Journal of the Korea Convergence Society
Vol. 10. No. 6, pp. 33-39, 2019
ISSN 2233-4890
https://doi.org/10.15207/JKCS.2019.10.6.033

비정형 보안 인텔리전스 보고서 기반 토픽 자동 추출 모델

허윤아[1], 이찬희[1], 김경민[1], 임희석[2*]
[1]고려대학교 컴퓨터학과 학생, [2]고려대학교 컴퓨터학과 교수

Topic Automatic Extraction Model based on Unstructured Security Intelligence Report

YunA Hur[1], Chanhee Lee[1], Gyeongmin Kim[1], HeuiSeok Lim[2*]
[1]Student, Department of Computer Science and Egineering, Korea University
[2]Professor, Department of Computer Science and Egineering, Korea University

요 약 지능형 사이버 공격 기법이 다양화됨에 따라 보안 침해 사건, 글로벌 범죄 등의 사건 발생이 증가하고 있다. 지능형 공격을 예측하고 대응하기 위해서는 공격 기법의 특성, 수법, 유형을 파악해야 한다. 이를 위해 수많은 보안 기업 회사에서는 다양한 공격 기법을 빠르게 파악하고 더 큰 피해를 막기 위해 보안 인텔리전스 보고서를 배포한다. 하지만 각 기업에서 배포하는 보고서에 대한 형식이 맞춰져 있지 않으며, 대량의 비정형 보안 인텔리전스 보고서가 배포되고 있다. 본 논문은 비정형한 보안 인텔리전스 보고서에 대한 문제점을 고려하여 정형화된 데이터로 추출하는 방안을 제안한다. 또한, 대량의 보안 인텔리전스 보고서를 파악하기 위해 소요되는 시간을 줄이고자 대량의 보고서를 주제별로 분류할 수 있는 보안 인텔리전스 보고서 토픽 자동 추출 모델을 제안한다.

주제어 : 보안, 인텔리전스 보고서, 분석시스템, 토픽 모델링, 분류 시스템

Abstract As cyber attack methods are becoming more intelligent, incidents such as security breaches and international crimes are increasing. In order to predict and respond to these cyber attacks, the characteristics, methods, and types of attack techniques should be identified. To this end, many security companies are publishing security intelligence reports to quickly identify various attack patterns and prevent further damage. However, the reports that each company distributes are not structured, yet, the number of published intelligence reports are ever-increasing. In this paper, we propose a method to extract structured data from unstructured security intelligence reports. We also propose an automatic intelligence report analysis system that divides a large volume of reports into sub-groups based on their topics, making the report analysis process more effective and efficient.

Key Words : Security, Intelligence Report, Analysis, Topic Modeling, Classification

1. 서론

최근 내·외부의 침입 공격 기법이 다양화되고 지능화됨에 따라 침해사고 발생이 증가하고 있다. 이와 같은 잦은 해킹 공격 때문에 정보의 저장, 수집, 검색, 송·수신을 진행하는 도중 정보의 유출, 변조 및 훼손이 되고 있다. 이러한 빈번한 정보보안 사고는 회사, 금융권, 방송국 등을 목표로 공격하며, 이로 인해 내부자의 정보유출뿐만

*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 April 24, 2019 Revised May 31, 2019
Accepted June 20, 2019 Published June 28, 2019

아니라 금전적 손실 및 기업 이미지가 하락할 수 있다.

이러한 공격을 대처하기 위해 수많은 보안 기업들은 최신 사이버 보안에 대한 취약점 및 위협 정보를 분석하고 실제 기업 내부에서 보안을 어떻게 인식하고 대처하는지에 대한 방법을 보고서로 작성한다[1]. 이를 보안 인텔리전스 보고서(Security Intelligence Report)라고 명칭 하며, 보안 인텔리전스 보고서는 매 분기, 연도별로 작성하여 제공한다. 하지만 보안 인텔리전스 보고서는 정형화된 양식이 없으므로 수많은 기업에서 비정형화된 보안 인텔리전스 보고서를 작성하고 있으며 다양한 형태의 보고서를 확인할 수 있다.

