Do You Really Follow Them? Automatic Detection of Credulous Twitter Users

Alessandro Balestrucci[1(✉)], Rocco De Nicola[2], Marinella Petrocchi[2,3],
and Catia Trubiani[1]

[1] Gran Sasso Science Institute, L'Aquila, Italy
{alessandro.balestrucci,catia.trubiani}@gssi.it
[2] IMT School for Advanced Studies, Lucca, Italy
rocco.denicola@imtlucca.it
[3] Institute of Informatics and Telematics (IIT-CNR), Pisa, Italy
marinella.petrocchi@iit.cnr.it

Abstract. Online Social Media represent a pervasive source of information able to reach a huge audience. Sadly, recent studies show how online social bots (automated, often malicious accounts, populating social networks and mimicking genuine users) are able to amplify the dissemination of (fake) information by orders of magnitude. Using Twitter as a benchmark, in this work we focus on what we define *credulous* users, i.e., human-operated accounts with a high percentage of bots among their *followings*. Being more exposed to the harmful activities of social bots, credulous users may run the risk of being more influenced than other users; even worse, although unknowingly, they could become spreaders of misleading information (e.g., by retweeting bots). We design and develop a supervised classifier to automatically recognize *credulous* users. The best tested configuration achieves an accuracy of 93.27% and AUC-ROC of 0.93, thus leading to positive and encouraging results.

Keywords: Twitter · Humans-bots interactions · Gullibility · Disinformation · Social networks · Data Mining · Supervised learning

1 Introduction

The diffusion of information on Social Media is often supported by automated accounts, controlled totally or in part by computer algorithms, called bots. Unfortunately, a dominant and worrisome use of automated accounts is far from being benign: malicious bots are purposely created to distribute spam, sponsor public characters and, ultimately, induce a bias within the public opinion [13]. Especially, their malicious activities are of high efficacy when performed on a

targeted audience [4,5] to, e.g., generate misconception or encourage hate campaigns [7]. Recent work in [24,29] demonstrate that bots are particularly active in spreading low credibility content and amplifying their significance. Moreover, human-operated accounts contribute to the diffusion of disinformation by, e.g., retweeting and/or liking fake content.

In a previous work [3], the authors shed light on so called *credulous* Twitter users assuming, with a harmless abuse of language, that they refer to human-operated accounts with a high percentage of bots as friends. Unlike [3], where the authors performed an analysis involving the friends of a set of human-operated accounts - a highly time consuming task - here we design and develop a classifier to find out credulous Twitter users, by considering a number of features that do not take the friendship with bots into account. Starting by considering a set of features commonly employed in the literature to detect bots [10,26], we end up with a lightweight classifier, in terms of costs for gathering the data needed for the feature engineering phase. The classification performance achieves very encouraging results – an accuracy of 93.27% and an AUC (Area Under the ROC curve) equal to 0.93.

We believe that automatically detecting credulous users is a promising line of research. Such an investigation could help researchers to: 1. better understand the characteristics of those users more polarized and/or more willing to be influenced; 2. unveil low-credibility and/or deceptive content and limite their online diffusion; 3. devise alternative strategies for bot detection by concentrating the analysis on the friends of credulous users; 4. improve the users' awareness about threats to data trustworthiness.

The following section presents the approach for the automatic detection of credulous Twitter users, while Sect. 3 presents the experimental results. Section 4 discusses the outcome and suggests further investigations. Section 5 presents related work in the area, arguing on the differences, the contributions and the novelty w.r.t. our work. Section 6 concludes the paper.

2 The Approach

2.1 Datasets

We consider three publicly available datasets[1]: `CR15` [10], `CR17` [12], and `VR17` [26]. From the merging of these three datasets, we obtain a unique labeled dataset (human-operated/bot) of 12,961 accounts - 7,165 bots and 5,796 humans. We use this dataset to train a bot detector, as described in Sect. 2.2. To this end, we use the Java Twitter API[2], and for each account we collect: tweets (up to 3,200), mentions (up to 100), IDs of friends and followers (up to 5k).

