# Automatic Detection of Bots – Comparative Review

**Nirdhum Narayan**

*School of Computing Science and Engineering, Galgotias University, Gautam Budha Nagar, Uttar Pradesh, 226001, India*

---

**Abstract-** *Automatic detection of bots is critical to every internet user. This is due to the increasing advancement in knowledge and approaches by malicious bots to attack over the internet. With the increase in the three V's of Big Data i.e. Volume, Velocity and Variety of data generated by user on internet or social networks, many new ways have been designed to collect and analyze the data. Many software are designed to perform automatic analyzing of data and provide the users with services of better quality. However, at the same time a lot of spiteful software have also been used to disperse or circulate false figures or statistics or information, which can lead to real-world ramifications. According to many surveys many small or big organizations are constantly under security threats, which not only costs millions or billions of dollars in damage and recovery, it often also severely affects their reputation. At present a large amount false information is being circulated online and on the social media causing real time consequences in many places. This paper will provide a definition of the bots and background information to determine why it is reliable to study the automatic bot detection approaches. It will further dwell on the methods used by different automatic detection systems to identify bots and segregate the good bots from the bad bots.*

***Key Words***: **Security management, Bot detection, Patterns, Malicious Bots, Social bots, Machine Learning.**

## 1.INTRODUCTION

The increasing advancement in technology has led to risk factors while using the internet and other technology devices. This has led to the creation of internet bots, which are software that runs automatically in the internet. Bots can either be positive or negative, depending with the intentions of the programmers who create them[14]. Majority of the internet bots are harmful to the computers and other variable uses. Some internet bots are helpful, for instance the google bots that help in the indexing of searchers over the internet. Since it is hard for a new user to identify how a good bot and bad bot appear in the internet and their technological devices, this study will demonstrate the algorithms of automatic detection[17].

While identifying the risks of the bots might be easy, it is also reliable to identify the convenient approaches for a person to detect the presence of bots in their devices. Some of the common approaches to knowing the presence of the bots include static approaches on the web where requests are made by the bad bots, challenge-based approach to detect the bots and the human users through the challenges presented to them, and behavior approaches to detect the behavior of the users and compare to the bots[19]. The different methods might not all be effective as machine learning approaches have been introduced to automatically detect bots on the internet and technological devices.

Since many people are continually channelling their businesses and activities to the online platforms, it is convenient to inform them of the presence and means of detecting the bots. Therefore, this essay will demonstrate the methodologies of analysing the presence of the bots while using the internet and technological devices.

## 2. CONCEPTS AND TERMINOLOGIES

Machine Learning[13] with network flow feature has also been used in detection of these bots, but flow based approach takes high computation overhead and do not completely capture the network pattern. Machine Learning is the perfect technique to automatically capture the normal behavior. In this process a baseline is first established for normal behavior. Then, the decision engine finds any deviation from normal and alert it as threat. Feature extraction is an important step before training ML and most common features are flow based for example: source IP, destination IP, numbers of packets sent or received etc., but flow based approach do not capture the topological structure of the communicational graph whereas use of Graph based approaches are the best for representation of communication in a network.

### 2.1. Bot

An Internet bot, web robot or simply bot, is a software application that runs tasks that are automated such as 'scripts' over the Internet. Typically, bots perform operations that are simple and repetitive. The most common use of bots is for 'web crawling', in which an automated script fetches, analyze and files information from web servers. According to surveys more than half of all web traffic is generated by bots. "While bots thrive for different sinister purposes, they exhibit a similar behavioral pattern when studied up-close"[1]. There are different kinds of bots present: Social bots, Instant Messaging (IM), Internet Relay Chat (IRC), Commercial bots, Malicious bots etc.

### 2.2. Botnet

A botnet is a number of Internet-connected devices, each of which is running one or more bots and is controlled using command and control software. Malicious Botnets can be used to perform Distributed Denial-of-Service attacks (DDoS), steal data, send spam, and allows the attacker to access the device and its connection.

### 2.3. Machine Learning

Machine Learning (ML)[13] is the study of Computer algorithms improved automatically through learning and experience. These algorithms build a mathematical model based on sample data in order to make decisions. Machine learning approach is divided into three main categories:

- Supervised Learning
- Unsupervised Learning
- Reinforcement learning

### 2.4. Feature Extraction

Feature extraction is a process of reducing the initial set of raw data by dimensionality reduction into more manageable group for processing. A large number of variables requiring a large amount of computing resources to process are characteristic of these large data sets. Feature extraction is the method that selects and combines these variables into features by effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. It can also reduce the amount of redundant data for a given analysis. It is extremely useful when the number of resources for processing are needed to be reduced but without loosing relevant information.

### 2.5. Sentiment Analysis

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and bio-metrics to systematically identify, extract, quantify, and study effective states and subjective information.

## 3. METHODS OF APPROACH

The method of detecting these kind of software also consist of detecting similar patterns over different points of communication, network flow based approach, using sentiment analysis, classification on the basis of contrast patterns, using click-stream sequence and a graph based approach which estimates the degree of distribution.

