JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 7, pp. 99-107, July 2024
https://doi.org/10.9708/jksci.2024.29.07.099

Analysis of trends in information security using LDA topic modeling

Se Young Yuk*, Hyun-Jong Cha**, Ah Reum Kang*

*Student, Dept. of Information Security, Pai Chai University, Daejeon, Korea
**Professor, Dept. of Software Engineering, Pai Chai University, Daejeon, Korea
*Professor, Dept. of Information Security, Pai Chai University, Daejeon, Korea

[Abstract]

In an environment where computer-related technologies are rapidly changing, cyber threats continue to emerge as they are advanced and diversified along with new technologies. Therefore, in this study, we would like to collect security-related news articles, conduct LDA topic modeling, and examine trends. To that end, news articles from January 2020 to August 2023 were collected and major topics were derived through LDA analysis. After that, the flow by topic was grasped and the main origin was analyzed. The analysis results show that ransomware attacks in 2021 and hacking of virtual asset exchanges in 2023 are major issues in the recent security sector. This allows you to check trends in security issues and see what research should be focused on in the future. It is also expected to be able to recognize the latest threats and support appropriate response strategies, contributing to the development of effective security measures.

▸ **Key words**: cyber security, LDA model, time series regression analysis, LDAvis, word cloud

[요 약]

컴퓨터 관련 기술이 급변하는 환경에서 사이버 위협들은 새로운 기술과 함께 고도화되고 다양화되어 지속해서 등장하고 있다. 이에 본 연구에서는 보안 관련 뉴스 기사를 수집해서 LDA 토픽 모델링을 진행해 동향을 살펴보고자 한다. 이를 위해 2020년 1월부터 2023년 8월까지의 뉴스 기사를 수집하였으며 LDA 분석을 통해 주요 토픽을 도출하였다. 이후 토픽별 흐름을 파악하고 주요 기점에 대해 분석하였다. 분석 결과를 통해 2021년의 랜섬웨어 공격과 2023년의 가상자산거래소 해킹이 최근 보안 분야에서 큰 이슈인 것을 파악할 수 있다. 이를 통해 보안 이슈에 대한 동향을 확인하고, 앞으로 어떤 연구에 집중해야 하는지 확인해 볼 수 있다. 또한 최신 위협을 인지하고, 적절한 대응 전략을 지원할 수 있으며 효과적인 보안 대책의 개발에 기여할 것으로 기대된다.

▸ **주제어**: 정보보안, LDA 모델링, 시계열 회귀 분석, LDAvis, 워드클라우드

• First Author: Se Young Yuk, Corresponding Author: Ah Reum Kang
*Se Young Yuk (dbrtpdud01@naver.com), Dept. of Information Security, Pai Chai University
**Hyun-Jong Cha (hjcha@pcu.ac.kr), Dept. of Software Engineering, Pai Chai University
*Ah Reum Kang (armk@pcu.ac.kr), Dept. of Information Security, Pai Chai University
• Received: 2024. 05. 08, Revised: 2024. 07. 02, Accepted: 2024. 07. 02.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업 혁명 이후, 인공지능(AI, Artificial Intelligence)과 사물인터넷(IoT, Internet of Things) 등의 기술들이 발전하고 있다. 이러한 기술들의 발전은 우리의 삶과 밀접하게 연결되어 있기 때문에 최근 개발된 ICT(Information and Communications Technologies) 기술들에 의해 다양한 이슈에 직면하게 된다. 그중에서도 정보보안 문제가 큰 쟁점이 되고 있다. 정보보안 분야에서 위협 요인으로 작용하는 것은 데이터 유출, 보안 취약점, 피싱 그리고 다양한 공격 패턴 등 여러 가지가 있다[1].

정보보안 동향과 관련된 논문으로는 논문의 키워드를 활용해 정보보안과 관련해 글로벌 연구 트렌드를 분석하는 연구[2]와 보안의 여러 분야 중에서 산업 보안의 동향을 분석한 연구[3] 등이 있다. 하지만 최근 몇 년간 실질적인 정보보안 분야의 동향에 대한 분석이 부족한 상황이다. 이에 본 논문에서는 최근 뉴스 기사를 활용해 정보보안의 최근 트렌드에 대해 분석해 볼 것이다. 이러한 동향 분석 연구는 진행 중인 연구가 올바른 방향으로 이루어지고 있는지 확인하기 위해 필수적인 부분이다. 또한 정보보안 문제들이 어떻게 변화하고 있는지 파악한다면 더욱 정확한 방향성을 가지고 연구를 진행할 수 있을 것이다.

뉴스는 현재의 이슈를 신속하고 정확하게 전달해 주는 주요 매체로써 현재 사회의 다양한 동향을 분석하는 데 필수적인 정보를 제공해 준다. 또 날짜별로 이루어진 자료이므로 특정 분야의 동향을 분석하거나 사회적 흐름, 트렌드를 확인하기 좋은 데이터이다. 이에 따라 뉴스를 이용해 동향을 분석하는 연구들이 늘어나고 있다[4][5].

본 연구에서는 뉴스를 데이터로 사용해 정보보안 분야의 동향을 분석하려고 한다. 2020년 1월 01일부터 2023년 8월 31일까지의 정보보안 분야의 뉴스 기사를 수집하여 LDA(Latent Dirichlet Allocation, 잠재 디리클레 할당) 토픽 모델링을 활용하여 동적 토픽 모델링을 수행하여 대표적인 토픽을 산출한다. 이를 통해 최근 정보보안의 이슈를 유형화한다. 또한 시계열 회귀 분석을 통해 정보보안 분야의 토픽별 동향을 확인해 보고자 한다.

