Journal of Digital Convergence
Vol. 17. No. 11, pp. 233-240, 2019
ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.11.233

비정형 Security Intelligence Report의 정형 정보 자동 추출

허윤아[1], 이찬희[1], 김경민[1], 조재춘[2], 임희석[3*]
[1]고려대학교 컴퓨터학과 학생, [2]상명대학교 스마트정보통신공학과 교수,
[3]고려대학교 컴퓨터학과 교수

An Automatically Extracting Formal Information from Unstructured Security Intelligence Report

Yuna Hur[1], Chanhee Lee[1], Gyeongmin Kim[1], Jaechoon Jo[2], Heuiseok Lim[3*]
[1]Student, Division of Computer Science and Engineering, Korea University
[2]Professor, Division of Smart Information Communication Engineering, Sangmyung University
[3]Professor, Division of Computer Science and Engineering, Korea University

요 약 사이버 공격을 예측하고 대응하기 위해서 수많은 보안 기업 회사에서는 공격기법의 특성, 수법 유형을 빠르게 파악하고, 이에 대한 Security Intelligence Report(SIR)들을 배포한다. 하지만 각 기업에서 배포하는 SIR들은 방대하며, 형식이 맞춰져 있지 않다. 본 논문은 대량의 비정형한 SIR들에서 정보를 추출하는데 소요되는 시간을 줄이고 효율적으로 파악하기 위해 SIR들에 대해 정형화하고 주요 정보를 추출하기 위해 5가지 분석기술이 적용된 프레임워크를 제안한다. SIR들의 데이터는 정답 라벨이 없기 때문에 비지도 학습방식을 통해 키워드 추출, 토픽 모델링, 문서 요약, 유사 문서 검색 총 4가지 분석기술을 제안한다. 마지막으로 SIR들에서 위협 정보 추출하기 위해 데이터를 구축하였으며, 개체명 인식 기술에 적용하여 IP, Domain/URL, Hash, Malware에 속하는 단어를 인식하고 그 단어가 어떤 유형에 속하는지 판단하는 분석기술을 포함한 총 5가지 분석기술이 적용된 프레임워크를 제안한다.

주제어 : 보안 위협, 정보 추출, 머신러닝, 딥러닝, 문서 분류

Abstract 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 distributed by each company are huge and unstructured. 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 SIRs efficiently. Since the SIRs data do not have the correct answer label, we propose four analysis techniques, Keyword Extraction, Topic Modeling, Summarization, and Document Similarity, through Unsupervised Learning. Finally, has built the data to extract threat information from SIRs, analysis applies to the Named Entity Recognition (NER) technology to recognize the words belonging to the IP, Domain/URL, Hash, Malware and determine if the word belongs to which type We propose a framework that applies a total of five analysis techniques, including technology.

Key Words : Threat Information, Information Extraction, Machine Learning, Deep Learning, Document Analysis

*This research is supported by Ministry of Culture, Sport and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research&Development Program 2017. (No. R2017030045).
*Corresponding Author : HeuiSeok Lim(limhseok@korea.ac.kr)
Received October 2, 2019 Revised October 30, 2019
Accepted November 20, 2019 Published November 28, 2019

1. 서론

사이버 공격 공격을 대처하기 위해 수많은 보안 기업들은 최신 사이버 보안에 대한 취약점 및 위협정보를 분석하고 실제 기업 내부에서 보안을 어떻게 인식하고 대처하는지에 대한 방법을 보고서로 작성한다[1]. 이를 보안 인텔리전스 보고서(SIR, Security Intelligence Report)라고 명칭 하며, 본 논문에서는 줄여서 SIR라고 명칭한다. SIR들은 매 분기, 연도별로 작성하여 제공한다. 하지만 SIR들은 정형화된 양식이 없으므로 수많은 기업에서 비정형화된 SIR들을 작성하고 있으며 일관성 없는 다양한 형태의 보고서를 확인할 수 있다. 이와 같이 비정형화된 많은 양의 보고서들이 생성되면 SIR들을 통일할 수 없기 때문에 핵심적인 정보를 추출하기 위한 많은 인력과 시간이 필요하다. 또한, 방대한 SIR들에서 사용자가 원하는 문서를 찾는 것도 시간이 많이 소요된다. 이처럼 방대한 비정형 SIR들을 정형화된 정보를 효율적으로 추출하고 분석할 수 있는 도구의 중요성이 주목받고 있다.

