Identification of Fake Profiles in Twitter Social Network

Mário Antunes
CIIC, School of Technology and Management, Polytechnic of Leiria
Leiria, Portugal
e-mail: mario.antunes@ipleiria.pt

Hugo Baptista
School of Technology and Management, Polytechnic of Leiria
Leiria, Portugal
e-mail: hfontainhas@hotmail.com

Baltazar Rodrigues
School of Technology and Management, Polytechnic of Leiria
Leiria, Portugal
e-mail: baltazar.rodrigues@ipleiria.pt

***Abstract*— Online social networks are being intensively used by millions of users, Twitter being one of the most popular, as a powerful source of information with impact on opinion and decision making. However, in Twitter as in other online social networks, not all the users are legitimate, and it is not easy to detect those accounts that correspond to fake profiles. In this work in progress paper, we propose a method to help practitioners to identify fake Twitter accounts, by calculating the "fake probability" based on a weighted parameter set collected from public Twitter accounts. The preliminary results obtained with a subset of an existing annotated dataset of Twitter accounts are promising and give confidence on using this method as a decision support system, to help practitioners to identify fake profiles.**

Keywords – online social networks; Twitter; fake profiles.

I. INTRODUCTION

The exponential growth of social networks and the role they play in today's society, both socially and in business, means that it is important to study in depth how they work. The popularity of these applications has led to the possibility of exposing confidential information, the spread of phishing and other cybercrime related activities.

One of the ways to enhance these cybercrime episodes is to use fake profiles, as the information conveyed through fake online social networks profiles may cause disastrous damage to individual and business entities. The cybercrime on online social networks is rising and the "real" authors are not usually punished [3]. These profiles are created for the purpose of anonymizing the account owner or to promote alienation, which challenges law enforcement to identify and trace the attacker.

Online social networks providers have implemented security mechanisms to mitigate these problems, by applying captchas or email validation, and also by requesting the mobile number to send a verification code. Research community is also aware of this issue and machine learning techniques have also been applied to mitigate the problem [10]-[12]. Despite the promising results obtained with some techniques, there are still limitations regarding the access to the public data and the lack of tools available to analyze the "fakeness probability" of an account.

A major challenge is to understand how fake profiles are created and how they work, in order to come up with a solution that may help and warn the users about a possible identified fake profile. The method proposed in this paper aims to contribute to identify a presumable Twitter fake profile, by calculating its fake probability. The developed method collects the values of a predefined set of parameters associated with a public profile, and further applies a weighted parameter set derived from the method published in [1], to calculate the likelihood of an account to be fake. The tests were based on published datasets with Twitter accounts classified as legitimate or associated to a fake profile. Other parameters, not previously mentioned, were added to the weighted parameter set, in order to enhance the results. A comparison between both approaches revealed the usefulness of these new parameters. This methodology and the algorithm may also be extended to other social networks, since the limitations on accessing and using the available Application Programming Interface (API) may be overcome.

The results obtained are promising both in performance and in the level of assertiveness. The datasets used for the tests have provided good indications regarding the level of accuracy to identify accounts related with fake profiles.

The paper is organized as follows: Section II describes the state of art related with the subject; Section III depicts the architecture and the methodology defined to process the Twitter profiles. Section IV describes the tests setup and the datasets used; Section V presents the results; and finally, in Section VI we delineate the conclusions and present actions for future work.

II. BACKGROUND

This work is based on the method proposed by El Azab et al. [1]. The methodology presented was designed for Twitter, but it can be extended to other social networks. According to [1], the Twitter account analysis is based on a set of parameters and a corresponding assigned weight.

The authors' approach proposes the use of as few parameters as possible. Firstly, the factors that categorize a profile as fake were found. Secondly, a classification algorithm that uses the factors previously found to classify an account as corresponding to a fake profile was applied.

Other works have proposed different parameters set, as depicted in Table 1 [1][2][3]. It was found that unnecessarily high number of parameters were used, many of them were not used by social network users or had default values. However, more important than the number of parameters is their relevance in a fake profile detection scenario. In [1], the initial set was of 22 parameters (Table 2).

 ISBN: 978-1-61208-796-2

TABLE I. PARAMETER SET PROPOSED BY DIFFERENT RESEARCHERS (ADAPTED FROM [1]).