본 연구의 구성은 2장에서 연구에 대한 이해를 위해 LDA 토픽 모델링과 적정 토픽 수 결정에 관해 설명한다. 3장에서는 연구 수행 과정을 설명하고 4장에서 연구 결과를 기술한다. 마지막으로 5장에서 결론과 향후 연구 계획에 관해 서술한다.

II. Preliminaries

1. LDA Topic Modeling

토픽 모델링은 문서의 주제를 추론하는 통계적인 모델링의 기법으로 텍스트 마이닝 기법 중 하나이다. 토픽 모델링은 LDA(Latent Dirichlet Allocation), LSA(Latent Semantic Analysis), HDP(Hierarchical Dirichlet Process) 등 다양한 알고리즘 및 분석기법이 있다. LDA는 다양한 주제로 구성된 문서들을 주제별 단어의 분포를 추론하는 데 사용하며 주로 뉴스 기사, 학술논문 데이터, 고객 리뷰 데이터를 분석하여 주요 주제를 파악할 때 사용한다. LSA는 문서 간의 유사성을 분석하고, 쿼리와 관련된 문서를 검색할 때 사용한다. HDP는 새로운 문서가 추가되는 데이터를 동적으로 주제를 조정해야 할 때 사용한다. HDP는 LDA와 달리 데이터 내부의 구조와 패턴을 유연하게 파악할 수 있다. 텍스트 데이터의 구조나 분석하려는 목적에 따라 각 특징에 맞는 알고리즘을 사용한다.

LDA 토픽 모델링은 토픽 모델링의 대표적인 알고리즘이다. 다양한 주제를 가진 문서들에서 각 문서의 토픽 분포와 각 토픽의 단어 분포를 확률 분포에 기반하여 토픽을 추출하는 확률적 생성 모델이다. 최근 동향을 분석하는 연구에서 LDA 토픽 모델링을 많이 사용하고 있다. 스마트시티 관련 학술 연구 동향 분석하는 연구에서 LDA 토픽 모델링을 이용하였고[6], 운송 시스템의 특허 기술 동향 분석 등의 최근 동향을 분석하는 연구에서도 LDA 토픽 모델링을 사용하였다[7].

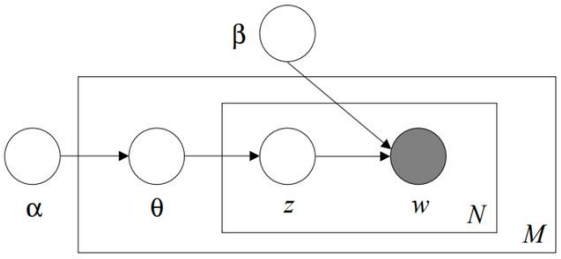

Fig. 1. Graphical model representation of LDA

Fig. 1은 LDA의 그래픽 모델 표현이다. 모델에서 알파(α)와 베타(β)를 포함한 바깥쪽 판은 문서를 나타낸다. 그리고 M과 N을 포함한 안쪽 판은 문서 내에서 반복되는 주제와 단어의 선택을 나타낸다. M은 전체 문서의 개수를 나타내고, N은 문서에 포함된 단어의 총개수를 나타낸다. 알파(α)는 토픽들의 분포에 대한 파라미터로, 각 문서가 어떤 토픽에 집중될지를 결정한다. 베타(β)는 토픽들의 단어 분포를 나타내며, 각 토픽이 어떤 단어들로 이루어져

있는지를 결정한다. 세타(θ)는 문서의 토픽 분포를 나타내며, 문서마다 각 토픽에 속할 확률 분포를 나타낸다. z는 LDA 모델에서 사용되는 변수이며 단어가 어떤 토픽에 속하는지 나타낸다. w는 문서 내의 특정 위치에 있는 단어를 나타내며, 이 변수는 각 단어가 어떤 토픽에 속하는지를 나타낸다[8].

2. Determining the number of topics

토픽 모델링을 수행할 때 가장 중요한 것은 적정 토픽 수를 정하는 것이다. 적정 토픽 수를 결정하는 방법으로는 혼잡도(perplexity), 응집도(coherence)가 있다[9]. 혼잡도를 활용해 토픽 수를 결정하는 방법은 2015년 8월에 European Journal of Operational Research (EJOR)에서 제안된 방법이다[10]. 혼잡도는 모델이 주어진 데이터를 얼마나 잘 설명하는지를 나타내는 척도이다. perplexity 값이 낮을수록 더 좋은 모델임을 의미한다. 국내 지역 지리교육 연구 동향 분석을 위해 LDA 토픽모델링을 진행했을 때 Perplexity 값을 활용하였다[11]. 응집도인 Coherence Score를 활용해서 토픽 수를 결정하는 방법은 2010년 06월에 북미 계산언어학 협회(NACCL, North American Conference on Chinese Linguistics) 연례 학회 처음 제안된 방법이다[12]. 이는 주어진 토픽이 얼마나 의미 있는지 측정하고 각 토픽의 단어들이 얼마나 서로 관련성이 있는지를 나타낸다. Coherence Score가 높을수록 해당 토픽이 의미 있는 주제로 이루어져 있음을 나타낸다. 뉴스 기사를 통해 치매 관련 신체활동에 대한 이슈 분석[13]과 스마트 제조에 대한 특허 기술 동향의 분석[14]과 같은 연구에서 적정 토픽 수를 결정할 때 Coherence Score를 사용하여 토픽 모델링을 진행하였다.

