Identification of Credulous Users on Twitter

Alessandro Balestrucci
Gran Sasso Science Institute
L'Aquila, Italy
alessandro.balestrucci@gssi.it

Rocco De Nicola
IMT School for Advanced Studies
Lucca, Italy
rocco.denicola@imtlucca.it

Omar Inverso
Gran Sasso Science Institute
L'Aquila, Italy
omar.inverso@gssi.it

Catia Trubiani
Gran Sasso Science Institute
L'Aquila, Italy
catia.trubiani@gssi.it

ABSTRACT

Social networks can quickly propagate information to large audiences and can be used to spread fake news or to provide false figures of popularity. Social bots, i.e., software robots that automatically interact with human users and produce content under a fictive identity, are used for such harmful activities. In this paper, we study the relationship between bots and genuine human users with the aim of identifying those "credulous" users who are particularly exposed, and unintentionally contribute, to the activities planned by a network of bots. Spotting credulous users is useful to service providers to display warnings on their dashboards, scan their activities for early signs of attacks, or take more active measures to prevent or limit the negative effects of their activities.

Here we aim at identifying credulous users on Twitter starting from those involved in any social relation with a bot. To that end, we rely on an existing bot detector along with its dataset of genuine users and bots that we extend with additional information about the friends of each genuine user. To single out credulous users out of genuine ones, we study the effectiveness of different metrics or combinations thereof. We see this as a first step towards singling out features that can be used to detect credulous users without resorting to the expensive analysis of the nature of their friends.

KEYWORDS

Social networks, Bots, Twitter, User Analysis

ACM Reference Format:
Alessandro Balestrucci, Rocco De Nicola, Omar Inverso, and Catia Trubiani. 2019. Identification of Credulous Users on Twitter. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297280.3297486

1 INTRODUCTION

Popular social networking services, such as Twitter and Facebook, are rapidly evolving as pervasive means of information where news are quickly propagated to large audiences. In fact, domestic users of such services can keep up with the news effortlessly, while routinely checking out their own social channels of interest. They are, however, also at risk of being misinformed, because there is no control on the reliability of the news themselves.

As a matter of fact, the diffusion and propagation of deliberately misleading information for harmful purposes is quite recurring in social networks. Its considerable impact on the economy, the society, or even the democracy, represents a major concern, even more so in view of the uncontrolled speed of propagation and of the size of the targeted audience.

Social bots, i.e. software robots that automatically interact with human users and produce content under a fictive identity, are at the source of such harmful activities. Bots are wide-spread, with recent estimates indicating that up to 15% of Twitter users are actually bots [20, 41], and can be so powerful to the point of heavily influencing public opinion [19, 26, 27]. This has motivated a vast body of work on bot recognition in social media [4, 21, 35, 38]. In particular, recent approaches to automated bot detection on Twitter rely on directly observing specific features, such as the ratio of friends over followers of a registered user, the quantity or frequency of their interactions, the expressiveness of their comments, the presence of a name, face photo, address, biography or any additional information on the profile, and many more [7, 11, 13].

Bot detection is undoubtedly essential in contrasting the negative effects of malicious activities in social networks. On the other hand, the role of humans in this matter does not seem to have received as much attention. Yet recent findings do hint at some connection between genuine users and harmful activities on social networks. Among the other things, it has been observed that the majority of genuine users normally do not check the reliability of articles from social media, and many even share these articles [16]. Now, depending on the activities of their contacts, these users may well end up contributing actively, although unknowingly, to spreading harmful content. These considerations motivate further effort to identify significant categories of genuine users that play an active part in this business, their distinctive features or behavioral profile. However, due to the nontrivial classification of human behaviour, the problem of analyzing genuine users is quite different from bot recognition. Known social interaction mechanisms such as opinion changes and social influence have previously been studied [14, 28, 37], also some attempts have been made to categorize social network users [40].

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297486

In this paper, we deliberately draw attention to those users of social networks who are particularly exposed to the malicious activities planned by bot networks. For simplicity, by abuse of language we refer to them as *credulous* users. In particular, we acknowledge the existence of such class of users, and provide a technique to automatically rank them on the basis of their *gullibility*. Identification is then achieved by selecting the highest-ranked elements. Spotting credulous users can be useful, for instance, to service providers to display warnings on their dashboards to reduce the risk of their involvement in harmful activities (e.g., spreading fake news, hate speeches, etc.), or to hold up some of their activities (e.g., content re-posting) in order to slow down the propagation of malicious information.

Although our intuition is general, here we focus on Twitter, a popular platform at the time of writing. Our starting point is an existing bot detector along with its dataset of genuine users and bots [41]. We adopt the mentioned dataset as an initial seed for crawling fresh data (i.e., account information, tweets, and mentions). The crawling removes outdated information and at the same time restores information that was previously removed due to restrictions imposed by the service provider; in addition, as required by our data analysis procedure, we extend it to the friends of the genuine users in the initial dataset.

We then revise the bot detector to adopt a unique decision model, selected after training and testing multiple machine learning algorithms, evaluating their prediction accuracy with respect to the ground-truth provided by the instances in [41]. We feed the extended dataset to the revised bot detector, classifying as bots or humans the friends of the genuine users in the initial dataset. Finally, we rank the users by combining different metrics, including the ratio of bots over humans in their list of friends, and their seniority, i.e., the time they were created.

