EasyChair Preprint

№ 7509

# Exploring Open Source Information for Cyber Threat Intelligence

Madhavi Netke, Sarita Patil and Manjushree Mahajan



February 26, 2022

*MANUSCRIPT*

# (Exploring Open Source Information for Cyber Threat Intelligence.)

Principle Author[a:] : Madhavi Manoj Netke,

Co-author[a] : Prof.Sarita Patil.

Co-author[b :] Prof.Manjushree Mahajan

[a]*Department Name : Computer Dept*

*University Name: Savitribai Phule Pune University*

*College name : G.H.Raisoni college of Engineering and management , wagholi, pune*

*City : Pune ,*

**ARTICLE HISTORY**

Received:
Revised:
Accepted:

*DOI:*

**Abstract:**Cyberspace is one of the most complex systems ever developed by humans; many people use it daily, yet few comprehend it. Use Social media cannot replace security specialists. to examine certain attacks in detail, such as de-

detecting network irregularities, worms, and port scanning besides However, analysing social media data can uncover new cyber threat and security threat patterns in- comprising carding and data theft. We used AI- ing the proposed system's cyber danger. Ideal design is produced by training a Twitter cyber-Threat dataset with SVM, NB, DT, RF, ANN. Preferable model for defining cyberthreats and their types.

**Keywords: Machine Learning , Support vector machine , Naïve bayes classifier,Artificial neural network,Decision tree classifier,Random Forest,Threats etc.**

## 1. INTRODUCTION

Professionals may exploit the huge amount of data available in cyberspace to secure a company's network design and prevent invasions. Social media cannot replace some types of threats, such as malware, port scanning, and network traffic abnormalities. Monitoring social media data can help spot new cyber attack trends and security problems including breaches, carding, and hijacking. Academics have utilised Twitter to acquire critical information on terrorist activities, natural disaster outbreaks, and strange events. As a consequence, Twitter is an excellent resource for staying current on the most recent cyber security breakthroughs. This article will examine notable data breaches, as well as a map depicting the geographical distribution of cyber attacks in the United States and a list of popular attack strategies. Finally, we developed a system for detecting cyber-threat-related texts in the deep and surface webs using socio-personal and technological indicators, with the goal of providing artifacts for cyber security specialists and law enforcement agencies to use in preventative and prosecution methods.

## 2. Problem Statement

Social media cannot take the role of security specialists conducting in-depth investigations into certain types of assaults, such as difficult-to-understand network traffic, worms, and port scans. In contrast, analysing social media data can aid in the discovery of new patterns of cyber hazard and security risk, such as data theft, carding, and hacking.

## 3. LITERATURE SURVEY

Victor Adewopo, Bilal Gonen, Festus Adewopo.[1] Cyberspace is one of the most complicated human-made systems, and many people utilise it daily. However, most users are unaware of it. Historically, cyber attacks were mainly random to entice unwary victims. More data suggests that hacker forums and individuals share cyber attack knowledge. This article suggests using open source information from the surface web (Twitter) and deep web hacker forums to find cyber-related content. Our method can give cyber security experts and law enforcement agencies with correct data to build control and containment plans for cyber attacks involving deep web threats and surface web threats. We evaluated over recorded incidents in the Privacy Rights Clearinghouse (PRC) Chronology of Data Breaches. Finally, we recommend geospatial cyber attack risk profiling as a potential study field. The index includes phrases like cyber attack, Deep web, cyber security, and cyber threat.

