Detection of Botnet and DDoS using Network traffic analysis and Machine Learning Algorithms

Rajesh Babu Yallamanda
Rama devi Gunnam
Nissankara Lakshmi Prasanna, (nssankaralakshmi@proton.me)

Research Article

Keywords: Botnet, DDOS attack, Naïve Bayes, SVM, Network Traffic Analysis, Machine Learning

Posted Date: October 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3397184/v1

Abstract

The tremendous growth of the Internet brought many cyber-attacks and cybercrimes. This led to continuous development of new types of cyber-attacks, attacking tools and attacking techniques which allowed the attackers or adversaries to penetrate or get into more compounded or well-managed environments. Botnet is a real threat as it is a group of compromised or hacked Internet-connected devices. Each of those compromised devices or systems are injected with malware and they are controlled from a remote and random location without the prior hint or knowledge of the device's rightful owner and can be used to perform DDOS attack. The model proposed here detects or predicts both botnet and DDoS attack with the highest or maximum accuracy. Highest accuracy is obtained with the selected feature and with a chosen classification algorithm. Naïve Bayes and SVM classification algorithms are used for achieving high levels of accuracy. The model detects mixed high-rate, low-rate DDOS attacks and Botnet through Network Traffic Analysis and Machine Learning.

I. Introduction

A cyber security threat is a spiteful or malicious act that attempts to cause damage or harm to data, steal data, or destroy the structure of digital life. Cyber-attacks include various threats that cause harm to a system such as computer viruses or trojans, data breaches, Ransomware attacks, Systems getting hacked, Denial of Service (DoS) attacks etc. Cyber-attacks and cybercrimes are frequently increasing and growing in terms of both the number of attacks and level, or intensity of damage caused to its sufferers.

Attackers build their targets varying from single individuals to small or medium sized companies and even big/giant companies. Every year it looks like the number of attacks of attacks are getting bigger and more and with that it results in a bigger number of attacks that penetrate through or defeat the security of excessively big and large companies. It affects information and data security, business continuity and most importantly customer's trust. Cyber-attacks can cause wide variety of damage such as electrical blackouts, data theft, exposure of sensitive people and government data, compromise of medical records, breaches on national security secrets, paralyze system and phone networks making data not available etc. It can be said without a doubt that cyber threats or cyber-attacks can cause huge damage to the functioning of life.

Big companies and small companies combined spend more than hundreds of billions of dollars each year on cyber security. Among the most dangerous and effective attacks are DoS and DDoS (Distributed Denial of Service) attacks. DDoS attack has changed the most common and normal attack mode which is peer-to-peer. Distributed denial of service attack does not use statistical rule. It is extremely challenging to separate attack or typical conduct just through the sorts of conventions or administrations and it is one reason the dispersed refusal of-administration attack is not viewed as simple to recognize [1].

A set of controlled bot or botnet is compromised internet connected devices such as PCs, cellphones or IOT gadgets. A bot continues to run over the Internet. Computer emerges from "Robot", and most of the

bots on the internet or network are controlled by a master bot. Botnets [2] are generally used to perform particular selected assignments, such as DDoS attack, transmission of malware, spam sending, theft of sensitive data and phishing attacks.

Botnet or bots use only truly little or insignificant amounts of processing or computing power to avoid damaging normal device functions and notifying or alerting the user. There is a different type of botnet [3] varying from simple in nature to overly complex and advanced type that are even designed or programmed in a way to update their behavior in order to prevent them from being detected by cyber security software such as firewalls, antivirus etc. Normal users or victims using the connected devices or systems are completely unaware that they're connected devices are compromised or infected and being controlled by cyber criminals randomly from remote location.

Bots that are compromised generally are waiting for the bot master to in-struct them or communicate with them to perform various malicious actions such as a DDoS attack or for spam dissemination. With the introduction of botnets, the landscape for DDoS attack has been completely changed and the challenge for cyber security has also increased. DDoS attack generally requires lot of systems to send multiple requests at same time or similar time intervals and introduction of botnet to do DDoS attack enables malicious adversaries to drastically expand the size and reach of potential attacks. It also allows those attacks to be executed with a lot of precision and control which makes it virtually impossible to trace it back to the original attacker.

Botnet makes the use of DNS, HTTP etc protocol to communicate with compromised bots. DNS and HTTP based traffic in a network is generally considered to be legitimate traffic by most of the firewalls. Things like these are taken as an advantage by the botnet master to get through or pass the organization firewall. This is how the malicious software's get downloaded into the computers of users.