We experiment with our ranking mechanism as follows. After selecting the topmost credulous users from the gullibility list, we calculate the *efficacy* expressed in terms of the number of detected bots over analyzed users. We repeat this process by varying two parameters. First, we modify the cutoffs for selecting the credulous users from the ranked list; this is helpful to verify if the ratio of bot friends decreases when considering a larger set of users as credulous ones. Second, we build two more datasets of different sizes by randomly selecting users from the initial dataset; this allows to test our technique on multiple initial sets of users. We consistently observe over all combination of the above two factors, a greater efficacy for smaller cutoffs, thus substantiating the validity of our proposed gullibility ranking. Note that, in our case, it is not feasible to apply standard measures such as *precision*, *recall*, and *F-measure* [18, 30], because of the absence in the literature of existing datasets for the considered class of users.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our approach for identifying credulous users on Twitter. Section 4 reports the experimental results that help us to compare the impact of different metrics on the quality of the decision procedures about gullibility. Section 5 concludes the paper and provides future research directions.

2 RELATED WORK

The work presented in this paper mainly concerns the following three research directions related to social networks: (i) bot detection, (ii) spreading of misinformation, (iii) analysis of human behaviour. Hereafter we discuss the related work along these three lines of research.

2.1 Bot detection

Different approaches have been recently proposed to build tools (bot detectors) to discern automated accounts from human ones.

A bot detector for Twitter that applies a complex machine learning algorithm on elaborate features, such as the reposting rate as well as temporal patterns, was proposed in [16]. An important, aside, finding of this work is that a very high percentage of Twitter users (86%) do not check the reliability of articles from social media they receive, and 27% of them share un-checked news. These figures represented for us good motivations to single out credulous users and to possibly warn them.

Over a thousand features are considered by another bot detection technique for Twitter based on machine learning [41]. This technique clusters the detected bots into different subclasses of spammers and promoters. In our work, we have used both their bot detector, which is publicly available as a web service, and their dataset [41] as starting points of our procedure for identifying credulous users.

In [24], different machine learning algorithms are analyzed in order to build a classification model for bots by considering the published contents and the publishing behavior of the users. This work points out that humans exhibit complex timing behavior, as opposed to the more regular one of bots. Similar situations are described in [5], where temporal patterns are extracted from the activity time series of bots. We plan to extend our revised bot detector to take into account also the evolution of behaviors over time.

Some statistical techniques have been adopted for investigating bot-like behavior on Twitter in [31]. This approach ranks the accounts on the basis of the number of specific features that are relevant to specific bot types, like fake followers and content polluters. Our approach could benefit from this work by considering the distinctive features of the different kinds of bots.

In [4], a detection technique is described that relies on the analysis of user activities on Twitter and Facebook (i.e., posting, sharing, liking, tweeting, retweeting, and deleting) and their correlation. Highly-synchronous activities are considered abnormal, hence generated by bots. Our work does not currently consider users activities, but also these can be useful for identifying credulous users. Retweeting fake news, for instance, might be a plausible symptom of gullibility.

2.2 Misinformation in social networks

The importance of understanding the propagation of misinformation in Twitter has been highlighted in [25], where a quantitative analysis of tweets during the Ebola crisis was performed, and the users were categorized according to whether they received any news about the virus, and, if so, whether they retweeted it.

Methods for evaluating the credibility of published information have also been proposed, such as one that relies on two main features (i.e., topic and source) for ranking tweets and estimating their reliability [23]. Approaches of this kind are somehow complementary to ours, as they might be used to assess the reliability of the users producing tweets and to rate the gullibility of their friends.

Several models of diffusion of hoaxes in social networks have been provided. A known modeling framework [39] focuses on the analysis of fact-checkers that associate to the news their probability of being fake, and thus can remove the ones with a high probability. Our goal is not to detect or delete fakes, but only to provide warnings about suspicious spreaders.

Another model aimed instead at detection, based on supervised learning, was proposed to detect suspicious behavioral patterns on Twitter [2]. This technique, however, makes no distinction as to whether the misinformation is intentionally spread or not. In contrast, we are interested only in credulous users that propagate fake news unintentionally.

Special attention has been dedicated to the problem of spreading political misinformation on social media [6, 34, 36, 43]. For instance, in [34] text analysis is combined with temporal patterns. Shared hyperlinks represent the main means for circulating false rumors. This technique could be combined with ours to refine our analysis, by considering the actual content of the news published by credulous users, and checking the presence of hyperlinks to unreliable websites. As another example, in [36], Twitter data related to the 2016 USA elections are analyzed according to specific features, such as the number of followers and friends, the used links per day, the retweets per day, with the aim of identifying malicious activities and specifically those by bots.

2.3 Analysis of human behaviour

Sentiment analysis and bot detection have been combined in [15], where several sentiment-related factors emerged as possible indicators for bots in Twitter. A scoring engine assigns to a topic in a tweet a value ranging from -1 to +1 to denote maximally negative and positive attitude, respectively. This score is exploited to discern humans from bots; humans are observed to express a stronger positive sentiment than bots.

