifying fully automated accounts is challenging (Rauchfleisch & Kaiser, 2020).

Tracking the evolution of bots and human behavior online, researchers have shown that bots are increasingly aligned with human activity trends, suggesting that some have grown more sophisticated (Luceri et al., 2019) While it is widely known that there are large numbers of inauthentic accounts online (Varol et al., 2017) estimate that 9% to 15% of all active Twitter accounts are bots) the best strategy for detecting bots remains an open question (Howard et al., 2018). Against this backdrop, several methods have been introduced to tackle this problem, including posting pattern analysis (Bastos & Mercea, 2019; Santini et al., 2021), botnet detection

(Mazza et al., 2019), machine learning classifiers (Davis et al., 2016; Sayyadiharikandeh et al., 2020; Varol et al., 2017), as well as deep learning approaches that combine traditional handcrafted features with embeddings from temporal activity patterns and posted user content (Mou & Lee, 2020).

Among different techniques and trained classifiers, Botometer (formerly BotOrNot), developed at Indiana University, has established itself as a standard bot detection tool in the social sciences, particularly for communication studies. As a matter of research, Grimme et al. (2018) demonstrated that Botometer could not classify fully automated bot accounts and hybrid profiles created precisely by authors in their experiment. Rauchfleisch and Kaiser (2020) have also shown in their study that Botometer scores were imprecise when it comes to estimating bots in a language other than English (i.e., German) and that, over time, Botometer thresholds are prone to variance, leading to false negatives (bots being classified as humans) and false positives (humans being classified as bots). Fernquist et al. (2018) also demonstrated that Botometer fails with non-English language tweets. The authors reported that their own supervised classifier trained in Swedish tweets outperforms Botometer for Swedish General Election analysis.

Despite the fact that bot detection is a difficult machine learning task, particularly because the evolving sophistication of bots generates a constant need for updating the decision patterns of computational classifiers with improvements and retraining (Yan et al., 2020), few researchers have explicitly discussed the diagnostic ability of Botometer and similar tools. A recent research trend is to test the limits of current bot detection frameworks in an adversarial setting (Yan et al., 2020). For example, Cresci et al. (2019) proposed the use of evolutionary algorithms to improve social bot skills and Grimme et al. (2018) employed a hybrid approach involving automatic and manual actions to improve a supervised bot detection system. Yang et al. (2019) argue that the technology race between methods to develop sophisticated bots and effective countermeasures makes it necessary to update the training data and features of detection tools.

The goal of this paper is to follow this research agenda by proposing a supervised machine learning classifier, called Gotcha, based on a subset of features adopted in Botometer's framework (Varol et al., 2017) and trained with a Portuguese dataset. Besides critically assessing both Botometer and Gotcha's performance, we aim to understand how our classifier behaves and evolves over time. We question the effects of enlarging the annotated ground truth dataset and perform tests to evaluate the generalization capabilities of the retrained model. Our ultimate purpose is to discuss the role of computational methods in social science research focusing on understanding the limitations and potential of machine learning systems to identify bots on social media platforms.

RESEARCH QUESTIONS

Considering the prominent use of Botometer in computational social science studies worldwide as a classifier of automated accounts, its general diagnostic ability in detecting non-English language automated tweets is our primary concern. The tool faces pitfalls when classifying hybrid automated accounts (Grimme et al., 2018), dealing with non-English language tweets (Echeverría et al., 2018) and identifying updated and sophisticated bots (Cresci et al., 2017). As

Fernquist et al. (2018) indicate, there might be differences in the performance of Botometer and other trained classification models regarding language and topics.

> *RQ1: How accurate is Botometer in classifying a specific dataset in Brazilian Portuguese, concerning tweets about the Amazon Forest Fires?*

We aim at developing a detection system based on a supervised machine-learning classifier, that we named Gotcha. Thus, we evaluate its performance and compare it to Botometer, assessing whether there are differences between languages and methods:

> *RQ2: With regard to the precision and recall diagnostic ability, what is the performance of a machine learning system (Gotcha) based on the Botometer framework but trained to classify accounts in Brazilian Portuguese when compared to Botometer?*

Previous research has shown that bot classification scores are not stable over time (Grimme et al., 2018). Therefore, we are interested in not only measuring Gotcha's performance once but in tracking its performance over a period of time. For this, we analyze datasets related to the same issue (Amazon Forest Fires) in different time periods and investigate how Gotcha performs:

> *RQ3: How stable are Gotcha classifications over time on the same topic (Amazon Forest Fire crisis in 2019 and 2020)?*

In order to approach Gotcha's performance decay over time, we follow suggestions made in previous research (Yang et al., 2019) and enrich the previously-trained model using the new annotations. Thus, we investigate the effects of increasing and updating annotated datasets:

> *RQ4: What is the accuracy and precision of Gotcha if new annotations and training data are included to update the classifier parameters for a new dataset in Brazilian Portuguese on the same topic?*

In answering these four questions, we want to shine a light on some of the issues related to bot detection, to indicate the limitations of Botometer classification and to discuss what consequences this has for social scientists conducting research on bots.

DATA & METHODS

The datasets of Twitter accounts have been extracted by monitoring the terms and keywords associated with recent events in Brazil, ranging from 2018 to 2021. We collect the data using the Standard Tier of the Twitter API, performing periodic requests for each of the search terms using the */search* endpoint, in order to monitor real-time evolution of events. Afterwards, we randomly selected a subset of profiles considering the acquired data in each period of time and manually annotated the selected accounts.

We briefly present the 7 datasets used to train and test the proposed framework. The monitored period of each event, as well as summary of statistics regarding the number of tweets and distinct users collected, are summarized in Table 1.