이와 같이 비정형화된 많은 양의 보고서들이 생성되면 보안 인텔리전스 보고서를 통일할 수 없기 때문에 핵심적인 주제를 추출하기 위해 많은 인력과 시간이 필요하다. 또한, 방대한 보안 인텔리전스 보고서에서 사용자가 원하는 문서를 찾는 것도 시간이 많이 소요된다. 이처럼 방대한 보안 인텔리전스 보고서가 있을 때 이 문서들이 어떤 주제를 갖는지 효과적으로 추출하고 해당 주제를 대표할 수 있는 단어로 표현할 수 있는 토픽 모델링의 중요성이 주목받고 있다[2,3].

토픽 모델링(Topic Modeling)이란 비정형 데이터 중 텍스트로 이루어진 문서에 대해 각 문서의 내용이 어떤 주제들을 포함하고 있는지 파악할 수 있는 확률 모델이다. 토픽 모델링에서 의미론적 분석을 위한 방법으로 대표적인 알고리즘 기법인 LDA(Latent Dirichlet Allocation)이 있다. LDA 알고리즘은 한 문서에 나타난 단어들이 어떤 단어들의 군집 속에 있는가에 따라 잠재된 주제를 파악한다[4,5]. 토픽 모델링은 정보, 네트워크, 텍스트 마이닝과 같은 자료에서 유의미한 구조를 파악할 때 유용하게 사용된다.

본 논문에서는 PDF 파일로 만들어진 보안 인텔리전스 보고서를 바탕으로 연구를 진행하였으며, 이에 본 논문은 PDF 내의 텍스트를 추출할 때의 문제점을 고려하여 정리된 문서를 생성하고 LDA(Latent Dirichlet Allocation)알고리즘을 활용하여 효율적으로 주제를 추출한다.

본 논문은 다음과 같이 구성되어 있다[6]. 2장에서는 토픽 모델링에 대하여 정의한다. 3장에서는 비정형 보안 인텔리전스 보고서를 정형화하기 위해 사용했던 방법에 대해 논한다. 4장에서는 3장에서 데이터 설명 및 정형화된 보안 인텔리전스 보고서를 기반으로 전처리 과정을 설명하며 토픽 자동 추출 모델에 대해 논한다. 대표적으로 Bag-of-Words와 LDA (Latent Dirichlet Allocation)를 적용하는 방법을 살펴본다. 5장에서는 본 연구에서 진행된 모델에 대한 실험 결과를 논하였으며, 6장에서는 본 논문에 대해 결론 및 향후 연구 방향에 대해 논한다.

2. Topic Modeling

토픽 모델링은 텍스트 기반의 방대한 문서 내에서 맥락과 관련된 주제를 찾아내는 알고리즘이며, 결과적으로 n개의 주제 중에 각 문서가 어떤 주제 포함되는지의 비율과 주제에 포함된 단어들의 분포를 알 수 있다[2]. 이러한 특징의 토픽 모델링은 다양한 분야의 문헌 자료를 기반으로 연구의 분석 도구로 사용되어 왔다[7]. 토픽 모델링에서 사용되는 대표적인 알고리즘은 Blei et al의 Latent Dirichlet Allocation 알고리즘으로 각 어휘가 어떤 어휘들의 군집에 포함되어 있는지에 따라 주제를 추론하는 확률 기반의 알고리즘이다[8]. 강범일 et al.은 신문 기사를 기반으로 토픽 모델링 기법을 이용하였으며, 신문 기사에 보도되는 기업이나 기관의 정치적, 사회적 이슈 및 기업 마케팅에 대해 분석하고 주제를 파악하여 신문 기사 매체별 보도의 정파성을 분석하였다[7]. 김형지 et al.(2018)은 스마트 폰 중독과 관련된 뉴스 기사를 2010년부터 2018년 3월까지 수집하여 토픽을 추출하였고, 스마트 폰 중독과 관련된 주요 토픽은 12개를 추출하였으며, 년도별 사회적 이슈가 어떻게 변화하였는지도 살펴보았다[9]. 또한, Nektaria Potha et. al.는 저자 검증 방법(Author Verification Method)에서 토픽 모델링을 모델의 일부로 사용하여 내·외부 문서 기반 프로파일 저자 검증 시스템을 보여준다[10]. 따라서 본 논문은 비정형화된 보안 인텔리전스 보고서를 정형화하여 토픽 모델링 기법을 통해 토픽별 보안 분야를 정의한다.