Behavioural profiling has also been adopted for detecting anomalous behaviour or human behaviour. It has been observed that fast changes in the reactions of Facebook users to content posted by other users correspond to possible sources of anomalous behaviour [33].

In [1] a method for identifying human behaviour in social networks is proposed. The experimentation is conducted on Facebook and the approach is based on building graphs that represent the expected behaviours in terms of sequence of activities. Behavioural patterns that do not reflect any of the known models are associated to likely malevolent purposes. We expect that similar techniques can be used for Twitter.

Another way to single out genuine accounts is based on comparing the similarity of sequences of activities (i.e., comment, like, share, mention, etc.) of groups of users. The main intuition is that groups of bots seem more likely to share common patterns, whereas genuine users exhibit a more heterogeneous behavior [10].

We plan to extend the evaluation of our approach on further datasets of humans and bots obtained according to the above techniques.

3 APPROACH

Our approach for the identification of credulous users on Twitter consists of two separate processes.

The purpose of the first process is to produce a refined decision model for bot detection. This is built upon an existing bot detector and a dataset of Twitter users [41], where each user is associated to a label indicating whether it is a genuine one (in other words, a human user) or a bot. During this process, data crawling from Twitter is performed to refresh the initial dataset. The updated information is then converted into a set of representative user features. We then use multiple subsets of these features to experiment different machine learning algorithms [17]. At the end, we select the best combination (of features and algorithm) to obtain an improved decision model for bot detection. Section 3.1 describes this process in detail.

The second process achieves our final aim of identifying credulous users. We start from the refreshed dataset, this time only considering the users previously labeled as humans, and extend it with additional information about the friends of those users. We then exploit the revised decision model (obtained by the first process) to label each friend as a bot or a human. We also introduce a set of rules to discern whether a genuine user is a credulous one. This process is described in detail in Section 3.2.

3.1 Revisited bot detection

Our study starts by considering a publicly-available supervised dataset of Twitter users along with a bot detector trained on it [41] (see Figure 1).

Bot detectors have the tendency to gradually become obsolete and to loose precision, because bots evolve continuously [29], and existing datasets degenerate as time goes by, for instance due to suspended accounts. To partially overcome this issue, we decided to derive an improved decision model after refreshing the initial dataset. We run different machine learning algorithms on different sets of features, in order to compare their prediction accuracy. We eventually adopt as our refined decision model the best performing alternative among the considered ones.

Due to policy restrictions[1], the initial dataset we relied on only contains the user IDs and the associated labels indicating whether a given account corresponds to a human user or a bot. These IDs represent Twitter accounts, which we refer to as *basic users*. We implement *data crawling* on top of the Twitter API[2] to retrieve further information on those basic users (see ① in Figure 1).

Our crawler fetches the following data for each basic user:

- the *tweets*: the content in form of text, photos, etc. published by the user on his or her main page);
- the *mentions*: the tweets not published by the user, but where the user has been tagged by other users;

[1] **Twitter Developer Policy**: https://goo.gl/BiAG16
[2] **Twitter API**: https://goo.gl/2FXfi5

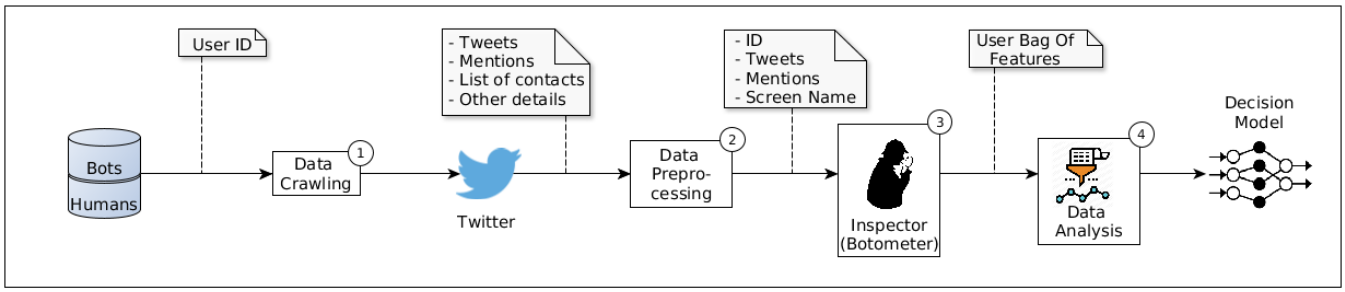

Figure 1: Revised Bot Detection.

- the *list of contacts*: the list of the IDs of the users involved in any social relations with the considered user, i.e., *followers* and *friends* of a given user;
- *other details*: the *screen name*, the *description*, the *status count*, and other public information about a Twitter account.

During this step we also filter out from the initial dataset all the entries that are no longer valid, such as suspended or deleted user accounts.

After refreshing the initial dataset, the *data preprocessing* step (see ② in Figure 1) transforms the user data into a suitable format for querying the *inspector*. Requests to the inspector include the following user's data: *ID*, *tweets*, *mentions*, and *screen name*. The inspector returns as output the probability of the user being a bot, along with a *bag of features*, i.e., a representative set of feature-value pairs (see ③ in Figure 1). For this task we rely on the Botometer web service[3].