The *elections_2018* dataset consists of accounts which were posted during the last Brazilian General Elections (2018), between August and October, 2018, using search terms associated with the profiles of each candidate running for president. The *amazon_2019* dataset contains accounts monitored during the "Fire Season" in the Brazilian Amazon in 2019, searching for hashtags like "#SOSAmazon", "#AmazonRainforest" and "#PrayForAmazonia". The *sleeping_giants* dataset is composed of accounts that acted in the disinformation campaign against the Sleeping Giants Brasil organization in May 2020. The *marielle* dataset was generated from monitoring the "#FederalizaçãoNão" hashtag on Twitter (May 2020), as part of the campaign against Federalizing the investigation into the murder of Marielle Franco, a Rio de Janeiro councilwoman. The *soros* dataset, related to the #StopSoros campaign on Twitter, was taken from the Portuguese tweets extracted by monitoring the campaign against Hungarian investor and philanthropist George Soros on 12 August 2020. The *globolixo* dataset was collected during the #GloboLixo campaign against the most popular TV station in Brazil, in Jan 2020. Finally, the amazon_2020 dataset was collected during the 2020 "Fire Season" in the Brazilian Amazon. The monitored period of each event, as well as summary of statistics regarding the number of tweets and distinct users collected, are summarized in Table 1.

Table 1: Data Acquisition Results from 2018 to 2020

Dataset	Start Date	End Date	Number of tweets	Number of distinct users
elections_2018	2018-08-05	2018-11-04	26,350,192.00	2,231,522.00
amazon_2019	2019-08-23	2019-09-30	1,709,161.00	688,476.00
marielle	2020-05-21	2020-06-01	218,648.00	120,985.00
sleeping_giants	2020-05-19	2020-05-23	490,020.00	127,335.00
soros	2020-08-12	2020-08-13	310,872.00	47,631.00
globolixo	2021-01-02	2021-01-04	60,947.00	23,442.00
amazon_2020	2020-06-19	2020-10-07	358,494.00	192,544.00

Criteria for manual annotation

We performed manual annotations of a subset of accounts in order to provide a large and reliable ground-truth dataset, as the manual identification of accounts is useful to aid the training of supervised learning algorithms (Alvisi et al., 2013; Ferrara et al., 2016). An interpretative observational analysis of the bots' profiles was carried out by at least two coders, and reliability was guaranteed by a third analysis. Wang et al. (2014) have demonstrated that expert annotators are efficient in detecting social bot accounts, indicating that a majority voting protocol, in which a profile is categorized based on multiple human analyses, exhibits a near-zero false positive rate.

The profile analysis aimed at identifying computational routines based on observable traces of the account characteristics and activity. Regarding the profile information, the following criteria were considered in the expert categorization of the accounts: (a) account creation date; (b) profile

description; (c) profile image content and authenticity; and (d) pursuit of anonymity and untraceability. In regard to the accounts' posting behavior, coders examined: (a) amount and content of tweets posted by the account; (b) amount and content of retweets posted by the account; (c) production of human-like and original content (i.e., personal comments and opinions on posts and use of natural language); (d) participation in coordinated campaigns and botnets (i.e. adherence to trending topics and suspicious followers and interactions); (e) use of automation service or platform; (f) posting patterns regarding time period, content, source, theme and interactions. The number of human and bot accounts annotated in each dataset are summarized in Table 2:

Table 2: Manually annotated datasets for the Gotcha Framework

Dataset	humans	bots
elections_2018	320	159
amazon_2019	76	118
sleeping_giants	36	61
Marielle franco	154	42
George Soros	108	97
globolixo	120	70
amazon_2020	108	100
	922	**647**

Training Methodology

For each labeled account, we collect its 200 most recent tweets at the time, as well as the 100 most recent posts mentioning the analyzed user, which corresponds to the maximum number of tweets that one can gather when performing a single request for the corresponding endpoints (statuses/user_timeline and users/lookup) in the Twitter API Standard Tier.

During the training phase, we performed cross validation in order to choose the best hyperparameter setting, using Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) as our classifier, which gives the probability of an account being a bot. The candidate models were evaluated according to the ROC AUC metric, and we considered the model exhibiting the higher mean ROC AUC score across the folds as the best model.

In order to refine our model, we incrementally augment the dataset used as the training set, as new events arose and subsequently new datasets were collected and labeled, yielding updated versions of the proposed framework. We split each new dataset into training and test sets, in a stratified fashion, i.e. preserving the proportion of bots and humans in each one. We used 80% of the labeled accounts to augment the existing training set, and considered the remaining 20% of accounts as the test set.

Thus, our framework was capable of achieving a better generalization performance in cross-domain scenarios, as it was being continuously refined and exposed to automated accounts with more sophisticated behaviors, and was able to identify previously unseen types of bots.

When designing a new version of the proposed framework, we augmented the existing training set (which comprises all the training sets from the previous events), and used the test set of the actual event to evaluate the generalization of the new model. Thus, we could better assess the suitability of the model in detecting the automated behaviors that were present in the scenario under analysis.

The proposed framework

The Gotcha classifier is a bot detection framework based on the Botometer framework (Davis et al., 2016) that leverages 1150 characteristics and different types of information for a given user (Varol et al., 2017) to quantify its degree of automation, originally trained to classify users who publish in English. Our model, developed for the analysis of users that post in Portuguese, uses a subset of 525 characteristics based on those used in the Botometer model (Dong & Liu, 2018). It excludes the groups of Network features, which are the most expensive to compute (Mazza et al., 2019), as well as the Sentiment features, which are specific for the English language. The authors also recommend the models to be trained without the aforementioned set of features on different languages (Dong & Liu, 2018). Similar to the Botometer model, the Gotcha features are divided into 4 groups: User, Friends, Content and Temporal Pattern, summarized on Table 3:

Table 3: 525 features of the Gotcha framework

User features	**Friend features**
Screen name length	Number of distinct languages for each group
Number of digits in screen name	Entropy of language use for each group
User name length	Fraction of users with default profile for each group
Default profile	Fraction of users with default profile image for each group
Default picture	(*) Account age distribution
Account age	(*) Number of friends distribution
Number of unique profile descriptions	(*) Number of followers distribution
(*) Number of friends distribution	(*) Number of tweets distribution
(*) Number of followers distribution	(*) Description length distribution
(*) Number of favorites distribution	**Content features**
(*) Profile description length distribution	(*) Number of words in a tweet
(**) Number of friends	(*, ***) Frequency of POS tags in a tweet
(**) Number of followers	(*, ***) Proportion of POS tags in a tweet
(**) Number of favorites	**Temporal features**
Number of tweets (per hour and total)	(*) Time between two consecutive tweets
Number of retweets (per hour and total)	(*) Time between two consecutive retweets

User features	**Friend features**
Screen name length	Number of distinct languages for each group
Number of digits in screen name	Entropy of language use for each group
Number of mentions (per hour and total)	(*) Time between two consecutive mentions
Number of replies (per hour and total)	
(*) For each distribution, the following eight statistics are computed and used as an individual feature: min, max, median, mean, std. deviation, skewness, kurtosis and entropy.	
(**) The signal-noise ratio and the relative change values are computed.	
(***) The following POS tags are considered: verbs, nouns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, wh-, pronouns and the "others" category (such as coordinating conjunctions and cardinal numbers.	