3. 연구 방법

본 연구에서는 국내·외 수많은 보안 기업들에서 작성하는 영문 보안 인텔리전스 보고서를 분석 대상으로 하였으며, 다양한 파일의 보안 인텔리전스 보고서 중 PDF 문서가 대부분이었다. 본 논문에서는 제안하는 방법은 txt, 한글, word 문서도 변환할 수 있으며, 그 중 변환의 제약이 많은 PDF 문서들을 분석하였다. 분석을 진행하기 위해서는 문자열로 이루어진 본문을 각 PDF 파일로부터 추출하는 과정이 선행되어야 한다.

3.1 PDF 문서 변환

PDF 문서는 단락, 문장, 본문 등의 구분이 없으며, 각 글자의 글씨체, 크기와 위치 정보만 담겨 있다. 따라서 PDF 문서를 분석하여 텍스트를 일관성 있게 추출하고, 기계학습 모델에 사용할 수 있도록 이를 문장 단위로 구분하고 토큰화하는 과정이 선행되어야 한다.

본 연구에서는 문자열로 이루어진 PDF 문서들을 분석하기 위해 각 PDF 문서마다 TXT 파일로 추출하는 연구를 진행하였다. 하지만 PDF 문서에서 추출할 때 이미지 내에 있는 글자의 처리는 본 연구 범위에서 벗어나므로 PDF 내에 순수 텍스트로 존재하는 내용만을 추출의 대상으로 연구하였다. 즉, 본 연구는 PDF 문서를 입력으로 받아 해당 문서의 본문을 토큰화된 문장 단위로 추출하였다.

3.1.1 PDF 텍스트의 노이즈 제거

PDF 문서 내에 존재하는 페이지 번호, 주석, 워터마크 등이 텍스트 추출 시 노이즈로 작용한다. Fig. 1과 Table 1과 같이 PDF 문서의 일부와 이를 단순히 텍스트로 추출했을 시의 결과물 예시가 있다. 예시에서 확인할 수 있는 바와 같이, 쪽 번호, 글자 크기, 위치 등이 무시되며, 줄 구분도 문장 단위가 아닌 단순히 PDF 상의 위치에 따라 이루어져 있다. 이러한 문제들로 인해 PDF로부터 단순히 텍스트만 추출하는 방법으로는 PDF의 본문을 효과적으로 파악하기 어렵다.

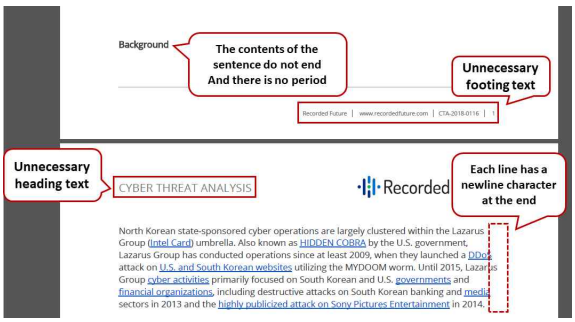

Fig. 1. Example of problematic PDF file when extracting text

Table 1. When a PDF document is simply extracted as text

Background **Recorded Future ⏐ www.recordedfuture.com ⏐ CTA-2018-0116 ⏐ 1** **CYBER THREAT ANALYSIS** North Korean state-sponsored cyber operations are largely clustered within the Lazarus Group (Intel Card) umbrella. Also known as HIDDEN COBRA by the U.S. government, Lazarus Group has conducted operations since at least 2009, when they launched a DDoS attack on U.S. and South Korean websites utilizing the MYDOOM worm. Until 2015, Lazarus Group cyber activities primarily focused on South Korean and U.S. governments and financial organizations , including destructive attacks on South Korean banking and media sectors in 2013 and the highly publicized attack on Sony Pictures Entertainment in 2014.