To obtain our revisited bot detector we perform *data analysis* (see ④ in Figure 1). We thus compare the prediction accuracy in human-bot classification of different machine learning algorithms on multiple subsets of the features. The combination (of features and algorithm) showing the highest accuracy becomes our revised bot detector, i.e., the *decision model* of Figure 1.

The prediction accuracy results are shown in Table 1. The columns of the table list the considered algorithms: C4.5 [32], based on decision trees, random forests (RF) [3], RIPPER (R) [8], and neural networks (NN) based on the multilayer perceptron model with back propagation [22]. We selected these algorithms because they are well-known in the literature [11, 13, 41, 42]. Moreover, the random forest algorithm has proven to be a rather accurate classifier for bot detection [41]. For our tests, we used the implementations available in the Weka tool-suite [17].

To determine the probability of a Twitter account being a bot, the bot detector relies on six categories of features [41]:

(1) *user-based*: the number of friends and followers, the number of tweets produced by the users, profile description and settings;
(2) *friends*: the used language, local time, popularity, etc., extracted from followers-friends (i.e., retweeting, mentioning, being retweeted, and being mentioned);
(3) *network*: the different types of communication (i.e., retweet, mention, and hashtag) weighted considering the frequency of interactions or co-occurrences;
(4) *temporal*: the user activity (e.g., production of tweets) over different time intervals;
(5) *content*: the type of natural language, the length and entropy of the text being tweeted;
(6) *sentiment*: the attitude or mood of a conversation, e.g., arousal, valence, and dominance scores.

Botometer produces as output the so-called *english score* (ES) that relies on the six categories above and the *universal score* (US) that ignores sentiment and content features, being them English-specific. We did perform four different evaluations reported in Table 1. The first three rows of the table refer to the outcome of the experiments based on the two scores separately and on their combination. The fourth, more effective, experiment considers not only the values assigned by Botometer to the above six categories but also the numbers of tweets and mentions separately. These eight features constitute our *bag of features*. The values in the table represent the achieved prediction accuracy (expressed as a percentage) of models validated by means of the 10-fold cross validation. The highest value of 82.26% (highlighted in the table) is obtained by using neural networks trained with the bag of features.

Table 1: Prediction accuracy of the new decision models.

	C4.5	RF	R	NN
ES	78.00	77.96	78.78	79.70
US	77.60	77.05	77.41	78.05
ES+US	78.23	73.02	78.65	79.83
Bag of features	81.16	81.71	81.30	**82.26**

The decision model so obtained is used in the next process to determine whether a Twitter account, more precisely a given friend of a basic user, is a bot or a genuine one.

3.2 Identification of credulous users

The process of identifying credulous users is shown in Figure 2. It starts by considering the human users of the same dataset [41] previously used in Section 3.1. This time, however, we are interested in analyzing the friends of these humans. Twitter provides two types of social relationships: the *friends* are the users followed by a basic user; the *followers* are the users following a basic user. We limit ourselves to reasoning about the friends of the basic users, though. The friends are, in fact, more significant to our purposes, because it takes an active action in order for a user to become friends with someone.

[3] **Botometer web interface**: https://goo.gl/uyhG5c

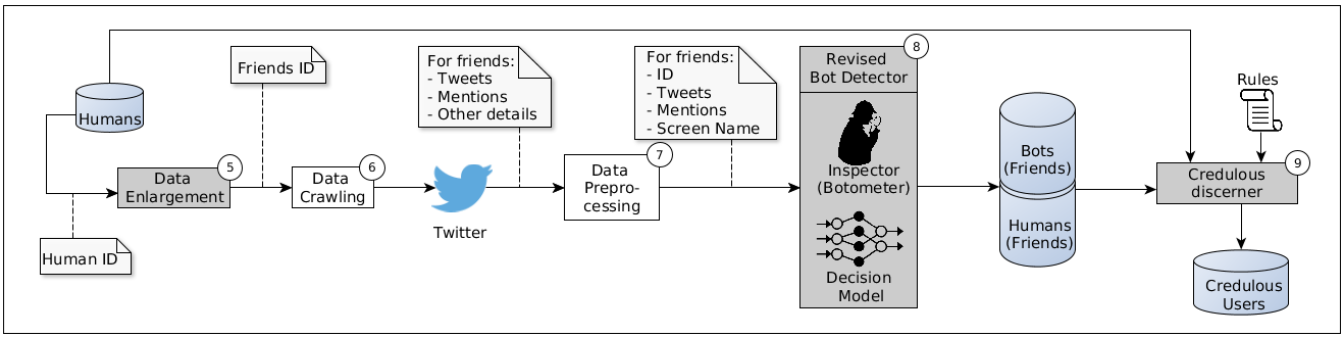

Figure 2: Identification of credulous users.