Widely used methods that can be used in detecting bots are usually Anomaly based. We can distinguish two main families of detection techniques in web bots:
1.) Behavioral detection.
2.) Fingerprinting-based detection.

There are many different approaches to detect a bot, some of those methods consists:
1.) Graph Based Approach
2.) Sentiment Analysis
3.) Clickstream Sequence
4.) Network Flow Summaries
5.) Contrast Pattern-Based Classification

Automatic bot detection methods have been enhanced due to the increased knowledge and skills of developers who produce their bots and match them to look like good bots. Google bots are the most attacked and has exposed several users to security threats while using their internet of devices[23]. Therefore, the automatic detection methods to be discussed in this section will provide analysis of the different approaches to identify the good bots, bad bots, their differences, and to prevent the bad bots from establishing any form of threats on devices.

### 3.1. Using Signals

The use of signals occur through extracting hundreds of information from a single request followed by addition of supplementary information to understand the nature of the bots. While operating a computer, there are different signals that come from websites and in form of ads on the internet[25]. The signal information is collected to identify the proprietor of the information. Some of the data used in this process include the enrichment of the source API. This is due to the fact that advanced bots also use similar browsers to the attacked browsers. This simplifies the process, by detailing the nature of such bots, whether they are google bots, which are good bots, or they have hijacked an IP for the google bots to pose as good bots[25].

The data collected through the requests made from the internet are integrated to come up with the source and effects of the bots to the internet and device of an individual. This detection process is swift and uses the signals to communicate whether the proprietors have malicious intent, and the extent of damage it would do to a device and different pieces of information[25]. Upon detection, the supplementary information is used to prevent any further operation of the bots on the devices and the internet and from sending any further requests. The use of signals to detect bots is normally used by browsers to avoid any damages that might be done to their clients while operating on the internet.

### 3.2. Verified Bot Authentication

The process of creating a bot requires authentication from google. There are a set of rules that must be followed to authenticate a single bot, which also limits the number of servers they can send their requests. Google bots meet the set criteria, however, the growing knowledge of the hackers have enabled them to use other methods to reach out to the consumers of internet services[24]. This is by attaching their bots to different IP addresses and replicating as google bots. A recent data indicated that more than 30% of the bots that claim to be google bots are actually not what they claim[24]. In this sense, it is hard to determine whether they have benefits or are mischievous. This presents the need of verified bot authentication.

A verified bot authentication method identifies the verification requirements for every bot to determine whether google has verified them. The bots that are not verified by google are singled-out and users requested to block or avoid clicking at the websites. Different approaches used by the developers of these bots can be identified through the verification bot authentication which will enhance bot detection on the computer and smart devices[22]. This method is reliable for detection of the advanced bots that appear as good bots but have other malicious intent on the internet and devices of users. The verified bot authentication is offered by different companies that have the intentions of enhancing the online and device security of users to prevent any forms of bots appearing and maliciously causing harm to the information and devices used by different users[22].

### 3.3. Signature-Based Detection

This detection approach is only application for the least advanced bots. It is because of the approaches that it uses to collect data regarding the bots and use them to establish a defense for the system. The approach establishes information that include the specific incoming patterns that are used by websites for chrome and the chrome headers. The patterns might entail the HTTP pattern, fingerprint, and mobile fingerprint which comes through SDK[16]. This approach uses this information by reliably identifying the lack of flow of signatures for the websites and links sent as requests to users, and identify whether they are good bots or bad bots. This approach has enhanced the security of several computer users by enabling them identify the patterns for future use against being exposed to threats by the malicious bots[20]. The google chrome have specific headers, which when not represented by the links sent as request, should be easily detected through this automatic detection approach. Nonetheless, it is hard to use this approach on more advanced bots. The hackers have devised approaches of forging the signatures, so that they are not detectable on the computers, neither can the users identify them easily[18]. This requires the more advanced approach of machine learning approach to automatic identification of bots.

### 3.4. Machine Learning

Machine learning is a branch of artificial intelligence that assist with analysis of data and automates model building without any significant human intervention. This was introduced to assist in the increase of both security and reliance to improve the data collection, analysis, and management without human intervention[15]. In bots detection, machine learning conducts variable tasks to detect bots. The first task is feature extraction by identifying the required features, and those where the bots should be identified. The information is translated into machine language, whereby they can consistently identify the patterns. This moves to the second step of stacked ensemble approach where massive data are segmented through predictions and features such as the headers, IP reputation, and mouse movements[21]. The analysis of the information will determine whether it is a human or bot and illustrate whether it is a good or bad bot to the system. The detection process is long, but it is shortened through the fast assembly and differentiation of data by prediction by the artificial intelligence.

## 4. CONCLUSION

Bots have become a normal occurrence while using the internet. It is only right to gain advantage of the methods of detection and prevention of impacts of the malicious bots. This paper discovers the different automated approaches to bots detection. The information accumulates to four major approaches that include using signals, verified bots authentication, signature based detection, and machine learning. The information in this paper is reliable to improving security and detection against the malicious bots.

