Using Natural Language Processing and Random Forests to Detect Twitter Bots

David Paul-Pena
ECE, Tandon-NYU
Brooklyn, New York
DPaul@nyu.edu

Michael Shamouilian
ECE, Tandon-NYU
Brooklyn, New York
Mike.Sha@nyu.edu

***Abstract*—This paper serves as a project proposal to describe our proposed method for detecting bot accounts on Twitter. The proposed methods to use are Random Forests and Ensemble SVM for classification and word2vec and Twitter API for feature extraction. The 2 methods will be compared and analyzed based on results from implementing and testing both architectures.**

***Index Terms*—Machine Learning, Natural Language Processing, Twitter**

I. Introduction

With millions of twitter accounts and billions of tweets a year, Twitter has gained great popularity among social media savvy individuals, companies and robots. A robot user is a user controlled by a bot which is progremmed to execute commands in an automated fashion. Each user is designated by a user name with a @ pre-fix. All three types of users exploit the basic functionalities of Twitter to connect and communicate with others.

These basic functionalities include Tweet, posting a message, Reply, answering a tweet, Retweet, sharing another users tweet, Like, designating the user likes a tweet, Hashtag, method of organizing and collecting based on a phrase or word, and Mention, including the username of another user in a tweet. For more information on the functionalities and details of how to execute such actions see [1].

Many tutorials exist on the Internet boasting the simplicity of how to create a robot account on twitter. These bots can be simple such as retweeting certain tweets, sending reminders to a user, tweeting random facts or words or jokes of the day. However, not all bots are benign, some are malicious. The next section will describe the types of malicious bots and their effects.

This paper is concerned with testing and comparing different machine learning algorithms for detecting twitter bots. The next section will describe the motivation for detecting bot accounts. The related works section describes previous research efforts in creating twitter bot detectors. The data section will describe the data set collected and the method behind collecting the dataset. The algorithms used section will provide a brief overview of the machine learning algorithms used and the code will be provided in the code section. The results section will have the results generated from Random Forests and EnsembleSVM, and they will be analyzed in the evaluation section. The video link section will provide a link to a video representation of this paper on YouTube. Finally the conclusion section we describe the next steps and improvements that can be made.

II. Motivation

While some twitter bots are harmless, and may in fact be helpful, such as having the bot tweet to a user to take your medication on time, many can be harmful. The 5 major harmful types of twitter bots are SPAM and SPIM bots which flood users with tweets to click on a harmful link. Zombie bots which are compromised accounts aggregated to launch harmful attacks in twitter. Malicious File-Sharing bots answer peoples tweets with links to files wanted but with malicious code injected into the files. Malicious Chatterbots cover as a dating service to talk with users in an attempt to discover personal facts and financial information from the user. And Fraud bots is the broadest category covering most simple bots that are meant for financial gain, such as bots which generate false clicks for ads. For more information on these types of bots see [2].

While it should be noted that all of these attacks can be perpetrated by a single human user, bots can do it on very large scales with a little human effort, greatly magnifying the negative effects. All of these bots fill the twitter space with harmful tweets that ruin the user experience for others and brings a bad name to twitter. Having detectors that can find these bots, the bots can then be reviewed and disabled to restore order to the twitter space. Therefore the main motivation in detecting twitter bots, and therefore this paper, is to help reduce the number of victims of these aforementioned malicious activities.

III. Related Work

Previous attempts have been made to classify Twitter accounts as bots or not. NLP techniques were using to extract tweet sentiments, such as contradiction rank, positive sentiment strength, and average topic sentiment, together with quantitative metrics such as tweet frequency were used in [3] to determine if an account is bot or not using different classifiers (Naive Bayes, SVM, Random Forest, boosting...) getting an Area Under the ROC curve (AUROC) of 0.73. While classification was not the main focus of the paper, they were able to improve the detection compared to previous work.

Some previous attempts were made in [4] analyzing the account relationship with other users, together with three content features extracted from the most recent 20 tweets to classify the account. Neural Networks, decision trees, and Bayesian classifiers were compared and the overall accuracy was 91.7% achieved with the Bayesian classifier.

In [5], Chu et.al. uses Random Forest for the classification of Twitter accounts in bot, human or cyborg, a mix of bot and human. They use three distinct sets of features for the classification. Entropy component based on the timing between tweets, where fixed time intervals was found to be indicative of a bot account, Spam Detection component, where the account's tweets are examined for spam content based on a modification on Bayesian classification using CRM114 (check paper for details), and an Accounts Property component where all of the standard features from the twitter API are used, i.e. account registration date, as well as links within tweets are check for the validation of the links, i.e. not a link to spam. 500,000 user accounts were collected. For ground truth classification, some accounts were classified by hand and others by having a Turing Tester converse with an account for 5 minutes. On average the paper reached 96% accuracy in detecting the 3 classes.

