The 3rd International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2019)
December 17-19, 2019. Phu Quoc, Vietnam

# An Analytical Framework for Automatically Extracting Formal Information from Unstructured Security Intelligence Report

Yuna Hur[1], Chanhee Lee[1], Gyeongmin Kim[1], Kinam Park[1], Heuiseok Lim[1],

[1] Korea University, Department of Computer Science and Engineering,
Anam-Ro. 145, Seongbuk-Gu, Seoul, South Korea
{yj72722, chanhee0222, totoro4007, spknn, limhseok}@korea.ac.kr

**Abstract.** As intrusion attack techniques become more intelligent and sophisticated, security incidents are increasing. In order to predict and respond to cyber attacks, a number of security companies quickly identify the methods, types and characteristics of attack techniques and are publishing Security Intelligence Reports(SIRs) on them. However, the SIRs are not formatted for each security company, and a large number of unstructured SIRs are publishing ever-increasing. In this paper, we propose a framework that uses five analytic techniques to formulate a report and extract key information in order to reduce the time required to extract information on large unstructured reports (SIRs) efficiently.

**Keywords:** Security Threat Information, Keyword Extraction, Topic Modeling, Summarization, Document Similarity, Named Entity Recognition

## 1 Introduction

Many security companies analyze vulnerability and threat information about the latest cyber security incidents and report security information and countermeasures against it[1]. This is called the Security Intelligence Report (SIR), and it is prepared for each quarter and year. Since large quantities of informal reports are generated in this way, it is impossible to unify the SIRs[2,3]. Therefore, it takes a lot of manpower and time to extract key information.

In this paper, we formalize the various file formats of SIR, and propose a method to formalize PDF2TXT considering the problem of transforming from a PDF document to text. Based on this, we propose a system for automatically monitoring and analyzing each SIR using unsupervised learning to extract the formalized information in order to quickly and efficiently provide the information desired by the user. Five different methods are applied to analytical techniques: Keyword Extraction Tool, for extracting important words; Topic Extraction Tool, which probabilistically judges which topic belongs to which SIR; Summarization, which summarizes documents about corresponding SIR in several sentences;, and Document Similarity, to identify similar SIRs to the target SIR. Finally, there is Named Entity Recognition, which

defines the threat information and applies the supervised learning method to automatically extract words judged as threat information from the document.

# 2 System Architecture

## 2.1 Dataset

The data used in this study are publicly available data related to APT (Advanced Persistent Threat) for cyber threat related hackers and incidents, and were collected by crawling PDF documents from 13 websites including github and blog. The SIRs are from 2008 to 2018, and 581 files were used for the study.

## 2.2 SIR Automated Analysis Framework

Many domestic and foreign companies are creating unstructured Security Intelligence Reports (SIR). This paper aims to extract meaningful information from atypical mass SIRs. Fig. When the SIR input comes, the document can be converted into text, and then the document can be identified through the five results using five analysis techniques. SIR extracted text from strings to analyze documents in various file formats. Most of the file format of SIR was PDF, and there were various problems when converting PDF into text. In this paper, we applied the previously studied document conversion (Doc2Txt) technique to reduce the error of document to text conversion [4]. In addition, correct label data must be constructed to extract structured information from unstructured large amounts of SIR. In this paper, the time and cost of data construction is limited, and the keyword extraction, topic modeling, document summary, and similarity document retrieval techniques use the Unsupervised Learning method.

# 3 Result

In this paper, we developed five analytical techniques to extract meaningful information from a large amount of Security Intelligence Report (SIR): Keyword Extraction, Topic Modeling, Summarization, Document Similarity, and Named Entity Recognition(NER). Since there is no label for the model except Named Entity Recognition, it is used to randomly validate the "FTA 1011 Follow up" document in the SIR. As a result of Keyword Extraction, the top five words are "netsat", "netui3", "designateds", "drive", and "setup35". have. Topic Modeling is also categorized into three topics: Topic1 can be classified as security enhancement, Topic2 as network security, and Topic3 as "Attack Type". When entering the "FTA 1011 Follow up" document for verification, Topic 2 was extracted to match the subject matter of the document. The "FTA 1011 Follow up" document used for verification in the Summarization included a description of the Malware system and its solutions, netsat.exe and netui3.dll. When confirming the results of the analysis technique of this model, it was confirmed that the core contents of the

document were included. In addition, when checking the visualization(t-SNE) of Document Similarity, it provided a way to introduce and defend against malware[5]. The third closest SIR, "New_killdisk," also introduces a new attack technique. In the case of NER, the extracted threat information was repeatedly trained 50 epochs of the same system to confirm that it contains various threat information for each word. The standard deviation was 1.16. This is shown in Table 1 below.

**Table 1.** In the 50 epochs, the highest performance figure, lowest performance figure, and average performance figure

| F1-Score | Best Score of 50 Epochs | Worst Score of 50 Epochs | Overall Average Score |
|---|---|---|---|
| Best Score of 50 Epochs | 79.3% | 72.3% | 73.3% |
| Worst Score of 50 Epochs | 77.4% | 66.4% | |

## 4 Conclusion

In this paper, we proposed keyword extraction, topic modeling, document summary, and similarity document search analysis based on unsupervised learning method. In order to evaluate the four automatically extracted techniques, we evaluated them based on randomly selected documents and extracted meaningful information for each analysis technique. In addition, the NER analytical technology established the correct label for SIRs and conducted the supervised learning method, and applied quantitative evaluation. The accuracy of automatically detecting and extracting threat information from the SIR was 73.3%. Through this, SIR can easily process the threat information visually and extract the threat information with high accuracy without any human processing. Through this study, it is expected to be an effective platform for providing security information on security issues because it is easy to search and search through short time SIR by extracting efficient information by using five analysis tools based on large unstructured SIR.

**Acknowledgments.** This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189 ).

## References

1. Alina Opera, Zhou Li, Robin Norris and Kevin Bowers. 2018. MADE: Security Analytics for Enterprise Threat Detection. ACSAC 34, (2018, December). 124-136. DOI: https://doi.org/10.1145/3274694.3274710

2. Endgame. (2016). Using Deep Learning To Detect DGAs. [Online] https://www.endgame.com/blog/technical-blog/using-deep-learning-detect-dgas
3. Amazon. (2018). GuardDuty Intelligent Threat Detection AWS. [Online]. https://aws.amazon.com/guardduty.
4. Y. A Hur, C. H. Lee, G. M. Kim & H. S. Lim. (2019). Topic Automatic Extraction Model based on Unstructured Security Intelligence Report. Journal of the Korea Convergence Society, 10(6), 33-39.
5. L.V. D. Maaten, and G. Hinton. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.