Machine Learning and Bots Detection on Twitter

Norberto De Almeida Andrade, Giuliano Carlo Rainatto, Fontamara Lima, Genésio Renovato, Denis Gustavo Espacacherch Paschoal

Abstract: ***The social networking sites digital Become Increasingly popular, they also Attract the attention of spammers. This article, Twitter, the popular micro-blogging service, is an example of the studied bot detection on digital social networking sites. Machine learning is considered to regular spam robots Distinguish. To facilitate the detection of spam, there are three aspects, the number of friends, number of followers and users. Data from all groups are extracted to Twitter. Three features have been added in 20 most recent user tweets. A set of current data is collected from the Twitter Object-telescope itself as it is Necessary to use two different methods. Evaluation experiments have increased the risk of error on Twitter.***

Keywords: Twitter, Machine Learning, Bots, TrustRank

1. Introduction

The digital social networking sites are becoming more popular every day, such as Facebook, Twitter and LinkedIn. Among all these sites, Twitter is one of the most studied because of its huge intereação between network users. Lately, the exponential increase in spam has become a growing problem on Twitter and other online social networking sites (Bakshy, 2011).

Spammers use Twitter as a tool to post multiple duplicate updates containing malicious links, abusing the response function to post unsolicited messages to users, and hijack threads trend. Spammers also put offensive terms in the Twitter trending topics displayed on the main Twitter page several times, forcing Twitter to remove offensive terms (Tavares & Faisal, 2013).

Twitter tried several ways to fight spam, which includes the addition of a "report as spam to their service and cleanliness of suspicious accounts. However, legitimate Twitter users complain that their accounts and Twitters are getting arrested in the anti-spam actions. Twitter recently admitted the accidental suspension of the accounts as a result of spam cleanup effort (Brito, 2013).

In this article, spam suspicious behavior are studied. The goal is to apply methods of machine learning to distinguish robots normal spam. Posteriorly the article is organized as follows. In Section 2, related works are discussed. In Section 3, new features based on proposed content and graphics to facilitate spam bots detection. Bayesian classification method is applied in Section 4 for detecting spam on Twitter. Section 5 presents two data collection methods. They are also conducted experiments to evaluate the performance of the detection system.

2. Surveys

Spam detection has been studied for a long time. One of the first research focuses on detecting spam email and spam detection Web (Agarwal, 2011). Sahami et al. (1998) proposed a Bayesian approach to filter spam emails. The results of the experiment show that the classifier has better performance considering resources beyond plain text e-mail messages. Currently e-mail spam filtering and has a very mature technique (Aiello, 2014). Bayesian spam filtersemails are implemented both customers and modern e-mail servers. The Web is huge, it changes rapidly and spreads on computers distributed geographically, for this reason is a significant challenge to detect spam web (Davis, 2016).

The TrustRankalgorithmis proposed to compute the confidence score for a Web graphic. Based on scores calculated where good pages with higher scores are spam pages that can be filtered in the results of search engine (Haustein, 2016).

Authors based on the Web link structure proposed a measure Spam Mass to identify spam links (Chavoshi, HamooniandMueen, 2016). A graphic model driven Web is proposed in Brito et al. (2013). The authors apply classification algorithms for directed graphs to detect real-world links spam. In Ferrara et al. (2016), both based features links as content-based features are proposed. The basic decision tree classifier is implemented to classify spam. In Tavares andFaisal (2013), semi-supervised learning algorithms are proposed to increase the performance of a classifier that needs only small amountoflabeled samples.For spam detection in other applications, Davis (2016) presents an approach to detect spam calls through IP telephony call SPIT in VoIP system. As Clark (2016) the popular methods of semi-supervised learning, an improved algorithm called MPCK-Means is proposed. In Varol (2017), the author collects three sets of Twitter network data: user behaviors, geographic pattern of growth and current size of the network are studied.

