Journal of the Korea Industrial Information Systems Research Vol.27 No.5, Oct. 2022 :37-48
http://dx.doi.org/10.9723/jksiis.2022.27.5.037 ISSN:1229-3741

텍스트마이닝을 활용한 정보보호 키워드 기반 소셜미디어 빅데이터 분석

(Social Media Bigdata Analysis Based on Information Security Keyword Using Text Mining)

정 진 명[1], 박 영 호[2]*
(JinMyeong Chung and YoungHo Park)

요 약 디지털 기술의 발전으로 사회적 이슈들이 SNS와 같은 디지털 기반 플랫폼을 통해서 소통되고 여론을 형성하기도 한다. 본 연구에서는 소셜미디어를 통해서 공유되고 있는 정보보호 이슈관련 여론을 살펴보기 위하여 대표적인 단문 소셜네트워크서비스인 트위터 빅데이터 분석을 진행하였다. 2021년 1년간 14개 정보보호 관련 키워드를 중심으로 데이터를 수집한 후, 데이터마이닝 기술을 활용하여 용어 빈도(TF)분석과 피어슨 계수를 활용한 상관분석을 통해 키워드간의 상관관계를 밝혔다. 또한 잠재적 확률기반 LDA 토픽모델링을 실시하여 정보보호분야에 많은 관심을 받았던 6개의 주요 토픽을 도출하였다. 이러한 결과는 관련 산업의 전략수립이나, 정부 정책수립 시 주요 키워드를 도출하는 기초데이터로 활용될 수 있을 것으로 기대된다.

핵심주제어: 정보보호, 텍스트마이닝, 토픽모델링, 소셜미디어

Abstract With development of Digital Technology, social issues are communicated through digital-based platform such as SNS and form public opinion. This study attempted to analyze big data from Twitter, a world-renowned social network service, and find out the public opinion. After collecting Twitter data based on 14 keywords for 1 year in 2021, analyzed the term-frequency and relationship among keyword documents with pearson correlation coefficient using Data-mining Technology. Furthermore, the 6 main topics that on the center of information security field in 2021 were derived through topic modeling using the LDA(Latent Dirichlet Allocation) technique. These results are expected to be used as basic data especially finding key agenda when establishing strategies for the next step related industries or establishing government policies.

Keywords: Information security, Text mining, Topic modeling, Social media

1. 서 론

인터넷은 언제 어디서나 사용 가능하고 사회 경제적으로 큰 영향을 주는 사회 기반 플랫폼으로 자리 잡았다. 최근에는 스마트폰 보급 확대,

* Corresponding Author: parkyh@knu.ac.kr
Manuscript received June 27, 2022 / revised August 30, 2022 / accepted October 12, 2022
1) 경북대학교 정보보호학과, 제1저자
2) 경북대학교 전자공학부, 교신저자

클라우드서비스 및 사물인터넷 등 새로운 기술과 맞물려 인터넷의 기능과 영향력은 더욱 높아지고 있다. 정보기술의 발달은 다양한 생활의 편의성을 진작시켰지만, 반대로 사생활의 노출이나 개인정보 유출, 시스템 해킹을 통한 데이터 유출이나 대량의 트래픽 유발에 따른 서비스 중단 등의 사고 발생으로 사회적인 이슈들을 양산하고 있다. Park(2012)는 사회적으로 발생하는 정보보호 이슈들은 전통적인 언론매체를 통해서 생산되고 공유되었으나, 최근에는 인터넷과 스마트폰의 발달로 새로운 소통의 패러다임으로 부상한 SNS를 통해 공유되면서 기존 정치·사회적 관계를 변모시키고 있다고 하였다. SNS를 통해서 형성되는 여론의 영향력이 정치인이나 정책관계자들의 관심을 끌면서, 그와 관련하여 다양한 연구도 진행되어 왔으며, 매스미디어에 대칭되는 성격으로서의 소셜미디어 빅데이터에 대한 분석이 활발하게 이루어지고 있다. 정보보호 이슈들 또한 소셜미디어를 통하여 활발하게 정보가 생산 및 공유되고 있다. 그러나 정보의 정보보호 정책 수립 시에는 글로벌 동향이나, 최신 기술, 산업적 관점의 보고서 혹은 매스미디어의 정형화된 데이터를 기초데이터로 수립하기 때문에 대중의 시각에서 바라본 관점이 누락 될 수 있다. 따라서 그에 데이터 분석을 통한 정보보호 정책 수립을 위한 주요 이슈를 발굴하여 기초데이터로 활용할 필요가 있다. 본 논문에서는 정보보호 키워드를 중심으로 SNS상에 공유되는 사용자들의 콘텐츠를 수집하고, 텍스트마이닝을 통해서 대중들의 관심을 많이 받았던 이슈들을 분석하였다. 이를 위해서 대표적 단문메시지 SNS인 트위터로부터 1년간 정보보호 키워드를 포함한 문서를 수집하였으며, 이를 바탕으로 주요 키워드 출현빈도, 키워드 그룹간의 연관관계 분석 및 토픽모델링을 통해 1년 동안 정보보호 관련 주요 이슈들을 군집화하였다. 이 결과는 대중들의 관심을 받았던 주요 토픽들을 도출함으로써 정보보호 산업의 향후 트렌드를 예측하거나 정부의 정책 방향을 수립하는데 도움이 될 것이다.