User Features: The features in this group are derived from the user metadata being analyzed, such as the number of characters and number of digits in the username, whether the username uses standard cover and profile photos and the age of the account. Additionally, the user's 200 most recent tweets are collected to extract features such as the total and per hour number of original tweets, retweets and mentions. We also compute features using the metadata from the accounts that interact with the user under analysis, including descriptive statistics on the distribution of their followers and friends.

Friend Features: These features are computed based on the data from interacting users, which first are divided into 4 groups: **(1)** accounts that the analyzed user retweets; **(2)** accounts it mentions; **(3)** accounts that retweet the analyzed user and **(4)** the profiles that mention the analyzed user. For each of them, we extract features such as the number of different languages used in tweets from that group, as well as the fraction of users who use cover photo and standard profile pictures. We also compute features using the age of each account, the number of friends, followers, tweets and the number of characters in each profile description.

Content Features: In order to characterize the linguistic patterns used by the analyzed user, we compute the number of words in each of its 200 most recent posts, as well as the entropy of word usage. Furthermore, each post is submitted to a POS tagging algorithm (Loper et al., 2009), a Natural Language Processing task to identify the grammatical class, such as pronouns and adverbs, of each word in a given text. Nine different classes are considered and, for each of them, we compute statistics on the absolute and relative frequency distributions.

Temporal Pattern Features: Finally, to analyze the user's temporal post patterns, its posts are separated into 3 groups: **(1)** user tweets; **(2)** retweets and **(3)** mentions. For each of them, the tweets are arranged in chronological order and the time intervals between two consecutive activities are calculated. We compute some descriptive statistics using the interval distributions for each group.

The first version of the Gotcha framework was trained using the whole elections_2018 dataset, and evaluated on the amazon_2019 dataset. The generalization performance in the first version, as well as a comparison with existing methodologies, will be presented in a later section. We reference this version as **Gotcha_v1** throughout the following sections.

The framework has been constantly updated as new accounts are labeled, in order to ensure a good performance across different contexts. As previously reported in the literature, automated accounts with unseen characteristics can easily deceive bot detection systems (Echeverría et al., 2018), as they are unable to recognize the newer activity patterns of more sophisticated bots. Furthermore, the arms race between researchers and bot developers is intensifying, since as new detection methods are introduced, the bot developers update the *modus operandi* of their accounts, in order to evade the most recent systems (Cresci et al., 2017).

This reported behavior has motivated the periodical updates of the Gotcha classifier, as we observed a similar drop in the performance of our model when applied in new scenarios in our use cases, even when we analyzed the same topic in different periods of time.

We report these results in the next section, where we show how the current version of the Gotcha framework performed when evaluating accounts from the same context, one year later (amazon_2019 and amazon_2020). We reference this version as **Gotcha_01-2021**, which has been trained using all the aforementioned datasets but the amazon_2020 one. We also evaluate the impact of the refinement procedure on the Gotcha performance in the amazon_2020 test set, after being trained using the training set of all datasets, generating its latest version, the **Gotcha_07-2021**.

RESULTS

Evaluation on the Amazon 2019 dataset

In order to address Research Questions 1 and 2, we evaluated the **Gotcha_v1** using the **amazon_2019** dataset and compared its performance with the predictions of Botometer. We consider the accuracy, precision, recall, F1 score and ROC AUC as our evaluation metrics. Table 4 presents the results of the two methodologies:

Table 4: Performance evaluation on the amazon_2019 dataset

	Precision	Recall	F1	Accuracy	ROC AUC
Gotcha_v1	**0.67**	**0.85**	**0.75**	**0.66**	**0.68**
Botometer	0.58	0.48	0.53	0.48	0.54

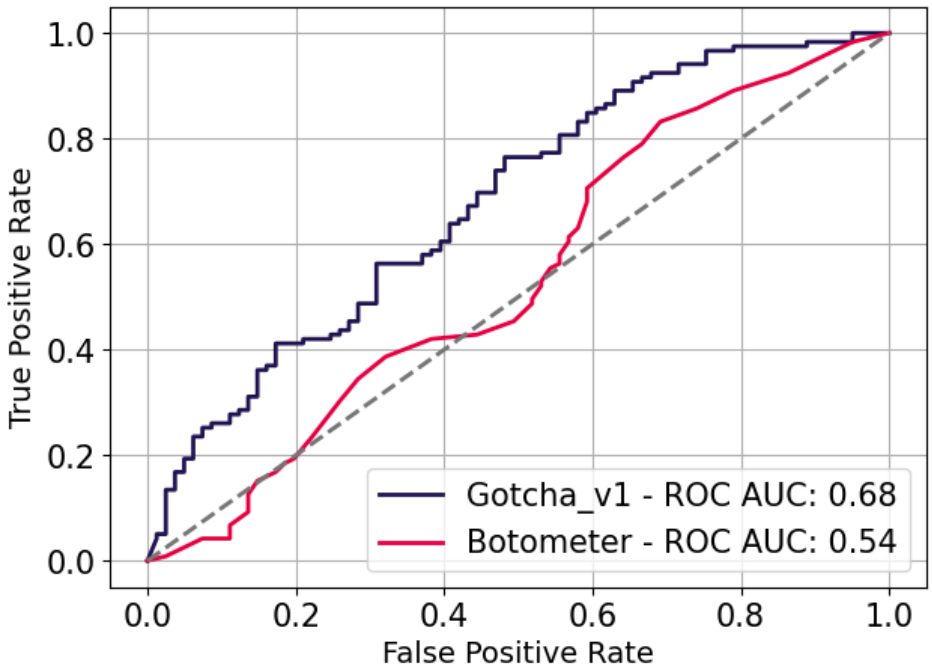

Figure 1: ROC curve for Gotcha_v1 and Botometer frameworks in the *amazon_2019* dataset.