Let us denote by U the set of basic users in the initial dataset, H_b the set of basic humans in U, and B_b the set of basic bots in U. We additionally denote by $F(h)$ the set of users that are friends of a basic human h, and by $F(H_b)$ the union set of friends over all the basic humans in H_b. Therefore:

$$U = H_b \cup B_b \ , \ H_b \cap B_b = \varnothing$$
$$F(h) = \{f \mid h \in H_b \wedge f \ is \ friend \ to \ h\}$$
$$F(H_b) = \bigcup_{h \in H_b} F(h)$$

To prevent the overall number of friends $|F(H_b)|$ from growing excessively, we only consider the subset H'_b of H_b whose users have at most 400 friends on Twitter:

$$H'_b = \{h \in H_b : |F(h)| \leq 400\}.$$

The *data enlargement* step (see ⑤ in Figure 2) consists in building, for this restricted set of basic human users, the overall list of friends to consider, i.e., $F(H'_b)$. The subsequent *data crawling* step (see ⑥ in Figure 2) is performed on the list of users $F(H'_b)$ produced above. This step is similar to the crawling step described in Section 3.1, except that here the list of contacts is not fetched. We perform a *data preprocessing* step that prepares the requests to our revised bot detector (see ⑦ in Figure 2). Specifically, information related to the friends of a genuine user (i.e., account details, tweets and mentions) is given as input to the bot detector, that returns as output their features, exactly as explained in Section 3.1. We consider two additional features (i.e., number of tweets and mentions). The resulting instance is processed by our revisited bot detector (see ⑧ in Figure 2). In this way, we compute a prediction p (i.e., 0 for humans, and 1 for bots) for each newly fetched user. At the end, we are able to derive the set $BF(h)$ of *bot friends* for every given human user h, and the overall set $B(H'_b)$ of *bots* for the whole set of basic humans users:

$$BF(h) = \{f \in F(h) : h \in H'_b \ and \ p(f) = 1\}$$
$$B(H'_b) = \bigcup_{h \in H'_b} BF(h)$$

In the following, we discuss the rules to rank the human users in H'_b. The rules determine each a separate ranked lists of users, and are eventually combined to build the set of credulous users.

The first rule (R_1) calculates for a given basic human user the ratio between the number of bots among its friends over the total number of friends. Intuitively, this captures the observation that a user with a high number of bots in the list of friends is more likely to be influenced. This is expressed as follows:

$$R_1 = \frac{|BF(h)|}{|F(h)|}, \ h \in H'_b$$

The second rule (R_2) ranks the users according to the normalized ratio between the number of bot friends and the overall number of friends. Normalization is introduced to capture cases where humans have a high ratio of bots over their friends, but the actual number of friends is low in comparison to other users of the same dataset. The rule is:

$$R_2 = \overbrace{\frac{|BF(h)|}{|BF(h_{maxB})|}}^{bot \ normalization} * \overbrace{\frac{|F(h)|}{|F(h_{maxF}|)}}^{friend \ normalization}$$
$$h, h_{maxB}, h_{maxF} \in H'_b$$

where h_{maxB} represents the human with the highest number of bots among its friends, and h_{maxF} denotes the human with the highest number of friends.

The third rule (R_3) aims at giving relevance to the *seniority*, or *experience* of a user. Intuitively, more experienced users tend to select their friends with more care. For each basic user, the rule calculates the ratio between the value calculated by R_1 over the age (in months) of the account, denoted as age_m:

$$R_3 = \frac{R_1}{age_m(h)}, \ h \in H'_b$$

The fourth rule (R_4) considers the normalized relation between the number of bots, friends, and age. The idea in this case is to capture the increased ability of younger accounts to effectively filter out more bots. Specifically:

$$R_4 = R_2 * \overbrace{\frac{age_m(h)}{|age_m(h_{maxA})|}}^{age \ normalization}, \ h, h_{maxA} \in H'_b$$

where h_{maxA} represents the eldest human in H'_b.

On the basis of the above rules, we obtain four ranked lists of the users in H'_b in descending order. We additionally combine the four rules to understand which characteristics are more relevant. We firstly study the usefulness of normalization:

R_{13} considers the set of users selected by both R_1 and R_3. The idea is to prioritize users with a considerable amount of bots among their friends, but also take into account the percentage of bots with respect to their seniority. These two rules do not include normalized factors.

R_{24} considers the set of users selected by both R_2 and R_4. The rationale is to prioritize the normalization related to bots, number of friends, and age of a basic human. These two rules embed normalized factors in their specification.

We also investigate the usefulness of considering the age of the user accounts:

R_{12} considers the users selected by both R_1 and R_2. By doing so, we prioritize the information on the number of bots and friends, intentionally excluding the age.

R_{34} considers the set of users selected by both R_3 and R_4. Here we jointly consider the number of bots, the number of friends, and the age.

Finally, we combine all the four rules as follows:

R_{1234} considers the users jointly selected by all the rules combined together, so to observe the highest-ranked users with respect to all the provided rules.

The *credulous discerner* step (see ⑨ in Figure 2) applies all the rules defined in this section and produces ranked lists of users. Note that selecting the topmost users from these lists yields different sets of credulous users. It is worth to remark that this process is not a decision model to classify *credulous* users. It can be rather considered as a set of rules that contribute to understand if a human is more exposed to bots than others. This represents a first direction for building a preliminary dataset of *credulous* users. In the next section, we apply these rules to our dataset and discuss the experimental results.

4 EXPERIMENTAL RESULTS

The main purpose of our experimentation is to validate our gullibility ranking. During our experiments, however, we noticed some other interesting findings, that we shall point out as we go along.