II. Review of Literature

Most of the research on different botnet and DDoS detection techniques separately. Botnet and DDoS specific types are investigated and have been used in most of the research. Botnet identification procedures or approaches can be characterized into three methodologies have based location, network-based discovery and hybrid-based recognition. Also, DDoS attack location can be arranged into low-rate identification and high-rate detection.

Huang at et.al 2013, [4] has proposed a powerful bot have location arrange-ment that is typically founded on network disappointment following in a client or a host during brief period. Vahdaniet and Etemad.al [5], by using host analysis address the detection of centralized botnet from C&C. The methodology that they propose is based on real traffic analysis of the incoming and the outgoing host.

Zhao et al. in 2013 [6] suggest a P2P approach, which involved the detection or identification of botnets with the help of network traffic analysis in a network. Their proposed concept or study is to pick 12 characteristics based on that extraction of flow behavior for a limited period of time. G.Kirubavathi and

R.Anitha et al. in 2016 [7] propose general botnet identification fit for identifying various kinds of botnet. This approach dissects the progression of organization traffic during time spans consistently.

Hybrid-based detection is an inter relationship between the network traffic analysis and detection method and host traffic analysis or detection method. Zeng et al.in 2014 [8] proposed a detection technique to identify unique botnet that consolidates both discovery strategies. Abdullah et al. in 2014 [9], proposed a method for taking care of or managing the P2P botnet architecture. They proposed a solution which is similar to Zeng's method depending on the combination of both detection methodology. File systems such as the registry and log file are analyzed in the case of host based.

Low-rate DDoS attacks are a type of DDoS attack in which an attacker or an adversary sends huge attack packets which are the same as the legitimate packet. Because of such nature Low-rate DDoS attacks have become a real threat in the current cyber space. Du and Abe et.al in 2008 [10] proposed technique that uses entropy metric based on IP packet size, this metric can identify DDoS attacks either low-rate or high-rate. Jadhav and Patil et al. in 2013 [11] proposed a plan or technique that elaborates an entropy-based strategy to identify low-rate DDoS attacks.

Haining Wang, Cheng Jin and Kang G.Shin et.al 2007[12], performed a de-tailed study on the various defense methodologies or schemes for the allegedly spoofed DDoS attacks. They found out that each scheme has its own benefits and limitations. Another filtering technique based on hop-count is proposed in [13]. In this case, an attacker fabricates fields of the TCP/IP header to launch a spoofed DDoS attack.

These issues are addressed separately. The current study proposes a model to detect mixed high-rate, low-rate DDOS attacks and Botnet through Network Traffic Analysis and Machine Learning. The goal of the project is obtained by using Support Vector Machine and Naïve Bayes classification algorithm. Machine Learning approach is used to predict or detect a DDOS and botnet attack with the highest or maximum accuracy. Cyber security threat is a spiteful or a malicious act that attempts to cause damage or harm to data, steal data, or destroy the structure.

III. Proposed Methodology

A. *Datasets description*

CCIDS2017 is the data set used for this analysis. This data set is used to train algorithms and to classify DDOS and Botnet Attacks based on its unique features. The informational collection contains objective port, different bundle attributes, for example, parcel length, stream span, header length and so on, attack mark on the off chance that the traffic is DDoS or botnet traffic and different TCP banners. Elements, for example, bundle size, parcel length, stream span, forward bundle, in reverse bundle and other different bundle ascribes are utilized to decide DDoS attack. DDOS traffic has high stream rate, greater bundle size than ordinary parcel size, more bundle length than the typical bundle length and so on.

CCIDS2017 data set also includes the results of the network traffic analysis that was performed using CIC Flow Meter with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack. The “CICIDS2017” dataset contains data with more than 80 network flow features. Random splits are performed over data to divide it into 80% of the training data set, 20% of test data set.

B. *Feature Selection and dimensionality reduction*

Data features selection has a huge influence and impact on the performance of any machine learning based model. Selection of an important or significant set of features and cleaning of the raw data are critical steps in designing a model. The various methods include Principal Component Analysis, Linear Discriminant Analysis and Generalized Discriminant Analysis. Dimensionality reduction may be both linear and non-linear, depending upon the method used. Feature Selection picks certain features that most relate to your predictive variable or performance you are interested in, automatically or manually. Selecting correct and helpful features assist in reducing Overfitting, training time and improve accuracy.