본 논문에서 토픽 모델링을 진행하기 위해 적정 토픽 수를 Perplexity와 Coherence Score를 사용하여 결정하였다.

3. LDAvis

LDAvis는 토픽과 용어의 관계를 파악해 시각화하여 해석하는 방법으로 2014년 12월에 Journal of the Royal Statistical Society Series A(Statistics in Society)에서 제안된 방법이다[15]. LDAvis의 레이아웃은 왼쪽에는 각 토픽을 나타내는 원과 오른쪽에는 용어 막대그래프가 있다. 원의 크기는 토픽의 유 빈도를 나타내며 크기가 클수록 해당 토픽이 더 중요하고 빈도가 높다. 막대그래프는 선택된 토픽을 해석하는데 가장 유용한 용어를 나타내며 용어의 상대적 빈도와 특정 토픽에서의 빈도를 나타낸다. 막대그래프에서 빨간 막대는 특정 토픽에서 용어의 빈도를 나타내고 회색 막대는 전체 토픽에서 용어의 빈도를 나타낸다. LDAvis에서 사용되는 매개변수는 람다(λ)로 용어의 중요성을 조절하는 역할을 해준다. 람다 값이 1에 가까울수록 빨간 막대가 넓어지고 회색 막대가 좁아진다. 반대로 람다 값이 0에 가까울수록 빨간 막대가 좁아지고 빨간 막대가 넓어진다. 소비자학 분야의 연구 동향을 분석하여 LDA 알고리즘을 사용한 후 시각화하기 위해 LDAvis를 사용했다[16]. 국내 기록관리학의 연구 동향을 파악하기 위해 토픽모델링을 수행하였고 시각화 도구인 LDAvis로 토픽별 거리를 가시적으로 표현하였다[17].

III. The Proposed Scheme

1. System Overview

Fig. 2는 제안하는 연구의 구성도를 보여준다. 본 논문에서는 보안 관련 뉴스를 데이터로 활용하였다. 데이터는 2020년 01월부터 2023년 08월까지 모든 언론사의 뉴스를 수집하였다. 수집한 전체 데이터는 13,390건이다. 수집한 데이터는 필터링과 전처리 과정을 거쳐 분석에 사용된다.

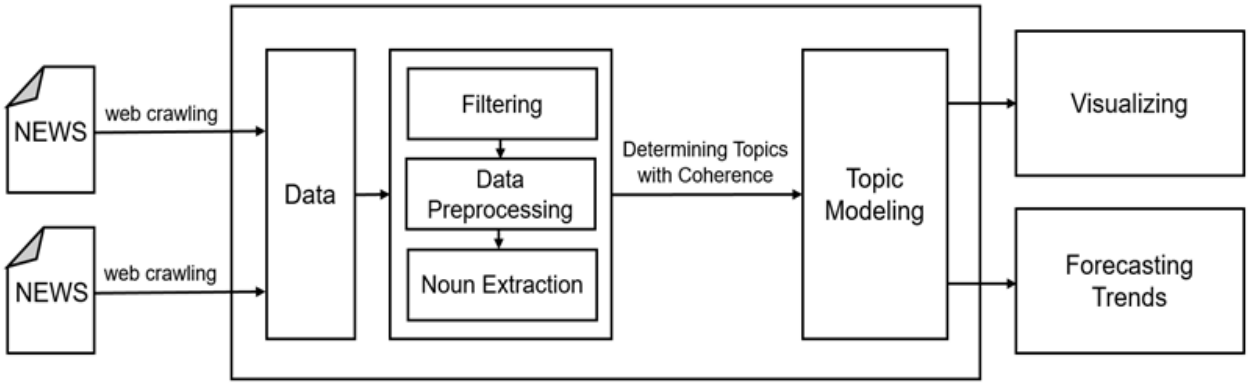

Fig. 2. System configuration diagram

필터링은 중복된 데이터와 비어있는 데이터는 제외하였다. 또한 전처리 과정을 거쳐 2글자 이상의 명사만 추출하였다. 이후 270개 이상의 기사에 등장하면서 기사 전체에서 80퍼센트 이하를 차지하는 단어만 남기도록 하였다. 최종적으로 추출된 874개의 단어를 이용하여 분석을 수행한다. 마지막으로 LDA 모델링을 이용해 토픽 모델링을 진행하였고 LDAvis를 이용해 시각화하였다. 이를 통해 정보보안에 대한 최근 동향을 파악하였다.

2. Experiment

2.1 Experimental Environment

실험을 위해 사용한 컴퓨터 환경과 라이브러리의 버전은 Table 1과 같다. Windows 10 기반으로 개발을 위해 Python 3.9.12를 사용하였고, LDA 모델 학습을 위해 Gensim 4.3.2 라이브러리를 사용하였다. 또한 시각화를 위해 pyLDAvis 3.4.1 라이브러리를 사용하여 구현하였다.

Table 1. System environment

CPU	AMD Ryzen 7 3700X 8-Core Processor
RAM	32.0GB
OS	Windows 10
Version	Python 3.9.12
	Gensim 4.3.2
	pyLDAvis 3.4.1

2.2 Data Collection

신뢰성이 있는 뉴스 기사를 선정하기 위해서 국내 최대 포털 사이트인 네이버에 올라오는 국내외 주요 언론사의 뉴스 기사로 선택하였다. 뉴스 중에서 보안/해킹 분야의 뉴스 기사를 대상으로 하였다. 2020년 1월 1일부터 2023년 8월 31일까지의 뉴스 기사를 날짜별로 뉴스 기사를 크롤링하기 위해 파이썬에서 requests 라이브러리와 Beautiful Soup 라이브러리를 활용하였다. 먼저 각 날짜의 뉴스 링크를 수집해 html을 파싱하여 뉴스 기사를 크롤링하였다. 데이터는 날짜와 기사 내용으로 구분하였다. 이 과정을 통해 수집한 보안/해킹 분야 뉴스 기사 데이터는 총 13,390개이다. 이를 활용해 정보보안 분야의 동향을 분석할 예정이다.

