TWITTER BOT DETECTOR

VISHWAJEET RAI[1] Research Student[1]
YASHDEEP SINGH[2] Research Student[2]
FARAZ CHISHTI[3] Assistant Professor[3]

Abstract: - According to a report by CNBC about 40 million accounts on Twitter are bots and it has been estimated that up to 50% of the activity on Twitter is from these bots. Their motive is to advertise a product, distribute spam, or alter public opinion. Then it becomes extremely necessary to detect bots to protect genuine users from misinformation and malicious intents. Several types of research have been carried out in past years, but current algorithms still lag in performance and are not much efficient in determining whether the account is genuine or not. The idea behind this project is to build a Binary Classifier that identifies a given user as "bot" or "Human".

This application is a web browser-based plug-in that would give a score to a given account based on 12 features that are verification status, followers, friends to follower's ratio, frequency of tweets, re-tweet count per tweet Tweets variance, etc. Based on that score the end-user application would identify the account. It is our opinion that an application like this is sorely needed for a Twitter user for his own safety and for better confidentiality of data. The approach is to identify bots, of any type based on features that are acquired from user profile metadata and accumulated statics based on timeline data (tweet history).

1. INTRODUCTION

Twitter a Social networking website and a microblogging service. It is based on "tweets" which are 140 characters long messages and is used by various celebrities, authors, journalists, media, etc. It is used for various purposes from being in touch with people of your interests to the company's promotion.500 million tweets are sending every day for various purposes. But the downside is that about 15% of these accounts are not genuine accounts according to a recent done by from computer scientists at Indiana University and the University of Southern California. Before we go into the downsides of the rise of bots online, it's important to take a step back and realize that the mere existence of bot technology isn't inherently evil or malicious.
Indeed, when technologists talk about "bots" in the broad sense, they're really just referring to any software application that can execute commands, reply to messages, or perform routine tasks on the internet automatically, or, at the very least, with minimal human intervention.

"Twitter bot" is the term used for the anomaly software that controls a Twitter account via the Twitter API. This bot software may autonomously perform actions such as tweeting, re-tweeting, liking, following, unfollowing, or direct messaging other accounts.

Detecting these bots have been of interest to academics. Indiana University has developed an application called Bolometer previously known as BotOrNot, which scores Twitter accounts based on their odds of being a "Twitterbot" Negative implication that goes far beyond the political. For instance, that massive bot population throws the entire nature of Twitter marketing – its efficacy, its reach, and its purpose – into question for countless brands using the platform for social marketing media marketing. While it may be all but impossible to ever fully purge its network of malicious bots, Twitter has been taking some key strides toward downplaying their negative effects in recent weeks, unveiling a handful of new security features, including the ability to block accounts without profile photos or verified email addresses. There are 12 features based on which we give a score to an account, to find out whether it is a genuine one or not.

2. LITERATURE REVIEW

The literature review includes the current information including substantive findings, theoretical and methodological contributions to the Twitter bot detection.

R. Gorwa[1] wrote an article on Twitter bot. According to him "Twitter has a serious bot problem, and Wikipedia might have the solution". His main concern was toward the brainwashing of the viewer's mind by bots. Twitter Bot play a vital role in spreading unchecked news which can be false or misleading information and affect the judgment of human.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu [2] *wrote an article "**Fake News Detection on Social Media: A Data Mining Perspective".*** According to the article" Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media on the other hand, it enables the widespread of fake news" [2].

S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi [3]

Wrote an article **"Fame for sale: efficient detection of fake Twitter followers".** Their main concern was fake followers, Fake followers are those accounts which are specifically created to increase the number of followers of a target account. As fake followers may change concept like influence, popularity on the twitter which can impact on economy, politics, and society.

Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, Alessandro Flammini [4] wrote an article "**Online Human-Bot Interactions: Detection, Estimation, and Characterization [4]".** According to this article, Increasing evidence suggests that a growing amount of social media content is generated by autonomous entities known as social bots. In this work, we present a framework to detect such entities on Twitter [4].

E. Tacchini, G. Ballarin, M. L. D. Vedova, S. Moret, and L. de Alfaro[5] wrote an article **"Some Like it Hoax: Automated Fake News Detection in Social Networks [5]".**

3. OBJECTIVE

The objective is to build a Binary Classifier that identifies a given user as "bot" or "Human".