Benevenuto et al. [2]	Gurajala et al. [8]	Stringhini et al. [9]
• Number of followers • Number of followees • Followers / followees ration • Number of wweets • Age of the user account • Number of times the user was mentioned • Number of times the user was replied to • Number of times the user replied someone • Number of followees of the user's followers • Number of tweets received from followees • Existence of spam words on screen name • Minimum time between tweets • Maximum time between tweets • Average time between tweets • Median time between tweets • Number of tweets posted per day • Number of tweets posted per week	• Numer of followers • Identification • Friends count • Account verified • Date of creation • General description • Location • Account is updated • URL of profile image • Screen name	• Following / Followers ratio • URL ratio • Similarity among the messages sent by a user. • Friend Choice between screen names • Number of messages sent by a profile • Spammers that send less than 20 messages • Number of friends of a profile

After running the following five learning algorithms with k-fold cross-validation, against a dataset based on "the Fake project" [5], namely: Random Forest, Decision Tree, Naïve Bayes, Neural Network and Support Vector Machine, 19 parameters were chosen. By applying a gain measure algorithm, a weight for each parameter was calculated.

TABLE II. THE INITIAL PARAMETERS SET [1].

Attributes	Weight
The account has at least 30 followers	0.53
The account has been geo-localized	0.85
It has been included in another user's favourites	0.85
It has used a hashtag in at least one tweet	0.96
It has logged into Twitter using an iPhone	0.917
It was mentioned by a twitter user	1
It has written at least 50 tweets	0.01
It has been included in another user's list	0.45
Number of followers and friends' ratio	0.5
User have at least one favourite list	0.17
the profile contains a name	0.0
the profile contains an image	0.0
the profile contains a biography	0.0
the profile contains a URL	0.0
it writes tweets that have punctuation	0.0
it has logged into Twitter using an iPhone	0.0
it has logged into Twitter using an Android device	0.0
the profile contains a physical address	0.0
it has logged into twitter.com website	0.0
it is connected with Foursquare	N/A
it is connected with Instagram	N/A
it has logged into Twitter through different clients	N/A

By applying a comprehensive task list to choose the best parameters set [4][6], a list of the ten most relevant was obtained, which should be used to identify an account as fake [1]. From this list, the parameters whose weight is above 50% and which contribute heavily to the calculation, were identified. The seven parameters obtained, and their corresponding weight are the following [1]:

- The account has at least 30 followers. 0.53
- The account has been geo-located: 0.85
- It has been included in user's favorites: 0.85
- It has used a hashtag in at least one tweet: 0.85
- It has logged into Twitter using an iPhone: 0.96
- It was mentioned by a Twitter user: 1
- Numbers of followers and friends' ratio: 0.5

This set of seven parameters, and its corresponding values were tested in the proposed method described in Section III and benchmarked with other sets, with eight, ten, and eleven parameters.

III. PROPOSED METHOD

The algorithm receives a profile name ("screen name") or a set of profiles and processes them through Twitter API, available at [13]. The first step is to identify if an account with the "screen name" provided exists. Then, for each parameter, the method queries the available profile parameters through the Twitter API. After processing all the parameters, the probability of fakeness of the account is calculated, according to the weight value of each parameter, as described in Section II.

Tests have also been done for Facebook and Instagram, but due to the successive restrictions of the corresponding API, it has become inviable. Some restrictions were related

 ISBN: 978-1-61208-796-2

with General Data Protection Regulation (GDPR) and other with successive privacy breaches that were exploited and addressed by various companies. Another limitation found on Twitter has to do with the privacy settings that the user can select. If the user limits access to the data by defining it as private, it becomes impossible to calculate the level of fakeness of a profile.

IV. DATASET

To perform the tests, two datasets, each one with 100 accounts, both in .CSV format, with the screen names of users to search, were prepared. One dataset has only genuine accounts (not fake) and another has accounts previously classified as fake. The datasets were collected from the My Information Bubble (MIB) Project [7] and a summary is shown in Table 3.