The identification of credulous users follows the approach presented in [3]. To this end, we need to detect the amount of bots which are friends of the 5,796 human-operated accounts. Due to the rate limits of the Twitter APIs and to the

---

[1] Bot Repository Datasets: https://goo.gl/87Kzcr.
[2] Twitter API: https://goo.gl/njcjr1.

huge amount of friends possibly belonging to these human-operated accounts, we consider only those accounts with a list of friends lower than or equal to 400 [3]. This leads to a dataset of 2,838 human-operated accounts, namely `Humans2Consider` hereafter. By crawling the data related to their friends, we overall acquire information related to 421,121 Twitter accounts.

2.2   Bot Detection

A bot detection phase is required to discriminate bots and genuine accounts in the dataset of selected friends. The literature offers a plethora of successful approaches [13]; however, also due to the capabilities of evolved spambots to evade detection [12], the performances of the diverse techniques degenerate over time [19]. Furthermore, some bot detectors are available online, but not fully usable due to restrictions in their terms of use, see, e.g., [8]. To overcome these issues, we design and develop a supervised approach, which mixes features from popular scientific work and novel features here introduced.

Regarding the features, we consider two sets. The first one derives from Botometer [26], a popular bot detector[3]. In addition to the original Botometer features [26], we also include: the CAP[4] (Complete Automation Probability) score, the Scores[5], the number of tweets and mentions; we call *Botometer+* this augmented set of features. The second feature set is inherited from [10], where a classifier was designed to detect fake Twitter followers. We use almost all their *ClassA* features[6], except the one about duplicated pictures, because it was not possible for us to verify whether the same profile picture was used twice; we call *ClassA−* this reduced set of features. The conjunction of the two sets of features is referred in the following as *ALL_features*.

We use 19 learning algorithms to train our classifier (with a 10-fold cross validation) and we compare their classification capabilities with respect to the three feature sets (*Botometer+*, *ClassA−* and *ALL_features*). The classification performances are evaluated according to: percentage of *accuracy*, *precision*, *recall*, F-measure (*F1*), and Area Under the ROC Curve (*AUC*). On the most accurate classifier, Hyper-Parameter tuning is performed. The tuned classifier is then used to label the friends of the `Humans2Consider` dataset (see Sect. 2.1).

2.3   Identification of Credulous Twitter Users

The identification of credulous users can be performed with multiple strategies, since there are various aspects that may contribute to spot those users more exposed to the malicious activities of bots. In our previous work [3], we introduced a set of rules to discern whether a genuine user is a credulous one. These rules allow to rank users by relying on the ratio of bots over the user's list

---

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

[4] Complete Automation Probability: https://tinyurl.com/yxp3wqzh.

[5] English/Universal Score: https://tinyurl.com/y2skbmqc.

[6] ClassA features require only information available in the profile of the account [10].

of friends. Here, we inherit these rules to rank the users in our dataset (see Sect. 2.1), but further ranking strategies can be also considered. Our goal is to build a *ground truth* of credulous users to derive an assessed characterization of these accounts. Applying the approach defined in [3], we identified as *credulous* 316 users in `Humans2Consider`.This constitutes the input data for the next step. We note that the approach in [3] is very expensive in terms of data gathering. For example, for the investigated dataset, it requires 421k users' account information and 833 million of tweets.

2.4   Classification of Credulous Twitter Users

Goal of this phase is to build a decision model to automatically classify a Twitter account as credulous or not. As ground-truth, we consider the 316 accounts identified as credulous according to the process described in Sect. 2.3.

We experiment the same learning algorithms and the same feature sets considered in Sect. 2.2, with 10 cross-fold validation. However, for credulous users classification, the learning algorithms take as input a very unbalanced dataset: we have 2,838 human-operated accounts (see Sect. 2.1) and, among them, 316 have been identified as credulous accounts (see Sect. 2.3). To avoid working with unbalanced datasets, we split the sets of not credulous users into smaller portions, equal to the number of credulous users. We randomly select a number of not credulous users equal to the number of credulous ones; then, we unify these instances in a new dataset (hereinafter referred to as *fold*). Then, we repeat this process on previously un-selected sets, until there are no more not credulous instances. Such procedure has been inspired by the *under-sampling iteration* methodology, for strongly unbalanced datasets [18]. Each learning algorithm is trained on each fold. To evaluate the classification performances on the whole dataset, and not just on individual folds, we compute the average of the single performance values, for each evaluation metric.