Priyanka Ranade, Sudip Mittal, Anupam Joshi and Karuna Joshi.[2] The Internet's multilingual structure hampers the cybersecurity community's concerted efforts to mine threat intelligence from OSINT data on the web strategically. OSINT sources such as social media, blogs, and dark web vulnerability storefronts are available in a variety of languages, which makes security analysts' employment more difficult. who are unable of inferences from intelligence in languages they do not understand Third-party translation engines are becoming more powerful, but they are still in their early stages. unsuitable for private security situations To begin, due to privacy and confidentiality standards, sensitive intelligence is not permitted as an input to third-party engines. Furthermore, third-party engines generate broad translations that are often lacking in specificity. Terminology for cyber security We address these troubles in this study and describe our approach for processing threat intelligence across new languages. We develop a neural network-based system that receives cyber security data in many regional dialects and provides the appropriate English translation. Translation into English can then be deciphered by an analyst and used as input to an AI-based cyber-defense system capable of taking corrective action We have used this as a proof of concept. built a pipeline that converts Russian threats into English, RDF, and victories representations Translations on our network are optimized. Data on cyber security, especially.

Ying Dong1, Wenbo Guo2,4, Yueqi Chen2,4, Xinyu Xing2,4,Yuqing Zhang1, and Gang Wang3.[3] - Public vulnerability databases like CVE and NVD have been quite successful in promoting vulnerability disclosure and remediation. However, As databases amass huge amounts of data, concerns about their quality and consistency grow. We propose an automated system VIEM to detect inconsistencies between fully standardised NVD unstructured CVE descriptions and related vulnerability reports. VIEM allows us to measure information consistency on a vast scale, and gives the community with a tool to maintain the CVE/NVD databases. VIEM extracts programme names and versions from unstructured text. We offer personalised in order for VIEM to recognise previously unseen software names and versions using deep learning NER and RE. depending on syntax and context Ground-truth testing reveals the system's accuracy (0.941 precision and 0.993 recall). We use VIEM to look at 78,296 CVE IDs and 70,569 vulnerability reports during the last 20 years. Our findings imply that Versions of software are common. With time, only 59.82 of the vulnerability reports/CVE summaries strictly match the specified NVD data. Case studies indicate NVD's inaccurate information about susceptible software versions.

Ariel Rodriguez1,a) Koji Okamura1,b) 4] The Internet is constantly shifting, leading to the establishment of a plethora of new data sources that may be leveraged to acquire insight into the cyber threat landscape and, as a response, better prepare for cyberattacks. In light of something like this, we describe an end-to-end real-time cyber situational awareness system that tries to retrieve securityrelevant information from Twitter.com. This system classifies and processes information. based on sentiment analysis and data analytics techniques, collects the data retrieved and delivers real-time cyber situational awareness information. This investigation will aid security analysts in assessing the amount of cyber risk in their business quickly and efficiently, allowing them to take proactive steps to plan and prepare for future attacks before they occur.

Masashi KADOGUCHI, Shota HAYASHI, Masaki HASHIMOTO,Akira OTSUKA [5] Cyber attacks methodologies have become increasingly sophisticated in recent years, making it more difficult to resist an attack, even if a kind of defence existed. Some kind of countermeasure is taken. It is critical to have a prediction of cyber attacks, suitable precautions, and effective use of cyber intelligence that permits these activities in order to successfully handle this circumstance. Malicious hackers share variety of information through certain groups, such as the dark web, demonstrating that cyberspace has a significant amount of intelligence. This research concentrates on dark web forums and provides a way for retrieving forums that contain valuable information or intelligence from large numbers of forums and identifying attributes of each topic using machine learning, natural language processing, and other methodologies. We will be able to grasp the situation using this method. growing cyberthreats and implement appropriate countermeasures against malicious activity.

## 4.ALGORITHM

The Support Vector Machine, or SVM, is a well-known Supervised Learning technique that can be used to solve classification and regression problems. However, it is mostly used in Machine Learning to solve classification problems. The goal of the SVM algorithm is to find the best line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be easily placed in the right category in the future. The optimal choice boundary is referred to as a hyper plane. SVM is used to select the extreme points/vectors that will help create the

hyper plane. The algorithm is known as a Support Vector Machine, and support vectors are the most extreme examples. Consider the diagram below, which depicts how a decision boundary or hyper plane is used to categories two distinct groups.