TABLE III. DATASETS INFORMATION FROM MIB PROJECT.

group name	Description	acc	tweets
genuine acc (2011)	Verified human operated accounts	3,474	8,377,522
social spambots #1 (2012)	retweets of an Italian political candidate	991	1,610,176
social spambots #2 (2014)	spammers of paid apps for mobile devices	3,457	428,542
social spambots #3 (2011)	spammers of products on sale at Amazon.com	464	1,418,626
traditional spambots #1 (2009)	training set of spammers used by Yang, et al. [14]	1,000	145,094
traditional spambots #2 (2014)	spammers of scam URLs	100	74,957
traditional spambots #3 (2013)	automated accounts spamming job offers	433	5,794,931
traditional spambots #4 (2009)	automated accounts of spamming job offer.	1,128	133,311
fake followers (2012)	accounts that inflate the number of followers of another account	3,351	196,027

The dataset used, which includes 100 examples of each class (fake and genuine) is a subset of the vast datasets of Twitter accounts, available at MIB. The examples used in our experiments were identified as being related with active Twitter accounts.

V. RESULTS

Table 4 illustrates the average of fake percentage for both genuine and fake accounts presented in the dataset, by calculating the parameters values with 7, 8, 10 and 11 parameters. That is, for genuine accounts the percentage of fake is expected to be low. However, for the fake accounts, the probability of fakeness should be high.

The average results obtained for each dataset were as follows. For the genuine accounts dataset, we have obtained 52.92% of fake probability with 11 parameters, 55.88% with 10 parameters, 41.38% with 8 parameters and, the best result, 33,59% with 7 parameters. For the fake accounts dataset, the best result was obtained with an average of 87.73%, by using the set with 11 parameters, 86.5% with 10 parameters, 83.13% with the 8 parameters and 80.71% with 7. This means that in the dataset with only genuine accounts, the lower the fake probability (33.59%) is, the better chance the account has to be genuine. Otherwise, in the dataset of fake accounts, the higher the fake probability (87.73%) is, the better chance the account has to be fake.

TABLE IV. AVERAGE OF FAKE PERCENTAGE.

	11 par.	10 par.	8 par.	7 par.
Genuine accounts	52,92	55,88	41,38	33,59
Fake accounts	87,73	86,50	83,13	80,71

In order to better understand the values obtained, the "fake probability" was sliced into six stripes, between 40% and 90% in intervals of 10%. The underpinning idea is to have a closer precision regarding the level of assertiveness which is intended to be considered in the analysis. In a decision support system approach, this analysis could tune the level of confidence of the decision maker.

Table 5 represents the number of accounts on each percentual range, for the dataset of genuine accounts. Analysing the table, it is possible to observe that, with 11 parameters, 9 accounts have a high probability of being fake (90%). This means that, in a dataset with genuine accounts, almost all of them have a low probability of being fake. For the same parameters set, it is also possible to observe that 96 accounts have a fake probability below 50%, which may infer they are legitimate.

TABLE V. RESULTS IN PREDEFINED INTERVALS FOR GENUINE ACCOUNTS

%	>=40	>=50	>=60	>=70	>=80	>=90
11 par.	96	66	41	24	15	9
10 par.	70	45	29	22	15	9
8 par.	64	41	29	19	11	7
7 par.	45	29	20	19	10	7

However, with a threshold >= 80% the number of false positives increases to 15, and with a percentage >= 70%, increases to 24 accounts misclassified.

The range of values with the largest difference is between 60% and 70%. For the 10 parameters set with a threshold of 40%, we obtained 70 accounts, with a value >= 50% the value decrease to 45, and with a value of >60% we registered 29 accounts. Finally, we obtained 22 accounts that have a fake probability >=70%, with 80% the value decreases to 15, and with a value >=90% only 9 were classified as fake. In this case, the interval with the largest decrease is between 40% and 50%.

Table 6 represents the values obtained for the fake dataset, identifying the accounts in each fake interval. For a fake threshold of 70% all accounts are classified as fake except for the set of 7 parameters which has classified 97 accounts as fake. For 11 parameters set and for a threshold < 90%, 99 accounts were classified as fake, and only one account was misclassified. Even for the threshold of 90%, that is a high level of sensitivity, the method classifies 40% of the accounts correctly.

 ISBN: 978-1-61208-796-2

TABLE VII. RESULTS IN PREDEFINED INTERVALS FOR FAKE ACCOUNTS

%	>=40	>=50	>=60	>=70	>=80	>=90
11 par.	100	100	100	100	99	40
10 par.	100	100	100	100	97	40
8 par.	100	100	100	100	97	33
7 par.	100	100	100	97	96	33

Considering the data represented in Table 5, for values >=80%, and in Table 6 for values >=90%, comparing the values of the different parameter sets, we may infer that some parameters have no impact on the overall values obtained in the experiments.

