On the Stability of BotHunter Scores

Lynnette Hui Xian Ng[1*], Dawn C. Robertson[1], Kathleen Carley[1]

1 CASOS Center, Institute for Software Research, Carnegie Mellon University,
Pittsburgh, Pennsylvania, United States
{huixiann,drobert,carley}@andrew.cmu.edu

Abstract

To combat the spread of harmful information through the use of bots, a series of social media bot detection algorithms have been developed. The stability of the scores and classification produced from these bot detection algorithm is of importance to avoid false positives and misclassification. In this study, we established the stability of the BotHunter algorithm developed by Carnegie Mellon University. We recommend using the BotHunter bot probability scores with the following parameters: (a) a threshold value of 0.70 for bot classification; (b) at least 20-50 tweets per agent.

Introduction

To combat the spread of harmful information through the use of bots, a series of social media bot detection algorithms have been developed [1]. These algorithms range from using text modelling methods [2–4] to account features like screen names or network information [5–7] to tweet meta-features like posting behavior [8–10].

Two commonly used detection platforms are Botometer and BotHunter. Botometer uses a supervised ensemble classification based on 1150 features extracted for each Twitter agent [11, 12]. The BotHunter algorithm is developed by Carnegie Mellon University [13]. It classifies agents using a supervised random forest method with a multi-tiered approach, each approach making use of more features as before, from content to user to network features.

In this study, we access the stability of the BotHunter algorithm and characterize its predictions on suspended accounts. We analyze 5000 Twitter *agents* and their BotHunter scores, across a 150 day timeframe. We hope to answer the following research questions:

RQ1: What is a good threshold for a consistent BotHunter score?

RQ2: What is the minimum number of tweets for a consistent BotHunter score?

Data and Methods

We collected tweets of 5000 Twitter agents on a daily basis for 150 days in September 2020. We also collect the agents' profile information, such as number of followers and locations. To detect and label bots , this study relied on the BotHunter platform developed by Carnegie Mellon University.

We analyze the volatility of the BotHunter bot probability score value through: volume volatility, which investigates how the scores change across the number of tweets;

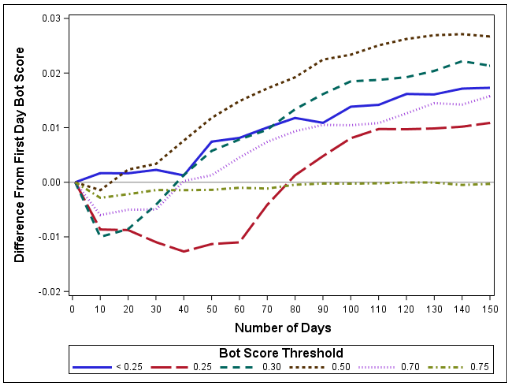

Fig 1. Evolution of bot scores from first day's bot scores across time

and threshold volatility, which investigates how different threshold values applied to the probability scores affect the BotHunter classification.

To investigate volume volatility, we processed the BotHunter algorithm in increasing number of tweets. We analyzed the bot-probability scores of each agent, from one tweet till the total number of tweets we collected from the agent.

To investigate threshold volatility, we specify 5 threshold values: [0.25, 0.30, 0.50, 0.70, 0.75] and looked at bot percentages across our analysis timeframe. At each threshold value, a BotHunter bot probability score larger than the threshold is classified as a bot, and a BotHunter score lower than the threshold is classified as a non-bot.

Results and Discussion

Stability of bot scores. In general, the mean difference in the initial bot score derived from one day's worth of tweets and subsequent bot scores across time do not fluctuate significantly. Figure 1 show the trend of bot scores across time as compared to the first day's scores. Even though the difference in bot scores start to increase across time, the difference is very small, usually not significant enough for a change in bot classification. These results point to the stability of the BotHunter algorithm over time and number of tweets.

Threshold-Based Analysis. Selecting a threshold value is crucial to any study to balance the number of false positives. As thresholds tighten, lesser percentage of agents are initially classified as bots. At the 0.70 threshold, a typical threshold value used for studies with BotHunter, almost half the agents were initially classified as bots, alluding to the importance of a stable bot detection algorithm. Previous studies use a threshold value of 0.60-0.70 [14–16], and our observations support it as a stable threshold.

Table 1 (in Appendix) presents the proportion of agents that fall within a threshold band on its first day's bot score against its final bot score. Based on the data, we recommend a 0.70 threshold value for BotHunter classification as it differentiates agents best.

Comparing the difference of bot scores of an agent's first tweet against an increasing number of tweets show that the mean difference increases as volume increases. As visualized in Figure 2, for a good bot score, an average of 20 tweets should be used. This is where the graph has the steepest increase. For an absolute stable score, the bot scores plateau around 400 tweets, though the difference is a hundredth of a decimal point which is unlikely to affect bot classification.

