Bot-hunter: A Tiered Approach to Detecting & Characterizing Automated Activity on Twitter

David M. Beskow and Kathleen M. Carley

School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
Email: dbeskow@andrew.cmu.edu and kathleen.carley@cs.cmu.edu

Abstract. As malicious automated agents, or bots, are increasingly used to manipulate the global marketplace of information and beliefs, their detection, characterization, and at times neutralization is an important aspect of a national security operations. Unhindered, these information campaigns, assisted by automated agents, can begin slowly changing a society and its norms. Within this context, we seek to lay the groundwork for *bot-hunter*, a Tiered Approach to bot detection and characterization, while simultaneously presenting an event based method for annotating data.

1 Introduction

Although social media bots can create positive effects, a subset of malicious bots have recently gained widespread notoriety for their intervention and manipulation of the marketplace of information, ideas, and beliefs [13,7,15,17]. This subset of malicous bots are involved in propaganda [16], suppression of dissent [21], and network infiltration/manipulation [11,5]. While their communication is often less sophisticated and nuanced than human dialogue, their advantage is the ability to conduct timely informational transactions effortlessly at the speed of algorithms. This advantage has led to a variety of creative autonomous agents deployed for beneficial as well as harmful effects. While their purpose, characteristics, and "puppet masters" vary widely, they are undeniably present and active. Their effect, while difficult if not impossible to measure, is tangible.

Detecting/neutralizing these malicious automated agents is a part of an emerging area of research that has recently been called Social Cyber Security [8]. According to Carley et al [8], the focus of this emerging discipline is *...to characterize, understand, and forecast cyber-mediated changes in human behavior, social, cultural and political outcomes, and to build the cyber-infrastructure needed for society to persist in its essential character in a cyber-mediated information environment under changing conditions, actual or imminent social cyber-threats.* Under this umbrella of Social Cyber Security, our research hopes to weld together and expand several existing bot detection/characterization methods, specifically focusing on the Twitter Micro-blogging platform.

This paper lays the foundation for a tiered supervised machine learning approach to bot detection and characterization. Additionally, it highlights the novel use of event oriented bot annotation. We believe that, while leveraging traditional machine learning models, our efforts to build training data are as important if not more important than our efforts to improve models and feature space.

Our research has identified several tiers of Twitter data collection and related machine learning features and models. The constraints of data availability and rate limiting associated with the Twitter API [2] create these Tiers, which are summarized in Table 1. These tiers are cumulative (i.e. Tier 2 includes features from Tier 0 and 1). There is a trade off between richness of data and the computation time to extract, build features, and classify accounts. We recognized the need to have a multi-tiered approach that provides capability at each of these tiers. The model selected will depend largely on the data available as well as whether the specific use case requires high accuracy or high volume. A *Tier 1* model is designed for characterizing bot activity in large data streams, while a *Tier 3* model would be applied to problems that require high accuracy on a limited number of active accounts (i.e. not suspended).

Table 1: Four *tiers* of Twitter data collection to support account classification

Tier	Description	Focus	Collect/process Time per 250 Accounts	# of Data Entities (i.e. tweets)
Tier 0	Tweet text only	Semantics	N/A**	1
Tier 1	Account + 1 Tweet	Account Meta-data	∼ 1.9 seconds	2
Tier 2	Account + Timeline	Temporal patterns	∼ 3.7 minutes	200+
Tier 3	Account + Timeline + Friends Timeline	Network patterns	∼ 20 hours	50,000+

While numerous research efforts have attempted to exploit pieces and parts of this data spectrum, few have attempted to create a comprehensive approach that covers all tiers. The closest effort that we've seen is the Botometer effort discussed later in this paper. While offering an robust model through an accessible API, it is only offered at Tier 2, meaning high volume classification is computationally expensive. Additionally, if does not exploit the rich network features available at Tier 3. This paper seeks to lay the groundwork for this comprehensive Tiered Approach, discuss event focused data annotation, as well as build and evaluate a Tier 1 model. Future research for the *bot-hunter* suite of models will specifically focus on the elusive third Tier.