3   Experimental Results

All the experiments are performed with Weka [28], a tool providing the implementation of several machine learning algorithms. In the following, we present the main results obtained for bot detection and credulous classification, all the details are publicly available: https://tinyurl.com/y4l632g5.

The first column of Tables 1 and 2 shows the set of features considered for learning (i.e., *ALL_features*, *Botometer+*, *ClassA−*, see Sect. 2.2). The second column reports a subset of the adopted machine learning algorithms whose name is abbreviated according to the Weka's notation and reported in the following:

> IBk: K-nearest neighbours [1], NB: Naive Bayes [17], SMO: Sequential Minimal Optimization [22], JRip: RIPPER [9], MLP: Multi-Layer Perceptron [21], RF: Random Forest [6], REP: Reduced-Error Pruning [23], 1R [15]

The remaining columns report the evaluation metrics mentioned above.

Table 1. Results for bot detection

	alg	accuracy	precision	recall	F1	AUC
			evaluation metrics			
	SMO	98.04	0.98	0.98	0.98	0.98
ALL_features	JRip	97.92	0.99	0.98	0.98	0.99
	RF	**98.33**	**0.99**	**0.98**	**0.98**	**1.00**
	SMO	97.64	0.98	0.98	0.98	0.98
Botometer+	JRip	97.61	0.98	0.97	0.98	0.98
	RF	**97.97**	**0.98**	**0.98**	**0.98**	**1.00**
	IBk	91.03	0.91	0.93	0.92	0.91
ClassA-	JRip	94.38	0.96	0.94	0.95	0.96
	RF	**95.84**	**0.98**	**0.95**	**0.96**	**0.99**

Table 2. Results for credulous detection

	alg	accuracy	precision	recall	F1	AUC
			evaluation metrics			
	IBk	89.69	0.74	0.73	0.90	0.96
ALL_features	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
	REP	93.07	0.99	0.88	0.93	0.94
	IBk	65.03	0.61	0.60	0.63	0.70
Botometer+	JRip	66.42	0.67	0.67	0.66	0.67
	RF	**67.81**	**0.68**	**0.69**	**0.68**	**0.73**
	IBk	92.59	0.74	0.73	0.92	0.97
ClassA-	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
	REP	93.09	0.98	0.88	0.93	0.95

Regarding bot detection, Table 1 shows that all the machine learning algorithms well behave, regardless of the feature set. Random Forest is the one that performs best. When the set *ALL_features* is used (see the shaded line in Table 1): accuracy = 98.33%, F1 = 0.98 and AUC = 1.00; and after the tuning phase, we obtain a final accuracy = 98.41%.

Table 2 shows that *ALL_features* and *ClassA−* have good and quite similar classification performances, contrary to *Botometer+*. Both *ALL_features* and *ClassA−* demonstrate their efficacy to discriminate credulous users. On the contrary, the *Botometer+*'s features properly work for bot detection tasks only.

Going into deeper details, in Table 2 we can notice that the 1R algorithm obtains the best accuracy percentage (93.27% with $\sigma = 3.22$) and F-score (0.93), but not the highest AUC (0.93). It is worth noting that the values of the 1R algorithm are exactly the same when considering *ALL_features* and *ClassA−*. This means that the algorithm selects *ClassA−*'s features only, the ones from *Botometer+* are useless in this case. This is a relevant result since we recall that *ClassA−* features refer to the profile of accounts and it is less expensive to collect them.