In this study, we have tested the BotHunter algorithm on a sampled Twitter dataset and established the stability of the BotHunter classification. However, due to the

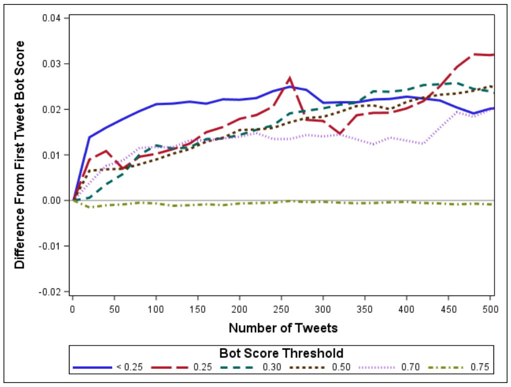

Fig 2. Evolution of bot scores from first tweet's bot scores

random sampling of Twitter agents, data biasness may incur.

Conclusion

In this study, we study the stability of a used bot-detection platform, BotHunter. Bot classification is bound by threshold values, which we investigate 5 key values. Our results show that: For effective bot classification using the BotHunter algorithm, we recommend having at least 20-50 tweets for the agent in study. We recommend a 0.70 bot classification threshold.

Acknowledgments

The research for this paper was supported in part by the Knight Foundation and the Office of Naval Research grant N000141812106 and by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the US Government.

Threshold/ **Final bot score**	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
≤0.25	1.32	1.08	1.30	1.00	0.90	0.18	0.18	0.14	0.10	0.04	0.06			0.02						
0.25		0.04	0.04	0.06	0.28	0.56	0.10	0.04	0.02	0.02	0.06	0.04								
0.30	0.06	0.06	0.20	0.08	0.08	0.46	1.18	2.06	2.02	2.00	0.92	0.60	0.54	0.22	0.12	0.10	0.02	0.02		
0.50	0.02	0.04	0.02		0.04	0.02	0.02	0.10	0.30	0.70	2.54	2.88	4.04	4.30	2.36	1.20	1.68	0.26	0.08	
0.70									0.06	0.04	0.04	0.18	0.22	0.68	3.12	1.18	0.98	0.30	0.08	
0.75			0.02				0.02		0.02	0.08	0.12	0.18	0.46	1.82	5.64	9.96	13.86	15.14	7.88	

Table 1. Proportion of agents that fall within a threshold band on its first day bot score vs its final bot score. We recommend the threshold value of 0.70 for BotHunter classification as it differentiates agents best.

References

1. Orabi M, Mouheb D, Al Aghbari Z, Kamel I. Detection of Bots in Social Media: A Systematic Review. Information Processing Management. 2020;57(4):102250. doi:https://doi.org/10.1016/j.ipm.2020.102250.

2. Wei F, Nguyen UT. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In: 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). IEEE; 2019. p. 101–109.

3. Kudugunta S, Ferrara E. Deep neural networks for bot detection. Information Sciences. 2018;467:312–322.

4. Kosmajac D, Keselj V. Twitter Bot Detection using Diversity Measures. In: Proceedings of the 3rd International Conference on Natural Language and Speech Processing. Trento, Italy: Association for Computational Linguistics; 2019. p. 1–8. Available from: `https://www.aclweb.org/anthology/W19-7401`.

5. Beskow DM, Carley KM. Its all in a name: detecting and labeling bots by their name. Computational and Mathematical Organization Theory. 2019;25(1):24–35.

6. Beskow DM, Carley KM. You are known by your friends: Leveraging network metrics for bot detection in twitter. In: Open Source Intelligence and Cyber Crime. Springer; 2020. p. 53–88.

7. Minnich A, Chavoshi N, Koutra D, Mueen A. BotWalk: Efficient adaptive exploration of Twitter bot networks. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017; 2017. p. 467–474.

8. Chavoshi N, Hamooni H, Mueen A. DeBot: Twitter Bot Detection via Warped Correlation. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). Los Alamitos, CA, USA: IEEE Computer Society; 2016. p. 817–822. Available from: `https://doi.ieeecomputersociety.org/10.1109/ICDM.2016.0096`.

9. Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In: Proceedings of the 10th ACM Conference on Web Science; 2019. p. 183–192.

10. Chu Z, Gianvecchio S, Wang H, Jajodia S. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? IEEE Transactions on Dependable and Secure Computing. 2012;9(6):811–824. doi:10.1109/TDSC.2012.75.

11. Sayyadiharikandeh M, Varol O, Yang KC, Flammini A, Menczer F. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In: Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management. CIKM '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 2725–2732. Available from: `https://doi.org/10.1145/3340531.3412698`.

12. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: Detection, estimation, and characterization. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11; 2017.

13. Beskow DM, Carley KM. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In: Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. vol. 3; 2018. p. 3.

14. Beskow DM, Carley KM. Agent Based Simulation of Bot Disinformation Maneuvers in Twitter. In: Proceedings of the Winter Simulation Conference. WSC '19. IEEE Press; 2019. p. 750–761.

15. Uyheng J, Magelinski T, Villa-Cox R, Sowa C, Carley KM. Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. Computational and Mathematical Organization Theory. 2020;26(4):465–483.

16. Joshua Uyheng LHXN, Carley K. Bot Activity in the 2020 Singaporean Elections: A Social Cybersecurity Analysis; 2020. Available from: http://sbp-brims.org/2020/proceedings/papers/working-papers/SBP-BRiMS_2020_paper_84.pdf.