2 Data

In this section we present an event-oriented approach to data annotation. Rather than use *honey pots* [14] or *suspended accounts* [4] to annotate bot accounts, as past efforts have done, our effort focused on forensic analysis and data collection related to reported bot events. Given certain bot intimidation attacks, if the collection is performed properly, it becomes easy to label certain accounts as likely automated.

We decided to focus on a known and publicized bot attack against the Atlantic Council Digital Forensic Labs (DFR Lab), and tangentially against the NATO Public Affairs Office [18]. This attack primarily occurred between 28 August and 30 August 2017. On 28 August the Daily Beast posted an article about an alleged intimidation attack against the DFR Lab Bot Research Team. When DFR Labs and NATO Press shared this news article on Twitter, they were immediately harassed by thousands of bots. This bot attack was also accompanied by targeted intimidation against DFR Lab employees both on and off line [18].

Using the unique phrases and hashtags that were amplified in this event, we were conducted a targeted query on the Twitter REST API to collect the data for this event. Both the DFR Lab Twitter Account (@DFRLab) and the NATO Public Affairs Account (@NATOPress) respectively average 130 and 43 interactions (retweet, follow, like, etc.) on their accounts on a daily basis. At the height of the bot attack they were averaging 6,000 interactions per hour. As illustrated in Figure 1, we were able to label 99% of this activity as bot activity, which we confirmed manually with random sampling. This method provided us with approximately 19,221 accounts that we could easily label as a bot. In April 2018 (8 months after the incident) we checked on these accounts, and found that of the 19,221 original accounts, 18,360 had been suspended, and 12 had been deleted by the user. This means that 95.5% of those accounts are now suspended, validating our targeted collection strategy.

We found that a number of accounts used a randomly generated alphanumeric 15 character strings for the screen name such as **Wy3wU4HegLlvHgC** (not an real account). Our team has observed this in other bot attacks [15], and separately used this phenomenon to annotate a large bot training set [6]. Additionally, we found that 60% of our bot accounts had a profile image. Conducting a reverse Google lookup of the image, we found that many Twitter Accounts used the same profile image. This seems to provide evidence that these accounts are mass produced with the same stock photo or hijacked profile image. We found that the mean age of *bot* accounts (3.9 years) was statistically greater than the mean age of *human* accounts (3.36 years) with $p.value = 2.2e^{-16}$.

In exploring the account time-line (account history), we found that 43.5% of their statuses were reteets. Content involved a wide variety of topics (sports, news, advertisement, explicit content, etc) and was from a variety of languages. The random nature of this content leads us to hypothesize that the actors behind this attack are either using or imitating “bot-for-hire” accounts which post content to support a variety of actors that hire them. These observations reflect a change in bot tactics from what researchers have observed in the past. These