본 논문에서는 SIR들에서의 다양한 파일 형식을 정형화된 텍스트로 바꾸기 위해 이전 연구에서 개발한 문서 변환할 수 있는 PDF 문서 변환인 Doc2Txt(Document to Text)를 적용한다[2]. 본 기술은 다양한 확장자를 갖는 문서에 대해 변환이 가능하다. 그 중 PDF 문서에서 텍스트로 변형될 때의 문제점을 기반으로 정형화시킨다. 또한, 본 연구에서 적용한 데이터는 정답 라벨(label)이 존재하지 않기 때문에 비지도 학습(Unsupervised Learning)과 지도 학습(Supervised Learning)을 이용하여 각 SIR에 대해 자동으로 분석하는 프레임워크를 제안하다.

분석 기술에는 5가지 기술을 적용한다. 중요한 단어를 추출하는 키워드 추출 기술, SIR이 어떤 토픽에 속하는지 확률적으로 판단하는 토픽 모델링 기술, SIR에 대해 문서를 요약하여 몇 문장으로 표현하는 문서 요약 기술, SIR과 내용이 비슷한 SIR를 파악하기 위한 유사도 문서 검색 기술이 있으며, 4가지 기술은 정답 라벨이 없기 때문에 비지도 학습을 적용하였다. 마지막으로 위협 정보를 정의하며, SIR에서 위협 정보로 판단되는 단어를 자동으로 추출하는 개체명 인식(Named Entity Recognition) 기술은 정답 라벨을 구축하여 지도 학습 방법을 적용하였다. 제안된 기술을 통해 사용자가 원하는 정보를 빠르고 효율적으로 정보를 추출할 수 있는 분석 프레임워크를 제안한다.

2. 관련 연구

각 분야에서의 문서나 보고서에 관해 분석한 선행연구는 다양하다. 의학 분야에서 보고서를 분석하는 tool도 있으며, Saeed Hassanpour et al.(2017)은 무릎 MRI 촬영 결과에 대한 의사 소견의 데이터를 기반으로 정상인지 비정상인지를 자동으로 판별하는 모델을 제안하였다[3]. 무릎 MIR 의사소견서의 용어 및 패턴을 파악하기 위해 NLP 기술을 사용하였고 이를 정상 또는 비정상으로 분류하기 위해 머신러닝 기법인 SVM(Support Vector Machine)을 사용하였다. 보안에서도 기계학습 기법을 적용하여 공격 log를 분석하는 등 관련된 많은 연구가 있다. 하지만 보안 분야에서는 정해진 명칭에 대해 탐지하고 분석하는 선행연구가 많다. Alina Oprea et al.(2018)은 악의적인 공격인 malware 의 다양성으로 인해 기업은 백신, 방화벽 등을 통해 공격을 탐지며, web proxy log를 분석하여 MADE 시스템을 통해 공격을 탐지하고 위험 정도에 따라 우선순위를 지정한다[4]. MADE 시스템은 기계학습을 활용하여 악의적인 domain을 예측한 확률을 기반으로 우선순위를 매긴다[4]. 이 외에도 자동화 분석 모델은 악성 도메인 탐지 분석[5], 비정상적인 계정 액세스 탐지[6] 등과 같은 보안 응용 프로그램에 적용되었다. 본 논문은 대량의 보안 인텔리전스 보고서를 쉽게 파악하고 분석하기 위해 보안 특성상 분석하기 어려운 부분을 처리하기 위해 전처리 과정에 집중하였으며, 기계학습과 딥러닝을 통해 SIR들을 분석하고 이에 따라 효율적으로 원하는 보고서를 검색하고 찾을 수 있도록 초점을 맞추었다.

3. SIR 자동 분석 프레임워크

위협정보에 대해 국내·외 수많은 기업에서는 비정형된 SIR(Security Intelligence Report)들을 작성하고 있다. 본 논문에서는 비정형한 대량의 SIR들에서 유의미한 정보를 추출하는 것을 목표로 한다. Fig. 1.과 같이 SIR 입력이 들어오면 문서를 텍스트로 변환한 후 5가지 분석 기술을 사용하여 5가지 결과를 통해 문서를 파악할 수 있다. SIR은 다양한 파일 형식으로 이루어져 있는 문서들을 분석하기 위해 문자열로 이루어진 본문을 추출하였다. 그중 대부분 SIR의 파일 형식은 PDF였으며, PDF를 텍스트로 변형할 때 다양한 문제점이 있었다. 본 논문에서는 이전에 연구한 문서 변환(Doc2Txt) 기술를 적용하

여 문서에서 텍스트로 변환하는 오류를 줄였다[2]. 또한, 비정형한 대량의 SIR에서 정형한 정보를 추출하기 위해 정답 라벨(label) 데이터를 구축해야 한다. 본 논문에서는 데이터 구축하는 시간과 비용이 제한되어 키워드 추출, 토픽 모델링, 문서 요약, 유사도 문서 검색 기술은 비지도 학습(Unsupervised Learning) 방법을 적용하였으며, 개체명 인식(Named Entity Recognition) 모델의 경우 데이터를 구축하여 지도학습(Supervised Learning) 방법을 이용한 모델을 제안한다.