SVMs are classified into two types:
SVM Linear: Linear SVM is a classifier for linearly separable data, which means that if a dataset can be classified into two classes using a single straight line, it is linearly separable data, and the classifier is called Linear SVM.

SVM (non-linear): Non-linear SVM is used to classify non-linearly separated data, which means that if a dataset cannot be classified using a straight line, it is non-linear data, and the classifier used is Non-linear SVM.

Computing the (soft-margin) SVM classifier amounts to minimizing an expression of the form

$$\left[\frac{1}{n}\sum_{i=0}^{n}\max(0,1-y_i(w.x_i-b))\right]+\lambda||w||^2$$

We focus on the soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly-classifiable input data.

A.Artificial Neural Network (ANN) Algorithm:-

Fully interconnected multi-layer neural networks, ANNs can be shown in the illustration below. Each of these layers consists of an input layer and several hidden layers, as well as an output layer. Each layer's nodes are linked to each other via these connections. We can make the network deeper by increasing the number of layers that are stored [6]

**1. The ANN algorithm's steps**

- Step 1: Defining a Sequential model is the first step.
- Step 2: Apply a sigmoid activation function to a dense layer.
- Step 3: Use an optimizer and a loss function to compile the model.
- Step 4: Analyze the data and fit the model to it.

The weighted total of the inputs is computed by the artificial neural network, which also incorporates a bias. A transfer function is used to express this computation.

$$\sum_{i=0}^{n} wi * Xi + b$$

The weighted total is used as an input to an activation function to generate the output. Activation functions determine if a node should fire or not. The only ones who make it to the output layer are those who are fired. Depending on the work at hand, there are a variety of activation functions that can be used.

B.Random Forest Algorithm:-

Unsupervised classification is accomplished using the random forest technique. In keeping with its name, this procedure results in a dense forest. The size of a forest appears to be inversely related to the number of trees in the forest. Similarly, the random forest classifier gets increasingly accurate as the number of trees in the forest grows. Knowing how a decision tree works should allow you to guess how accurate the results will be. Some decision tree algorithm rules will be developed as a result of the objective-feature training. The test dataset may be predicted using the same set of rules [8] [10].

The Random Forest level is the sum of all the trees in the forest. The total number of trees is divided by the sum of the feature's importance rating on each tree:

fi sub(i) = the value of a feature I ni

sub(j)= node j's importance

$$RFfi_i = \frac{\sum_{jealltrees} normfi_{ij}}{T}$$

T = total number of trees

RFfi sub(i)= the importance of feature I estimated from all trees in the Random Forest model

normfi sub(ij)= the normalized feature importance for I in tree j

C.Naive Bayes Algorithm:-

The Naive Bayes algorithm, which is a supervised learning method based on the Bayes Theorem, is being used to solve classification problems. This approach is frequently used in text classes with a considerable volume of dynamic data in the training set. The Naive Bayes Classification algorithm is one of the most feasible and elegant approaches for quickly creating prediction models. Spam filtering, sentiment

analysis, and article classification all use the Naive Bayes algorithm. The phrase "Naive Bayes Algorithm" is made up of two words: Naive Bayes [1]. It's naïve, as the name implies, because it believes certain characteristics exist regardless of other conditions. This reddish-orange, spherical, and soft fruit is referred to as a fruit in addition to being named an apple because of the following characteristics: Each characteristic, while not mutually exclusive, contributes to the fruit's identification as an apple .

The Bayes theorem allows you to calculate the posterior probability, P(c|x), from P(c), P(x), and P(x|c) using P(c), P(x), and P(x|c). The Naive Bayes classifier assumes that the impact of a predictor's value (x) on a given class (c) is independent of the values of other predictors. Class conditional independence is the term for this assumption.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

D.Decision Tree Algorithm:-

DTA is an unique method that may be adapted to different of predictive modeling scenarios. To generate decision trees, an algorithmic mechanism that splits the data set according to conditions can be adopted. Decision settings are 1 the most powerful supervised algorithms. They is used for both regression and classification. The decisionmaking nodes split the data and leave the outcomes. A binary tree illustrates a person's age, eating habits, and habits [3]. "Information Gain" refers to splitting data using entropy. A dataset's entropy diminishes when split on an attribute.