본 연구에서는 이를 극복하기 위해 우선 PDF 문서를 HTML 문서로 변환하였다. HTML로 변환된 문서에는 글씨체 및 글씨 크기 등의 정보가 담겨 있으므로, 이를 바탕으로 노이즈를 줄일 수 있다. 또한, HTML의
 태그를 통해 더 상세한 줄 바꾼 정보를 파악할 수 있다.

문서 내 노이즈를 제거하기 위해 크게 3가지 문제점이 있었다. 첫 번째는 머리말 또는 꼬리말, 쪽 번호, 문서 작성자 등 문서 내에서 텍스트로 인식되는 문제였다. 문서 내에 있는 텍스트를 제외하기 위해 html 코드를 확인하고 전체 문서에서 일반적인 글자 크기가 아닌 것이 5% 미만일 때의 문자들을 제거하였다. 또한 다양한 PDF 파일 형식으로 작성하다 보니 똑같은 문자임에도 불구하고 다른 문자로 인식하는 문제점이 있었다. 이는 유니코드 문자표를 정리하여 함수처리를 통해 하나의 문자로 통일하였다.

두 번째 문제점은 문장별로 줄 구분이 되어있지 않았고 한 줄마다 줄 바꿈이 들어가 있어 문장의 경계를 알 수 없었다. html에서
은 줄 바꿈을 의미하는데 새로운 문단의 시작일 수 있지만 같은 문장이 다음 쪽으로 이어지는 문장일 수도 있다. 또한 하이픈(-)은 공간상 긴 단어를 둘로 분리할 때 쓰이기도 하지만 하이픈이 포함된 단어도 있다. 이를 해결하기 위해 문장 끝의 단어와 그다음 줄 첫 단어를 이어 쓰고, 띄어 쓰고, 하이픈도 넣어봤을 때 영문 위키피디아에 있는 단어의 등장 빈도를 비교하여 적용하였다.

마지막으로 PDF 파일은 문장 경계 구분이 없다. 위 Fig. 1과 같이 “Background” 뒤에 마침표가 없고 줄 바꿈만 포함되어 있다. 이처럼 제목, 부제목 등에는 마침표 없이 줄 바꿈으로 다음 단락과 구분이 안 되는 경우가 있다. 또한, 보안 인텔리전스 보고서의 특성상 URL과 IP 주소가 많이 나오며 마침표가 포함되어 있으므로 URL과 IP 주소를 정규식을 통해 구분하여 해당 마침표를 다른 기호로 사용한 다음 마침표 기반 문장 경계 인식 모델을 활용하여 문장의 끝을 구분하였다. 구분된 결과는 다음 Table 2와 같다.

Table 2. This is an example of extracting the same PDF document by the method developed in this task

Background North Korean state-sponsored cyber operations are largely clustered within the Lazarus Group (Intel Card) umbrella . Also known as HIDDEN COBRA by the U.S. government , Lazarus Group has conducted operations since at least 2009 , when they launched a DDoS attack on U.S. and South Korean websites utilizing the MYDOOM worm . Until 2015 , Lazarus Group cyber activities primarily focused on South Korean and U.S. governments and financial organizations , including destructive attacks on South Korean banking and media sectors in 2013 and the highly publicized attack on Sony Pictures Entertainment in 2014 .

4. 보안 인텔리전스 보고서를 기반한 토픽 자동 추출 모델

보안 위협을 줄이기 위해 다양한 기업에서 비정형화된 인텔리전스 보고서를 작성하고 있으며, 본 연구에서는 PDF 파일로 작성된 비정형화된 보안 인텔리전스 보고서를 분석하여 TXT 파일로 변환하였다. 이와 같이 비정형화된 보안 인텔리전스 보고서를 활용하여 유의미한 구조를 발견하고 주제별로 분류할 수 있는 모델을 제안한다.

