# An Overview of Cyber Threat Intelligence Platform and Role of Artificial Intelligence and Machine Learning

**Chapter** *in* Lecture Notes in Computer Science · December 2020

DOI: 10.1007/978-3-030-65610-2_5

CITATIONS
8

READS
796

**2 authors:**

Abir Dutta
Sharda University
**2** PUBLICATIONS **29** CITATIONS
SEE PROFILE

Shri Kant
Sharda University
**41** PUBLICATIONS **232** CITATIONS
SEE PROFILE

# An Overview of Cyber Threat Intelligence Platform and Role of Artificial Intelligence and Machine Learning

Abir Dutta(✉) and Shri Kant(✉)

Research and Technology Development Center, Sharda University, Noida, UP, India
abir_wbsetcl@yahoo.com, shrikant.ojha@gmail.com

**Abstract.** Ever enhancing computational capability of digital system along with upgraded tactics, technology and procedure (TTPs) enforced by the cybercriminals, does not match to the conventional security mechanism for detection of intrusion and prevention of threat in current cyber security landscape. Integration of artificial intelligence, machine learning and cyber threat intelligence platform with the signature-based threat detection models like intrusion detection system (IDS), SNORT, security information and event management (SIEM) which are being primarily implemented in the network for continuous analysis of the indicator of compromise (IoC) becomes inevitable, for prompt identification of true events and subsequent mitigation of the threat. In this paper, author illustrated the approach to integrate artificial intelligence and machine learning with the cyber threat intelligence for the collection of actionable threat intelligence from various sources like dark web, hacker's forum, hacker's assets, honeypot, etc. Furthermore, the application of threat intelligence in the aspect of cyber security has been discussed in this paper. Finally, a model has been proposed for generating actionable threat intelligence implementing a supervised machine learning approach employing Naïve Bayes classifier.

**Keywords:** Cyber threat intelligence · Artificial intelligence · Machine learning · Cyber security · Threat

## 1 Introduction

According to the recent survey of IT Governance of UK, around 6 billion data security breach has been recorded in the first quarter of 2020 and in the recent COVID-19 pandemic outburst this figure is boosting significantly due to espousing of "work from home" model by the organizations without implementing sufficient counter measures to deceive the attacks. Threat changes its nature of function and the structure very frequently in the domain of cyber security and advanced attack technique such as advanced persistent threats (combining both "multi-vector" and "multi-staged"), are adopted by the cyber criminals to attack the victim continuously in order to infiltrate the network and remains undetected for a period of time and finally filtrate out the targeted information without causing mutilation of the network. However, integration of artificial


S. Kanhere et al. (Eds.): ICISS 2020, LNCS 12553, pp. 81–86, 2020.
https://doi.org/10.1007/978-3-030-65610-2_5

intelligence, machine learning and deep leaning with cyber threat intelligence generates an automated framework to extract intelligence from various sources and blending this actionable threat intelligence to existing security mechanism to perform quickly, accurately, effectively and efficiently.

Residue of this paper is structured as follows. Section 2 specify the survey of present cyber threat intelligence status, Sect. 3 demonstrates the GAP identified in the literature review, Sect. 4 describes the Role of AI, ML and DL in cyber threat intelligence, while Sect. 5 illustrate author's proposal for implementation of AI and ML in cyber threat intelligence platform, Sect. 6 represents the future scope of work and finally, Sect. 7 provides the conclusion of the research.

## 2 Overview of Cyber Threat Intelligence

Several definitions of cyber threat intelligence have been exposition across the literature. However, the most comprehensive definition of CTI is "cyber threat intelligence is any evidence-based knowledge about threats that can inform decisions, with the aim of preventing an attack or shortening the window between compromise and detection".

Radical shift of attack technology placed the traditional signature based threat detection model into severe challenging position, which generates the necessity of combining of cyber threat intelligence platform and existing communication methodology. Jorge Buzzio Garcí et al. 2019 focuses on implementation of actionable threat data feed collected from CTI to enhance the software defined networks (SDN) security. Threat Intelligence are gathered by commissioning collective intelligence framework (CIF version 3) which receives "feed" type from internal or external sources of known attacks and CIF primarily stored the information like IP addresses, domains and URLs, of suspected activities and elaborate the way of preventing malicious traffic using physical testbed consisting of five components such as CIF server, SDN controller and application, OpenFlow switch and hosts.

In cyber threat intelligence platform STIX format has been extended [1] to facilitate the interpretation of critical patterns. Extension of STIX allows marking the feature of an object and these features represents relationship among various objects.