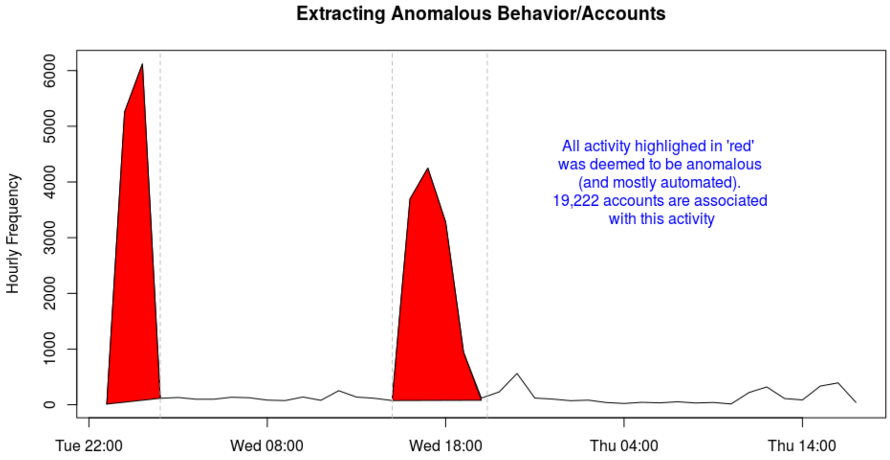

Fig. 1: Separating bot event anomalous behavior from normal daily behavior

same observations and hypothesis have also been made by leaders at DFR Labs [18]. Many of the accounts demonstrated unique temporal patterns strongly associated with automated accounts. These accounts would launch a "volley" of 24-26 tweets (all re-tweets) once every 24-26 hours.

In order to train a model, we also needed accounts that we could tag as "human", and not automated. We used the Twitter Streaming API to collect a sample of *normal* Twitter data, intentionally collecting both weekend and weekday data. From this data we randomly selected 70,000 accounts to tag as *human* Twitter accounts that we can sample from for training. Past research has estimated that 5-8% of twitter accounts are automated [20]. If this is true, then we mis-labeled a small amount of our accounts as *human*. We assessed that this was an acceptable amount of noise in the data, but that it will undoubtedly limit the performance of supervised learning models that train and test on the data.

2.1 Feature Engineering

Our feature engineering started with one driving constraint; we wanted to limit features to those available in *Tier 1*, namely the basic Twitter JSON. Most researcher who collect Twitter data, whether from the REST API or the Streaming API, will collect basic Twitter JSON data. This JSON represents each individual tweet (or status), meta-data associated with the Tweet, and the user object and user profile data [1]. This JSON data is relatively easy to collect, and both the Streaming and REST API's can provide $\sim 8K$ tweets per minute. Given this driving constraint, we developed the features described in Table 2.

Table 2: Features by Tier 1 Model

Source	User Attributes	Network Attributes	Content	Timing
Tier 1	screen name length	# of friends	Is last status retweet?	account age
	default profile image?	# of followers	same language?	avg tweets per day
	screen name entropy	# of favorites	hashtags last status	
	has location?		mentions last status	
	total tweets		last status sensitive?	
	source (binned)		*bot* reference?	

Our team has not found any other team that has used *screen name entropy*, *same language?*, and *bot reference* for a Tier 1 model. The *screen name entropy* feature leverages Shannon string entropy of the user screen name. Shannon entropy is defined in 1, where p_i is the normalized count for each character found in the string.

$$H\left(A\right) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{1}$$

The *same language* feature represents whether or not the user object and the status object have the same language. The *bot reference* feature represents whether or not the account describes itself as a bot ('bot' in name/description).

3 Models and Performance

We evaluated several traditional supervised machine learning models on the feature space described above. We chose to evaluate Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Trees, and Random Forest models. All of these models have been used in previous bot research attempts.

The baseline model performance for all models is provided in Table 3. Given that the Random Forest model performed best (as found in other similar research [20]), we achieved $AUC = 0.994$ with tuning. Random Forest will support Tiers 1-3 of the *bot-hunter* framework. SVM is used for the Tier 0 model as described in [6].

Table 3: Baseline Model Performance

Model	Accuracy	AUC	Precision	Recall	Kappa
Nave Bayes	0.611	0.885	0.563	0.989	0.220
SVM	0.660	0.921	1.000	0.319	0.319
Logisitic Regression	0.907	0.950	0.859	0.973	0.816
J48 Decision Tree	0.963	0.963	0.960	0.966	0.925
Random Forest	0.981	0.994	0.996	0.966	0.962

4 Related Work

Many efforts have attempted to classify bots, leveraging supervised and unsupervised machine learning as well as crowd-sourcing, community detection, and correlated accounts [3]. The first deliberate detection of automated accounts on the Twitter Platform began in 2010 when [9] conducted three-class classification (human, bot, cyborg) using an ensemble model. Other early works investigated automated accounts from the perspective of spam and spam prevention [23,12,19]. In 2011, a team from Texas A&M became the first team to leverage *honey pots* to annotate bots [14] . These *honey pots* used bots that generate nonsensical content, designed only to attract other bots.