C. *Classification algorithms*

Scikit-learn, the accepted Python library is utilized for machine get the hang of ing, matplotlib for plotting, and OpenCV for stacking and preprocessing pictures in the dataset. Directed learning to identify Botnet and DDOS attacks is applied. Two Machine learning calculations Naïve Bayes and Support Vector Machine for information characterization were explored and tried. The calculations chosen depended on their powerful execution and execution of organization security.

SVC, NuSVC and LinearSVC are classes fit for performing double and multi-class characterization on a dataset. SVC is utilized for this situation. GaussianNB carries out the Gaussian Naive Bayes calculation for arrangement. The probability of the elements is thought to be Gaussian. The classification_report() function will return all behavior of the model such as precision, recall, f1-score and support. It also returns micro average, macro average, weighted average and samples average. The accura-cy_score() will return the accuracy which the model is based on the provided argu-ments such as training accuracy and testing accuracy.

D. *Network Packet Analysis*

Characteristics of high-rate DDOS attacks is that most of the packets are of ab-normally big size, frequent flow rate etc making it easier to detect in compared to low rate ddos attack. Network packet analysis can be implemented in order to determine botnet attacks. Traffic to and from specific ports such as 53, 80 and 6697 which have DNS, http and irc traffic incoming and outgoing can be used to determine botnet traffic. “DomainName Service (DNS) is a significant help on web which allows the resolutionof have names, or space names to IP address and IP address to area names. Observing and examining DNS question information can tell the presence of malicious practices in really look at framework, since a piece of the DNS inquiry information may be created by botnets. Additionally, IRC and HTTP traffic can likewise be

very much like DNS traffic to decide botnet. This study mainly emphasizes the difference between packet properties of attack packets and legitimate packets.

E. *Model Overview*

The model depends on elements, for example, parcel streams, port numbers, traffic examination and so on. The proposed DDOS and Botnet model uses data set that has two classes of data traffic. The very first step in the proposed model is to define the set of features/functions and classes that are initially intended to be used. Pandas' library is used to load the dataset. Once the CSV file is loaded, data from the columns are spitted into input and output variables. The data containing all the features and labels will be stored in a 2-dimensional array. The first dimension contains all selected rows, and the second dimension contains all selected columns. The 2-dimensional array can then be split into two arrays by picking subsets of columns using the standard NumPy slice operator. Loaded dataset split as 85% of data for training and 15% data for testing.

Initially the model which is proposed is trained. Training set which contains data of 800 unique data. Labelling the data is performed for the ones that are either DDOS traffic or Botnet Traffic. Training stage is basically a stage where in both the algorithms SVM and Naïve Bayes algorithm-based computations are made use of to learn them about the classifiers. Next, we go to the assessment methodology through which the AI calculation is picked for the utilization in proposed location or expectation model in light of the most or higher grouping exactness acquired.

During the second phase of the model the packet flows, ports etc are observed and classified in the first stage to decide whether the traffic is DDOS traffic or botnet traffic. DNS traffic in port 53, HTTP traffic in port 80 and IRC traffic in 6667 is used with different packet properties in order to classify botnet attack. Packet size, Packet length, Packet frequency and other packet properties are used in order to classify DDOS attack.

IV. Results

The model is experimented using two classification algorithms Naïve Bayes and Support Vector Machine in order to predict Botnet and DDoS attack. The Fig. 1 and Fig. 2 depicts the performance measures for Naïve Bayes algorithm and SVM. The Fig. 3 and Fig. 4. depicts Confusion Matrix and Classification performance of the detection model using training set of both algorithms.

75.3% is least accuracy produced by the Naïve Bayes algorithm and 99.68% is highest accuracy of SVM. DDoS attack and botnet attacks are detected by mechine learning approach. 99.68% accuracy was obtained as a result which is maximum. The accuracy can be obtained with the set of features that is selected and classification algorithm that is chosen.

V. Conclusion and Future Scope

The review proposed a model to deal with DDoS and Botnet attacks by using mechine learning algorithms. The CCIDS2017 informational index is utilized in the review to distinguish botnet and DDoS attack. Dataset has 80 elements out of which 20 were chosen. The calculations, which were applied to the informational collection, are Support Vector Machine and Naïve Bayes. The exploratory outcomes showed that the model involving SVM calculation has more prominent exactness in contrast with model utilizing Naïve Bayes. Contrasting every one of the outcomes showed that SVM has the improved effectiveness of 99.68%. An exploration with new highlights to further develop the location precision of the proposed model is likewise essential for future work. The future upgrades/work likewise includes order of sort of DDOS attacks and botnet attacks in view of the various banners, administrations in ports and convention types.