## 3 GAP Identified

Scope of experiment is still expected in this domain such as- manual labelling of information is error prone and time consuming activity that can be replaced by providing an automated window to upload the relevant intelligence [2], so that the portal will automatically sense the threat intelligence and tag appropriate information to minimize the effort of incident responder and extract the optimal combat mechanism. Furthermore, developing a multi-layer cyber threat intelligence ontology can be explored in future studies. Multilayer CTI ontology can be constructed by explaining formal definition and lexicon; inclusion of abstract layer of CTI in the ontology, constraints must be exposition properly to facilitate underlying Web Ontology Language (OWL) with the analytical capabilities.

## 4 Role of AI and ML in CTI Platform

Integration of machine learning in cyber security domain helps to build a superior and reliable model [3–7] pertaining to the malware detection, spam classification and network intrusion identification. Implementation of artificial intelligence at various phase of cyber threat intelligence like tactical intelligence and operational intelligence describe that, tactical threat intelligence is suitable for "Multi-Agent" system where as operational CTI is applicable for "Recurrent Neural Network".

To transform the overwhelming threat intelligence generated by structured and unstructured sources to actionable data feed is an utmost crucial aspect of CTI perspective. At the same time eradication of spurious threat intelligence is also expected for building an effective and accurate intelligence platform.

## 5 Proposed Model for Using AI and ML with Cyber Threat Intelligence Domain

### 5.1 Outline of the Proposed Model

In this section we proposed a machine learning approach to extract actionable threat intelligence from various sources. Phases of the model generation is depicted in Fig. 1.

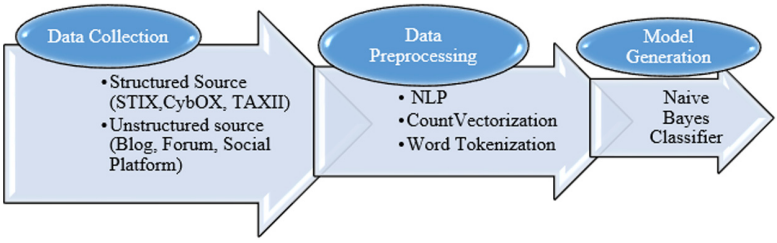
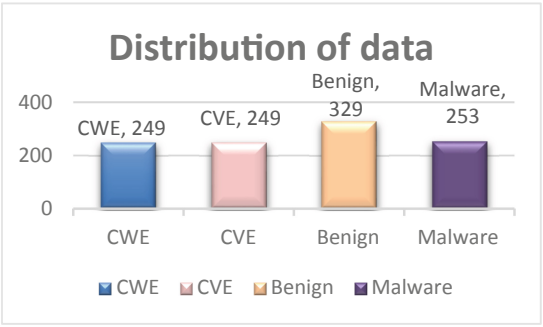



**Fig. 1.** Depicts various phases of proposed architecture

### 5.2 Data Accumulation

Cyber threat intelligence is generated from various sources like structured sources (STIX, CybOX, TAXII) as well as from unstructured sources (hacker's forum, Blog, blacklists). In this experiment we have collected latest CVE entries of 2020 and CWE list v4.1 along with malware dataset and Goodware dataset from 'kaggle' web portal to construct our dataset. In CVE list, entries are in.xml format and we have filtered the "title' and "description" attributes merely under the 'vulnerability" tag by using ElementTree XML API of python discarding all other attributes like header, references etc. from the intended dataset and converted the extracted information of CVE list into CSV format. In addition to that we have downloaded the latest CWE list version 4.1 in and extracted the "description" attribute from the CWE list. Furthermore, we have collected few malware description as well as few benign descriptions from 'kaggle' web portal to construct our final dataset in CSV format with approximately 1100 dataset.




Finally, extracted attribute is labelled as- 1 for the sentiment related to the vulnerability, weakness or malware and 0 for the sentiment related to the benign posts in order to prepare the pre-processed dataset for implementation of ML classifier.