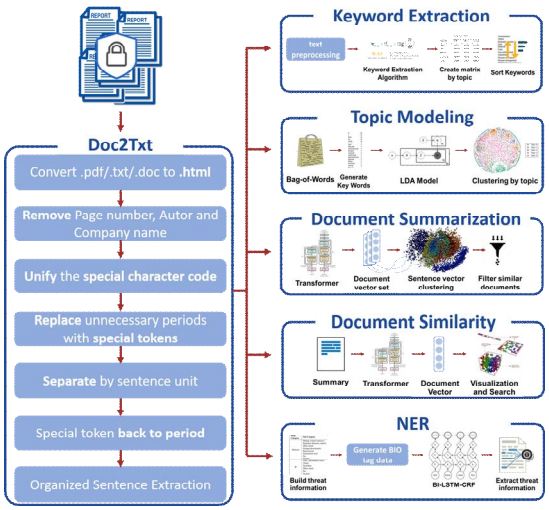

Fig. 1. An overal frame work for efficiently extracting key information about a SIRs

3.1 DataSet

본 논문에 사용된 데이터는 사이버 위협 관련 해커나 사건들에 대한 APT(Advanced Persistent Threat)와 관련된 것으로 공개적으로 사용할 수 있으며, github, 블로그 등 총 13개의 웹사이트에서 PDF 문서를 크롤링하였다. 이에 대한 보안 인텔리전스 보고서는 2008년부터 2108년까지의 파일이고 581개의 파일을 연구에 사용하였다. 추가로 FireEye에 있는 SIR 36개의 문서는 일부 실험 평가를 위해 사용하였다.

3.2 키워드 추출 및 토픽 모델링 분석 기술

본 논문에서 SIR들에 대한 자동 분석 프레임워크를 생성하기 위해 이전 연구에서 개발한 키워드 추출과 토픽 모델링을 적용하였다. 토픽 모델링이란 전체 문서 집합에서 n가지 주제를 추출한 후 새로운 문서가 입력될 때 어떤 주제에 속하는지 판단하는 확률적 통계모델이다. 본 모델은 총 600개의 SIR을 활용하여 문서 내 단어를 추출한 후 추출된 토큰(Token)을 분석 목적에 맞게 가공하였다. 모든 SIR을 토큰화하여 가공된 토큰은 약 20,879,712개였으며, 그중 중복되지 않고 유일한 토큰 수는 약 250,226개이다. 토큰화된 문서 집합에서 후보 키워드를 추출하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 모델을 이용하여 단어에 대한 가중치 계산을 통한 중요도를 평가한 상위 10개의 키워드를 추출하였다[7]. 토픽 모델링에서는 전체 문서에 대한 단어 사전을 생성하며, 이를 위해 Bag-of-Words를 활용하여 출현 빈도와 단어 간의 관계를 고려하여 상위 5000개 단어에 대한 단어 사전을 생성하였다. 생성된 단어 사전과 SIR을 Latent Dirichlet Allocation 모델을 적용하여 토픽 수가 3개, 10개, 15개일 때 각 토픽에 대한 해당 단어를 추출하여 적합한 토픽의 개수를 정하였다[8].

3.3 문서 요약 분석기술

3.3.1 딥러닝을 활용한 문장 벡터 생성

본 논문에서 활용한 SIR들에 대한 데이터가 적기 때문에 적은 양의 데이터로 문장 임베딩을 생성하는 효과적인 모델인 전이 학습(Transformer Learning)을 적용하였다[9,10,11]. 이는 도메인과 상관없이 다양한 문장에 대한 임베딩 벡터를 생성하기 위해 정답 라벨이 있는 Stanford Natural Language Inference (SNLI) Corpus와 정답이 없는 Web News, 위키피디아(Wikipedia), Web QA(Question-Answer)를 학습시켰으며, 성능을 향상시켰다. 인코더는 소문자 Penn Treebank 토큰화를 입력받아 512차원의 문장 임베딩 벡터를 생성한다[12]. 문장 임베딩은 의미론적 텍스트 유사성에서 문장 수준의 의미 유사성 점수를 계산하는데 사용된다. 본 연구에서는 매우 큰 딥러닝 모델을 사용하므로 대량의 메모리가 요구된다. 하지만 필요에 따라 batch_size를 조절함으로써 메모리 사용량을 변경할 수 있으며, 작은 batch_size는 메모리 사용량을 줄이는 대신 실행 속도를 약간 증가시킨다.