2. 선행연구

2.1 여론 의제

최근 디지털 기술의 발달에 따라 각종 여론 또한 디지털 기술이 구현된 플랫폼을 통해서 보급되는 경향이 커지고 있다. 한 언론사의 기사만을 구독하지 않고 인터넷 포털 사이트의 뉴스메뉴를 통해서 같은 동일 주제에 대한 다양한 언론사의 기사를 검색할 수 있으며, SNS를 통해서 대중의 여론이 형성되고 공유되기도 한다. 사회 이슈에 대한 국민의 여론을 기존의 매스미디어가 아닌 디지털 기술을 활용한 SNS를 통해서 직접적으로 확인할 수 있다는 장점 때문에 정치인들이나 정책입안자들이 SNS를 통해 형성되는 여론 의제에 대하여 관심을 가지게 되었으며, 따라서 자연어처리 및 분석에 관한 다양한 연구도 진행되어 왔다. Ku(2002)가 진행한 2000년 미국 대통령 선거 기간에 기존 뉴스미디어 의제와 공중 의제에 관한 웹사이트 캠페인의 영향에 대한 연구는, 정치인으로 한정되긴 했지만, 특정 웹사이트가 기존의 뉴스 미디어에 대하여 의제설정기능을 가지고 있음을 통계적으로 나타냈다는 데 의의가 있다. 이와 유사한 연구로 Yoon et. al.(2003)은 인터넷 웹사이트가 특정한 논쟁적 이슈 및 그 속성에 대한 공중 의제 설정과의 연관성에 관하여 연구하였는데, 공중 의제 설정의 단계에서 안티사이트에 많이 노출될수록 이슈로 부각된다는 가정을 검증하였다는 특징이 있다. Park(2005)는 인터넷 자유 게시판의 저널리즘적 특성과 사회적 영향력을 분석하여 뉴스 의제가 웹사이트뿐만 아니라 인터넷 게시판을 비롯하여 소셜네트워크 서비스에서도 형성될 수 있는 가능성을 보여줬다고 할 수 있다. Kim et. al.(2011)은 소셜미디어인 트위터가 리트윗 기능을 통하여 뉴스를 재매개하는 역할을 하며, 나아가서 뉴스를 매개하고 선정하는 과정에서 참여자들의 뉴스 선택 혹은 새로운 정보를 추가하는 뉴스미디어로서의 트위터를 연구하였다는 의의를 가진다. 그 밖에 Choi et. al.(0211)은 정치적 성격의 트위터와 신문/방송뉴스의 의제를 비교하여 트위터가 정치적인 이슈에 대해서는 신문/방송과 핵심의제를 공유한다는 것을 밝혔으며, Park(2012)는 나아가서 트위터 정치적 이슈

에 대하여 여론형성에 참여하는 행태에 대한 연구를 실시하여 무상급식 주민투표와 10·26 재보선 기간의 여론 형성에 영향력을 행사한다는 결론을 내렸다.

위의 연구들은 SNS 특히 140자 미만 단문과 링크를 통한 정보를 생산 및 공유하는 트위터가 각종 사회적인 이슈에 대하여 뉴스 의제를 형성하는 데에 참여하고 있으며, 사회적인 영향력을 가지고 있다는 사실을 인정하고 있다. 따라서 이슈를 달리하여 정보보호 관련 이슈들에 대한 여론을 파악하기 위해서 트위터 데이터를 수집하여 분석하는 것이 유의미하다고 판단할 수 있을 것이다. 다만, 기존의 연구에서는 정치적 이슈와 같은 특정 주제에 대한 뉴스 혹은 뉴스 검색 상위 키워드가 트위터를 통해서 매개되거나 재매개 되는 현상에 대한 연구를 하거나, 논문 초록의 키워드 분석을 통한 연구 트렌드를 분석하는 것에 주를 이루었다고 할 수 있다. 정부의 정책에 관련된 이슈들이 트위터와 같은 소셜미디어 서비스에 어떻게 반영되고 있는지에 관한 연구는 미미하다. 따라서 본 연구에서는 정부의 정책 문서들로부터 추출한 정보보호 분야 키워드들이 트위터라는 매체를 통해서 어떻게 의제를 형성했는지 파악해보고자 한다.

2.2 텍스트마이닝

텍스트마이닝은 빅데이터를 분석하는 기술인 데이터 마이닝의 한 부분으로 다음과 같이 자연어처리 프로세스를 거쳐 진행된다. 먼저 크롤링 등 다양한 방법으로 텍스트를 수집한 후, 사전처리를 진행한다. 데이터 사전 처리 방법은 각종 도구나 라이브러리, 연구자의 의도에 따라서 다르게 구성할 수 있다. 특수문자와 분석 목적과 관련이 없는 문자의 제외처리가 필요하며, 비속어나 유의어들의 통합작업도 필요하다. 수집된 텍스트에 대하여 사전 처리를 끝낸 후에 본격적인 분석을 수행하게 되는데, Kim et. al(2017)은 텍스트 분석 기술에 대하여 크게 빈도분석과 군집화, 그리고 분류로 나누어 설명하였는데, 문서의 용어 벡터화에 가장 널리 사용되는 방법은 용어의 출현 빈도를 기반으로 한 TF-IDF(term frequency - inverse document frequency) 방법이다. 빈도분석을 활용하는 방법으로는 워드클라우드, 동시 출현 관계를 반영한 네트워크 분석, 그리고 시계열 출현 빈도를 반영한 추세분석 등을 할 수 있다. 출현 빈도와 군집화 기법을 복합적으로 반영한 기법으로는 대표적으로 토픽모델링이 있는데, 문서와 단어의 출현 가능성에 대한 확률로 토픽 군집을 만드는 방식이라고 할 수 있다. 분류 기술은 문서를 사전에 정의된 범주에 따라서 자동 할당하는 기술로 감성분석(sentimental analysis)도 정의된 감정사전에 따라서 감정점수를 산정하여 긍정과 부정을 분류하는 특성상 분류기술로 구분되고 있다.