### 5.3 Feature Extraction and Language Processing

Collected information is in text format which required conversion to the suitable data format for the acceptance of ML Algorithm. Natural language processing approach is employed on the raw dataset such as remove duplicate word, punctuation and stop words to extract the clean text words. Thereafter, clean text is tokenized to form the lemmas and converted text to matrix of token counts using Bag of Words, count vectorization technique to extract the potential keywords related to the threat and non-threat as tabulated here.

| Keywords | Category |
|---|---|
| denialofservice, exploitation, Out-of-bounds, unauthenticated, memory corruption, privilege, destroywindow , Uncontrolled, vulnerability, firmware, | Threat |
| ReadFile, Sleep, GetSystemDirectoryW, HeapFree, CreateDirectoryW, CopyFileW, FindClose | Non-threat |

### 5.4 Machine Learning Classifier

Several algorithms are there for natural language processing and text analysis; deep learning classifiers such as CNN, RNN are suitable for text analysis. In this experiment we have adopted Naïve-Bayes classifier for extracting high level threat intelligence from the dataset, considering 70% data for training dataset and 30% data for test dataset. Naïve Bayes algorithm is used for classification problem specially for text classification. In this model one feature is independent of existence of another feature i.e. each feature contributes to the prediction without having correlation. Text vector is the data feed for this model to train the model, followed by testing the model to evaluate the performance of the model and finally, predict true events (malware or threat) for unknown data.

### 5.5 Performance of the Proposed Model

We have evaluated different performance metrics such as accuracy, precision and f1-score of the model against the training as well as test dataset using Naïve Bayes classifier and accuracy of the model reflects as 98.2% and 96.6% for the training dataset and test dataset respectively as furnished below

- Accuracy of the model for training dataset

```
              precision    recall  f1-score

           0       1.00      0.94      0.97
           1       0.98      1.00      0.99

    accuracy                           0.98
   macro avg       0.99      0.97      0.98
weighted avg       0.98      0.98      0.98

Confusion Matrix:
 [[210  13]
 [  0 535]]
Accuracy NB Train Dataset:
 0.9828496042216359
```

- Accuracy of the model for test dataset

```
              precision    recall  f1-score

           0       1.00      0.90      0.95
           1       0.95      1.00      0.98

    accuracy                           0.97
   macro avg       0.98      0.95      0.96
weighted avg       0.97      0.97      0.97

Confusion Matrix:
 [[ 97  11]
 [  0 217]]
Accuracy NB Test Dataset:
 0.9661538461538461
```

## 6 Future Scope of the Proposed Model

Although, the proposed model illustrated in the earlier section is based on the Naïve Bayes classifier, an identical experiment may be carried out in future by employing recurrent neural network (RNN) of deep learning methodology in order to measure the prediction accuracy level of the alternately developed model and subsequently a comparison may be drawn on the different evaluation metrics. Moreover, Scope of this model can be extended to design an automated framework that will interact the underlying standards dictionary in sustainable fashion to retrain the model in continuous mode at the same time integrate the intelligence with the signature based security mechanism seamlessly to arrest the zero-day vulnerability of security architecture.

## 7 Conclusion

Rapid evolving of security realm in cyberspace compelled IoC to change its nature significantly. In this research we have identified several issues, challenges and opportunities of threat intelligence and through rigorous survey it has emerged that, cyber threat intelligence platform is still at its infant stage and profound scope still exists to be uncovered. However, threat intelligence is much more organized nowadays with due support of structured standards such as STIX, TAXII and more others. Perhaps, CTI demands more inputs to develop a systematic and streamlined ontology within the cyber threat intelligence.

Integration of artificial intelligence, machine learning with cyber threat intelligence assists to deceive the cyber threat automated and accurately with less computational

time. Here we have developed a machine leaning based model implementing Naive Bayes classifier to extract the potential threat intelligence from structured data source and predict the threat with more than 96% of accuracy level.

## References

1. Ussath, M., et al.: Pushing the limits of cyber threat intelligence: extending STIX. Springer Conference Paper Information Technology New Generations, pp. 213–225 (2016)
2. Ghazi, Y., et al.: A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources. IEEE-2018 International Conference on Frontiers of Information Technology (FIT), pp. 129–134 (2018)
3. Kim, I., et al.: Cyber threat detection based on artificial neural networks using event profiles IEEE Access, **7**, 165607–165626 (2019)
4. Buczak, A.L., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tutor. **18**(2), 1153–1176 (2015)
5. Liu, H., Lang, B.: Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. MDPI (2019)
6. Raad Abbas, A., et al.: Detection of phishing websites using machine learning. Springer Nature Singapore Pte Ltd. 2020, Lecture Notes, vol 1989, pp. 1307–1314 (2018)
7. Bhanu Prakash, B., et al.: An integrated approach to network intrusion detection and prevention using KNN. Springer Nature Singapore Pte Ltd. 2020, Lecture Notes Vol-89, pp. 43–51 (2020)