We rank the genuine accounts according to the rules defined in Section 3 and generate a different ranked list for each rule. We thus obtain different sets of credulous users by selecting the topmost elements of these ranked lists. To quantify the usefulness of these sets, we define a measure of *efficacy* as the ratio between the number of detected bots over the total number of friends for the considered set of credulous users. We recall that these sets represent a first source of knowledge to further investigate the features of the user accounts, and single out credulous users.

Having applied the selection criteria from Section 3.2, we obtain from the dataset [41] 754 human users to be considered. This is our dataset D_1. Furthermore, in order to check that the obtained results are not dependent on the specific dataset, we build from D_1 two smaller portions, D_2 and D_3, by randomly extracting a half and of a quarter of the elements of D_1, respectively.

We performed a preliminary investigation on these three datasets by measuring the efficacy for all the humans, without distinguishing the credulous ones. The 754 human users in D_1 turned out to have about 126k friends, 17k of which were marked as bots, leading to an efficacy of 0.14; D_2 includes 377 humans, 65k friends, and 8k bots, for an efficacy of 0.13; D_3 has 188 humans, exposing 35k friends, and 4k bots, hence the efficacy is 0.13. An interesting thing to observe here is that all these values confirm that the claim about roughly 15% of Twitter users being bots [20, 41] also holds for the induced network of friends in the considered dataset. Another thing worth noticing is that our approach is expensive, as it required scanning about 126k user accounts in order to analyze the gullibility of only 754 genuine users.

Table 2 reports our experimental results. The datasets are reported in the leftmost side of the table, where columns *size* and *id* report the number of users and their identifier, respectively. The table is arranged into four sections, each corresponding to a different group of experiments. The first one (specific rules) investigates the efficacy associated to the evaluation of the four rules of Section 3.2 in isolation. To calculate the efficacy, we introduced some cutoffs on the number of genuine users to be considered as credulous users. For example, for the D_1 dataset, we set three cutoff values to 200, 150, and 100 (shown in column *cred* of Table 2). By setting the cutoff to the topmost 200 credulous users, the four rules yield an efficacy of 0.275, 0.197, 0.265, and 0.189, respectively. We remind that these values represent the ratio between the amount of analyzed credulous bots friends over the total number of their friends. This means that, for example, by applying rule R_1 with the largest cutoff, 27% of these credulous friends turn out to be bots.

We can observe an increasing trend in the efficacy values for smaller cutoffs (i.e., 150 and 100). For example, in Table 2 we can see that the efficacy values are 0.275, 0.303, and 0.344 when considering 200, 150, and 100 credulous, respectively. We also observe similar trends for all the other specific rules (i.e., R_2, R_3, R_4). Among all the efficacy values for the four specific rules, it is possible to see that the best values are obtained by considering rule R_1 with a cutoff value of 100. Using R_3 leads to similar values.

After considering the different rules separately, we study some possible combinations of them. We select the topmost credulous users (again with cutoffs at 200, 150 and 100 elements) from the ranked lists separately produced by the specific rules, and then considering the credulous users obtained by intersecting all these different lists.

In the second group of experiments (normalization), we investigate the effect of normalization and compare the efficacy of the conjunction of rules R_1 and R_3 (denoted by R_{13} in Table 2) that do not use normalization and the conjunction of the remaining two rules R_2 and R_4 (denoted by R_{24} in Table 2) that instead rely on normalization. We have that the best efficacy values are obtained in the absence of normalization. For example, with the initial dataset of 200 credulous users, R_{13} is calculated on a set of 152 credulous users, with an efficacy of 0.296. With the same dataset, R_{24} is instead calculated on a set of 174 credulous users. In this case, the achieved efficacy of 0.2 is sensibly lower than R_{13}. It is worth noticing that the number of credulous users is larger when comparing R_{13} w.r.t. R_{24} because two heterogeneous datasets of Twitter accounts are built, however efficacy still reflects the same trend observed before, i.e., more permissive cutoffs lead to lower efficacy.

Table 2: Experimental results.

Datasets		Specific rules					Normalization				Seniority				All rules	
		cred	R_1	R_2	R_3	R_4	R_{13}		R_{24}		R_{12}		R_{34}		R_{1234}	
id	size		eff.				cred	eff.	cred	eff.	cred	eff.	cred	eff.	cred	eff.
D_1	754	200	0.275	0.197	0.265	0.189	152	0.296	174	0.200	101	0.271	67	0.284	63	0.294
		150	0.303	0.217	0.289	0.204	114	0.308	125	0.218	66	0.303	42	0.304	37	0.320
		100	0.344	0.236	0.324	0.228	71	0.367	82	0.243	36	0.337	23	0.350	19	0.370
D_2	377	100	0.273	0.183	0.270	0.173	78	0.302	86	0.184	45	0.273	28	0.299	27	0.304
		75	0.305	0.199	0.291	0.183	59	0.324	59	0.202	31	0.303	18	0.315	17	0.323
		50	0.350	0.224	0.326	0.214	37	0.368	42	0.228	18	0.344	13	0.342	9	0.378
D_3	188	48	0.276	0.197	0.264	0.191	36	0.304	40	0.204	24	0.278	18	0.288	16	0.309
		36	0.308	0.223	0.293	0.213	31	0.315	31	0.221	18	0.309	13	0.315	13	0.315
		24	0.353	0.247	0.333	0.233	18	0.368	21	0.248	9	0.368	8	0.340	6	0.376