4 Discussion

The results in Table 2 show the capability of our approach to automatically discriminate those Twitter users with a large number of bots as friends, namely credulous, *without explicitly considering the features of the latter, which would imply a very high cost in terms of data gathering.* To better understand this point, we recall that the approach in [3] for the identification of credulous users needs to crawl a large amount of data, due to the necessity of extending the analysis to the friends of a Twitter account. In the specific case under investigation, this means to retrieve information for more than 400k user accounts, 11 millions of tweet mentions, and more than 820 millions of tweets. As opposite, the credulous detector here proposed requires to gather the profile information of 2,838 accounts only. The classification performances are really promising, with the best accuracy 93.27%, best F1 0.93, best AUC 0.93. We remark that such results have been achieved by relying on so called *ClassA−* features only, i.e., features extracted from the account profile. It is peculiar how the features useful to discriminate credulous genuine accounts are features belonging to the account profile only. This preliminary result calls for three further investigations: 1. to compare the range of values assumed by these features when detecting credulous accounts with the one assumed to detect social bots (as in [10]); 2. to explore the reason why more complex features (such as the ones of Botometer) do not seem to give good results to find credulous users; 3. to perform a deeper analysis on the importance of each specific feature when discriminating credulous users, by means, e.g., of Principal Component Analysis [28]).

Finally, even if the design of a bot detector is not the primary goal of this paper, but only a mean through which we obtain the ground-truth for training the credulous user classifier, we notice that, compared to the performances reported in [12,29], our bot detector achieves very good classification performances. This strengthens the robustness of the ground truth obtained in Sect. 2.3, since the friends' nature evaluation is assessed by means of a very accurate classifier.

5 Related Work

Our work is related to all those approaches that investigate peculiar features of social networks users. We discuss the ones we find more relevant for our approach, with the caveat that the presented literature review is far from being exhaustive.

A survey on users' behaviour in social networks is proposed in [16]: it is remarked that the recipients of shared information should be chosen, in a more precautionary way, by taking into account more real-life relationships and less virtual links. Our approach works exactly in the direction of enhancing the awareness of users, by classifying the ones more exposed to attacks of social bots.

Information spreading on Twitter is investigated in [20], where the authors demonstrate that the probability of spreading a given piece of information is higher when promoted by multiple sources. This supports our attempt to analyze the percentage of bots within the friends of human-operated Twitter accounts, as a symptom for being more tempted to disinformation.

In [2], human behaviour on Facebook is analyzed by building graphs that capture sequence of activities. Behavioural patterns that do not match any of the known benign models likely signal malicious objectives. Similarly, the realization of a classifier to automatically recognize credulous users is the first step to derive their sequence of activities and, hopefully, peculiar behavioral patterns.

In [11,14], a behavioural analysis of bots and humans on Twitter is performed, to draw fundamental differences between the two groups. Specifically, the former demonstrates how, despite a higher level of synchronization characterizing bot accounts, the human behaviour on Twitter is far from being random. The latter defines a 'credibility score' as a measure of how many tweets by bots are present in the timeline of an account. Our work supports the discrimination of credulous users and it may lead to a deeper characterization of human accounts.

To the best of our knowledge, few research explores ways to automatically recognize those Twitter users susceptible to attacks of social bots or exposed to disinformation. A notable example in [27] builds on interactions (mentions, replies, retweets and friendship) between genuine and bot accounts, to obtain a ground truth of users susceptible to social bots. Then, similar to our approach, different learning algorithms have been adopted to train a classifier. Contrary to their approach, the current work is able to classify users close to social bots with lightweight features, all computed from data available in the user's profile. Another brand new line of research is the detection of users susceptible to fake news. Work in [25] monitors the replies of Twitter users to a priori known fake news, in order to tag the same users as vulnerable to disinformation or not. Then, a supervised classification task is launched, to train a model able to classify gullible users, according to content-, user-, and network-based features.

6   Conclusions

Inspired by recent literature that shows how disinformation is not only promoted by social bots but also emphasized by genuine peers, in this work we proposed a supervised classification engine to discriminate *credulous* users, i.e., human-operated accounts with a high percentage of bots as friends. The classifier achieves very good performances and avoids a heavy feature engineering and extraction phase. Further research efforts will be devoted to investigate the

behaviour of credulous users, as well as the posted content, to know more about their peculiarities and the quality of information they contribute to diffuse.