Gotcha_v1 outperforms Botomoter, exhibiting the best overall performance across all analyzed metrics. We observe a substantial prevalence of our model in the recall score, indicating a superior capability in identifying the automated users in the analyzed scenario. These results provide further evidence that the performance of bot classifiers can drop considerably in cross-domain scenarios. Furthermore, our framework, which has been trained using accounts from a dataset of Portuguese tweets, outperformed the Botometer in all metrics by a large margin.

Performance comparison on unanimous annotation cases

Our reported results considered the whole labeled dataset. Additionally, we performed an evaluation considering the subset of accounts where all annotators were unanimous in their codification, which corresponds to 67% (130) of the total number of accounts. As previously reported in the literature, some types of bots are able to masterfully imitate human behavior on social networks, making it difficult to identify that they are automated (Cresci et al., 2017, 2018). This could affect the annotation process, causing a split decision on the class of some accounts. Table 5 presents the results for evaluating the subset of unanimous-labeled accounts:

Table 5: Performance evaluation on the subset of unanimous-labeled accounts for the amazon_2019 dataset

	Precision	Recall	F1	Accuracy	ROC AUC
Gotcha_v1	**0.77**	**0.95**	**0.85**	**0.79**	**0.80**

Botometer	0.66	0.52	0.59	0.53	0.58

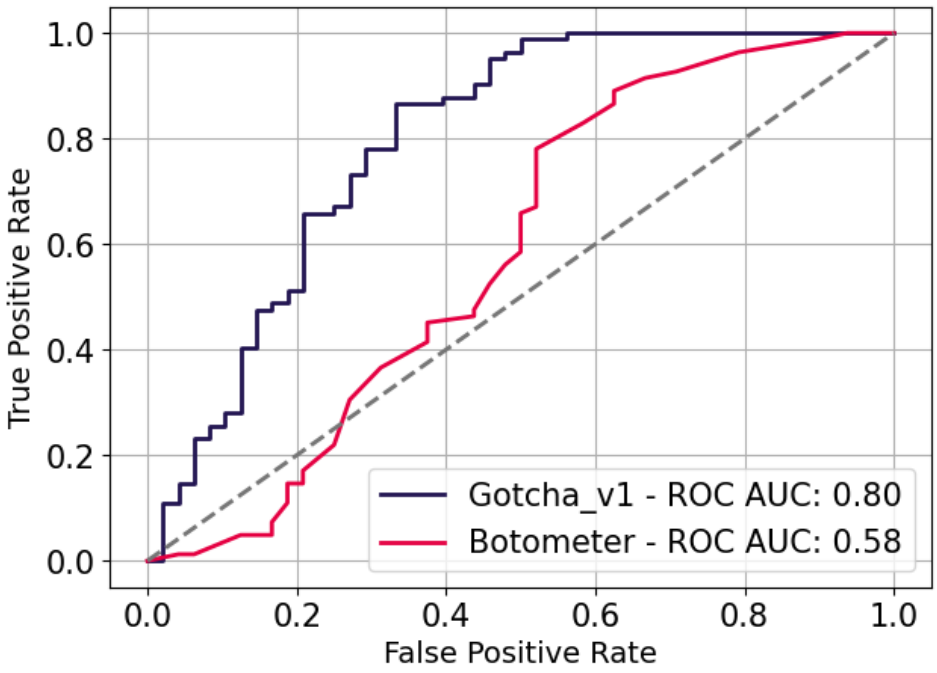

Figure 2: ROC curve for the Gotcha_v1 and Botometer frameworks considering the subset of unanimous-labeled accounts in the *amazon_2019* dataset.

As observed in the previous experiment, the **Gotcha_v1** performed consistently better than Botometer in identifying the bots that were active during the "Amazon Fire Season" in 2019. Moreover, the performance of both frameworks improved all the analyzed metrics, suggesting that the bot classifiers are better at distinguishing between authentic and automated accounts when the differences are clearer for the human annotators.

Evaluation of the Gotcha performance on the Amazon 2020 dataset

For Research Question 3, we evaluated the performance of **Gotcha_01-2021** on the test set of our newest amazon_2020 dataset. As mentioned in the previous section, this version had been trained using all training sets of labeled accounts available at the time, excluding the amazon_2020 training set, in order to evaluate the generalization capability of our most up-to-date model for identifying bots on the same topic ("Amazon Fire Season") one year later.

It is worth mentioning that bot behaviors detected in the amazon_2019 dataset were considered during the **Gotcha_01-2021** training phase. Thus, it was expected that this version of the framework would be able to identify the automated accounts if they behave in the same way as the previous year. We observed a precision of 0.53, recall of 0.56, F1-score of 0.54, accuracy of 0.55 and a ROC AUC of 0.64. These results indicate that the capability of the Gotcha framework has significantly dropped over time, as the framework was not able to perform as well as in the 2019 scenario. This result also suggests a shift in the *modus operandi* of the automated accounts

that were active in the same context during 2020. Thus, it would be desirable to refine the procedure in order to ensure the reliability of further bot analysis on this recent event.