2.3 Data Preprocessing

크롤링으로 수집된 뉴스 기사를 분석할 때 정확도를 높이기 위해 필터링과 전처리 과정을 거쳤다. 필터링 과정에서는 Article 열에서 중복된 행과 비어있는 행을 제거하는 필터링을 진행하였다. 필터링 전의 총 기사 개수는 13,390개이고 월평균 약 304개이다. 필터링 후의 총 기사 개수는 9,858개이고 월평균 약 224개의 기사를 수집하였다.

데이터를 정제하기 위해 파이썬에 있는 정규 표현식 re 모듈을 사용하여 한글을 제외한 모든 문자를 공백으로 대체하였다. 한글 형태소 분석기인 Konlpy 라이브러리를 사용하며 뉴스 기사의 명사만 추출하였다. 추출된 총 단어의 개수는 23,856개이다. 추출한 명사 중에서 계속 반복되지만 불필요한 단어인 뉴스, 기사, 기자라는 단어를 제외한다. 이후 9,858개의 기사 중에서 270개 이상의 기사에서 등장하면서 기사 전체에서 80퍼센트 이하를 차지하는 단어만을 남겼다. 이렇게 전처리 과정을 거친 단어는 총 874개이다. Fig. 3은 전처리 과정을 거치기 전과 후의 결과 중 일부를 보여주고 있다.

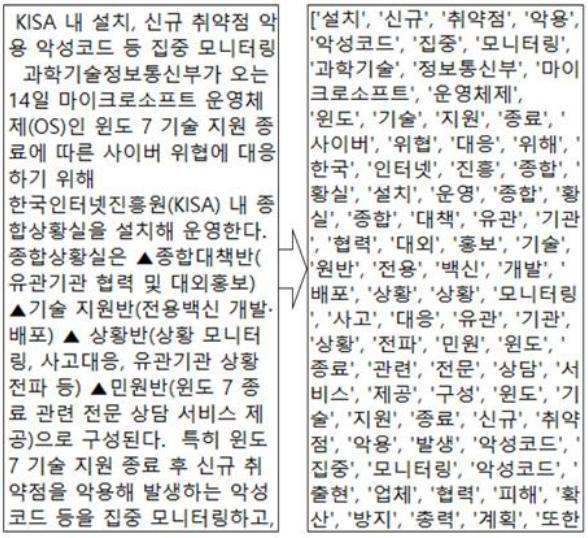

Fig. 3. Files before and after the preprocessing process

2.4 Determining the Number of Topics

LDA 토픽모델링을 실행하기 전에 먼저 적정 토픽 수를 결정했다. 적정 토픽 수를 결정하는 데에는 Perplexity와 Coherence Score를 사용하였다. 토픽 1개에서 10개까지의 Perplexity와 Coherence 값을 계산하였다. 계산된 결과 값을 맷플롯립 라이브러리를 활용해서 출력하였다. Fig. 4는 Perplexity의 결과를 보여주고 Fig. 5는 Coherence Score의 결과를 보여주고 있다.

토픽 1에서 10까지 중에서 Perplexity에서 결과 값이 작으면서 Coherence Score에서 결과 값이 높은 6을 적정 토픽 수로 결정하였다.