In the third group of experiments (seniority), we investigate the efficacy associated to rules R_1 and R_2 combined (denoted by R_{12} in Table 2) that do not consider the seniority of the user account w.r.t. rules R_3 and R_4 combined (denoted by R_{34} in Table 2) that, instead, consider the longevity of accounts. We notice that the best values of efficacy are obtained by the rules that take seniority into account. For example, with the initial dataset of 200 credulous users, R_{12} is calculated on a set of 101 credulous users, and the efficacy is 0.271. With the same dataset, R_{34} is instead calculated on a set of 67 credulous users, and the efficacy of 0.284 is only slightly larger than that of R_{13}. The efficacies for all entries of R_{12} and R_{34} are very similar, despite the considered number of credulous users that instead are much less for R_{34}. For example, considering the dataset of 150 credulous users, rule R_{12} selects 66 credulous users with an efficacy of 0.303, whereas rule R_{34} only considers 42 credulous users, but with basically the same efficacy. This suggests that considering seniority is useful for the datasets under analysis, but has a limited impact.

In the fourth group of experiments (all rules), we evaluate the combined efficacy of all the four rules. We select the first 200, 150 and 100 credulous users in ranked lists produced for the specific rules separately and then consider their intersection. This cuts down the number of selected credulous users to 63, 27, and 19, respectively. Such a large reduction in the size of the dataset, due to the intersection, shows that each specific rule has the effect of classifying as credulous users different genuine users. By observing in the table the column of related efficacy values, we can see that by combining all the four rules we obtain the best results. Furthermore, the observation about the increase in efficacy for smaller cutoffs is still valid. In conclusion, the larger efficacy for smaller cutoffs, consistently observed over all our experiments, substantiates the validity of our proposed gullibility ranking.

5 CONCLUSIONS AND FUTURE WORK

Our work was initially motivated by the observation that the gullibility of genuine users is an important factor in spreading malicious activities across social networks. Our primary goal was therefore to acknowledge the existence of, and draw attention to, the class of credulous social network users, whom unknowingly support the malicious activities of bot networks. More concretely, we have proposed a possible way of identifying Twitter users belonging to the above class of interest. In particular, we have described and implemented a method to rank genuine users on the basis of their gullibility, so to isolate the most credulous ones from the rest. We have reported an experimental evaluation on a publicly-available dataset that confirms the validity of our ranking mechanism.

Our contribution differs from most of the existing body of closely related work, where the main concern is instead detecting automated activities, such as those performed by bot networks, and the human component is largely neglected.

As a direct, practical advantage of our current technique, a service provider could identify the most credulous users and send them warnings or even precautionarily delay or block their activities. This would limit their unintentional contributions to any harmful activity, and thus indirectly dampen intended malicious effects, such as the viral spreading of fake news.

In the immediate future, we plan to experiment with our approach on other datasets, such as [12] and to use our current dataset of credulous users to train a new decision model, and then validate it on a different dataset. We also plan to extend our experimentation to take into account the activities of the users and their behaviour. In fact, we currently do not distinguish passive and active users, i.e., the ones that are intentionally spreading fake information.

Our current technique is expensive, as it requires scanning the (potentially many) social connections in order to compute the gullibility score of a given genuine user. It would be worthwhile to invest further effort in the development of an alternative decision model that can achieve a similar or greater precision without considering the social contacts of a user. To that end, it would be useful to determine relevant features that contribute to single out the credulous users, so that, instead of retrieving information on their friends, we can recognize them by only analyzing the user profile. An interesting direction for future research is to apply different methodologies to classify credulous users, e.g., approaches that only use account details such as class-A features [11], or behavioural analysis, like for instance DNA fingerprint [9].

An enhanced decision model might be useful to improve efficiency in bot detection. In fact, given the high density of bots among the friends of the credulous users, the decision model could be useful as a guidance for spotting more bots by only inspecting a restricted set of users (the friends of the credulous), rather than scanning the network exhaustively.

Conceptually, our approach is not limited to a specific platform, and in principle it should be possible to generalize it to other contexts, e.g., to Facebook, Google+, etc. However, we anticipate that every such attempt would likely result in a considerable engineering effort, due to possible issues such as policy restrictions of providers that make it difficult to find good datasets, and pose specific new challenges for fetching large bulks of relevant data, for interfacing with the different APIs to access those services, for dealing with the service request throttles, and the like.

ACKNOWLEDGMENTS

We would like to thank Marinella Petrocchi for fruitful discussions on this topic.

REFERENCES

[1] Flora Amato, Aniello Castiglione, Aniello De Santo, Vincenzo Moscato, Antonio Picariello, Fabio Persia, and Giancarlo Sperlí. 2018. Recognizing human behaviours in online social networks. *Computers & Security* 74 (2018), 355–370.

[2] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. 2015. A model for identifying misinformation in online social networks. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 473–482.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation.. In *Proceedings of the IEEE International Conference on Data Mining*. 817–822.

[5] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2017. Temporal patterns in bot activities. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1601–1606.