This application is a web browser-based plug-in that would give a score to a given account based on 12 features that are verification status, followers, friends to follower's ratio, frequency of tweets, re-tweet count per tweet Tweets variance, etc. Based on that score the end-user application would identify the account.

4. DATASET AND FEATURE

a.) DATA SOURCES

To develop our classification, we need a variety of publicly datasets consists of pre-labeled examples of genuine and bots account to train and cross-validate our model. all the datasets are accessible via Indiana University's Bot Repository project [9].

TABLE 1. TRAINING DISTRIBUTION [4], [5]

group name	description	account	Tweets	years
Genuine accounts	verified accounts that are humanoperated	3,474	8,377,522	2011
social spambots #1	retreaters of an Italian political candidate	991	1,610,176	2012
social spambots #2	spammers of paid apps for mobile devices	3,457	428,542	2014
social spambots #3	spammers of products on sale at Amazon.com	464	1,418,626	2011
traditional spambots #2	spammers of scam URLs	100	74,957	2014
traditional spambots #3	automated accounts spamming job offers	433	5,794,931	2013
traditional spambots #4	another group of automated accounts spamming job offers	1,128	133,311	2009
fake followers	simple accounts that inflate the number of followers of another account	3,351	196,027	2012

We used Training distribution (Table 1) to train and rotation estimation of our models.

TABLE 2. Test Distribution

Dataset	User	%	Tweets	Year
Human	1510	67	1437889	2017
Bot	735	33	667477	2017
Total	**2245**	**100**	**2105366**	

The Test distribution (Table 2) is approximate the current Twitter environment. Test distribution is used for model cross-validation as well as for calculating our results. We downloaded fresh tweets, retweets and user profile for each of the accounts.

b.) Preprocessing and Feature selection

Twelve features are generated by given users profile and information on tweet from the described dataset in section 3. A which are immediately available in Bot repository dataset/Twitter API or from derived statistics (Table 3). We used many directly available features, such as (1) followers count, (2) verification status, (3) favorites count and (4) friends count.

Friend to follower's ratio (5) is derived statics that has been used in literature for many years [3]. Some of our other features were calculated by aggregating over users' tweets, such as (6) numbers of mentions per tweets, (7) retweet count per tweet and (8) favorite count per tweet.Number of hashtags. All features by scaling to unit variance and centering the mean at zero is standardized to make them ready for the model fitting process.

c.) Raw data downloaded from twitter API and to get the desired features for each user. After dividing the results into 4 datasets (1) Train, (2) Train-dev, (3) Dev, and (4) Test, to make the train and train dev set split the example randomly from the training distribution by 80:20 ratio. Dev and test dataset were created by us from splitting examples of test distribution at a 50:50 ratio. Models are trained by using train datasets. Dev and train Dev are used for model cross-validation and test dataset is used to show the result.

TABLE 3. FEATURES

FEATURES	Description
Verification status	True if the user has an account that has been authenticated by Twitter
FOLLOWERS	count Number of followers the user has
FAVORITES	count Number of tweets the user has liked in the account's
FRIENDS	count Number of users this account is following
NUMBER OF MENTIONS PER TWEET	Number of tweets in sample
NUMBER OF HASHTAGS PER TWEET	Count of hashtags in tweet sample "#topic" / Number of tweets in sample
NUMBER OF URLS PER TWEET	Count of URLs in tweet sample
RETWEET COUNT PER TWEET	Count of retweets done by the account
FAVORITE COUNT PER TWEET	Count of favorites in the account
UNIQUE TWEET PLACES COUNT	Count of unique tweet places tagged
VARIANCE IN TWEET RATE	Variance of tweet rate (no of tweets / hour)

5. MODELS

To train our model we used trained set of examples around 10,000 accounts each with a twelve-dimensional feature vector.

We label each example account with binary number {0,1}, where 0 for bot and 1 for genuine users.