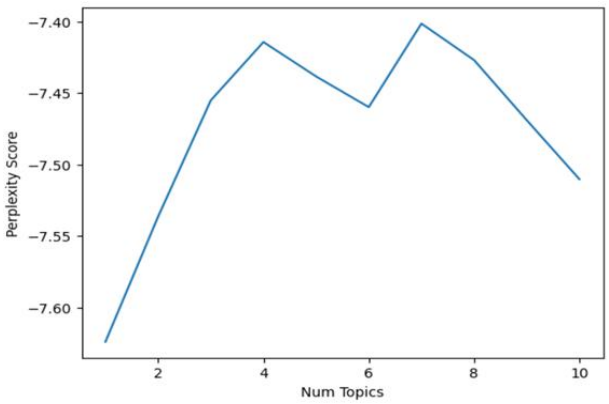

Fig. 4. Perplexity Score results graph

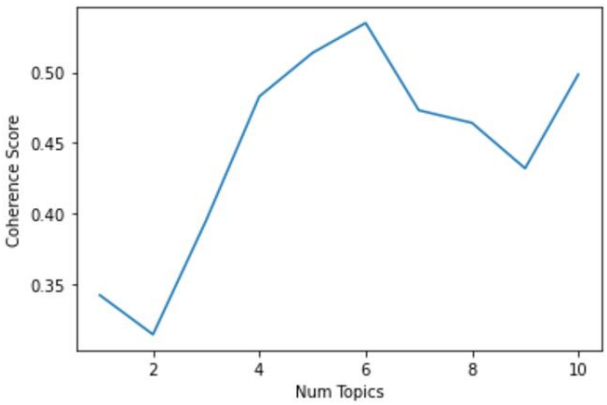

Fig. 5. Coherence Score results graph

2.5 LDA Topic Modeling

LDA 토픽 모델링을 시행할 때 이상치 데이터를 삭제하기 위해서 전처리 과정을 거친 단어 중 270개 이상의 문서에서 나오고 문서 전체 80퍼센트 이하로 차치하는 단어만 남기도록 하였다. 남은 단어들을 카운트 벡터로 변환하여 LDA 토픽 모델링을 하였다. LDA 토픽 모델링 결과를 시각적으로 확인하기 위해서 pyLDAvis 라이브러리를 사용하였다. Fig. 6는 LDA 토픽 모델링 후 pyLDAvis 라이브러리를 사용하여 나타낸 결과이다. 왼쪽 원은 6개를 통해서 토픽 간의 관계를 파악할 수 있다. Topic 1, Topic 2과 Topic 3은 보안 계열로 비슷한 토픽으로 원 사이의 거리가 가까웠고 Topic 4와 Topic 5는 해킹과 개인정보와 관련된 내용으로 일부 겹치는 내용이 있었다. 마지막으로 Topic 6은 암호화폐로 다른 토픽과는 주제가 달라 거리가 떨어져 있는 것을 확인 할 수 있었다. 왼쪽 그림에서는 주요 단어들이 문서 전체에 차지하는 비중을 확인 할 수 있었다.

IV. Experiment result

1. Topic Modeling Results

pyLDAvis 라이브러리를 이용한 토픽 모델링을 수행한 결과 토픽 6개 중에서 가장 비중이 큰 토픽은 Topic 1로 20.3%이고 다음으로는 Topic 2가 20.2%, Topic 3이 20%, Topic 4는 16.5%, Topic 5는 12.3%, Topic 6으로 10.7% 순으로 나타났다. Topic 1을 구성하는 키워드 중 유