VI. CONCLUSIONS

It is important to give people the knowledge needed to identify fake accounts on social networks. For the police investigators, in a digital forensics' perspective, this kind of solutions helps to better deal with cybercrime and malicious activity in online social networks.

This work in progress aims to contribute to identify fake profiles in online social networks. The work provides an additional resource that may help deciding about the veracity of a Twitter profile. The sensitivity of the decision was calculated by the probability intervals defined in the analysis. For instance, if an account shows a fake probability of 90%, it is possible to infer that it is strongly fake; being legitimate means that an account shows a fake probability of 40% or less. This method does not give a guarantee that an account is fake or genuine, but it gives an additional help on the overall final decision.

Besides the proposed methodology and the preliminary tests carried on, it was also evaluated the impact of the number of parameters extracted from the Twitter profiles. The development of a web application that incorporates the method and work described in this paper, is now being carried on. The web application should be available to those who aim to evaluate the legitimacy of a Twitter account.

The research is now focused on two major directions: i) to explore others API besides Twitter, such as Facebook, LinkedIn and Instagram. It is important to explore the various directions that may lead to obtain more information from the API, even using a paid version; ii) to work on the optimization of the parameters set and its continuous evaluation, by applying machine learning techniques for optimization. Finally, the parameters set can also be improved, not only in the weights but also in the selected parameters, as some of them are directed towards a specific scope (e.g., users that have a specific equipment, like iPhone) and some adjustments can be made in this subject.

REFERENCES

[1] A. El Azab, A. Idrees, M. Mahmoud and H. Hefny, "Fake Account Detection in Twitter Based on Minimum Weighted Feature set," International Journal of Computer, Electrical, Automation, Control and Information Engineering, World Academy of Science, Engineering and Technology. WASET, vol. 10, No. 1, pp.13-18, Nov. 2015.

[2] F. Benevenuto, G. Magno, T. Rodrigues and V. Almeida, "Detecting spammers on twitter. In Collaboration, electronic messaging", anti-abuse and spam conference (CEAS), vol. 6, No. 2010, pp. 12, 2010

[3] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," Neurocomputing, Elsevier vol. 159, pp. 27–34, Jul. 2015.

[4] D. Kagan, Y. Elovichi, and M. Fire, "Generic anomalous vertices detection utilizing a link prediction algorithm," Social Networks Analysis and Mining, vol. 8, no. 1, pp. 27, Dec. 2018.

[5] "The Fake Project.", http://wafi.iit.cnr.it/fake/fake/app/. [Retrieved: September, 2020].

[6] T. Yoshida, "Term weighting method based on information gain ratio for summarizing documents retrieved by IR systems", Journal of Natural Language Processing, vol.9, No.4, pp.3-32, 2001

[7] M. P. Fazzolari, "My Information Bubble project.", Available: http://mib.projects.iit.cnr.it/index.html. [Retrieved: September, 2020].

[8] S. Gurajala, J. S. White, B. Hudson, and J. Matthews, "Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in SMSociety , Toronto, ON, Canada, 2015

[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9, 2010

[10] A. Meligy, H. Ibrahim, and M. Torky, "Identity verification mechanism for detecting fake profiles in online social networks." *Int. J. Comput. Netw. Inf. Secur.(IJCNIS)*, vol.9, no. 1, pp.31-39, 2017

[11] A. K. Ojo, "Improved Model for Detecting Fake Profiles in Online Social Network: A Case Study of Twitter." Journal of Advances in Mathematics and Computer Science, vol. 33(4), pp.1-17, 2019

[12] S. R. Sahoo, and B. B. Gupta. "Hybrid approach for detection of malicious profiles in twitter." Computers & Electrical Engineering, vol. 76, pp. 65-81, 2019

[13] "Use cases, tutorials and documentation – Twitter developer"; https://developer.twitter.com/en. [Retrieved: September 2020]

[14] C. Yang, R. Harkreader and G. Gu, "Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers,", IEEE Transactions on Information Forensics and Security, vol. 8, no. 8, pp. 1280-1293, 2013

 ISBN: 978-1-61208-796-2