3.3.2 문장 클러스터링

본 논문의 문장 선택 단계에서는 유사한 의미의 문장 그룹을 찾기 위해 비지도 군집화 알고리즘인 K-means를 사용한다. K-means는 벡터 공간 모델에서 각 문장을 나타낸다. 따라서 각 문서는 자질 벡터로 표현된다. 본 논문에서는 앞서 기술된 전이 학습 기반 문장 임베딩 생성

모델로 문장 벡터로 구성한다. K-means는 중심(centroids)을 기반으로 하며, 이는 그룹 내 객체들의 평균으로 계산된 벡터 공간 모델상의 점이다. K-means 알고리즘은 각 객체를 가장 가까운 중심에 할당하고 새로운 중심을 다시 계산하는 과정을 반복함으로써 진행되며, 중심에 변화가 없으면 최종 군집화 결과가 나타난다. 이후 요약을 생성하기 위해 각 그룹에서 가장 대표적인 문장을 선택한다.

3.3.3 최적의 요약문 길이 선정

최적의 요약문 길이는 문서의 길이 및 종류에 따라 다를 수 있다. 문서의 길이가 너무 길 때 요약이 불충분하거나, 문서의 길이가 짧을 때 필요 이상으로 긴 요약을 생성하게 된다. 본 논문에서 문서 요약에 사용된 K-means 알고리즘의 결과에 따라 요약문의 길이를 정해야 한다. 따라서 본 논문에서는 최적의 요약문 길이를 판단하기 위해 K-means 알고리즘의 K의 개수를 변경하기 위해 Silhouette 점수를 적용하여 해당 문서에 맞게 요약을 수행하는 기능을 제안한다. K의 값은 최소 5개부터 최대 전체 문서의 10%까지의 범위 내에서 탐색 된다.

3.4 문서 유사도 검색

문서를 잠재 공간에 임베딩하면 벡터 공간 모델을 이용하여 문서 간의 유사도를 수치화할 수 있고, 이로부터 유사 문서 검색이 가능해진다. 본 논문에서는 문서 임베딩을 생성하기 위해 문서 요약에서 사용한 문장 임베딩 기술을 바탕으로 문서 요약 기술에 응용하였으며, t-SNE 알고리즘을 이용하여 문서 임베딩 공간의 차원을 축소하고, 이를 시각화하여 지능적으로 유사 문서를 탐색하는 방법을 제안한다[13]. 문서 공간의 차원 축소 및 시각화에는 텐서플로우의 Tensorboard 패키지를 이용하였다. TensorBoard는 웹 기반 UI를 기본 제공하며, 본 논문에서 각각의 점은 하나의 문서를 나타내며, 예시에서는 수집된 SIR들에 대해 594건을 시각화하였다.

3.5 Named Entity Recognition 분석 기술

SIR들을 기반으로 지도 학습(Supervised Learning)을 이용하여 비정형 위협정보를 자동으로 태깅하여 추출하는 기술 제안한다. 본 논문에서는 지도 학습 방법을 하기 위해 학습 데이터가 필요했으며, 보안학과에 재학 중인 5명의 연구원을 통해 태깅 데이터를 구축하였다. 아래 Table 1과 같이 태깅 class의 종류는 대분류는 총 4개로 구성되어 있으며 IP, Domain/URL, HASH, Malware가 있고 총 하위는 20개로 구성하였다. 아래 표는 SIR에서 추출해야 할 위협정보이다. 총 구축된 SIR 데이터는 총 608건이다. 하지만 수작업으로 데이터를 태깅하였기 때문에 학습 데이터는 일관성이 부족하며, 오류가 존재하였다. 본 논문에서는 위협정보 추출 데이터 후처리와 정규식을 통해 모델의 성능을 향상시켰다.