## REFERENCES

[1] Abbas Abou Daya, Mohammad A. Salahuddin , Noura Limam and Raouf Boutaba , "BotChase: Graph-Based Bot Detection Using Machine Learning", IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 17, NO. 1, MARCH 2020.

[2] C. Monica, N. Nagarathna, "Detection of Fake Tweets Using Sentiment Analysis", Springer Nature Singapore Pte Ltd 2020, 18 March 2020.

[3] Rashmi Ranjan Rout, Greeshma Lingam, D. V. L. N. Somayajulu, "Detection of Malicious Social Bots Using

Learning Automata With URL Features in Twitter Network", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, April 18, 2020.

[4] PEINING SHI1, ZHIYONG ZHANG AND KIM-KWANG RAYMOND CHOO, "Detecting Malicious Social Bots Based on Clickstream Sequences", VOLUME 7, February 26, 2019.

[5] Huy-Trung Nguyen, Quoc-Dung Ngo, Van-Hoang Le, "A novel graph-based approach for IoT botnet detection", Springer-Verlag GmbH Germany, part of Springer Nature 2019 , 23 October 2019.

[6] Abdurrahman Pektas, Tankut Acarman, "Deep learning to detect botnet via network flow summaries", The Natural Computing Applications Forum 2018, 23 June 2018.

[7] Long Mai, Dong Kun Noh, "Cluster Ensemble with Link-Based Approach for Botnet Detection", Springer Science+Business Media, LLC 2017, 3 October 2017.

[8] OCTAVIO LOYOLA-GONZÁLEZ, RAUL MONROY, JORGE RODRIGUEZ,
ARMANDO LOPEZ-CUEVAS, JAVIER ISRAEL MATA-SÁNCHEZ, "Contrast Pattern-Based Classification for Bot Detection on Twitter", Volume 7, April 16 2019.

[9] ESTÉE VAN DER WALT, JAN ELOFF, "Using Machine Learning to Detect Fake Identities: Bots vs Humans", IEEE Access.

[10] Jing Wang, Ioannis Ch. Paschalidis, "Botnet Detection based on Anomaly and Community Detection ", IEEE Transaction.

[11] AHMED A. AWAD, SAMIR G. SAYED, SAMEH A. SALEM, "Collaborative Framework for Early Detection of RAT-Bots Attacks", IEEE Access, June 13, 2019.

[12] Kamal Alieyan, Ammar ALmomani, Ahmad Manasrah, Mohammed M. Kadhum, "A survey of botnet detection based on DNS", The Natural Computing Applications Forum 2015, 16 November 2015.

[13] R. Boutaba et al., "A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities," J.Internet Services Appl., vol. 9, no. 1, pp. 1–99, 2018.

[14] Balestrucci, A., De Nicola, R., Petrocchi, M., & Trubiani, C. (2019, November). Do you really follow them? automatic detection of credulous Twitter users. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 402-410). Springer, Cham.

[15] Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Automatic Detection of Fake News Spreaders Using BERT. In CLEF.

[16] Cabri, A., Suchacka, G., Rovetta, S., & Masulli, F. (2018, June). Online web bot detection using a sequential classification approach. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 1536-1540). IEEE.

[17] Castillo, S., Allende-Cid, H., Palma, W., Alfaro, R., Ramos, H. S., Gonzalez, C. & Santander, P. (2019, July). Detection of bots and cyborgs in Twitter: A study on the Chilean presidential election in 2017. In International Conference on Human-Computer Interaction (pp. 311-323). Springer, Cham.

[18] Chen, C. L., Ku, C. C., Deng, Y. Y., & Tsaur, W. J. (2018, April). Automatic detection for online games bot with app. In 2018 Third International Conference on Fog and Mobile Edge Computing (FMEC) (pp. 289-294). IEEE.

[19] Davoudi, A., Klein, A. Z., Sarker, A., & Gonzalez-Hernandez, G. (2020). Towards automatic bot detection in Twitter for health-related tasks. AMIA Summits on Translational Science Proceedings, 2020, 136.

[20] Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May). Automatic online fake news detection combining content and social signals. In 2018 22nd conference of open innovations association (FRUCT) (pp. 272-279). IEEE.

[21] Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. PloS one, 15(10), e0241045.

[22] Sayyadiharikandeh, M., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2020, October). Detection of novel social bots by ensembles of specialized classifiers. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2725-2732).

[23] Talukder, S., & Carbunar, B. (2018, June). Abusniff: Automatic detection and defenses against abusive facebook friends. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1).

[24] Tao, J., Xu, J., Gong, L., Li, Y., Fan, C., & Zhao, Z. (2018, July). Nguard: A game bot detection framework for netease mmorpgs. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 811-820).

[25] Valencia, A., Adorno, H. G., Rhodes, C., Pineda, G. F., Cappellato, L., Ferro, N.,... & Müller, H. (2019, September). Bots and gender identification based on stylometry of tweet minimal structure and n-grams model. In CLEF.