To address Research Question 4, we augmented the existing training set of bot and humans with the training accounts extracted from the amazon_2020 dataset, retraining our framework using this new augmented dataset. After the refinement procedure, the newest Gotcha version (referenced as **Gotcha_07-2021**), exhibited a precision of 0.8, a recall of 0.67, a F1-score of 0.73, an accuracy of 0.76 and a ROC AUC of 0.84 in the amazon_2020 test set, achieving a much better performance than its predecessor. Table 6 summarizes the evaluation metrics for both models:

Table 6: Effect of the refinement procedure on the test set of the amazon_2020 dataset

	Precision	Recall	F1	Accuracy	ROC AUC
Gotcha_01-2021	0.53	0.56	0.54	0.55	0.64
Gotcha_07-2021	**0.8**	**0.67**	**0.73**	**0.76**	**0.84**

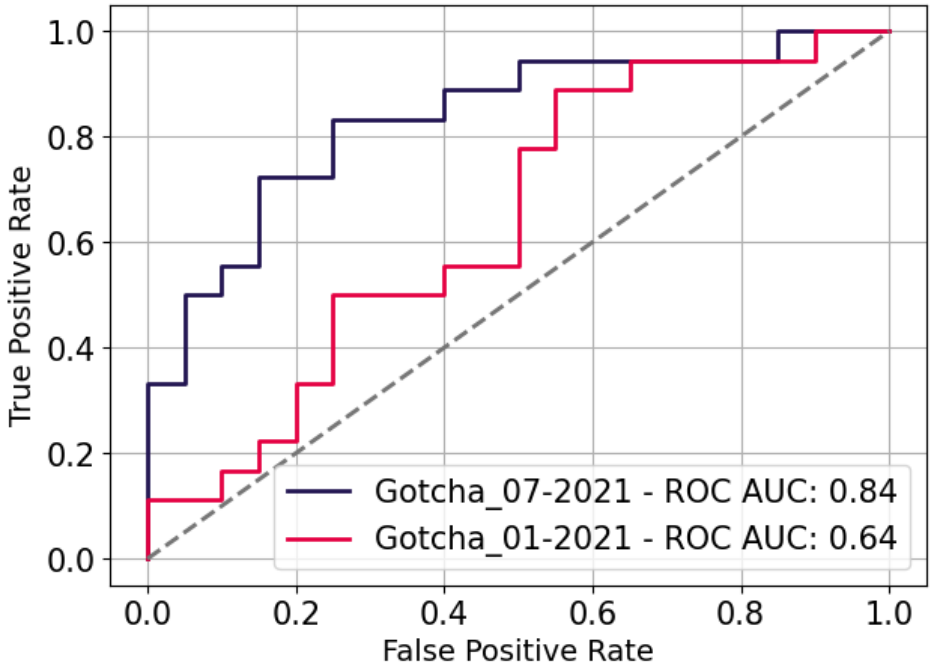

Figure 3: ROC curve for the Gotcha_07-2021 and Gotcha_01-2021 frameworks in the *amazon_2020* test set.

The observed results indicate that the refinement procedure is beneficial for improving the classifier's ability to identify automated users in newer contexts, as we observed considerable improvements with the refined mode in all analyzed metrics l. We argue that this refinement should even be mandatory, as the behavior of the automated accounts is constantly evolving to evade the current detection systems. Furthermore, refining the procedure would also be desirable for ensuring analysis reliability using classifier predictions, as this could be used to assess the

framework's generalization performance, as well as how the framework performs when exposed to possible new bot behavior.

Model Interpretation

In order to better understand the decision process of the latest version of the Gotcha classifier, we applied the SHAP methodology for model interpretability (Lundberg et al., 2020) on **Gotcha_07-2021**. Figure 4 shows the results:

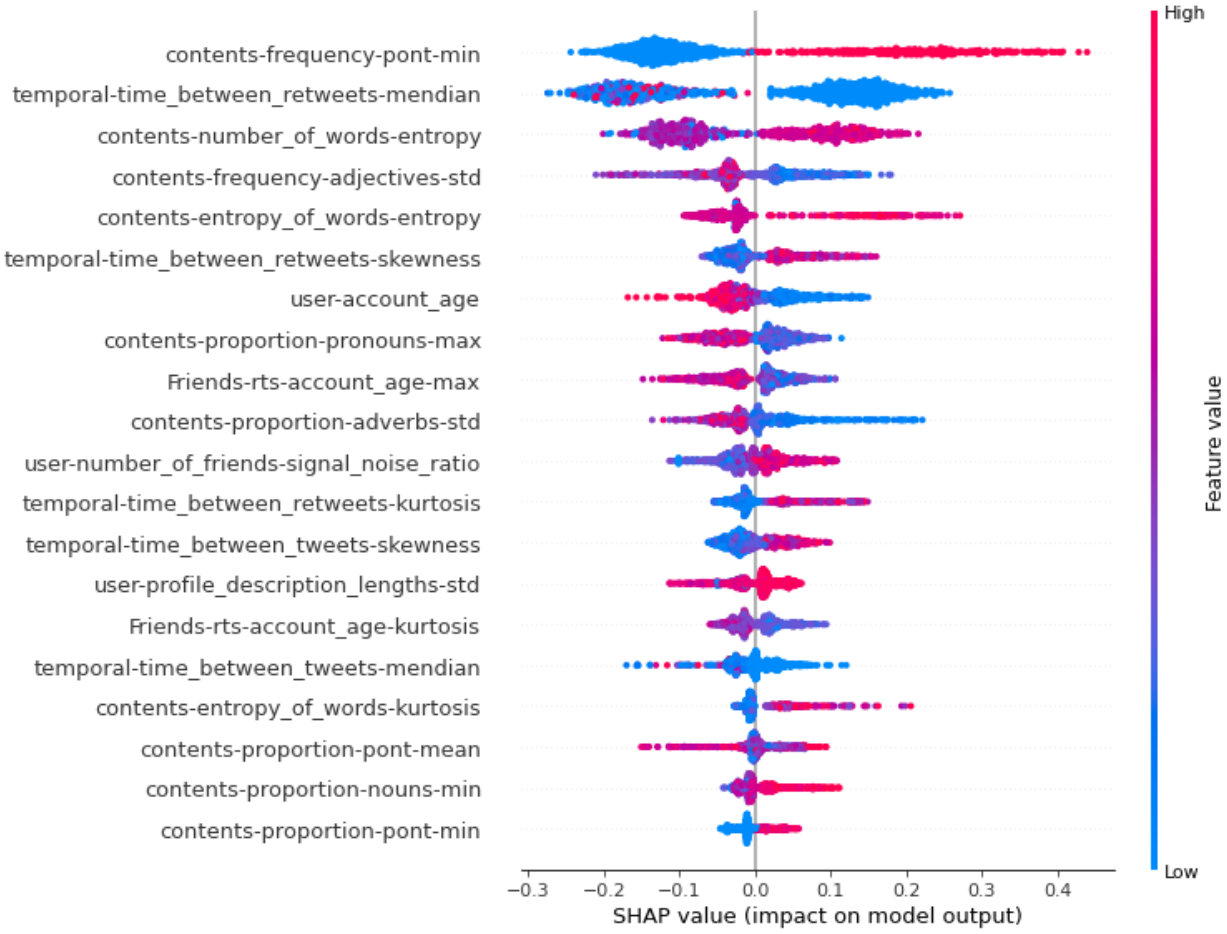

Figure 4: Summary plot of the SHAP values for the *Gotcha_07-2021* model. The values of each feature, for each sample in the training set, are represented as a point in the plot. A hue towards red indicates a relatively high value for this feature, according to the empirical feature distribution in the training set, whereas a bluer hue indicates a relatively low value. The values to the left of the central y-axis are for classifying an account as a human, while the values to the right are for classifying the profile as a bot.

Figure 4 indicates, for example, a relatively low value for the median time interval between retweets contributes towards classifying an account as bot, suggesting that the analyzed profile could be an automated user retweeting content in a short period of time. Furthermore, we observe that recently created accounts, exhibiting a low account_age value, are prone to be

classified as bots. These users could correspond to automated accounts recently introduced to social media platforms to meet new demands, such as specific campaigns or events.

Additionally, we observe that features derived from the content are also important during the classification process, such as the entropy of the word usage distribution, the frequency of adjectives and nouns. Interestingly, one of the most important content features is the standard deviation of the usage of adverbs by accounts. A low value for this feature contributes to assigning bot-like behavior, while a higher value represents an account with stronger signals of authenticity. This behavior could be associated with a more diverse usa of language by real users, as opposed to bot accounts which might act together to increase the popularity of the same topic or person, potentially sharing the same restricted set of posts.

LIMITATIONS

There are two aspects of our approach that require careful consideration: the first is related to the features used to develop our model and the second is regarding the selection of study cases. When we selected the features for our classifier based on the previous work of Botometer (Varol et al., 2017), we put aside the group of sentiment analysis and some of the Network characteristics. The sentiment features require the development of other classifiers which are very language dependent, warranting a separate study. Also, some network features require a large number of requests to the Twitter API. Since the request rate is limited, it would take a prohibitively long time to gather all the data needed for these features. Although we consider that these characteristics can contribute to the bot detection task, we decided to leave them out of the analysis for the time being.

For composing our annotated dataset, we sampled tweets posted during several events since 2018. The selected events, however, are all of a political nature. On the one hand, this gives us the ability to track and study the evolution of the behavior of bots on political subjects. On the other hand, this selection could skew the analysis towards these types of bots. New studies are therefore needed to evaluate the performance of our approach on datasets that include different types of messages.

DISCUSSION

Many position papers and special issues in reputed scientific publications such as Nature (i.e., ‘The Powers and Perils of Using Digital Data to Understand Human Behaviour’, 2021) indicate that the development of computational methods for social behavior analysis is an important and urgent agenda for a growing community of social science scholars (e.g., Alvarez, 2016; boyd & Crawford, 2012; Huberman, 2012; Lazer et al., 2018; Parks, 2014; Shah et al., 2015; Trilling, 2017). The thirst for methodological innovation in social science is also due to the enduring crisis that has characterized most of the existing techniques widely used in the field.

Indeed, computational social science studies not only involve large and complex datasets of digital traces and other “organic” data (Veltri, 2019) but require specific algorithmic solutions to analyze them (van Atteveldt & Peng, 2018). Such algorithms can provide instruments for testing existent theories on social and human behavior. However, despite social scientists enthusiasm with this emerging subfield, researchers should consider the fact that a method executed with a

machine learning tool, or an algorithm, does not make it a “computational method” (van Atteveldt & Peng, 2018) and does not guarantee its validity for answering social science questions.

Our findings indicate that due to the mutable nature of bots, supervised machine learning tools for bot detection need to be constantly updated with expert coding to perform well. Even specialists need to constantly revise the concept of what a bot is and how they behave on social media through empirical observation, considering it is an ever-changing object of analysis.

An urgent challenge for bot detection lies in questioning whether algorithms become more generalist and skillful over time at identifying automated profiles which tweet about different topics. In other words, it is important to understand how efficient old features are at detecting new bots which are moving targets and test the ability of AI classifiers to identify bots that act on different topics and agendas. Moreover, different supervised and unsupervised bot classifiers consider a wide variety of account features in their models, ranging from the simplest to the most expensive ones, extracted from the raw data obtainable through the Twitter Public API. Despite a few recent studies that have started to compare bot classifier performance (Yan et al., 2020), more evidence is needed to understand which set of features is more relevant for detecting novel bots in different languages and contexts.

Another predominantly neglected problem in bot detection studies is the transparency of different tools concerning system reliability and manual training models. r Tools do not document how their algorithms were manually trained, which parameters were considered by coders and how controversial annotations were considered to train the model. The role of human perception and the social ramifications of bots, particularly the political ones, have not been completely elucidated yet (Yan et al., 2020). Investigating the effectiveness of annotation parameters and understanding bias on bot perception is vital for advancing our understanding of the digital deception phenomena and for helping the academic community to devise reputable AI classifiers and countermeasures.

It should be noted that specialists also present bias in their interpretations, performing a degree of false positives and false negatives in their bot coding. Neither classical methods of manual coding nor machine learning classifiers guarantee the validity or reliability of detecting bots. In this sense, computational methods do not replace the existing social science methodological approaches, but rather complement them and vice-versa. On the one hand, the classifier model needs retraining for optimal performance. On the other hand, machine learning analysis can indicate coder bias and be used for testing manual protocols, helping to counteract the “myth of the trained coder” (Weber et al., 2015). This type of cross-discipline work is intrinsically challenging, calling for the constant development of interdisciplinary collaborations and new research tools, considering that traditional social science methods (especially qualitative approaches) can contribute to the development, calibration, and validation of computational methods (van Atteveldt & Peng, 2018).