2.3 토픽모델링

토픽모델링은 텍스트마이닝의 한 범주로 구분되는 것이 보통이다. Kim et. al(2017)의 연구에서도 군집을 만드는 기술로 분류되고 있으며, 문서와 토픽, 단어와 토픽, 문서와 단어를 통찰하는 확률적 추정의 이론적 배경을 이해하는 것이 필요하다. 토픽은 특정 시간이나 플랫폼에서 많이 이슈화되는 키워드를 의미하는데, 토픽모델링은 문서 집합에서 발생하는 토픽들을 발견하기 위한 확률적 모델로서 직관적으로 문서들에 분포된 단어들의 확률적인 발생 가능성을 토픽으로 추출하는 방식이다. 대표적인 토픽모델링으로 LDA(latent dirichlet allocation)가 있는데, 본 논문에서는 Blei(2012)가 제시한 기본적인 LDA 모델링의 개념을 활용한다. 문서에 분포된 토픽들의 비율과 문서 내 단어들의 분포 관계에 따라서 토픽을 추출하며, 그 개념은 다음 Fig. 1과 같다.

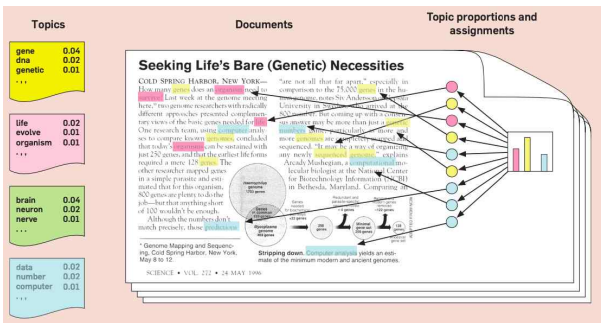

Fig. 1 The Intuition of LDA(Blei, 2012).

문서 내에 색상별로 블록 처리된 단어들은 왼쪽의 색상별 단어 그룹으로 묶게 되어 하나의 토픽그룹으로 분류될 수 있다. 가능한 많은 단어를 블록 처리하다 보면 이 문서들이 주로 어떤 토픽들로 구성되는지 특징을 찾을 수 있다. 오른쪽 히스토그램의 비율로 토픽이 문서를 구성하고 있다는 가정으로 각 토픽과 관계가 있는 단어를 해당하는 토픽으로 연결하는 순서로 모델링이 진행된다. 이 토픽들은 전체 문서에 분포되어 있는 집합이며 LDA 모델링은 이와 같은 그룹의 특성을 찾아내는 문서들의 통계적 모델이라고 할 수 있다. 다음 Fig. 2는 LDA 모델링을 통해서 토픽을 추출하는 변수들과 메소드의 구성을 나타낸다.

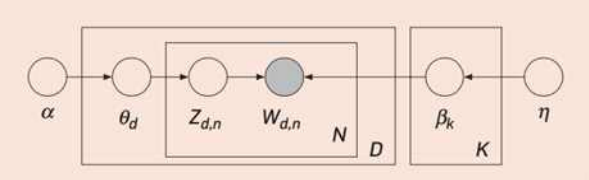

Fig. 2 LDA Model(Blei, 2012).

위 수식에서 사용자 지정변수는 α, β이며, 나머지 변수는 잠재 변수로써 드러나지 않는다. 각 노드는 랜덤변수로서 일반화 과정의 역할에 따라서 레이블링 된다. 우리가 관찰할 수 있는 노드는 단어와 문서의 집합뿐으로 음영표시가 되어 있는 동그라미이다. 동그라미는 변수를 나타내며 화살표의 시작은 조건, 화살표 종단은 결과를 나타낸다. LDA 모델링을 수식으로 나타내면 Fig. 3과 같다.

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, \omega_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(\omega_{d,n}|\beta_{1,K}, Z_{d,n}) \right)$$

Fig. 3 Equation of LDA(Blei, 2012).

$\Theta_{d,k}$ 는 d와 토픽 k비율과의 관계를 나타내며, $Z_{d,n}$ 는 d번째 문서의 n번째 단어와 토픽 k의 관계를 나타낸다. β는 사용자가 지정하는 변수이며, 나머지 변수들은 모두 잠재 변수이다.