3. Features

The resources extracted for spam detection include three features based on graphic and three features based on content. As a social networking site, Twitter allows users to create their own social graph (Botta, MoatandPreis, 2015). Threegraphics-basedfeatures are extracted from the Twitter social graph to capture the "next" relationship between users. Twitter also allows users to broadcast short messages in 140 characters, known as "tweet" to friends or followers (Haustein, 2016). In thisresearch are extraidoso three features based on content of the 20 most recent tweets of users.

Based features Graphic

Next, one of the most important functions and unique Twitter. Users can build their own social network by following friends and allowing others to follow themon Twitter (Agarwal, 2011). It'spossiblemonitor the accounts of his friends to get your updates automatically on your Twitter homepage when it logs in. Your friends can send your private messages, direct messages calls, if you follow them (Bakshy, 2011). Spammers use thefollowing function to draw the attention of legitimate users following your bills, since Twitter will send a notification by email when someone follows your account. Twitter considers this a spam bot, if this account has a small number of followers compared to the amount of peopleyou "follow" (Brito, 2013). Threegraphics-basedfeatures are the number of friends, the number of followers and the followers rate is extracted to detect spam on Twitter. If one follows your account, it will become one of his followers. If you follow someone's account, then it becomes one of your friends. The number of friends and the number of followers is extracted for each individual Twitter account(Chavoshi, HAMOONI and MUEEN, 2016). As Varol (2017) moreover, thefollowers rate is calculated based on the number of followers and the number of friends. Leave N_{fo} denote the number of followers, N_{fr} denote the number of friends and r_{ff} denote the proportion of followers. To normalize the follower ratio, this feature is defined as the ratio between the number of people you are following and the number of people who are following you. So:

$$r_{ff} = \frac{N_{fo}}{N_{fo} + N_{fr}}$$

Obviously, if the number of followers is relatively small compared to the amount of people you are following, the proportion of followers is relatively small and close to zero (Haustein (2016). At thesame time, the probability of the associated account to be spam is High (Varol, 2017). Based features Twitter content are introduced. Three characteristics are analyzed: the number of duplicate tweets, the number of HTTP links and the number of replies / mentions are drawn from the 20 responses of the latest tweets (Chavoshi, HamooniandMueen, 2016). First, anaccountcan be considered a spam to publish duplicate content on an account. A sample of Twitter spam page is shown in Table 1 of the analyzed Robots. Usuallynotlegitimateusers will post duplicate updates. Tweets duplicates are detected by measuring the Levenshtein distance (also known as edit distance) between two different tweets posted by thesameaccount (AbokhodairYooand McDonald, 2015). The Levenshteindistanceis defined as the minimum cost of transforming a chain together by means of a sequence of editing operations, including deletion, insertion and replacement of individual symbols (Ferrara, 2016). The distanceis zero if and only if the two are identical tweets. To avoid detection and spam different accounts, spam robots typically include @usernames in your duplicate tweets (Davis, 2016). Whenthedistances are calculated between Levenshtein different tweets, I clean the data excluding @replies, #topic and slinks HTTP. In other words, the answer / mention, topic and link information are ignored when capturing the duplicate tweet, instead, only the content of the tweets is considered (Haustein, 2016). Secondly, spam botstrying to post malicious links in your tweets to entice users to click. Twitter only allows you to post a message in 140 characters, some URL shortening services and applications such as bit.ly, it has become popular to meet the requirements (BRITO, 2013). The URL shortenerobscures the target address and as a result, facilitates spam accounts on matches, phishing or affiliate hiding. So Twitter considers this a spam factor if your tweets consist mainly of links, and not personalupdates (andFaisal& Tavares, 2013).