References

1. Tripathi, Nikhil & Mehtre, Babu, DoS and DDoS Attacks: Impact, Analysis and Countermeasures. 1–6 (2013).
2. Yogita Barse, Dr. Sonali Tidke, A Study on BOTNET Attacks and Detection Techniques, IOSR Journal of Electrical and Electronics Engineering e-ISSN: 2278 – 1676,p-ISSN: 2320–3331, Volume 15, Issue 3 Ser.
3. Anwar, Shahid, A Review Paper on Botnet and Botnet Detection Techniques in Cloud Computing, ISCI 2014 – IEEE Symposium on Computers & Informatic.
4. C.-Y. Huang, Effective bot host detection based on network Failure models, Computer Networks, vol. 57, no. 2, pp. 514–525, 2013.
5. F. Etemad and P. Vahdani, Real-time botnet command and control characterization at the host leve, in Proceedings of the Sixth International Symposium on Telecommunications, pp. 1005–1009, Tehran, Iran, November 2012.
6. D. Zhao, I. Traore, B. Sayed et al., Botnet detection based on traffic behavior analysis and flow intervals, Computers & Security, vol. 39, pp. 2–16, 2013.
7. vG. Kirubavathi and R. Anitha, Botnet detection via mining of traffic flow characteristics, Computers & Electrical Engineering, vol. 50, pp. 91–101, 2016.
8. J. He, Y. Yang, X. Wang, Y. Zeng, and C. Tang, PeerSorter: classifying generic P2P traffic in real-time, in Proceedings of the 2014 IEEE 17th International Conference on Com-putational Science and Engineering, pp. 605–613, Chengdu, China, December 2014.
9. R. Abdullah, M. Faizal, and Z. Noh, Tracing the P2P botnets behaviours via hybrid analysis approach, European Journal Scientific Research, vol. 118, no. 1, pp. 75–85, 2014.
10. Du R, Lu Kv, Petritsch C, Liu P, Ganss R, Passegue E, Song H, Vandenberg S, Johnson Rs, Werb Z, Bergers G. Hif1alpha Induces The Recruitment Of Bone Marrow-Derived Vascular Modulatory Cells To Regulate Tumor Angiogenesis And Invasion. Cancer Cell. 2008;13:206–220.

11. P. N.Jadhav and B. M. Patil, Low-rate DDOS Attack Detection using Optimal Objective Entropy Method, International Journal of Computer Applications, vol. 78, no. 3, pp. 33–38, 2013.
12. Jin, Cheng & Wang, Haining. (2003). Hop-Count Filtering: An Effective Defense Against Spoofed DDoS Traffic. 10.1145/948109.948116.
13. P. Indu, 2 Shalom Elza Joseph, 3M.C. Sreelakshmi and 4T. RemyaNair, Enhancement of HOP Count Filtering MechanismAn ANTI-IP Spoofing Technique, International Journal of Pure and Applied Mathematics Volume 114 No. 12 2017, 51–58 ISSN: 1311–8080 (printed version); ISSN: 1314–3395 (on-line version) url: http://www.ijpam.eu Special Issue.

Figures

```
cm is  [[ 82  79]
 [  0 159]]
             precision    recall  f1-score   support

          0       1.00      0.51      0.67       161
          1       0.67      1.00      0.80       159

avg / total       0.84      0.75      0.74       320

0.753125
```

Figure 1

Result Obtained for Naïve Bayes algorithm

```
99.6875
classification report
             precision    recall  f1-score   support

          0       0.99      1.00      1.00       161
          1       1.00      0.99      1.00       159

avg / total       1.00      1.00      1.00       320
```

Figure 2

Result Obtained for SVM

Machine Learning Algorithms	True Positive (TP)	False Negative (FN)	False Positive (FP)	True Negative (TN)
Naïve Bayes	82	79	0	159
SVM	161	1	0	158

Figure 3

Confusion Matrix

Machine Learning Algorithms	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
Naïve Bayes	75.3125	74	84	75
SVM	99.68	100	100	100

Figure 4

The Performance of Classification