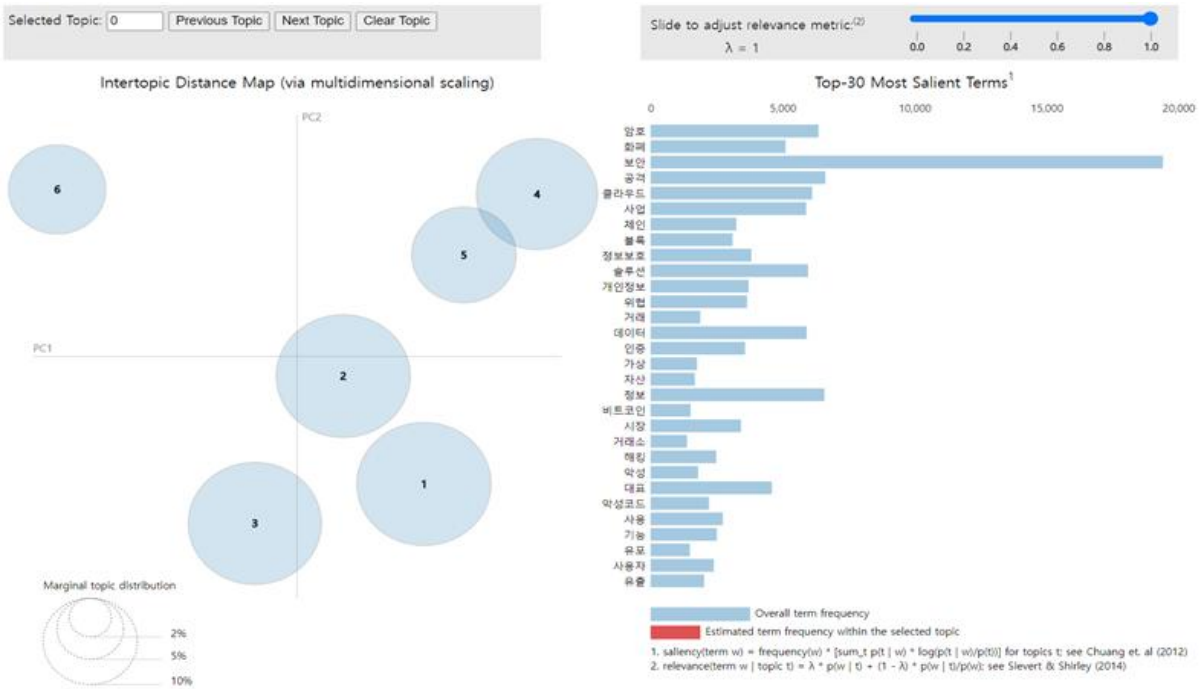

Fig. 6. pyLDAvis result

의미한 것은 ‘보안(Security)’, ‘클라우드(Cloud)’, ‘솔루션(Solution)’, ‘위협(Threat)’ 등으로 구성되었다. 이 단어와 기사 내용을 확인해 봤을 때 클라우드 보안(Cloud Security)이라고 주제를 명명하였다. Topic 2는 ‘정보 보호(Information Protection)’, ‘데이터(Data)’, ‘개인 정보(Personal information)’, ‘기업(Corporation)’, ‘산업(Industry)’, ‘금융(Finance)’, ‘센터(Center)’, ‘기관(Institution)’ 등의 단어로 구성되어 있어 주제를 사회의 사이버 보안(Cyber Security in Society)이라고 명명하였다. Topic 3은 ‘보안(Security)’, ‘기술(Skill)’, ‘서비스(Service)’, ‘솔루션(Solution)’, ‘개발(Development)’, ‘제품(Product)’ 등의 단어로 구성되어 있고 주제명은 보안 솔루션(Security Solutions)으로 명명하였다. Topic 4는 ‘공격(Attack)’, ‘해킹(Hacking)’, ‘악성 코드(Malicious Code)’, ‘유포(Spread)’, ‘랜섬(Ransom)’, ‘웨어(Ware)’, ‘해커(Hacker)’ 등 사이버 공격 관련된 단어로 구성이 되어 있어 주제를 사이버 공격(Cyber attack)으로 명명하였다. Topic 5는 ‘인증(Certificate)’, ‘서비스(Service)’, ‘정보(Information)’, ‘사용(Use)’, ‘보안(Security)’, ‘유출(Leakage)’, ‘개인 정보(Personal Information)’, ‘사용자(User)’ 등으로 이루어져 있고 개인의 개인정보와 관련된 내용들로 이루어져 있어 개인 정보 유출(Personal Information Leakage)로 주제명을 명명하였다. Topic 6은 ‘암호(Code)’, ‘화폐(Currency)’, ‘체인(Chain)’, ‘블록(Block)’, ‘거래(Deal)’, ‘가상(Virtual)’, ‘자산(Property)’, ‘비트코인(Bitcoin)’, ‘거래소(Exchange)’ 등으로 이루어져 있어 암호화폐(Cryptocurrency)로 주제명을 명명하였다. Topic과 주제명, 토픽이 이루어진 단어를 정리해서 Table 2로 나타냈다. 그리고 각 단어의 빈도를 확인하기 위해 워드 클라우드를 활용해 Fig. 7로 나타냈다.

Table 2. Themes and keywords per topic

Topic 1 (20.3%)	Theme	Cloud Security
	Keyword	Security, Cloud, Solution, Threat
Topic 2 (20.2%)	Theme	Cyber Security in Society
	Keyword	Information Protection, Data, Personal information, Corporation, Industry, Finance, Center, Institution
Topic 3 (20%)	Theme	Security Solutions
	Keyword	Security, Skill, Service, Solution, Development, Product
Topic 4 (16.5%)	Theme	Cyberattack
	Keyword	Attack, Hacking, Malicious Code, Spread, Ransom, Ware, Hacker
Topic 5 (12.3%)	Theme	Personal Information Leakage
	Keyword	Certificate, Service, Information, Use, Security, Leakage, Personal Information, User
Topic 6 (10.7%)	Theme	Cryptocurrency
	Keyword	Code, Currency, Chain, Block, Deal, Virtual, Property, Bitcoin, Exchange

2. Time Series Regression Results

LDA 모델링 결과로 나온 6개의 토픽이 2020년 01월부터, 2023년 08월까지 변화를 확인하기 위해 시계열 회귀 분석을 진행하였다. 토픽이 시간의 흐름에 따라 얼마나 분포하는지 시각적으로 확인하기 위해서 맷플롯립 라이브러리를 사용하여 그래프로 나타냈다. 토픽별 시간의 흐름에 따른 변화를 Fig. 8로 나타냈다. x축에는 날짜가 나와 있고 y축에는 토픽의 비중이 나타나 있다. Fig. 7을 통해서 월별로 시간의 흐름에 따라서 토픽의 흐름을 알 수 있다.

시계열 회귀 분석에서 나온 결과를 통해서 토픽이 급상승한 구간에 대해 조사하였다. Fig. 7에서 급상승한 구간과 KISA(한국인터넷진흥원)의 사이버 위협 동향 보고서를 비

Fig. 7. Word Frequency Analysis Word Cloud

교해 보았다. 2021년 상반기 사이버 위협 동향 보고서에 따르면 2021년 상반기에 전체적으로 랜섬웨어에 대한 사이버 공격이 많았고 2021년 5월에 동유럽에 근거를 둔 사이버 범죄 집단 다크사이드(Dark Side) 랜섬웨어 공격 때문에 사회적 혼란이 있었다[18]. 이에 그래프에서도 Topic 4에서 2021년 5월부터 6월 사이인 (a) 구간의 그래프가 크게 상승한 것을 확인 할 수 있다. 또한 2023년 상반기 사이버 위협 동향 보고서를 확인해 봤을 때 4월에 가상자산거래소가 해킹당해서 약 204억 규모의 자산이 유출되었다[19]. 이외에도 크리덴셜 스터핑으로 인한 금전 피해가 지속되었다. 이에 그래프에서 2023년 4월에서 5월 사이인 (b) 구간에서 그래프가 상승한 것을 확인 할 수 있다.