Table 1: Analyzed Robots

Robot Name	content Posted	Home Activity
grammar	Warns of grammatical errors and agreement of the Portuguese language.	June 2015
_reclamejá	Reports complaints of brands and customer purchases.	January 2014
linda_como_divas	Gives fashion tips and trends of celebrities.	September 2016
o_brasilalegal	Jokes and memes acid humor.	March 2015
comestics_2018	Tips on makeup, cosmetics and beauty tutorials.	December 2017
vigaristas_insanos	funny phrases, puns and memes.	February 2018
politico_imundo	Affairs on policy in Brazil and the world with sarcasm.	May 2016
crush_now	Sharing stories about former relationships.	August 2015
tinder_tander	Relationships and people looking for a couple.	April 2014
fatos_e_boatos	Facts about international celebrities.	November 2017
mitou_sempre	Football humorous tone.	January 2015
fifth_harmony_never	International singers not worth following.	July 2016
Maria Cecilia	Tips on pregnant women, infants and mothers of children up to 5 years.	November 2015
miga_sua_loka	Feminism and homosexuality.	March 2016
amor_prosa_sexo_poesia	Literatures of prose and poetry.	September 2015
café_com_empreendedor	entrepreneurship tips and money.	April 2017
marielle_presente	Discussions about the death of Marielle.	December 2016
bolsomito_me_representa	Supporters of Bolsonaro.	February 2015
tudodiferentedetudo	Geeks, nerds and games.	June 2017
the_books_onthetable	Books, serials and spoilers.	March 2016

The number of links in an account is measured by the number of tweets containing HTTP links in the 20 most recent tweets user. If a tweet contains the string "http: //" "www." Or, this tweet is considered to contain a link. Third, the number of replies / mentions is extracted from the 20 most recent tweets answers. On Twitter, users can use the format @ + username + message to designate their message as a reply to another person (Clark, 2016). Youcanrespondto any person tweet on Twitter, no matter if they are your friends or not. You can also mention another username (@username) anywhere in the tweet, instead of just the beginning. Twitter collects all tweets that contain your username in the User @name format in your answer guide. You can see all replies made to you and mentions of your username. The response functions and reference are designed to help users discover another on Twitter. However, the spam account uses the service to draw other users' attention by sending replies and mentions unsolicited. Twitter also consider it as a factor to determine spam. The number replies and mentions in an account is measured by the number of tweets that contain the response "@" sign

Spam Bots Detection

In this section, applies different classification methods such as decision trees, neural networks, support vector machine and k nearest neighbors (nearest neighbors) for spam bots on Twitter (Varol, 2017). SecondBotta, Moatand Preis (2015) and Ferrara (2016), among these algorithms, Bayesian classifier performs the best for several reasons. First, the Bayesian classifier is robust as to noise. Another reason that the Bayesian classifier performs better is by class label account provided based on the specific user default. Davis (2016) reportsthatuma spam probability is calculated for each individual user based on their behavior, rather than providing a general rule. In addition, the Bayesian classifier is a simple sorting algorithm and very efficient.The Bayesian classifier based on the known Bayes' theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

The conditionalprobability P (Y | X) is also known as the posterior probability for Y, as opposed to a prior probability P (Y). Each Twitter account is regarded as a vector X using values as discussed in Section 3 (Chavoshi, HamooniandMueen, 2016). The vectors are classifiedintotwo Y classes: spam and non-spam. AccordingtoVarol (2017) toclassify a data record, with later probabilitycalculated for eachclass:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$$

Since P (X) is a normalization factor that is equal for all classes, we just need to maximize the numerator $P(Y)\prod_{i=1}^{d} P(X_i|Y)$ to the classification (Varol, 2017).

Data set ofexperiments

The set of data is collected using two methods. First uses the methods of the Twitter API to collect detailed user information. Second, a Web crawler is developed to extract 20 most recent tweets of users. The use of the API method public schedule collects information on 20 users unprotected that defined a custom user icon in real time. This method can randomly select 20 unprotected users who upgrade their status on Twitter recently. Later extracted details of the current user, such as IDs, screen name, location, etc. At thesame time also is used API methods of social graphs to collect information about friends and followers of the user, as the number of friends, the number of followers, the list of IDs of friends, followers IDs list and etc. APIs friends and Twitter followers can return a maximum of 5,000 users. And if a user has more than 5,000 friends or followers, you can extract only a list of friends or followers. Based on observation, the number of friends and followers of the majority of users do not exceed 5,000 friends or followers, then this restriction does not affect significantly method.