4.1 분석 데이터

본 연구에 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. 본 연구에서 사용된 보안 인텔리전스 보고서는 총 581개의 문서이다. 추가로 FireEye에 있는 보안 인텔리전스 보고서를 사용하였다. FireEye는 2015년부터 공격 그룹인 APT37를 추적하며 그 활동에 대한 세부 정보를 작성하여 TEMP.Reaper라는 명칭으로 보안 인텔리전스 보고서를 작성하고 있으며, 이에 관해 본 연구에서는 36개의 문서 중 일부는 실험 평가를 위해 사용하였다.

4.2 데이터 전처리

본 연구에서는 영문으로 작성한 보안 인텔리전스 보고서를 활용하였다. 영문 전처리에 다양한 방법이 있는데 숫자나 문장 부호 제거하거나 불용어 제거, 누락/제거된 표현 복원 등 여러 가지 방법을 사용할 수 있다. 이 중 본 연구에서는 영문 전처리를 하기 위해 대/소문자 구분이 필요 없으므로 대문자를 소문자로 통일하였다. 또한 문장 부호를 제거하였으며 불용어 처리도 진행하였다. 불용어란 문장 내에서 빈번하게 등장하지만, 의미를 구성하거나 문장을 분석하는 데 있어 큰 영향을 미치지 않을 필요 없는 어휘들을 의미한다[11]. 본 연구에서는 문서의 있는 단어를 정확하게 예측하는 모델링이므로, “is”, “an”, “not”, “the” 등과 같은 불용어를 처리하였다.

4.3 단어 모음

단어 모음(Bag-of-Words)는 문서에 포함된 단어들의 순서를 고려하지 않고, 문맥과 단어들의 관계 및 출현 빈도(Frequency)를 고려하여 단어들의 집합을 저장하고 관리하는 방법이다[12,13]. 본 연구에서는 전체 문서를

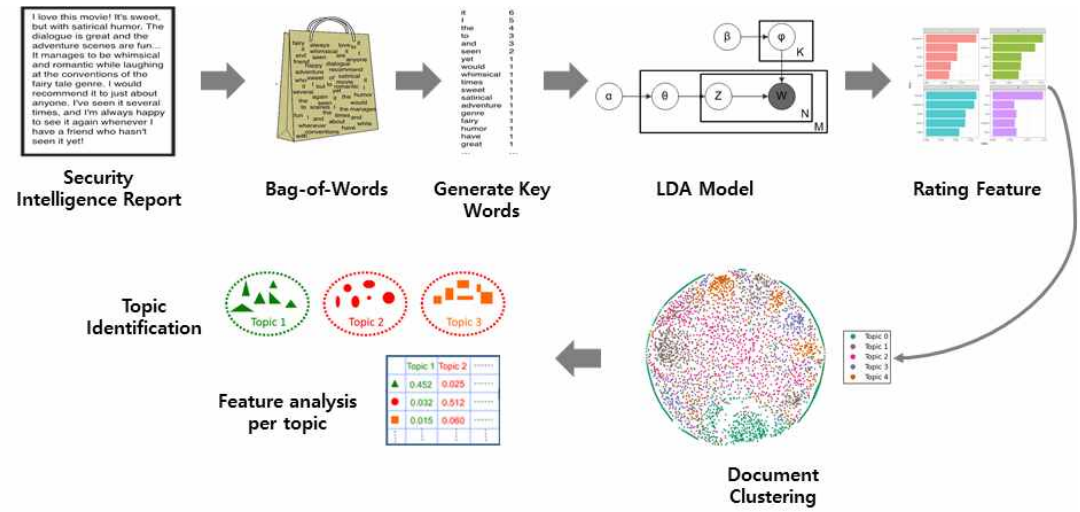

Fig. 2. Topic Modeling based on Security Intelligence Report

전처리된 단어들로 분리했으며, 단어 빈도의 수를 파악하여 벡터(vector)로 만드는 CountVectorizer를 사용하여 단어 모음집을 생성하였다. 본 모델에서는 전체 문서의 단어 모음집을 이용하여 상위 5000개의 단어를 단어 모음집으로 생성하여 모델을 진행하였다.