Aim was to make a function which can accurately predict the class of user i.e. {0,1}. To make such a function. We used logistic regression, gradient-boosted classification, by an MLP (multi-layer perceptron) neural network.

a) Logistic Regression

Logistic regression is used to predict whether our user is bot or genuine. Logistic regression fitting an "S" shaped line to the data where the curve goes from bot to genuine user. The curve can help to predict the probability that user is bot or not. b) Gradient-Boosted Classifier

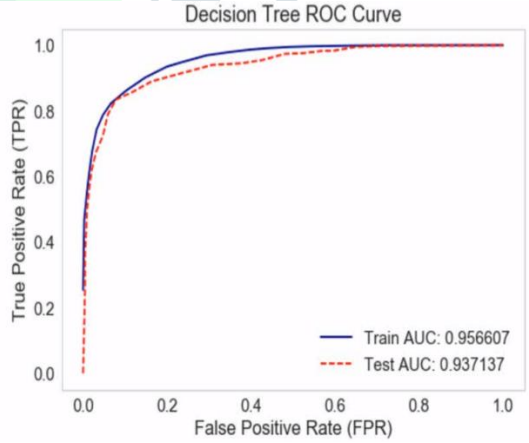
As we understood Logistic Regression divides the dataset or the feature space based on linear boundary separation. But Gradient Boosting differs as it solves problems other than stochastic gradient descent. It focuses on making a decision tree and uses a process called gradient tree boosting which tends to fit training set to the decision tree recursively.

a) Multi-layer perceptron classifier

Multi-layer perceptron classifiers are basically used in deep learning. It can also be called as "Hello World" of deep learning. Multi-layer, as the name signifies, has more than one perceptron it. A single perceptron has an input layer to receive the signal, an output layer which will make decisions based on the input fed to it. Between Input and Output there are an arbitrary number of hidden layers which are basically the real computational source of MLP. MLP are often used with supervised learning problems.

6. CONCLUSION

In conclusion, after comparing the performance of three different models (logistic regression, neural network, and gradient-boosted) on the problem of classifying a given Twitter user as "bot" or "human", Decision Tree classifier provides the highest accuracy(95%) and provides confident predictions. High precision denotes low false-positives (the probability of a user being identified as a bot when it is a non-bot is low). Thus, the models efficiently identify a non-bot account. But we get low recall score for other classifiers



Looking ahead in the future, we would like to collect our own dataset for further experimentation. Gathering more training examples could improve our distribution mismatch by making the Training distribution more like the real-world; It would also give us more control over how the "bot"/"human" labels were generated. In addition, we could create an English-only tweet corpus, which would enable the use of text-based features (E.g. tweet sentiment as in [5], topic extraction, words used). We would also be able to include friendship and

follower relationships between users in the dataset, to exploit bot community features as explored in [7]. Finally, in the future, we would like to turn our application prototype into a production-quality web plugin that could provide a real-time classification of accounts to Twitter users

.

7. REFERENCES

[1] R. Gorwa, "Twitter has a serious bot problem, and wikipedia
might have the solution," Quartz Media, October 23 2017.
[Online]. Available: https://qz.com/1108092/twitter-has-a-serious-botproblem-
and-wikipedia-might-have-the-solution/
[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on
social media: A data mining perspective," CoRR, vol. abs/1708.01967,
2017. [Online]. Available: http://arxiv.org/abs/1708.01967
[3] C. Yang, R. C. Harkreader, and G. Gu, Die Free or Live Hard?
Empirical Evaluation and New Design for Fighting Evolving Twitter
Spammers. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011,
pp. 318–337. [Online]. Available: https://doi.org/10.1007/978-3-642-
23644-01 7
[4] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and
M. Tesconi, "Fame for sale: efficient detection of fake twitter
followers," CoRR, vol. abs/1509.04098, 2015. [Online]. Available:
http://arxiv.org/abs/1509.04098
[5] ——, "The paradigm-shift of social spambots: Evidence, theories, and
tools for the arms race," CoRR, vol. abs/1701.03017, 2017. [Online].
Available: http://arxiv.org/abs/1701.03017
[6] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation,
and characterization," CoRR, vol. abs/1703.03107, 2017. [Online].
Available: http://arxiv.org/abs/1703.03107
[7] E. Tacchini, G. Ballarin, M. L. D. Vedova, S. Moret, and L. de Alfaro,
"Some like it hoax: Automated fake news detection in social
networks," CoRR, vol. abs/1704.07506, 2017. [Online]. Available:
http://arxiv.org/abs/1704.07506