[6] Shu-I Chiu and Kuo-Wei Hsu. 2018. Predicting Political Tendency of Posts on Facebook. In *International Conference on Software and Computer Applications (ICSCA)*. ACM, 110–114.

[7] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing* 9, 6 (2012), 811–824.

[8] William W. Cohen. 1995. Fast Effective Rule Induction. In *International Conference on Machine Learning*. 115–123.

[9] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the International Conference on World Wide Web Companion*. 963–972.

[10] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing* 15, 4 (2018), 561–576.

[11] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems* 80 (2015), 56–71.

[12] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. Social Fingerprinting: Detection of Spambot Groups Through DNA-Inspired Behavioral Modeling. *IEEE Trans. Dependable Sec. Comput.* 15, 4 (2018), 561–576.

[13] Isaac David, Oscar S Siordia, and Daniela Moctezuma. 2016. Features combination for the detection of malicious Twitter accounts. In *IEEE International Autumn Meeting on Power, Electronics and Computing*. 1–6.

[14] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2018. Polarization and fake news: Early warning of Potential misinformation targets. *arXiv preprint arXiv:1802.01400* (2018).

[15] John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots?. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 620–627.

[16] Phillip George Efthimion, Scott Payne, and Nicholas Proferes. 2018. Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots. *SMU Data Science Review* 1, 2 (2018), 5.

[17] Frank Eibe, MA Hall, and IH Witten. 2016. The WEKA Workbench. Online Appendix for" Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann* (2016).

[18] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[19] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Squicciarini. 2015. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 233–242.

[20] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[21] Zafar Gilani, Reza Farahbakhsh, and Jon Crowcroft. 2017. Do Bots impact Twitter activity?. In *Proceedings of the International Conference on World Wide Web Companion*. 781–782.

[22] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[23] Aditi Gupta and Ponnurangam Kumaraguru. 2014. Misinformation in Social Networks, Analyzing Twitter During Crisis Events. In *Encyclopedia of Social Network Analysis and Mining*. Springer, 922–931.

[24] Mufaddal Haidermota, Ashwini Pansare, and Drishit Mitra. 2018. Classifying twitter user as a bot or not and comparing different classification algorithms. *International Journal of Advanced Research in Computer Science* 9, 3 (2018).

[25] Fang Jin, Wei Wang, Liang Zhao, Edward R Dougherty, Yang Cao, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Misinformation propagation in the age of twitter. *IEEE Computer* 47, 12 (2014), 90–94.

[26] Kyumin Lee, Prithivi Tamilarasan, and James Caverlee. 2013. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

[27] Kyumin Lee, Steve Webb, and Hancheng Ge. 2015. Characterizing and automatically detecting crowdturfing in Fiverr and Twitter. *Social Network Analysis and Mining* 5, 1 (2015), 2.

[28] Filippo Menczer. 2016. The spread of misinformation in social media. In *Proceedings of the International Conference Companion on World Wide Web*. 717–717.

[29] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of Twitter bot networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 467–474.

[30] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[31] SiHua Qi, Lulwah AlKulaib, and David A Broniatowski. 2018. Detecting and Characterizing Bot-Like Behavior on Twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 228–232.

[32] Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

[33] PV Savyan and S Mary Saira Bhanu. 2017. Behaviour Profiling of Reactions in Facebook Posts for Anomaly Detection. In *2017 Ninth International Conference on Advanced Computing (ICoAC)*. IEEE, 220–226.

[34] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior* 83 (2018), 278–287.

[35] Neharika Singh and Madhumita Chatterjee. 2017. BotDefender: A Framework to Detect Bots in Online Social Media. *Journal of Network Communications and Emerging Technologies* 7, 9 (2017).

[36] Stefan Stieglitz, Florian Brachten, Davina Berthelé, Mira Schlaus, Chrissoula Venetopoulou, and Daniel Veutgen. 2017. Do Social Bots (Still) Act Different to Humans?–Comparing Metrics of Social Bots with Those of Humans. In *International Conference on Social Computing and Social Media*. Springer, 379–395.

[37] Pablo Suárez-Serrato, Margaret E Roberts, Clayton Davis, and Filippo Menczer. 2016. On the influence of social bots in online protests. In *Proceedings of the International Conference on Social Informatics*. 269–278.

[38] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, et al. 2016. The DARPA Twitter bot challenge. *arXiv preprint arXiv:1601.05140* (2016).

[39] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the International conference on World Wide Web (WWW)*. 977–982.

[40] Tayfun Tuna, Esra Akbas, Ahmet Aksoy, Muhammed Abdullah Canbaz, Umit Karabiyik, Bilal Gonen, and Ramazan Aygun. 2016. User characterization for online social networks. *Social Netw. Analys. Mining* 6, 1 (2016), 104:1–104:28. https://doi.org/10.1007/s13278-016-0412-3

[41] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

[42] Monika Verma and Sanjeev Sofat. 2014. Techniques to detect spammers in twitter-a survey. *International Journal of Computer Applications* 85, 10 (2014).

[43] Hyui Geon Yoon, Hyungjun Kim, Chang Ouk Kim, and Min Song. 2016. Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. *Journal of Informetrics* 10, 2 (2016), 634–644.