Table 1. Definition of tag category in security

Main Category	Sub Category	Tag Name
IP	Attack	ip.attack
	C&C, information source	ip.cncsvr
	DDos attack	ip.ddos
	Distribute	ip.distribute
	Normal	ip.normal
	Etc.	ip.unkown
Domain/URL	Codevia	url.Codevia
	C&C, information source	url.cncsvr
	Distribute	url.distribute
	Normal	url.normal
	Etc.	url.unknown
Hash	Hash	Hash
Malware	Backdoor(Remote control)	malware.backdoor
	Dropper(downloader)	malware.drop
	DDos attack	malware.ddos
	Information steal	malware.infosteal
	Mining(virtual currency)	malware.mining
	Ransomware	malware.ransom
	Normal	malware.normal
	Etc.	malware.unknown

3.5.1 학습 데이터와 후처리

태깅은 대분류 4개 소분류 20개로 이루어져 있으며, 문장 내에서 문맥 흐름상 고유한 의미를 나타내는 부분의 시작단어(Beginning, B-tag)로 표기하며 중간단어(Inside, I-tag)로 표기한다. 그 외의 태그 정보가 없는 단어는 Outside로 O라고 표기한다. 이와 같이 문서에 대한 위협정보를 정의하였으며, 총 SIR 608건의 data를 구축하였다. 하지만 Tagging data 구축에서 참여 인력의 실수로 B-tag와 I-tag 불일치, 오타 등의 문제점이 있었다. 이를 자동으로 파악하여 효과적으로 수정하기 위해 각 위협정보의 태그 빈도를 분석하였으며 등장빈도가 1% 미만인 tag들은 unknown으로 통합하였으며, 0.1% 미만인 것은 제거하였다. 그 결과 최종 tag 집합은 총 10

개로 적용하여 BIO tag는 총 21개로 구성되어 있다. 또한, 성능향상을 위해 regular expression을 통해 태그할 수 있는 CVE를 구분하여 tag를 추가하였다.

3.5.2 딥러닝을 활용한 위협정보 추출

본 논문에서는 deep learning 모델인 BI-LSTM(Long Short-Term Memory)-CRF(Conditional Random Field) network를 사용하였다[14]. BI-LSTM-CRF는 양방향 네트워크로 forward의 input feature와 backward의 input feature를 통해 학습하며 모든 hidden state를 거치며 양방향으로 학습한다. BI-LSTM-CRF에 들어가는 input feature의 단어 벡터에서 오타로 작성된 단어나 없는 단어는 unknown 벡터로 제외되기 때문에 오타가 없는 전체 단어로 했을 때 보다 성능이 떨어진다. 본 논문에서는 성능을 향상하고자 고빈도 단어를 n-gram으로 토큰화하여 빈도가 높은 단어에 대한 단어 벡터를 생성해주는 문자 단위의 feature를 추출하는 [15] 모델을 활용하였다. 추출된 feature는 input data로 들어가서 학습 데이터로 사용되었으며, 각 epoch에서 전체 학습 데이터로 mini-batch로 나누고 한 step에 한 개를 처리하도록 하였다. 그리고 BI-LSTM에 CRF layer를 추가하여 output 값에 따른 gradient 값을 계산하였다.

4. 실험 결과

본 논문에서는 사이버 위협과 관련된 SIR들을 기반으로 5가지 분석을 진행하였다. 5가지 분석기술 중 NER을 제외한 키워드 추출, 토픽 모델링, 문서 요약, 문서 유사도 검색을 비지도 학습기반으로 실험하였다. 비지도 학습으로 실험된 분석기술을 평가하기 위해 무작위로 SIR에서 “FTA 1011 Follow up” 문서를 검증하는 데 사용한다. 이 보고서는 일부 악성코드 기능에 대해 세부적으로 설명하고, 본사 구성 요소의 C2 기능에 대한 확장 된 세부 정보를 제공하며, 이 악성 프로그램의 방어적 탐지(netsat.exe, netui3.dll) 수단을 추가로 제공하며 공격기법인 malware에 대해 설명하고 있다. 본 문서를 통해 아래 5가지 분석기술을 평가 기준으로 잡았으며, 실험 결과는 아래와 같이 분석하였다.

4.1 키워드 추출 분석기술 실험 결과

키워드 추출 분석기술에서 약 580개의 SIR를 학습하였으며, 모든 SIR에 대해 정답 set이 존재하지 않아 비지도 학습 방법을 통해 개발하였다. 각 SIR마다 10개의 키워드를 추출하였으며, 검증하기 위해 “FTA 1011 Follow up”을 입력으로 사용하였다. 추출된 결과는 Table 2와 같다. 검증에 사용된 문서에 대한 상위 10개 키워드 추출 결과를 확인해보면, malware의 방어 방법인 netsat과 netui3은 상위 1, 2를 차지하는 것을 확인할 수 있다.