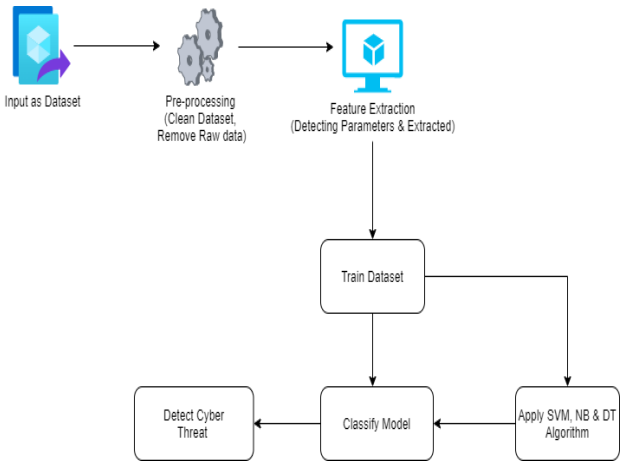
Gain (T, X) = Entropy (T) — Entropy (T, X)

• T stands for target variable.

• Entropy (T, X) = The entropy estimated after the data is split on feature X

• X = Feature to be split on

## V. SYSTEM ARCHITECTURE




Explanation –

1. Input as Dataset – Input as dataset , First Load dataset of Cyber Threat.
2. Data preparation is an important activity for cleaning data, eliminating raw data, missing values, and preparing it for a machine learning model, which enhances the model's accuracy and efficiency.
3. Feature Extraction – Feature extraction allows us to create new features by linearly combining old features. The new set of characteristics will have different values than the previous ones. The goal is to collect data with fewer features.
4. Classification – In classification, an algorithm learns from data and then uses what it has learnt to categories new observations. To put it another way, the training dataset is used to improve border conditions, which are then utilised to define each target class. SVM, NB, and DT are used here, along with a training dataset.
5. Output – Output is to Detect Cyber Threat

## VI. CONCLUSION

In this Proposed System, we present a unique approach of collecting data on kaggle website to analyse information about cyber threats and issues an early warning/detection system. Using only Twitter data for predicting cyber threats. A sentiment analysis on hacker forums to predict cyber threats. Machine learning algorithm used to detect cyber threats.

## VII.ACKNOWLEDGEMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

## VII.REFERENCES

1.Shengping Zhou Zi,Long Lianzhi ,Tan Hao Guo(“Automatic Identifification of Indicators of

Compromise using Neural-Based Sequence Labelling")Nanshan District, Shenzhen City, Guangdong Province, 518057, China

2.Priyanka Ranade, Sudip Mittal, Anupam Joshi and Karuna Joshi ("Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence ")University of Maryland, Baltimore County, Baltimore, MD 21250, USA Email: {priyankaranade, smittal1, joshi, karuna.joshi}@umbc.edu

3. Ying Dong1,Wenbo Guo2,, Yueqi Chen2,4, Xinyu Xing2,4, Yuqing Zhang1 , and Gang Wang3("Towards the Detection of Inconsistencies in Public Security Vulnerability Reports")4 JD Security Research Center, USA

4.Ariel Rodriguez1,a) Koji Okamura("Social Media Data Mining for Proactive Cyber Defense")

5.Masashi KADOGUCHI,Shota HAYASHI,Masaki HASHIMOTO,Akira OTSUKA("Exploring the Dark Web for Cyber Threat Intelligence using Machine Leaning").

6.Ba-Dung Le,Guanhua Wang,Mehwish Nasim,M. Ali Babar("Gathering Cyber Threat Intelligence from Twitter Using Novelty Classifification")

7.Manvi ,Ashutosh Dixit ,Komal Kumar Bhatia.("Design of an Ontology based Adaptive Crawler for Hidden Web")