토픽모델링이 텍스트 분석에서 활용될 수 있는 경우를 살펴보면 먼저, 특정 분야의 중심이 되는 주제의 변화에 관한 연구에 활용할 수 있다. 이와 관련된 연구로 Griffiths et. al(2004)는 2004년 미국 국립과학원회보에 게재된 초록에 토픽모델링을 적용하여 활발한 연구가 이루어지고 있는 주제와 연구가 감소하고 있는 주제를 밝히는 연구를 진행하였는데, PNAS 저널로부터 토픽을 추출하기 위하여 MCMC(markov chain monte carlo)알고리즘과 베이지안 모델링을 활용하였다. Wang et. al.(2006)은 시간에 따른 토픽의 변화 과정을 추적하는 연구에서 시간의 흐름에 따라 토픽이 어떻게 변화하는지 살펴볼 수 있는 TOT(Topics Over Time) 모델을 제시하고, 이를 통해 개인 이메일, 미국 대통령 연설문 등을 대상으로 모델을 적용하여 토픽의 시계열적 추이를 분석하였다. 토픽모델링은 소셜네트워크 분석(SNA, social network analysis)의 기법으로 사용하기도 하는데, Bae et al.(2013)은 2012년 한국 대통령 선거 관련 트위터를 분석하여 트위터 데이터의 토픽과 실제 기사 반영된 변화추이를 분석하였는데, 시각화를 통해서 단문인 트위터 문서의 토픽생성 한계를 시계열분석 및 시각화로 극복하고자 했다는 데에 의의가 있다. Cho et. al.(2018)은 토픽모델링 기법으로 개방형 혁신 주제에 관한 연구의 초록으로부터 주제들을 도출하여 정부의 정책 추진 방향과의 관계를 분석하였는데, 토픽으로 추출한 주제 간의 네트워크 분석을 통하여 상대적으로 비중이 높은 주제를 추출하여 정부의 정책방향과 비교를 시도하였다는 데에 의의가 있다. Park et. al.(2017)은 미국의 특허 문서 중 AI(artificial intelligence) 초록을 대상으로 토픽모델링 기법을 활용하여 핵심기술과 연도별 비중을 반영하여 기술 동향 분석 및 예측하였다. Choi et. al.(2020)은 다양한 파트너십에 대한 정책적 이슈를 파악하기 위해 파트너십 관련 단어를 포함하는 뉴스 기사와 댓글을 토픽모델링을 활용하여 분석하여 비교하였는데, 토픽의 수 결정을 위해 혼잡도(perplexity)와 조화평균(harmony

mean)을 고려하였다. Chung et. al.(2018)은 인공지능(A.I.)기술 관련 연구 활동 및 동향을 텍스트마이닝과 토픽모델링을 활용하여 분석하고 향후 연구 방향에 대한 시사점을 도출하였다. Park et. al.(2021)은 상담 성과에 관한 연구 동향을 분석하여 상담 성과 연구에 활용되는 주요 토픽을 도출해내기 위하여 중심성 분석과 토픽 모델링을 복합적으로 활용하였다. Lee et. al (2022)은 텍스트 자료에 대한 네트워크 분석과 토픽모델링을 실시하여 대학에서의 역량 기반 교육 관련 정보를 분석하고 시사점을 제시하였는데, 네트워크 분석과 토픽모델링의 개별 분석 및 비교분석을 하여 결과를 추출하였다는 데에 의미를 둘 수 있다. Lee(2020)은 '노인일자리' 관련 신문기사를 수집 및 분석하였는데, 토픽 모델링 기술을 깊이 있게 활용하여 결과를 추출하여 연도별 데이터 트렌드 결과 분석을 하였다.

토픽모델링을 수행할 수 있는 분석 도구로는 R, 자바 기반의 Mallet, 파이썬 기반의 Gensim 등이 있다. R은 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경으로서, 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있으며, 그중 토픽모델링에 사용하는 것은 LDA 관련 모델을 포함하고 있는 LDA 패키지이다. Mallet은 자바 기반의 패키지로, 통계적 자연어처리, 문서 분류, 군집화, 토픽모델링, 정보 추출 등 다양한 기계 학습을 텍스트에 적용할 수 있도록 한다. Mallet의 텍스트 인코딩에는 UTF-8을 기본으로 쓰고 있어, 한국어 처리가 가능하다는 특징을 가진다. Mallet의 토픽모델링 개발환경(toolkit)은 대규모 텍스트를 분석하는데 효율적이다. Gensim은 파이썬 기반의 공개 프로그램 라이브러리이며, 통계적 의미론 분석에 강점이 있는 도구라고 할 수 있다.

3. 연구 방법

3.1 정보보호 키워드 선정

소셜미디어에서 이슈화가 되고 있는 정보보호 관련 토픽을 발굴하기 위하여 먼저 데이터 수집을 위한 키워드 선정 작업을 시행하였다. 정부 정책의 반영을 목적으로 하기 위해서 정부에서 산출한 백서와 보고서 문서를 중심으로 사용 빈도수가 높은 키워드를 추출하였는데, 우리나라 국가 수준의 정보보호 관련 대표적인 기관인 국가정보원 및 한국인터넷진흥원에서 발간하는 연례 보고서로부터 추출하였으며, 그 결과는 Table 1과 같다.

Table 1 Information Security Keywords

Terms
ISMS(Information Security Management System), Cyber Attack, Data Protect, National Security, Information Security, DDoS, Security, Hacking, Hacking Incident, Information Leak, Infringement, Security Control, Vulnerability, (National) Security

3.2 트위터 데이터 수집

트위터는 2006년 3월에 서비스를 시작한 소셜미디어 서비스이다. 사용자는 한 번에 140자까지 콘텐츠를 적을 수 있으며, 메일의 전달 기능과 유사한 리트윗 기능을 보유하고 있다. 따라서 트위터 상에서 생성된 정보는 높은 전파속도를 가질 수 있다. 트위터의 또 하나의 장점은 개발자들에게 OpenAPI를 제공하여, 트위터에서 생산되는 데이터를 외부로 활용할 수 있도록 인터페이스를 제공한다는 것이다. 단, 최근 10일간의 데이터를 수집할 수 있으며, 10일 이내에서는 일 단위로 임의로 날짜를 선택 및 수집할 수 있다. 데이터수집에 매우 편리한 환경을 제공하고 있는 트위터는 데이터의 생성과 전파속도가 빠른 특징을 가지고 있어 비정형 빅데이터를 수집 및 활용한 연구에 많이 활용되고 있다. 본 논문에서는 2021년 1년간 데이터 수집을 진행하였으며, 최종적으로 총 36,958건의 트위터 데이터를 수집하였다.