4.4 토픽 모델링

토픽 모델링(Topic Modeling)이란 문서 집합에서 텍스트의 숨겨진 의미구조를 발견하기 위한 확률적 통계모델 중 하나로 추상적인 주제를 발견하기 위한 텍스트마이닝 기법의 하나다. 토픽 모델링에서 대표적인 확률 분포 알고리즘인 LDA(Latent Dirichlet Allocation) 알고리즘이 있다[14].

LDA 알고리즘은 Unsupervised Generative Topic Model로 토픽 집합을 가정하고, 문서 내 단어들을 확률적으로 계산하여 숨겨져 있는 주제들을 예측하는 알고리즘이다[15]. 토픽 모델은 주제별로 포함되는 특정 단어와 이를 기반으로 해당 문서가 어떤 주제에 포함되는지 분포를 확률을 통해 생성 모델을 만들면, 특정 문서가 들어왔을 때 해당 문서가 어떤 토픽에 해당하는지 확률을 계산할 수 있다.

본 연구는 LDA 알고리즘을 기반으로 전체 문서에서 전처리된 단어 모음집을 생성하였으며, 단어 모음집을 기반으로 LDA 모델의 학습 절차에 따라 학습하였다. 첫 번째로 전체 문서에서 나올 수 있는 토픽 수를 임의로 선정하였으며, 단어 모음집을 기반으로 해당 토픽과 관련된 단어들을 배치하였다. 두 번째로 하나의 문서에서 각각의 단어와 단어 모음집을 비교하여 해당 단어의 빈도수를 구하였다. 마지막으로 각 문서에서 단어 모음집과 일치하는 단어들의 빈도수와 디클레어(Dirichlet) 분포 확률을 이용하여 해당 문서에 맞는 각 토픽 별 확률이 나타난다.

5. 보안 인텔리전스 보고서를 기반한 토픽 자동 추출 모델링 실험 결과

본 연구는 사이버 위협과 관련된 보안 인텔리전스 보고서를 기반으로 토픽 추출을 진행하였다. 실험에 사용된 문서 파일은 모두 PDF 형식이며, 실험에 사용하기 위해 PDF 파일 내의 텍스트를 추출하여 정리한 TXT 파일 총 581개의 보안 인텔리전스 보고서를 학습하였다. 테스트에 사용된 보안 인텔리전스 보고서는 FireEye에 있는 보안 인텔리전스 보고서 중 택하여 실험하였다.

본 연구에서는 토픽을 임의로 3개를 선정하였으며 해당 토픽에 포함되는 중요 단어들의 결과를 10개씩 출력하였으며, 아래 Table 3와 같다. Topic 1에서는 ncw를 통해 Network Centric Warfare를 유추하며 군사력 연결을 통해 효율적인 전략을 나타내며, socket을 통해 암호화 프로토콜을 예측할 수 있고, 바이러스, 툴 회사 등을 통해 “암호화”, “보안 강화”“시스템 보안”를 하기 위한 토픽인 것을 확인할 수 있다. Topic 2에서는 노출, Relative Virtual Address, 보안 자격증 등을 통해 “네트워크 보안”으로 추측할 수 있다. Topic 3에서는 감지, 랜섬웨어 용어, 악성코드 종류, 비판, 결과 같은 단어를 통해 “공격”과 관련된 문서 내용을 유추할 수 있다.

Table 3. Topic by bag-of- words

Topic	words
Topic 1	ncw, sockets, recommend, subvert, nprotect, alpha , reader, inconsistencies, sentinelone, researcher
Topic 2	leaks, uploader, rva, cleanup, decompression, domain, repurpose, volatile, offshore, cpp
Topic 3	ahead, detect, efficiency, counterparts, host, website vpnfilter, criticism, outcomes, applying

위와 같이 토픽을 분류하였고 이에 맞게 총 581개 보안 인텔리전스 보고서를 학습하였다. 또한, 본 연구에서는 보안 인텔리전스 보고서를 기반을 둔 토픽 모델링을 테스트하기 위해 FireEye의 보고서 중 “APT17_Report.pdf”을 넣고 각 토픽 중 어떤 토픽에 더 가까운지 나타낸 수치를 시각화하였다. 시각화한 결과는 아래 Fig. 3과 같다.