References

1. Aha, D., Kibler, D.: Instance-based learning algorithms. Mach. Learn. **6**, 37–66 (1991)
2. Amato, F., et al.: Recognizing human behaviours in online social networks. Comput. Secur. **74**, 355–370 (2018)
3. Balestrucci, A., et al.: Identification of credulous users on Twitter. In: ACM Symposium of Applied Computing (2019)
4. Bastos, M.T., Mercea, D.: The Brexit botnet and user-generated hyperpartisan news. Soc. Sci. Comput. Rev. **37**(1), 38–54 (2019)
5. Bovet, A., Makse, H.A.: Influence of fake news in Twitter during the 2016 US presidential election. Nat. Commun. **10**(1), 7 (2019)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Chatzakou, D., et al.: Mean birds: detecting aggression and bullying on Twitter. In: ACM Web Science Conference, pp. 13–22 (2017)
8. Chavoshi, N., et al.: DeBot: Twitter bot detection via warped correlation. In: Data Mining, pp. 817–822 (2016)
9. Cohen, W.: Fast effective rule induction. In: Machine Learning, pp. 115–123 (1995)
10. Cresci, S., et al.: Fame for sale: efficient detection of fake Twitter followers. Decis. Support Syst. **80**, 56–71 (2015)
11. Cresci, S., et al.: Exploiting digital DNA for the analysis of similarities in Twitter behaviours. In: IEEE Data Science and Advanced Analytics, pp. 686–695 (2017)
12. Cresci, S., et al.: The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: 26th World Wide Web, Companion, pp. 963–972 (2017)
13. Ferrara, E., et al.: The rise of social bots. Commun. ACM **59**(7), 96–104 (2016)
14. Gilani, Z., et al.: A large-scale behavioural analysis of bots and humans on Twitter. ACM Trans. Web **13**(1), 7 (2019)
15. Holte, R.: Very simple classification rules perform well on most commonly used datasets. Mach. Learn. **11**, 63–91 (1993)
16. Jin, L., et al.: Understanding user behavior in online social networks: a survey. IEEE Commun. Mag. **51**(9), 144–150 (2013)
17. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: 11th Uncertainty in Artificial Intelligence, pp. 338–345 (1995)
18. Lee, J., et al.: An iterative undersampling of extremely imbalanced data using CSVM. In: Machine Vision, vol. 9445 (2014)
19. Minnich, A., et al.: BotWalk: efficient adaptive exploration of Twitter bot networks. In: ASONAM, pp. 467–474 (2017)
20. Mønsted, B., et al.: Evidence of complex contagion of information in socialmedia: an experiment using Twitter bots. PLoS ONE **12**(9), e0184148 (2017)
21. Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. IEEE Trans. Neural Networks **3**(5), 683–697 (1992)
22. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods - Support Vector Learning (1998)
23. Quinlan, J.R.: Simplifying decision trees. Int. J. Human Comput. Stud. **27**(3), 221–234 (1987)

24. Shao, C., et al.: The spread of low-credibility content by social bots. Nature Commun. **9**(1), 4787 (2018)
25. Shen, T.J., et al.: How gullible are you? Predicting susceptibility to fake news. In: Web Science, pp. 287–288 (2019)
26. Varol, O., et al.: Online human-bot interactions: detection, estimation, and characterization. In: 11th Web and Social Media, pp. 280–289 (2017)
27. Wagner, C., et al.: When social bots attack: modeling susceptibility of users in online social networks. In: Making Sense of Microposts, pp. 41–48 (2012)
28. Witten, I.H., et al.: Data Mining: Practical Machine Learning Tools and Techniques (2016)
29. Yang, K.C., et al.: Arming the public with artificial intelligence to counter social bots. Hum. Behav. Emerg. Technol. **1**(1), 48–61 (2019)