3.3 데이터 전처리

트위터에는 각종 특수문자를 사용하는 광고링크와 유의미하지 않은 복수의 트윗이나 리트윗 데이터들이 많이 있어서, 데이터 분석을 위해서는 허수 데이터를 필터링 하는 전처리가 필수적이다. 본 논문에서는 트위터의 특성인 정보의 재 매개 기능을 고려한 분석을 실시하기 위하여 전처리를 2 가지 방향으로 진행하였다. 먼저 빈도분석을 위해서는 기존의 한글 자연어처리를 위한 특수문자나 한글이 아닌 외국어와 특수문자, 단어 간 공백을 제거한 후 키워드를 중심으로 트위터 문서를 통합하고, 중복 트윗을 제거하지 않고 분석을 실시하였으며, 토픽모델링을 실시할 때에는 전처리 단계에서 중복트윗을 제거하였다.

본 논문에서 처리한 데이터 전처리 작업 과정은 Table 2와 같다.

Table 2 Data Pre-Processing

No	Procedure
①	Remove Letters and Characters
②	Remove space between words
③	Remove Duplicated Tweet
④	Remove Tweets less than 3 Words
⑤	Extract Nouns
⑥	Remove Bad Words
⑦	Substitute between words

본 논문에서는 명사 단어를 추출하여 단어의 출현빈도와 문서에서의 단어 분포 관계를 기반으로 하는 토픽모델링을 실시하기 때문에 데이터의 신뢰도를 높이기 위하여 길이가 3글자 이하로 짧은 트위터는 제거한다. 또한 명사를 추출할 때에도 1글자인 명사는 제외하여 데이터의 균일성을 확보하였다. 추출한 단어의 리스트를 확인하여 비속어와 은어는 수동으로 삭제하였으며, 조사가 있거나 유의한 단어들을 통일하는 작업을 실시하였다. 최종적으로 분석 대상이 되는 트위터 문서의 수는 Table 3과 같다.

Table 3 Documents after Data Mining

Title	Documents
DDos	4,323
National Security	1,450
Vulnerability	1,095
Data Protect	990
Information Security	683
Information Leak	634
Security	472
(National) Security	372
Hacking	346
Cyber Attack	210
Security Control	63
Hacking Incidents	46
ISMS	41
Infringement	30

3.4 토픽모델링

토픽모델링은 문서(코퍼스)와 단어들의 출현빈도 및 분포에서 토픽을 추출하는 방법론이다. 따라서 토픽모델링에는 기본적으로 단어의 출현빈도와 문서에 분포된 단어들의 분산은 깊은 관계가 있다. 앞서 데이터 전처리를 통하여 1차적으로 각각의 트위터로부터 명사들을 추출하였으며, 문서와 단어의 행렬인 DTM(document term matrix)을 구성하였고, 문서의 상관관계 분석 및 단어축적 빈도를 구하여 시각화하였다. 토픽모델링은 Gibbs 샘플링을 활용한 토픽모델링을 실시하였다. Gibbs 샘플링은 마르코프 체인 이론 기반의 변인을 최소화하는 방법으로 분석 대상 표본이 거대할 때 사용하는 방법이다.

최적의 토픽 수를 결정하기 위해서는 혼잡도(perplexity) 지수를 사용하여 최적의 토픽 수 범위를 산정하는데, 혼잡도(perplexity) 지수는 Graffis, et al(2004)의 논문을 기반으로 하며, 총 4개의 매트릭스로 모델링할 수 있다. Arun2010, CaoJuan2009 지수는 값이 낮을수록, 적합하고, Deveaud2014, Griffiths2004 지수는 값이 높을수록 최적의 토픽 수에 가까움을 의미한다. 가장 간단하게는 Griffiths2004 값이 1에 가까운 포인트 전후의 토픽 수 범위를 변화시키며 모델링하여 최적의 토픽수를 선정하여 분석한다. 본 논

문에서는 Choi et. al.(2020)가 제안한 방법을 활용하되, 혼잡도(perplexity) 지수를 4가지 모두 사용하여 적정 토픽 수 구간을 반복하여 토픽모델링을 실시하였다.

4. 연구 결과

4.1 빈도 분석

전처리를 통하여 명사를 추출하여 키워드별로 그룹핑된 문서에서의 단어의 출현 빈도, 그리고 문서들 간의 연관성에 대한 분석을 실시하였다. 먼저 수집한 데이터에서 명사를 추출하여 사용빈도가 높은 단어 상위 20개와 그 숫자를 합산하면 다음 Table 4와 같다.

Table 4 Top-20 Term Frequency

Title	TF(Term Frequency)
Hacking	27,362
DDoS	10,049
Attack	6,210
Request	5,761
DB	4,346
Proxy	3,453
Twitter	2,730
Secret	2,499
Number	2,497
Facebook	2,478
Location	2,460
Kakao Talk	2,404
Tracking	2,308
Smart Phone	2,117
Account	2,093
Star	1,897
Iphone	1,884
Phone Hacking	1,812
Instagram	1,578
Infringement	1,322

전체 단어 중에 해킹이 27,362건으로 가장 많은 빈도를 나타내고 있으며, 그 다음으로 디도스 10,049건, 공격 6,210건의 순서로 나타났다. 위 빈도수는 명사를 추출할 때 단어의 중복을 제외하지 않은 것이다. 중복된 단어를 제거 했을 때의 단어수와 제거하지 않았을 때의 단어 숫자 차이는 150,000 단어 이상 차이가 났다. 다음 단어 축적 빈도와 키워드 그룹간 연관관계, 그리고 토픽모델링 시에는 중복단어의 사용이 모델링 시에 편향적인 결과를 나타낼 수 있기 때문에 중복단어를 제외한 상태에서 분석을 실시하였다.