본 연구의 실험결과 Fig. 3과 같이 테스트 보고서를 넣었을 때 Topic 0(Topic 1)는 약 11%, Topic 1(Topic 2)는 약 68%, Topic 2(Topic 3)는 약 22%로 이 중에 Topic 1(Topic 2)에 토픽에 포함된다는 것을 알 수 있다.

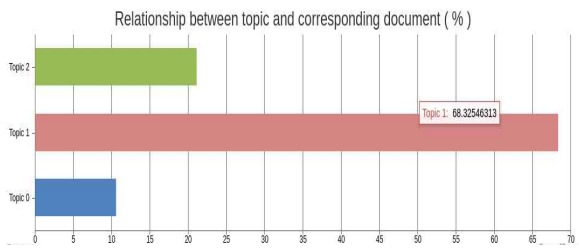

Fig. 3. Result of putting test document in Topic Modeling

본 연구는 각 보안 인텔리전스 보고서마다 주제에 대한 정답 셋(set) 없이 Unsupervised Learning으로 진행된 모델로서 정량적인 평가 방법 대신 만족도 평가를 통해 모델을 평가하였다. 1점에서 5점(1점: 매우 불만족, 2점: 불만족, 3점: 보통, 4점: 만족, 5점: 매우 만족)까지의 리커트 척도를 사용하여 계산하였다. 아래 Table 4는 보안 인텔리전스 보고서 토픽 자동 추출 모델에 대한 평가 문항이다.

Table 4. Security Intelligence Report Topic Automatic Extraction Model Satisfaction Evaluation Question

Number	Contents
1	Do you think the number of topics is properly classified?
2	Do the words for the classified topic seem appropriate?
3	Is the topic on this Security Intelligence Report appropriate?
4	How similar does the Security Intelligence Report match the topic results?

본 논문에서는 4명의 사람에게 랜덤으로 평가를 진행하였으며, 각 질문당 점수를 매기고 전체 질문 평가에 대한 평균을 내었다. Table 5는 평가 결과이다. 보안 인텔리전스 보고서 토픽 자동 추출 모델 만족도는 5점 만점에서 평균 4.06으로 적합한 결과가 나왔다.

Table 5. Security Intelligence Report Topic Automatic Extraction Model satisfaction

Person	Security Intelligence Report Topic Automatic Extraction Model satisfaction
Person 1	4
Person 2	4.5
Person 3	3.5
Person 4	4.25
Average	4.06

6. 결론

지능화된 다양한 공격 기법이 증가함에 따라 공격으로 인해 피해도 함께 증가하고 있다. 각 보안 기업에서는 피해를 줄이는 대응 방안으로 공격 기법을 파악하여 해결 방안을 제시하는 보안 인텔리전스 보고서를 배포하고 있다. 하지만 수많은 기업에서 배포한 보안 인텔리전스 보고서는 양식이 정해져 있지 않기 때문에, 대량의 비정형 보안 인텔리전스 보고서가 발행된다.

본 논문은 PDF 파일 형식 기반으로 비정형된 보안 인텔리전스 보고서를 크롤링하여 데이터를 수집하였으며, 이를 정형화된 파일로 변환하는 방안을 제안하였다. 또한 대량의 보고서를 수작업으로 토픽을 분류하기 위해 시간이 많이 소모되기 때문에 Bag-of-Words와 LDA 기법을 활용한 보안 인텔리전스 토픽 자동 추출 모델을 실험하였다. 본 모델은 정답 셋이 없으므로 만족도 조사를 통해 평가하였으며, 5점 만점에 4.06으로 좋은 평가를 얻을 수 있었다.