Table 2. Top 10 keywords extraction results for input document

Rank	Keyword
Top 1	netsat
Top 2	netui3
Top 3	designates
Top 4	drive
Top 5	setup35
Top 6	headquarters
Top 7	recycled
Top 8	directory
Top 9	exe
Top 10	copies

4.2 토픽 모델링 분석기술 실험결과

이전 연구를 기준으로 3가지 토픽을 기준으로 추출하였으며, 해당 토픽에 포함되는 중요 단어 7개씩 출력하였다. 그 결과는 아래 Table 3과 같다. Topic 1에 표현된 단어들을 통해 유추해 봤을 때 “보안 강화”, “암호화”, “시스템 보안”를 표현한 것을 확인할 수 있다. Topic 2에서는 leak, 보안 자격증(CCP), Relative Virtual Address 등을 통해 “네트워크 보안”으로 추측할 수 있다. Topic 3에서는 detect, criticism, 악성코드 종류 같은 단어를 통해 “공격 종류”로 유추할 수 있다.

Table 3. Related words for each topic

Topic	words
Topic 1	inconsistencies, ncw, subvert, sockets, nprotect, researcher, sentinelone
Topic 2	leaks, uploader, decompression, domain, volatile, cpp, cleanup
Topic 3	vpnfilter, ahead, counterparts, detect, host, criticism, website

위와 같이 토픽을 분류하였고 이에 맞게 총 581개 SIR을 학습하였다. 또한 비지도 학습 방법으로 훈련한 토픽 모델링을 평가하기 위해 “FTA 1011 Follow up”를

넣고 각 토픽 중 어떤 토픽에 더 가까운지 나타낸 수치를 시각화하였다. 시각화한 결과는 아래 Fig. 2와 같다.

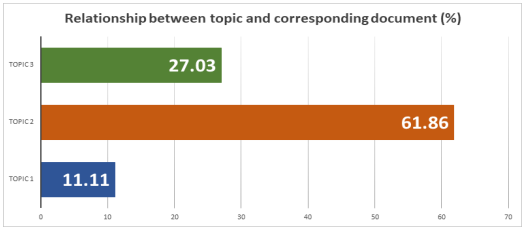

Fig. 2. Topic classified for test SIR

본 연구의 실험 결과는 테스트 보고서를 넣었을 때 Topic 1는 약 11%, Topic 2는 약 61%, Topic 3는 약 27%로 이 중에 Topic 2에 토픽에 포함된다는 것을 알 수 있다. 테스트에 쓰인 SIR는 악성코드 종류, 랜섬웨어와 같은 공격기법에 속하므로 본 결과가 어느 정도 일치하는 것을 보인다.

4.3 문서 요약 실험 결과

문서 요약 은 Wikipedia, SNLI 등을 바탕으로 훈련하여 모든 SIR에 대한 문장 임베딩 벡터를 추출하여 중요한 문장들을 요약하였다. “FTA 1011 Follow up”에 대한 요약된 결과는 아래와 같다.

Table 4. Summarized results for “FTA 1011 Follow UP” document

Document Summarization
The malware system consists of at least two Portable Executable (PE) files, one acting as a headquarters component and one acting as field unit or agent component. Analysis of the system relied on the availability of two files named netsat.exe and netui3.dll. However, using the malware 's behavior and determining the command file 's format via reverse engineering afforded the ability to test numerous assumptions about the malware 's intended use. designates any connected drive retrieve a directory listing designates a volume serial number collect data harvested from a targeted system did not execute This continues to suggest that intruders either have local or remote access to headquarters systems running netsat.exe or access to another application that automates remote C2 data/file retrieval.

하지만 각 SIR에서 요약된 정답 set이 존재하지 않으므로 비지도 학습 방법을 적용하였기 때문에 정량적인 평가가 어렵다. 검증에 사용된 “FTA 1011 Follow up” 문서의 내용에서 Malware 시스템에 대한 설명과 그에 대한 방안인 netsat.exe와 netui3.dll에 대한 설명이 포함되었다. 본 model의 분석 기법의 결과를 확인하였을 때 문서의 핵심 내용이 포함된 것을 확인하였다.