Another restriction of the Twitter API methods is the number of queries per hour. Currently, the rate limit for calls to the API is 150 requests per hour. To collect data from different time periods and avoid cluttering the Twitter Web servers, you need to track Twitter continuously and limit the request for 120 calls per hour. Although the Twitter API methods provide pure, there is no method to collect the recent tweets of an unauthorized user specific. The timeline API method can only return the latest update of 20 users unprotected (an update of a user). The method of the user timeline API can return the 20 most recent tweets posted only an authenticated user. Recent tweets postedby a user are important to extract resources based on content, as duplicate tweets. To resolve this problem, develops a Web crawler to collect the 20 most recent tweets from a specific user not protected based on user ID on Twitter. The extracted tweets are saved either as an XML file in a relational database. Finally, I collected a set of data for three weeks, from May 13 to June 7, 2018. The collection totaled 25,847 users, about 500,000 tweets and about 49 million followers / friends relationships are collected from publicly available data on Twitter.

4. Evaluation

To evaluate the method is labeled manually 500 Twitter user accounts for two classes: spam and not spam. Each user account is manually evaluated by reading the 20 most recent tweets posted by the user and checking the friends and followers of the user. The result shows that there is about 1% of spam account in the data set. The study shows that there are probably 3% of spam on Twitter. To simulate reality and avoid bias in the screening method, add up more spam data to the data set. As mentioned in Section 1, Twitter provides several methods for users to report spam, which includes sending a direct message to Twitter and click on the link "report for spam". The simplest method available audience is post a tweet in the format "@spam @username" where @username should mention spam account. He wondered "@spam" to collect an additional set of spam data. It turned out that this service is abused by fraud and spam. Only a small percentage of @spam tweets are reporting spam. Finally, the data set is mixed containing about 3% of spam. The assessment of the overall process is based on a set of measures commonly used in Machine Learning and Information Retrieval. Given a ranking algorithm C, it is considered a lossofmatrix:

	Prediction
	Spam Not Spam
true Spam	ab
no Spam	CD

Three measurements are considered in the evaluation experiments: precision, recall, and measurement accuracy is F. Q = a / (a + c) and the memory is R = a / (a + b). As F is defined as F = 2PR / (P + R). To evaluate the classification algorithms, we focus as F, it is a standard way to summarize precision and recall. Allforecastsreported in the survey are calculated using a crossover 10 validations. For each classifier, precision, recall and F measure are informed. Each classifier is tested 10 times, each time using the 9 of 10 partitions as training data and computing the losses array using the tenth partition as test data. The valuationmetrics are estimated average loss matrix. The evaluation results are shown in Table 2. The Bayesian classifier is the best overall performance compared to other algorithms.

Table 2: Assessment Rating

Sorter	Precision	Recall	Measure F
Decision tree	0677	0334	0432
Neural networks	1	0516	0591
Machines Support Vector	1	00:26	0.5
naïve Bayesian	0937	0937	0937

5. Conclusion

In this research, focused on the suspicious behavior of the bots spam in digital social networks. A popular microblogging service called Twitter, is studied as an example. A machine learning learning approach is proposed to identify the non-common spam bots. Basedonthe spam policy Twitter, graphics-based features and content-based features are extracted from the user's social graph and the latest tweets. The traditional classification algorithms are applied to detect spam suspicious behavior. A Web crawlerusing the Twitter API is designed to collect actual data from public information available on Twitter. Finally, we analyze the data set to evaluate the performance of the detection system. Several popular ranking algorithms are studied and evaluated. The results show that the Bayesian classifier provides better overall performance.