In 2014, Indiana University and the University of Southern California launched the *Bot or Not* online API service [10] (later rebranded as *Botometer*). This research used supervised machine learning algorithms on $1,150+$ features extracted from the user and time-line objects (Tier 2 model in our framework) trained on the Texas A&M dataset to help users evaluate whether or not an account is a bot [20]. The Botometer model is the primary tool leveraged for applied bot research today [22].

While these and other research efforts have created adequate models with specific subsets of the data, we have not found research at Tier 3, or research that combines these models into a comprehensive suite of tools.

5 Conclusions and Future Work

This paper lays the groundwork for a Tiered approach to bot detection while simultaneously presenting an event-based approach to bot data annotation. This Tiered approach with representative training data will allow researchers and organizations/agencies to choose from a selection of models based on their given needs and data. The performance of our Tier 1 model is comparable to the performance of other similar models (namely the *Botometer* algorithm), and is adequate as a baseline model. This specific model was developed for high volume, and is very helpful when trying to measure overall bot penetration in a large twitter collection/stream. Additionally, *bot-hunter* can run on existing data rather than requiring the Botometer API to recollect data that the researchers may already have.

Future research will build out the feature space and models for Tiers 2 and 3, with a focus on Tier 3 since network and content features from robust snowball sampling has not been previously explored. Additionally, we look to fuse several event based data sets with the diverse data we've collected through random string classification [6], as well as historical data such as the original Texas A&M data in order to build a robust and diverse training data set. Initial testing with various data sets shows evidence that data selection may be as important if not more important than feature engineering and model selection.

ACKNOWLEDGMENT

This work was supported in part by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative Award N000141812108, Office of Naval Research Minerva Awards N00014-13-1-0835/N00014-16-1-2324, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

References

1. Tweet object. `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`. Accessed: 2018-05-02.
2. Twitter rate limiting. `https://developer.twitter.com/en/docs/basics/rate-limiting`. Accessed: 2018-05-02.
3. Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. Malicious accounts: dark of the social networks. *Journal of Network and Computer Applications*, 79:41–67, 2017.
4. Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 15(5):475–491, 2016.
5. Matthew Benigni and Kathleen M Carley. From tweets to intelligence: Understanding the islamic jihad supporting community on twitter. In *Social, Cultural, and Behavioral Modeling: 9th International Conference, SBP-BRiMS 2016, Washington, DC, USA, June 28-July 1, 2016, Proceedings 9*, pages 346–355. Springer, 2016.
6. David Beskow and Kathleen M Carley. Using random string classification to filter and annotate automated accounts. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018.
7. Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. 2016.
8. Kathleen M Carley, Guido Cervone, Nitin Agarwal, and Huan Liu. Social cybersecurity. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018.
9. Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.
10. Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

11. Carlos Freitas, Fabricio Benevenuto, Saptarshi Ghosh, and Adriano Veloso. Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 25–32. ACM, 2015.
12. Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.
13. Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. *Browser Download This Paper*, 2016.
14. Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*, 2011.
15. Al Bawaba The Loop. Thousands of twitter bots are attempting to silence reporting on yemen. 2017.
16. Cristian Lumezanu, Nick Feamster, and Hans Klein. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
17. LM Neudert, B Kollanyi, and PN Howard. Junk news and bots during the german federal presidency election: What were german voters sharing over twitter?, 2017.
18. Benjamin Nimmo. #botspot: The intimidators, August 2017. [Online; posted 30 August 2017].
19. Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011.
20. Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017.
21. John-Paul Verkamp and Minaxi Gupta. Five incidents, one theme: Twitter spam as a weapon to drown voices of protest. In *FOCI*, 2013.
22. Stefan Wojcik, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. Bots in the twittersphere, Apr 2018.
23. Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2009.