4.4 유사 문서 검색 실험 결과

SIR들에서 각 SIR에 대하여 문서 임베딩 벡터로 변환하여 공간 모델을 적용하여 다차원 공간에 임베딩하였다. 이를 t-SNE 모델을 적용하여 공간을 축소하고 Tensorflow로 시각화 하였다. 이는 문서 간의 유사도를 수치화 할 수 있고 이로부터 유사 문서 검색이 가능하다. 아래 그림 3에서 각 점은 하나의 SIR를 의미하며, 이와 유사한 문서들이 유사도에 따라 내림차순으로 정렬되어 우측에 표시된다. Fig. 3와 같이 문서를 선택하면 유사 문서를 확인할 수 있다.

Fig. 3. Visualize and Search Similar Document Tools

유사 문서 검색 및 시각화의 경우 정답 set이 존재하지 않아 비지도 학습 방법을 적용하였기 때문에 정량적 평가가 어렵다. 검증을 위해 “FTA 1011 Follow up” 문서를 입력으로 넣었을 때 그림 3과 같이 상위 2개의 문서는 FTA와 관련된 SIR가 가까이 있었다. 또한 본 테스트 문서에 대한 내용을 보았을 때 malware에 대해 소개하고 방어할 수 있는 방안을 제공하였는데 3번째로 가까운 SIR인 “New_killdisk”에 대한 내용도 새로운 공격기법에 대한 소개하고 있다. 이와 같이 많은 문서를 학습시키지 않았으므로 SIR간의 거리가 있으며, 0.5정도 거리에 있는 경우 SIR의 비슷한 유형의 문서가 있는 것으로 확인하였다.

4.5 NER 실험 결과

SIR에서 비정형 위협정보를 정의하였으며, 이를 자동으로 인식하고 추출하였다. 지도 학습 방법으로 총 608개의 태그된 SIR를 학습하였다. SIR에서 비정형 위협 정보의 각 태그에 대한 성능을 아래 표와 같이 확인하였다.

Table 5. The precision and recall for each tag

Tag	Precision	Recall
url.unknown	67.7%	82.3%
url.normal	76.9%	39.4%
url.cncsvr	89.1%	34.7%
malware.unknown	63%	59%
malware.ransom	87.5%	77.8%
malware.infosteal	86.9%	71%
malware.drop	89.1%	72.5%
malware.backdoor	83.6%	74.9%
ip.unknown	85.8%	92.4%
hash	94.4%	91.3%
CVE	100%	100%

추출된 위협정보는 단어마다 다양한 위협정보를 포함할 수 있으므로 accuracy만으로 평가하기 어렵다. 성능을 평가하기 위해 NER에서 정량적 평가 방식인 F-Score를 적용하였다.

Deep Learning 시스템의 특성상, 랜덤으로 초기화되는 파라미터 값에 따라 모델의 성능의 차이가 있다. 따라서 본 논문에서는 동일한 시스템을 50 epoch 반복해서 학습시키고, 50 epoch학습된 모델의 F-Score 최종 성능에 대해 평균 값은 73.3%이였으며, 표준편차는 1.16이였다.

Table 6. In the 50 epochs, the highest performance figure and the lowest performance figure and the 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%	

5. 결론

본 논문에서 비지도 학습 방법을 기반으로 키워드 추출, 토픽 모델링, 문서 요약, 유사도 문서 검색 분석기술을 제안하였다. 자동으로 추출된 4가지 분석 기법을 평가하기 위해 랜덤으로 선택한 문서를 기반으로 평가하였으며, 각 분석 기법에 맞게 의미 있는 정보를 추출하였다. 또한 NER 분석기술은 SIR들에 대해 정답 라벨(label)을 구축하여 지도학습 방법으로 진행하였으며, 정량적인 평가를 적용하였다. SIR에서 자동으로 위협정보를 인지하고 추출하는 정확도는 73.3%의 결과가 나왔다. 이를 통해 SIR에서 위협정보를 시각적으로 쉽게 처리하며 사람이 직접 처리하지 않아도 높은 정확도의 위협정보 추출이 가능하다. 본 연구를 통해 대량의 비정형 SIR을 기반으로 5가지 분석 도구를 활용하여 효율적인 정보를 추출을 통해 탐색 및 검색에 용이하고 짧은 시간 SIR를 파악할 수 있으므로 보안 이슈에 대해 효율적인 정보 제공의 발판이 될 것을 기대한다.

REFERENCES

[1] M. E. Kuhl, J. Kistner, K. Costantini & M. Sudit. (2007). Cyber attack modeling and simulation for network security analysis. In Proceedings of the 39th Conference on Winter Simulation, 1180-1188.