4.2 키워드 상관관계 분석

수집된 트위터 데이터를 14개 키워드 그룹문서를 만들고, 먼저 축적빈도를 측정하였다. 단어의 축적 빈도를 측정하기 위해서는 특정 단어의 반복적인 사용으로 데이터가 편향되지 않기 위해서 중복된 트위터를 제거하였다. 중복된 트위터를 제거한 후의 축적 빈도는 Fig. 4와 같다.

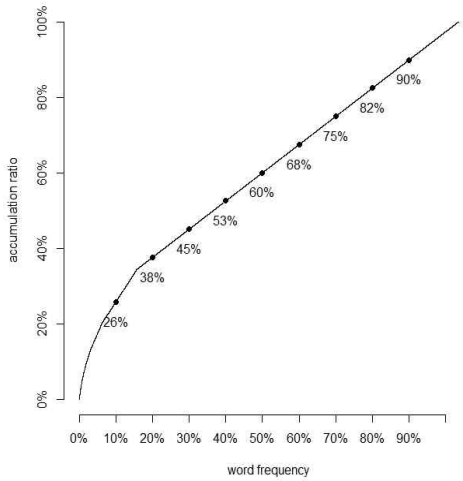

Fig. 4 Accumulated Frequency of Words.

단어축적 빈도는 전처리 완료된 데이터에서 단어별로 사용된 빈도를 축적하여 보여주는 것으로 Fig. 5 는 데이터가 특정 단어에 쏠리지 않고 고르게 사용되고 있음을 보여준다. 마찬가지로 명사를 추출하여 출현 빈도를 구하여 워드클라우드를 구성하면 다음 Fig. 5와 같다

Fig. 5 Wordcloud.

Fig. 5에서 볼 수 있듯이 글자의 크기나 분포 색감을 볼 때 비교적 균일하게 사용되었지만, 문서 중에 트위터, 사이버, 온라인, 암호화, 데이터, 디지털 등의 단어 출현이 다른 단어에 비하여 상대적으로 많은 것을 알 수 있다.

다음으로 키워드 그룹문서 간 상관관계를 피어슨 상관계수를 활용하여 구할 수 있다. 피어슨 상관계수는 한 문서에 word-i가 출현할 때 동시에 word-j가 출현할 확률을 뜻한다. 이 문서와 단어의 행렬을 도치시키면, doc-i가 등장할 때 동시에 doc-j가 등장할 확률을 구하여 키워드문서 간의 상관관계를 구할 수 있다. 14개의 키워드문서의 피어슨 상관관계를 분석해보면, 침해사고가 다른 키워드들과 상대적으로 깊은 연관관계를 맺고 있는 것으로 나타났다. 다른 키워드 문서들과 연관관계가 낮은 키워드는 안보, 국가안보, 정보보안, 정보유출 순으로 나타났다.

각 키워드 문서 간 피어슨 상관계수를 반영한 강한 연관관계를 나타낸 값은 Fig. 6과 같다.

(National) Security	⇔	Cyber Attack
Infringement	⇔	Vulnerability, Security, Security Control, ISMS, Information Security, Information Leak, DDos, Data Protect, Cyber Attack
Cyberattack	⇔	(National) Security, National Security
Vulnerability	⇔	ISMS
Information Leak	⇔	Data Protect
Hacking	⇔	Hacking Incident
DDoS	⇔	Hacking Incident, Infringement,

Fig. 6 Correlation among Keywords.

4.3 토픽모델링

본 연구에서는 R패키지의 LDA 토픽모델링 라이브러리를 사용하였다. 최적의 토픽수를 정하기 위하여 R패키지의 'ldatuning' 라이브러리를 활용하였는데, 혼잡도(perplexity) 지수를 측정하였다. 연구방법론에서 기술한 바와 같이 모델링을 실시할 때 1차적으로 2개부터 20개까지의 토픽수를 4개의 혼잡도 지표에 대하여 모델링을 실시하였다. 그 결과는 Fig. 7과 같다.

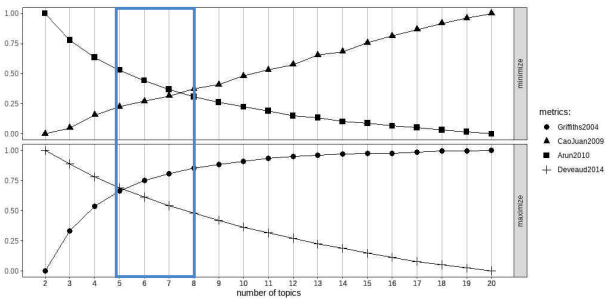

Fig. 7 Number of Topics.

최적의 토픽수를 구하기 위하여 토픽의 모델링 수는 2:20으로 20개까지 모델링을 설정하였더니 위 그림과 같이 5개에서 8개 사이에서 각 지표의 관점에서 최적의 토픽이 존재한다는 것을 알 수 있다. 따라서 토픽수를 5개에서 8개까지 증가시키면서 분석을 진행하였고 최종적으로 선정된 토픽 군집은 다음 Table 5와 같다.