The dbots etecção through machine learning to adapt their traffic to avoid non-human behavior. cybercriminals, media, politicians, fraudsters and even competitors use bots to their strategies, including promoting fake news (false news). Botscanperformspecific actions, such as generating duplicate accounts, participate in SPAM user-generated content, create fraudulent transactions or invade existing user accounts. Bots and non-human traffic are responsible for billions of dollars a year on advertising and click fraud. The proposal filtrar and detect the most sophisticated bots that mimic human behavior. Prevent and capture the abuse of login account, accounts and fraudulent purchases and other types of automated fraudulent behavior.

References

[1] Abokhodair, N., Yoo, D., & McDonald, D. W. (2015, February). Dissecting a social botnet: Growth, contentandinfluence in Twitter. In *Proceedingsofthe 18th ACM Conferenceon Computer SupportedCooperativeWork& Social Computing* (pp. 839-851). ACM.

[2] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., &Passonneau, R. (2011, June). Sentimentanalysisoftwitter data. In *Proceedingsofthe Workshop onLanguage in Social Media (LSM 2011)* (pp. 30-38).

[3] Aiello, L. M., Deplano, M., Schifanella, R., &Ruffo, G. (2012, May). People are strangewhenyou're a stranger: Impactandinfluenceofbotson social networks. In *SixthInternational AAAI ConferenceonWeblogsand Social Media*.

[4] Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone'saninfluencer: quantifyinginfluenceontwitter. In *Proceedingsofthefourth ACM internationalconferenceon Web searchand data mining* (pp. 65-74). ACM.

[5] Botta, F., Moat, H. S., &Preis, T. (2015). Quantifyingcrowdsizewith mobile phoneand Twitter data. *Royal Society open science*, *2*(5), 150162.

[6] Brito, F., Petiz, I., Salvador, P., Nogueira, A., & Rocha, E. (2013). Detecting social-network botsbasedonmultiscalebehavioralanalysis. In *Proc. 7th Int. Conf. Emerg. Secur. Inf., Syst. Technol.(SECURWARE)* (pp. 81-85).

[7] Chavoshi, N., Hamooni, H., &Mueen, A. (2016, November). Identifyingcorrelatedbots in twitter. In *InternationalConferenceon Social Informatics* (pp. 14-21). Springer, Cham.

[8] Clark, E. M., Williams, J. R., Jones, C. A., Galbraith, R. A., Danforth, C. M., &Dodds, P. S. (2016). Siftingroboticfromorganictext: a natural language approach for detectingautomationon Twitter. *JournalofComputational Science*, *16*, 1-7.

[9] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., &Menczer, F. (2016, April). Botornot: A system toevaluate social bots. In *Proceedingsofthe 25th InternationalConference Companion on World Wide Web* (pp. 273-274). International World Wide Web ConferencesSteeringCommittee.

[10] Ferrara, E., Varol, O., Davis, C., Menczer, F., &Flammini, A. (2016). The riseof social bots. *Communications ofthe ACM*, *59*(7), 96-104.

[11] Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., &Larivière, V. (2016). Tweets as impactindicators: Examiningtheimplicationsofautomated "bot" accountson T witter. *JournaloftheAssociation for Information Science and Technology*, *67*(1), 232-238.

[12] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach tofilteringjunk e-mail. In *Learning for TextCategorization: Papersfromthe 1998 workshop* (Vol. 62, pp. 98-105).

[13] Tavares, G., &Faisal, A. (2013). Scaling-lawsofhuman broadcast communication enabledistinctionbetweenhuman, corporateandrobottwitterusers. *PloSone*, *8*(7), e65774.

[14] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., &Flammini, A. (2017, May). Online human-botinteractions: Detection, estimation, andcharacterization. In *Eleventhinternational AAAI conferenceon web and social media*.