As we have demonstrated with our experiment, supervised methods for detecting bots have proven to be effective in many cases, but they do not perform well with datasets that present

features on which the system was not directly trained, such as language and topic. We also identified that the bot score changes over time for Portuguese datasets.

Moreover, further issues emerge from social media platforms that can impact bot classifier development and evolution, such as data collection and algorithm opacity. Regarding data collection, data access through so-called Application Programming Interfaces (APIs) does not provide researchers with comparable data (Theocharis & Jungherr, 2021), which can compromise the development and improvement of AI bot classifiers. Another drawback of Twitter's Public API was the lack of documentation concerning what and how much data can be collected (Morstatter et al., 2013), leading to constraints in terms of academic reproducibility. The fact that Twitter offered a Public API for years, which allowed researchers to collect samples of tweets, explains why it has become the most studied platform for bot detection as of 2023.

Other social media platforms such as Facebook, YouTube, LinkedIn and TikTok impose strong limitations on data collection especially on user features and behavior, hindering automation detection and malicious behavior identification. Platform data unavailability raises questions for social behavior analysis and limits the development of research on computational social science.

Concerning algorithm opacity, Twitter trending topics, a set of the platform's top terms updated in real-time, is also a black box that affects the identification of bots. The sole aim of fake accounts is to manipulate trending topic results and hinder countermeasure development. Twitter has proven vulnerable to this manipulation by small but coordinated user groups and by those who control automated accounts (Zhang et al., 2017). Bots have demonstrated an extraordinary ability to distort Twitter traffic, forcing chosen phrases and hashtags into the "trending" lists (Nimmo, 2019). However, Twitter's algorithm frequently changes without revealing its parameters and measurements for labeling trending terms. This lack of transparency favors the creation of bots, botnets and a complex range of techniques which amplify messages, producing the misleading appearance of a substantial organic movement that involves many thousands of people who are unable to perceive a public opinion influence operation (DiResta et al., 2019; Kollanyi et al., 2016).

Against this backdrop there is a strong social demand for machine learning tools with academic validation and credentials for digital social research. As we show in this study, social scientists should confirm the reliability of different tools created and tested only through the prism of computational studies before applying them to empirical social science research. We acknowledge that algorithm development has advanced the field of research on bot identification using machine-learning classifiers, inspiring researchers to develop their own tools. Although Botometer has moved on from becoming a somewhat reliable and widely accepted approximation for identifying bots on Twitter, we infer in our analysis that a successful computational model does not always guarantee reliable results, applicable to a specific real case. As bot detection relies on textual rather than numerical data, computational social science should also account for contextual dependent tools and analysis.

REFERENCES

Alvarez, R. M. (Ed.). (2016). *Computational Social Science: Discovery and Prediction* (Reprint edition). Cambridge University Press.

Alvisi, L., Clement, A., Epasto, A., Lattanzi, S., & Panconesi, A. (2013). SoK: The Evolution of Sybil Defense via Social Networks. *2013 IEEE Symposium on Security and Privacy*, 382–396. https://doi.org/10.1109/SP.2013.33

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 20:1-20:27. https://doi.org/10.1145/3274289

Barrett, B., & Kreiss, D. (2019). Platform transience: Changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, *8*(4), 1–22. https://doi.org/10.14763/2019.4.1446

Bastos, M., & Mercea, D. (2018). The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences*, *376*(2128), 20180003. https://doi.org/10.1098/rsta.2018.0003

Bastos, M., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Beatson, O., Gibson, R., Cunill, M. C., & Elliot, M. (2021). Automation on Twitter: Measuring the Effectiveness of Approaches to Bot Detection. *Social Science Computer Review*, 08944393211034991. https://doi.org/10.1177/08944393211034991

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Briant, E. (2021, February 10). The Grim Consequences of a Misleading Study on Disinformation. *Wired*. https://www.wired.com/story/opinion-the-grim-consequences-of-a-misleading-study-on-disinformation/

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., & Subramanian, D. (2018). An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter. *ArXiv:1804.05232 [Cs]*. http://arxiv.org/abs/1804.05232

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race.

Proceedings of the 26th International Conference on World Wide Web Companion, 963–972. https://doi.org/10.1145/3041021.3055135

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter. *Acm Transactions on the Web*, *13*(2), 11. https://doi.org/10.1145/3313184

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2018). From Reaction to Proaction: Unexplored Ways to the Detection of Evolving Spambots. *Companion Proceedings of the The Web Conference 2018*, 1469–1470. https://doi.org/10.1145/3184558.3191595

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274. https://doi.org/10.1145/2872518.2889302

DiResta, R. (2018, September 14). How Bots Ruined Clicktivism. *Wired*. https://www.wired.com/story/how-bots-ruined-clicktivism/

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2019). The Tactics & Tropes of the Internet Research Agency. *U.S. Senate Documents*. https://digitalcommons.unl.edu/senatedocs/2

Dong, G., & Liu, H. (2018). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

Echeverría, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Stringhini, G., & Zhou, S. (2018). LOBO: Evaluation of Generalization Deficiencies in Twitter Bot Classifiers. *Proceedings of the 34th Annual Computer Security Applications Conference*, 137–146. https://doi.org/10.1145/3274694.3274738

Fernquist, J., Kaati, L., & Schroeder, R. (2018). Political bots and the swedish general election. *2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018*, 124–129. https://doi.org/10.1109/ISI.2018.8587347

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Giansiracusa, N. (2021). Tools for Truth. In N. Giansiracusa (Ed.), *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More* (pp. 217–229). Apress. https://doi.org/10.1007/978-1-4842-7155-1_9

Grimme, C., Assenmacher, D., & Adam, L. (2018). Changing Perspectives: Is It Sufficient to Detect Social Bots? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10913 LNCS*, 445–461. https://doi.org/10.1007/978-3-319-91521-0_32

Guimarães, N., Figueira, Á., & Torgo, L. (2021). Towards a pragmatic detection of unreliable accounts on social networks. *Online Social Networks and Media*, *24*, 100152. https://doi.org/10.1016/j.osnem.2021.100152