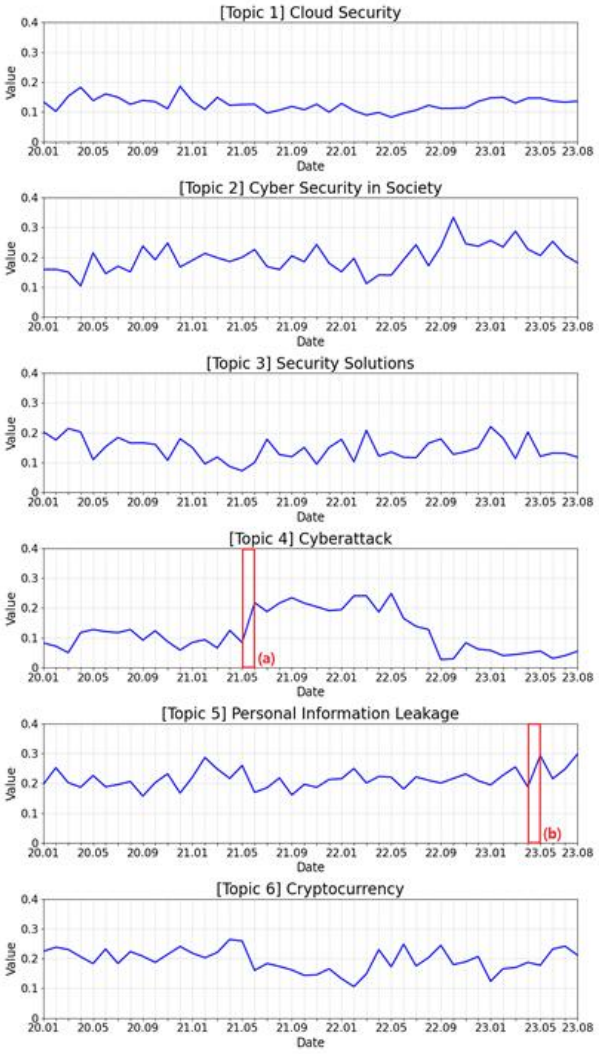

Fig. 8. Changes in the weight of topics over time

3. Comparison of research results

기존에 있는 토픽 모델링을 통해 동향을 분석하는 논문의 결과와 본 논문의 결과에 대해 비교하였다. Table 3은 연구 결과를 비교한 결과이다.

토픽모델링을 활용한 국내 지역 지리교육 연구 동향 분석[11] 논문에서는 1993년부터 2019년까지 발행된 논문 중 연구 목적과 부합하는 데이터로 LDA 토픽 모델링을 진행하였다. 적정 토픽 수를 결정하기 위해 Perplexity 값을 이용하였고 토픽모델링 결과는 워드 클라우드와 표를 통해 나타냈다. 토픽모델링을 활용한 주요국의 스마트 제조 기술 동향 분석[14] 논문에서는 특허 검색 서비스인 윕스온을 활용해서 1991년부터 2020년 사이에 미국과 유럽에서 출원된 특허 중 스마트 제조와 관련된 정보로 LDA 토픽 모델링을 진행하였다. 적정 토픽 수를 결정하기 위해서 coherence score 값을 활용하였다. LDA 토픽 모델링 결과는 표로 정리해서 나타냈다. 국내 기록관리학 연구 동향 분석을 위한 토픽모델링 기법 비교[17] 논문에서는 특정 논문지의 최초 발간 일부터 2016년까지 등재된 논문을 수집하여 LDA 토픽모델링과 HDP 토픽모델링을 진행하였다. 각각의 토픽모델링 결과를 LDAvis와 표로 나타내고 각 결과를 비교하였다, 본 논문에서는 2020년 01월부터 2023년 08월까지 뉴스 기사를 수집해 LDA 토픽모델링을 진행하였다. 적정 토픽 수를 결정하기 위해서 coherence score 값을 활용하였고 LDA 토픽모델링 결과를 시각화하

Table 3. Comparison of research results

Title	Data	Analytical Algorithm	Determining Topics	Visualization
Research Trends of Regional Geography Education Using Topic Modeling[11]	Thesis	LDA	Perplexity	Word Cloud, Table
Analysis of global trends on smart manufacturing technology using topic modeling[14]	Patent Data	LDA	coherence score	Table
Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea[17]	Thesis	LDA	X	LDAvis, Table
		HDP	X	LDAvis, Table
Proposed system	News	LDA	coherence score Perplexity	LDAvis, Table, Word Cloud

기 위해 LDAvis, 표, 워드 클라우드를 사용하였다.

본 연구에서는 최신 트렌드를 분석하기 위해 모든 언론사의 뉴스를 데이터로 사용하였다. 또한 토픽 수 결정하기 위해 2가지 기법을 모두 사용해 정확도를 높였고 결과를 여러 방법으로 도출하여 가동성을 높였다.

V. Conclusions

2020년 01월부터 2023년 08월까지 IT/보안 분야의 동향을 파악하기 위해 LDA 모델링과 시계열 회귀 분석을 사용하였다. 이 과정에서 LDA 모델링 결과 6개의 토픽이 도출되었으며, 시계열 회귀 분석을 통해 각 토픽의 시간적 변화를 살펴보았다. 또한, 토픽의 시간별 추세 중에서 특히 급상승한 구간에 해당하는 이슈들을 분석하여, 토픽의 흐름이 보안 이슈와 밀접하게 연관되어 있음을 확인할 수 있었다. 이 중에서도 2021년의 랜섬웨어 공격과 2023년의 가상자산거래소 해킹이 대표적인 보안 이슈로 확인되었다.

본 연구의 분석 결과를 통해 현재의 IT/보안 동향에 대한 이해를 높일 수 있다. 이를 통해 보안 전략을 개발하고 구현하는 데 도움이 될 것이다. 또한 정보보안 분야의 연구 방향을 결정할 때 기여할 것으로 기대된다.

하지만, 전처리 과정 후 한글로 된 명사 중에서 보기 불편한 단어가 있었고, 같은 의미를 가진 단어가 한글과 영어로 다르게 표현되었을 때 실험에 영향을 줄 수 있다는 문제점을 발견하였다. 따라서, 향후에는 같은 의미를 가진 다양한 단어를 통합할 수 있도록 동의어 처리와 표제어 추출을 활용해 하나의 의미를 가진 단어를 통합시켜 토픽 모델링의 정확성과 일관성이 높아지도록 개선할 예정이다. 또한, 여러 토픽이 포함된 기사에 대한 처리 방안을 연구하여 보다 신뢰할 수 있는 결과를 얻을 수 있도록 노력할 것이다.

ACKNOWLEDGEMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program(IITP-2024-RS-2022-00156334) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

REFERENCES

[1] Cho Sung Phil , "The Study on Threats of Information Security and Their Solutions in the Fourth Industrial Revolution," Korean Security Journal, no.51, pp.9-36, Jun 2017. 10.36623/kssa.2017.51.1

[2] CK academy, "Analysis of Global Research Trend on Information Security," Journal of the Korea Institute of Information and Communication Engineering, vol.19, no.5, pp.1110-1116, May 2015. 10.6109/jkiice.2015.19.5.1110

[3] Seung-Tae Yoo, Park Kyungseon, Lee Yoon-Seo, Hwang Seong-Jin and KIM Kang Seok, "Analysis of Domestic Industrial Security Trends using LDA Topic Modeling," Korean Journal of Industrial Security, vol.10, no.2, pp.79-103, Jan 2020. 10.33388/kais.2020.10.2.079

[4] Kyoungsik Na and Jisu Lee, "Trends of South Korea's Informatization and Libraries' Role Based on Newspaper Big Data," The Journal of the Korea Contents Association, vol.18, no.9, pp.14-33, Sept 2023. 10.5392/JKCA.2018.18.09.014

[5] Daemin Park, "Automated Time Series Content Analysis with News Big Data Analytics : Analyzing Sources and Quotes in One Million News Articles for 26 Years," Korean Journal of Journalism & Communication Studies, vol.60, no.5, pp.353-407, Oct 2016. 10.20879/kjjcs.2016.60.5.013

[6] Keon Chul Park and Chi Hyung Lee, "A Study on the Research Trends for Smart City using Topic Modeling," Journal of Internet Computing and Services, vol.20, no.3, pp.119-128, Jun 2019. 10.7472/jksii.2019.20.3.119

[7] Sung-Chan Jun, Han seongho and KIM SANG BAEK, "Technology Development Strategy of Piggyback Transportation System Using Topic Modeling Based on LDA Algorithm," Journal of The Korea Society of Computer and Information, vol.25, no.12, pp.261-270, Dec 2020. 10.9708/jksci.2020.25.12.261

[8] David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent dirichlet allocation." Journal of Machine Learning Research, vol.3, pp.993-1022, Mar 2003. 10.5555/944919.944937

[9] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling and Andreas Both, "Evaluating topic coherence measures," arXiv: Learning, Mar 2014. 10.48550/arXiv.1403.6397

[10] Lei Fang, "Congestion measurement in nonparametric analysis under the weakly disposable technology," European Journal of Operational Research, Vol.245, pp.203-208, Aug 2015. 10.1016/j.ejor.2015.03.001

[11] Gapcheol Kim and Hyunjong Noh, "Research Trends of Regional Geography Education Using Topic Modeling," Social Studies Education, vol.58, no.4, pp.49-67, Dec 2019. 10.37561/sse.2019.12.58.4.49

[12] David Newman, Jey Han Lau, Karl Grieser and Timothy Baldwin, Automatic Evaluation of Topic Coherence, Human Language

Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.100-108, Los Angeles, California, June 2010.

[13] Hyo-Jun Yun, Jae-Hyeon and Jiwun Yoon, "Introduction of topic modeling for extracting potential information from unstructured text data: Issue analysis on news article of dementia-related physical activity," vol.30, no.3, pp.501-512, Jul 2019. 10.24985/kjss.2019.30.3.501

[14] Yoonhwan Oh and Moon HyungBin, "Analysis of global trends on smart manufacturing technology using topic modeling," Journal of Korea Society of Industrial Information Systems, vol.27, no.4, pp.65-79, Aug 2022. 10.9723/jksiis.2022.27.4.065

[15] Sievert, Carson, and Kenneth Shirley, "LDAvis: A method for visualizing and interpreting topics," Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp.63-70, Baltimore, Maryland, USA, Jun 2014. 10.3115/v1/W14-3110

[16] KIM KEE OK, "Analysis of Research Trends in Consumer Science through Text Mining," Journal of Consumer Studies, vol.35, no.5, pp.19-47, Jan 2020. 10.35736/JCS.31.5.2

[17] PARK JUNHYEONG and HyoJung Oh, "Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: focused on LDA and HDP," Journal of Korean Library and Information Science Society, vol.48, no.4, pp.235-258, Dec 2017. 10.16981/kliss.48.201712.235

[18] Korea Internet & Security Agency, "Cyber Threat Trend Report for the Second Half of 2023", Korea Internet & Security Agency, Ministry of Science and ICT, pp.19-28, 2021.

[19] Korea Internet & Security Agency, "Cyber Threat Trend Report for the Second Half of 2023", Korea Internet & Security Agency, Ministry of Science and ICT, pp.4-9, 2023.

Authors

Se Young Yuk is currently pursuing the B.S. degree in the Department of Information Security at Pai Chai University in Daejeon, South Korea. She is interested in security, artificial intelligence and malware.

Hyun-Jong Cha received the M.S. and Ph.D. degree in Computer science and Defense Acquisition Program from Kwangwoon University, South Korea, in 2008 and 2014. He is a professor in the Department of Software Engineering at Pai Chai University in Daejeon, South Korea. His current research interests include information security, artificial intelligence, IoT, and blockchain.

Ah Reum Kang received the M.S. and Ph.D. degrees in information security from Korea University, South Korea, in 2012 and 2016. She is a professor in the Department of Information Security at Pai Chai University in Daejeon, South Korea. Her current research interests include security, artificial intelligence, malware, medical data analysis, and online game security.