More recently, [6] creates a new classification in terms of types of Twitter accounts, classifying the bot accounts into Consumption, Broadcast and Spam. Using numerical, categorical and series features to train Naive Bayes, Random Forest, SVM and Logistic Regression to achieve an accuracy of 84.32% on overall bot accounts using Random Forest.

IV. Data

The Dataset is comprised of 50 Twitter bot accounts and 50 legitimate Twitter human accounts. A bot account can be: 1) benign, automated tweets informing human accounts of different news, events, daily quotes, or any interesting information; or 2) a malicious bout account, that will only tweet SPAM links, or fake prizes. The selected bots accounts comprises both type of accounts.

To obtained the Dataset, a Python script has been used that first authenticate with the Python REST API using a registered Application Secret and Client Keys, and a second query that retrieves all the information of the 100 accounts in a single request. The JSON object is then parsed and converted into a CSV file.

The contents of the Dataset are the followings: 1) Twitter user ID as an integer, 2) Twitter user ID in a string format, 3) user name, 4) user location, 5) user filled description, 6) user profile URL, 7) number of followers, 8) number of friends (accounts that is following), 9) number of times listed, 10) creation date of the account, 11) number of favorite tweets, 12) a boolean indicating if it is a verified account, 13) number of tweets in the account, 14) language of the account, 15) last tweet sent, 16) a boolean indicating if the user is using the default profile, 17) a boolean indicating if the user is using the default profile image, 18) a boolean indicating if the user has an extended profile, 19) user nickname; and 20) a boolean indicating if the account is a bot or not.

To improve the accuracy of the classifier using Natural Language Processing Algorithms, the possibility of adding additional features, such as frequency of tweeting (extracted from the tweet history), and a larger collection of tweets from the accounts will be analyzed.

V. Algorithms Used

Additionally we are currently thinking about using Natural Language Processing (NLP) to help extract features from a specific twitter account. While twitter limits tweets to 140 characters, the structure of the tweet and the frequency with which certain key words appear in their tweets can provide good features.For example if the bot is a repetitive bot then its tweets will always have the same structure. For example a word of the day bot will have a word followed by a short definition, probably no hashtags or mentions. Where as people are more random and so their tweets will have a more random structure. Certain buzzwords like money, account, etc. may show up more frequently in bot accounts. Albeit people tend to repeat certain phrases, though these phrases should be different from the buzz words. One such proposed method is word2vec, which refers to a set of NLP algorithms which use deep learning which encode strings of words into lower sized vectors of features. This way the set of millions of tweets can be represented by only 10 to 20 feature points. Something similar [7] and using the tutorial provided by [8].

Once the feature are collected and processed, 2 different machine learning (ML) architecture are proposed to be used. The first is Random Forests, which use an ensemble of decision trees to classify the accounts as bot or not. The training data set is sub-sampled and fed to different decision trees with varying patterns and then the results are collated to predict the class. Decision Trees branch data based on the information gain and then come up with partitioning rules based on the feature values to classify the samples. More details are provided in the class notes on these. Instead of using just a single support vector machine (svm), EnsembleSVM can be used. It is like Random Forests for svm's instead of decision trees. Multiple instances of SVM's are trained with sub-parts of the training set and then with different parameters the classifications are collated. How an svm works can be seen in the class notes. Basically a hyper plane is created to segment the data features into the separate classes. For the specifics of EnsembleSVM look at [9].

References

[1] "Intro to twitter for business." [Online]. Available: https://business.twitter.com/en/basics/intro-twitter-for-business.html

[2] A. O'Donnell, "5 types of malicious bots and how to avoid them," Sep 2016. [Online]. Available: https://www.lifewire.com/what-are-malicious-bots-2487156

[3] K. V. . S. V. S. Dickerson, J. P., "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?" *Advances in Social Networks Analysis and Mining (ASONAM)*.

[4] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," 2010.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012.
[6] P. K. P. Richard J. Oentaryo, Arinto Murdopo and E.-P. Lim, "On profiling bots in social media."
[7] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
[8] Nadbor, "Ds lore." [Online]. Available: http://nadbordrozd.github.io/blog/2016/05/20/text-classification-with-word2vec/
[9] M. Claesen, F. De Smet, J. A. Suykens, and B. De Moor, "Ensemblesvm: a library for ensemble learning using support vector machines." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 141–145, 2014.