[2] 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. DOI : 10.15207/JKCS.2019.10.6.033

[3] S. Hassanpour, C. P. Langlotz, T. J. Amrhei, N. T. Befera & M. P. Lungren. (2017). Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology, 208(4),* 750-753. DOI: 10.2214/AJR.16.16128

[4] A. Opera, Z. Li, R. Norris & K. Bowers. (2018). MADE: Security Analytics for Enterprise Threat Detection. *In Proceedings of the 34th Annual Computer Security Applications Conference* (pp. 124-136). ACM. DOI: 10.1145/3274694.3274710

[5] Endgame. (2016). Using Deep Learning To Detect DGAs. [Online] https://www.endgame.com/blog/technical-blog/using-deep-learning-detect-dgas

[6] Amazon. (2018). GuardDuty Intelligent Threat Detection AWS. [Online]. https://aws.amazon.com/guardduty.

[7] J. H. Hur, J. H. Choi, J. H. Lee, J. B. Kim, & K. W. Rim. (2001). An Automatic Classification System of Korean Documents Using Weight for Keywords of Document and Word Cluster. *The KIPS Transactions: PartB, 8(5),* 447-454.

[8] T. K. Kim, H. R. Choi, & H. C. Lee. (2016). A Study on the Research Trends in Fintech using Topic Modeling. *Journal of the Korea Academia-Industrial cooperation Society, 17,* 11, 670-681. DOI: 10.5762/KAIS.2016.17.11.670

[9] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, & A. Bordes. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv preprint *arXiv:1705.02364.*

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L.

Jones, A. N. Gomez, Ł. Kaiser, & I. Polosukhin. (2017). Attention is all you need. *In Advances in neural information processing systems* (pp. 5998-6008).

[11] L. V. D. Maaten & G. Hinton. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9(Nov),* 2579-2605.

[12] D. Cer et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175.*

[13] L. V. D. Maaten & G. Hinton. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9(Nov),* 2579-2605.

[14] Z. Huang, W. Xu & K. Yu. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991.*

[15] C. H. Lee, Y. B. Kim, D. Y. Lee, & H. S. Lim. (2018). Character-Level Feature Extraction with Densely Connected Networks. *In Proceedings of the 27th International Conference on Computational Linguistics* (August. 2018). pp. 3228-3239. Conference Name:ACM Woodstock conference

허 윤 아(Yuna Hur) [정회원]

- 2016년 : 백석대학교 정보보호학과(공학학사)
- 2016년 ~ 현재 : 고려대학교 컴퓨터학과 석박사 통합과정
- 관심분야 : 인공지능, 자연어처리, 딥러닝, 정보추출
- E-Mail : yj72722@korea.ac.kr

이 찬 희(Chanhee Lee) [학생회원]

- 2016년 : 서강대학교 컴퓨터공학심화(학사)
- 2016년 ~현재 : 고려대학교 컴퓨터학과 석박사 통합 과정
- 관심분야 : 인공지능, 자연어처리, 딥러닝
- E-Mail : chanhee0222@korea.ac.kr

김 경 민(Gyeongmin Kim) [학생회원]

- 2017년 : 백석대학교 정보통신학부(공학학사)
- 2018년 ~ 현재 : 고려대학교 컴퓨터학과 석박사 통합과정
- 관심분야 : 딥 러닝, 자연어처리
- E-Mail : totoro4007@korea.ac.kr

조 재 춘(Jaechoon Jo) [정회원]

- 2010년 2월 : 제주대학교 컴퓨터교육과(이학사)
- 2012년 2월 : 고려대학교 컴퓨터교육과(이학석사)
- 2018년 2월 : 고려대학교 컴퓨터과(공학박사)
- 2018년 3월 ~ 2019년 2월 : 고려대학교 연구교수
- 2019년 3월 ~ 현재 : 상명대학교 공과대학 조교수
- 관심분야 : 컴퓨터교육, 자연어처리, 인공지능
- E-Mail : jae@smu.ac.kr

임 희 석(HeuiSeok Lim) [정회원]

- 1992년 : 고려대학교 컴퓨터학과(이학학사)
- 1994년 : 고려대학교 컴퓨터학과(이학석사)
- 1997년 : 고려대학교 컴퓨터학과(이학박사)
- 2008년 ~ 현재 : 고려대학교 정보대학 컴퓨터 학과 교수
- 관심분야 : 자연어처리, 인공지능, 기계학습, 정보검색
- E-Mail : limhseok@korea.ac.kr