Howard, P. N., & Kollanyi, B. (2016). *Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum* (SSRN Scholarly Paper ID 2798311). Social Science Research Network. https://doi.org/10.2139/ssrn.2798311

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology and Politics*, *15*(2), 81–93. https://doi.org/10.1080/19331681.2018.1448735

Huberman, B. A. (2012). Big data deserve a bigger audience. *Nature*, *482*(7385), 308–308. https://doi.org/10.1038/482308d

Kollanyi, B., Howard, P., & Woolley, S. (2016). *Bots and automation over Twitter during the U.S. election* (Data Memo No. 1; Computational Propaganda Research Project Working Paper Series, pp. 1–4). Oxford Internet Institute. https://regmedia.co.uk/2016/10/19/data-memo-first-presidential-debate.pdf

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science (New York, N.Y.)*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Loper, E., Klein, E., & Bird, S. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.

Luceri, L., Badawy, A., Deb, A., & Ferrara, E. (2019). Red bots do it better: Comparative analysis of social bot partisan behavior. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 1007–1012. https://doi.org/10.1145/3308560.3316735

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*. 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. *Proceedings of the 10th ACM Conference on Web Science*, 183–192. https://doi.org/10.1145/3292522.3326015

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0184148

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ArXiv:1306.5204 [Physics]*. http://arxiv.org/abs/1306.5204

Mou, G., & Lee, K. (2020). Malicious Bot Detection in Online Social Networks: Arming Handcrafted Features with Deep Learning. *SocInfo*. https://doi.org/10.1007/978-3-030-60975-7_17

Napoli, P. M. (2021). The platform beat: Algorithmic watchdogs in the disinformation age. *European Journal of Communication*, *36*(4), 376–390. https://doi.org/10.1177/02673231211028359

Nimmo, B. (2019). *Measuring Traffic Manipulation on Twitter* (Computational Propaganda Research Project Working Paper Series, p. 35). Oxford Internet Institute. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/01/Manipulating-Twitter-Traffic.pdf

Parks, M. R. (2014). Big Data in Communication Research: Its Contents and Discontents. *Journal of Communication*, *64*(2), 355–360. https://doi.org/10.1111/jcom.12090

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. *Proceedings of the 20th International Conference Companion on World Wide Web*, 249–252. https://doi.org/10.1145/1963192.1963301

Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PLOS ONE*, *15*(10), e0241045. https://doi.org/10.1371/journal.pone.0241045

Santini, R. M., Agostini, L., Barros, C. E., Carvalho, D., Centeno De Rezende, R., Salles, D. G., Seto, K., Terra, C., & Tucci, G. (2018). *Software Power as Soft Power. A Literature Review on Computational Propaganda Effects in Public Opinion and Political Process* [Data set]. University of Salento. https://doi.org/10.1285/i20356609v11i2p332

Santini, R. M., Salles, D., & Tucci, G. (2021, Spring). When machine behavior targets future voters: The use of social bots to test narratives for political campaigns in Brazil. *International Journal of Communication*.

Santini, R. M., Salles, D., Tucci, G., Ferreira, F., & Grael, F. (2020). Making up Audience: Media Bots and the Falsification of the Public Sphere. *Communication Studies*, *0*(0), 1–22. https://doi.org/10.1080/10510974.2020.1735466

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2020). Detection of Novel Social Bots by Ensembles of Specialized Classifiers. *CIKM*. https://doi.org/10.1145/3340531.3412698

Schuchard, R., Crooks, A. T., Stefanidis, A., & Croitoru, A. (2019). Bot stamina: Examining the influence and staying power of bots in online social networks. *Applied Network Science*, *4*(1), 55. https://doi.org/10.1007/s41109-019-0164-x

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. https://doi.org/10.1177/0002716215572084

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

The powers and perils of using digital data to understand human behaviour. (2021). *Nature*, *595*(7866), 149–150. https://doi.org/10.1038/d41586-021-01736-y

Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1–2), 1–22. https://doi.org/10.1080/10584609.2020.1833121

Trilling, D. (2017). Big Data, Analysis of. In *The International Encyclopedia of Communication Research Methods* (pp. 1–20). American Cancer Society. https://doi.org/10.1002/9781118901731.iecrm0014

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *ArXiv:1703.03107 [Cs]*. http://arxiv.org/abs/1703.03107

Veltri, G. A. (2019). *Digital Social Research* (1st edition). Polity.

Wang, G., Wang, T., Zheng, H., & Zhao, B. Y. (2014). *Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers*. 239–254. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/wang

Weber, R., Mangus, J., & Huskey, R. (2015). Brain Imaging in Communication Research: A Practical Guide to Understanding and Evaluating fMRI Studies. *Communication Methods and Measures*, *9*, 5–29. https://doi.org/10.1080/19312458.2014.999754

Yan, H. Y., Yang, K.-C., Menczer, F., & Shanahan, J. (2020). Asymmetrical perceptions of partisan political bots. *New Media & Society*, 1461444820942744. https://doi.org/10.1177/1461444820942744

Yang, C., Harkreader, R., & Gu, G. (2013). Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Transactions on Information Forensics and Security*, *8*(8), 1280–1293. https://doi.org/10.1109/TIFS.2013.2267732

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115

Zhang, Y., Ruan, X., Wang, H., Wang, H., & He, S. (2017). Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending. *IEEE Transactions on Information Forensics and Security*, *12*(1), 144–156. https://doi.org/10.1109/TIFS.2016.2604226

Zhdanova, M., & Orlova, D. (2017). *Computational Propaganda in Ukraine: Caught Between External Threats and Internal Challenges*. http://ekmair.ukma.edu.ua/handle/123456789/13179

CONFLICTS OF INTEREST

The authors report there are no competing interests to declare.

AUTHORS CONTRIBUTION

Rose Marie Santini: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration
Débora Salles: Data Curation, Investigation, Writing - Original Draft, Writing - Review & Editing
Fernando Ferreira: Methodology, Software, Validation, Formal analysis, Data Curation
Felipe Grael: Methodology, Supervision

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.