Table 5 Result of Topic Modeling

T1	Naver, Data, NIS, Story, AI, Internet, Security, Research Institute, Character, Program
T2	Encoding, BitCoin, Site, Project, Victim, RansomeWare, Democracy Party, Service, Facebook, Play
T3	RealTime, Cellular, LogIn, Criminal, Committee, Major Company, Software, Foreigner, Chosun Newspaper, Medium Company
T4	Event, Youtube, Cyber, Online, Car, Russia, Virus, Attacker, Lie, University
T5	Android, Korean, Possibility, Laptop, Voice, Smart, Last, Spokesman
T6	Solution, Specialist, Profile, People, Message, Email, Camera, Computer, Partner, Movie

본 논문에서는 Gibbs 샘플링으로 LDA 모델링을 실시하였다. Table 5의 토픽-단어의 결과는 beta값을 기준으로 모델링 결과를 나타낸 것이다. 토픽의 군집은 여러 테스트를 통해서 토픽별 특성이 보이는 6개 그룹을 선택하였다. 각 토픽 그룹별 주요 단어를 시각화하면 다음 Fig. 8과 같다.

Fig. 8 Terms with Beta Score of Topic.

앞서 DTM 행렬을 만들 때에 최초 데이터 수집 시 활용했던 14개 키워드로 그룹핑 하였다. 토픽 모델링 자체가 단어와 문서 그리고 토픽의 출현확률을 추정하는 작업이다. 따라서 각 키워드 그룹별로 토픽과의 출현확률을 산출하는 것은 유의미하다고 볼 수 있다. 최초 DTM을 만들 때의 키워드로 그룹핑이 되어 있으므로, 각 키워드 별로 각 토픽의 발생 확률을 산출하면 Table 6과 같다.

Table 6 Score of Latent Topic Possibility

Title	T1	T2	T3	T4	T5	T6
CyberAttack	0.14	0.26	0.07	0.15	0.22	0.16
Data Protect	0.13	0.24	0.13	0.09	0.18	0.21
DDoS	0.21	0.24	0.11	0.20	0.11	0.13
Hacking	0.13	0.21	0.17	0.17	0.13	0.19
Hacking Incident	0.09	0.16	0.21	0.17	0.18	0.18
Information Leak	0.16	0.19	0.17	0.15	0.13	0.20
Information Security	0.17	0.10	0.20	0.17	0.16	0.20

Infringement	0.14	0.24	0.18	0.13	0.18	0.13
ISMS	0.17	0.21	0.17	0.14	0.16	0.17
National Security	0.18	0.16	0.16	0.17	0.20	0.12
Security Control	0.16	0.28	0.12	0.18	0.14	0.12
Information Security	0.20	0.18	0.20	0.15	0.15	0.13
Vulnerability	0.17	0.19	0.17	0.12	0.19	0.16
(National) Security	0.17	0.13	0.16	0.24	0.14	0.16

위의 Table 6에서 각 토픽과 문서들과의 관계를 살펴보면, T1은 디도스공격과, T2는 사이버공격, 정보보호, 디도스공격, 침해사고, 보안관제 키워드와 연관되어 나타날 잠재적 확률이 높으며, T3는 해킹사고, 보안 및 정보보안, T4는 안보, T5는 사이버공격, T6는 정보보안 및 정보유출 문서와 잠재토픽으로서의 관계를 높이 가지고 있음을 알 수 있다.

5. 결 론

본 논문에서 연구한 이슈 및 키워드들은 일반인들의 접근성이 높지 않고 개념도 어려운 분야이다. 기존의 텍스트 분석 연구 논문에서 주로 연구되었던 뉴스 기사의 이슈나 연구 논문에서 키워드를 발췌하여 동향을 분석했었던 것과 달리 본 논문에서는 정부의 정책문서의 이슈를 키워드로 하여 트위터 데이터를 수집하고 텍스트 분석 및 토픽모델링을 통하여 대중이 재생산하는 토픽과 그 트렌드를 연구하였다는 것에 의의를 둘 수 있다. 트위터 데이터가 가지는 단문의 분석적인 한계에 대해서는 Bae et al.(2013)은 시각화와 시계열 분석을 통해 이를 극복하고자 하였는데, 본 논문에서는 처음부터 트위터 문서들을 수집한 키워드 중심으로 통합하여 말뭉치를(corpus)만들고, 분석을 실시하는 방법을 실시하였다. 토픽모델링 시에는 다시 트위터의 각 문서들로 DTM을 만들어 모델링을 실시하였다. 데이터 분석하기 위하여 공개 프로그램인 R 패키지 프로그램을 활용하여 정보보호 키워드 간의 상관관계와 주요 키워드의 문서 내 출현 빈도를 구하였다. 또한 문서 내 키워드의 출현 빈도를 산정하여 워드클라우드로 시각화하여 주요 키워드들을 한눈에 볼 수 있도록 나타내었다. 토픽모델링에서는 최적의 토픽 개수를 찾기 위해서 Choi et. al.(2020)의 방법을 활용하되, metrics 옵션을 4개 지수 모두 반영하여 검토구간을 최소화하여 최종적으로 분석에 적합한 6개 토픽그룹을 도출하였다. 또한 단어와 토픽, 문서와 토픽간의 잠재적 출현확률을 추정하였다.