본 연구를 통해 개발된 보안 인텔리전스 보고서 토픽 자동 추출 모델을 통해 보안 대응 방안을 좀 더 효율적으로 정보를 얻을 수 있는 발판이 될 것을 기대한다.

REFERENCES

[1] S. Y. Lee. (2018. 06. 18). Microsoft Announces Cyber Security Threat Report. *News of SecuN,* p. 1.

[2] T. K. Kim &H. R Choi &H. C. Lee. (2016). A Study on the Research Trends inFintech using Topic Modeling. *Journal of the Korea Academia-Industrial cooperation Society, 7(11),* 670-681.
DOI :10.5762/KAIS.2016.17.11.670

[3] L. Hong & B. D. Davison. (2010, July). Empirical study of topic modeling in twitter. *In Proceedings of the first workshop onsocial media analytics(ACM),* 80-88.

[4] N. C. Ho .(2016). An Illustrative Application of Topic Modeling Method to a Farmer's Diary. *INSTITUTE OFCROSS-CULTURAL STUDIES, 22(1),* 89-135.

[5] R. Krestel, P. Fankhauser & W. Nejdl. (2009, October). Latentdirichlet allocation for tag recommendation. *In Proceedings of the third ACM conference on Recommender systems,* 61-68.

[6] Y. A Hur, D. Y. Lee, K. K. Kim, W. H. Yu & H. S. Lim. (2017). A System for Automatic Classification of Traditional Culture Texts. *Journal of the Korea Convergence Society, 8(12),* 39-47.

[7] B. I. Kang, M. Song, W. Jho. (2013). A Study on Opinion Mining of News paper Texts based on Topic Modeling. *Journal of The Korean Society For Library And Information Science, 47(4),* 315-334.

[8] J. H. Bae, N. G. Han & M. Song (2014). Twitter Issue Tracking System by Topic Modeling Techniques. *Journal of Intelligence and Information System, 20(20),* 109-122.

[9] H. G Kim, S. U. Kim & S. T. Kim. (2018). Topic Modeling of Media Reports on Smartphone Addiction

– A Study on the Comparison of Government Policies between 2010 and 2018. K*orean Association for Braodcasting & Telecommunication Studies, 104,* 38–62.

[10] N. Potha & E. Stamatatos. (2019). Improving author verification based on topic modeling. *Journal of the Association for Information Science and Technology, 0(0),* 1–15.
DOI :10.1002/asi.24183

[11] H. H. Gill. (2018) The Study of Korean Stopwords list for Textmining, URIMALGEUL: *The Korean Language and Literature, 78,* 1–25.

[12] H. M. Wallach. (2006). Topic modeling: beyond bag–of–words. *In Proceedings of the 23rd international conference on Machinelearning(ACM),* 977–984.

[13] J. Yang, Y. G. Jiang, A. G. Hauptmann & C. W. Ngo. (2007). Evaluating bag–of–visual–words representations in scene classification. *In Proceedings of the international workshop on Workshop on multimedia information retrieval(ACM),* 197–206.

[14] D. M. Blei, A. Y. Ng & M. I. Jordan. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research, 3(Jan),* 993–1022.
DOI: 10.1162/jmlr.2003.3.4.–5.993

[15] Y. Guo, S. J. Barnes & Q. Jia. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation, *Tourism Management, 59*, 467–483.

허 윤 아(YunA Hur) [정회원]

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

이 찬 희(Chanhee Lee) [정회원]

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

김 경 민(Gyeongmin Kim) [정회원]

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

임 희 석(HeuiSeok Lim) [종신회원]

· 1992 고려대학교 컴퓨터학과(이학학사)
· 1994 고려대학교 컴퓨터학과(이학석사)
· 1997 고려대학교 컴퓨터학과(이학박사)
· 2008 ~ 현재 : 고려대학교 정보대학 컴퓨터 학과 교수
· 관심분야 : 자연어처리, 뇌신경 언어정보 처리
· E–Mail : limhseok@korea.ac.kr