본 논문에서의 함의점은 먼저 정보보호 키워드 간 상관관계에 따른 정책 반영 부분이 있을 것이다. 정부 보고서에서 추출된 이슈를 키워드로 검색하여 분석한 결과는 결국 정책입안자에게 향후 정책을 수립할 때 기초자료로 활용될 수 있다. 예를 들어 분석 결과 해킹사고 키워드가 다른 키워드들과 상관관계가 높다는 결과가 나왔는데, 이는 다른 키워드들의 출현확률과 해킹사고 키워드의 출현확률이 연관성이 높으며, 트위터 사용자들이 쉽게 다른 키워드들과 해킹사고 키워드를 함께 사용한다고 볼 수 있다. 또한 안보, 국가안보 키워드의 경우 다른 키워드들과 낮은 상관관계를 가진 결과로 나왔다. 이는 다른 정보보호 키워드와의 연관성이 낮아 정책에서도 독자적인 환경에 사용되어야 함을 의미한다. 이와 같이 뉴스나 연구의 트렌드 이슈 기반 분석 외에도 정부 정책을 수립하고 평가하는 데에도 텍스트마이닝 기반의 데이터 수집과 분석 기술을 활용할 수 있으며, 충분히 유의미한 결과를 도출해 낼 수 있을 것이다.

향후에는 트위터와 같은 소셜미디어서비스에서 유통되는 정보들의 긍정적인 부분과 부정적인 부분을 감성분석을 통해서 정책 수립 시 활용할 수 있는 매체별 혹은 서비스별 정책 데이터베이스 구축과 활용에 대한 연구 등을 고려해 볼 수 있을 것이다.

References

Bae, J. H., Son, J. E., and Song, M. (2013). Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques. *Journal of Intelligence and Information Systems,* 19(3), 141-156. https://doi.org/10.13088/jiis.2013.19.3.141

Cho, S. B., Shin, S. A., and Kang, D. S.(2018), A Study on the Research Trends on Open Innovation using Topic Modeling, *Informatin Policy*, 25(3), 52-74.

Cho, K. W., Han. N. Y.(2021). Research Trends on Emotional Labor in Korea using text mining. *Journal of the Korea Industrial Information Systems Research*, 26(6), 119-133.

Choi, H. Y., Lee, J. R., Jin, M. J.(2020). Intimate Partnerships and Family Policy in Korean News Articles and Comments: A Topic Model Analysis. *Family and Culture*, 32(4), 29-60.

Chung, M. S. & Lee, J. Y.(2018), Systemic Analysis of Research Activities and Trends Related to Artificial Intelligence(A.I.) Technology Based on Latent Dirichlet Allocation (LDA) Model). *Journal of the Korea Industrial Information Systems Research*, 23, 87-95.

Choi, J. H. and Han, D. S.(2011). A Study on the Correlation of Agendas between Politicians' Twitters and traditional News Media. *Journal of Communication*, 11(2), 501-532. https://doi.org/10.22693/NIAIP.2018.25.3.052

D. M. Blei(2012). Probabilistic Topic Model, *Communications of the ACM*, 55(4), 77-94

Kim, E. M. and Lee, J. H.(2011). The Diffusion of News through Twitter and the Emerging Media Ecosystem. *Korean Journal of Journalism & Communication Studies*, 55(6), 152-180.

Kim, N. G., Lee, D. H., Choi, H. C and William Xiu Shun, W.(2017). Investigations on Techniques and Applications of Text Analytics. *The Journal of Korean Institute of Communications and Information Sciences*, 42(2), 471-492.

Ku, G. T(2002). The Impact of Website Campaigning on Traditional News Media and Public Agenda: Based on Agenda-Setting. *Korean Journal of Journalism & Communication Studies*, 46(4), 46-75.

Lee, S. J., and Min, K. S.(2022), Intergrated Interpretation of Network Analysis and Topic Modeling in Text-mining: Focusing on College Competency-based Education. *Journal of Education Evaluation*, 35(1), 165-188.

Lee. S. J(2020), Topic Modeling of Newspaper Articles on Government 'Senior job program' via Latent Dirichlet Allocation. *Journal of Digital Convergence*, 18(10), pp. 537-546

Park, S. H.(2005). On the Journalistic Characteristics and Social Impacts of Internet Bulletin Board as a Public Opinion Space]. *Korea Regional Communication Research Association*, 5(3), 191-226.

Park, J. S,, Hong, S. G., and Kim, J. W.(2017), A Study on Science Technology Trend and Prediction Using Topic Modeling]. *Journal of the Korea Industrial Information Systems Research*, 22, 19-28.

Park, S. H.(2012). Critical Study on the Forming Public Opinion of SNS and Participation Behavior. *Korean Journal of Communication & Information,* 55-73.

Park, K. H., Lee, E. Y., and Yune, S. J.(2021), Counseling Outcomes Research Trend Analysis Using Topic Modeling - Focus on 「Korean Journal of Counseling, *Journal of digital convergence*, 19(11), 517-523.

T. Griffiths and M. Steyvers(2004), Probabilistic

Topic Models, *Proceedings of the National Academy of Sciences* Vol. 101 Issue suppl_1 Pages 5228-523

X. Wang, and A. McCallum(2006). Topics over Time: A Non-Markov Continuous -Time Model of Topical Tren. KDD '06: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 986. https://doi.org/10.1145/1150402

Yoon, T. I. and Shim, J. C.(2003). Agenda-Setting Effects of Controversial Websites. *Korean Journal of Journalism & Communication Studies*, 47(6), 194-219.

정 진 명(JinMyeong Chung)

- 고려대학교 컴퓨터학과 이학학사
- 고려대학교 정보보호 공학석사
- 경북대학교 정보보호학과 박사 수료
- (현재) 한국교육학술정보원 책임연구원
- 관심분야: 정보보호, 빅데이터, 교육정보화

박 영 호(YoungHo Park)

- 경북대학교 전자공학과 학사
- 경북대학교 전자공학과 석사
- 경북대학교 전자공학과 박사
- (현재) 경북대학교 전자공학부 교수
- 관심분야: 정보보호, 빅데이터