JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 5: 731~740, May. 2021

빅데이터 LDA 토픽 모델링을 활용한 국내 코로나19 대유행 기간 마스크 관련 언론 보도 및 태도 변화 분석

서예령[1] · 고금석[2] · 이재우[3*]

An analysis of the change in media's reports and attitudes about face masks during the COVID-19 pandemic in South Korea: a study using Big Data latent dirichlet allocation (LDA) topic modelling

Ye-Ryoung Suh[1] · Keumseok Peter Koh[2] · Jaewoo Lee[3*]

[1]Graduate Student, The Department of Security Convergence Science, Chung-Ang University, Seoul, 06974 Korea
[2]Assistant Professor, Department of Geography, University of Hong Kong, Hong Kong, China
[3*]Assistant Professor, The Department of Industrial Security, Chung-Ang University, Seoul, 06974 Korea

요 약

본 연구는 LDA 토픽모델링 분석을 적용하여 한국 내 세 번의 코로나19 대유행 시기를 기준으로 마스크와 관련된 뉴스 빅데이터를 수집, 분석하였다. 분석 결과 각 시기별 마스크라는 단어를 중심으로 언론보도가 마스크 정책과 관련된 주제에서 사건사고 위주로 바뀌어가는 것을 실증적으로 살펴볼 수 있었다. 즉 제1차 시기의 경우 마스크 생산과 공급이, 제2차 시기에서는 마스크 착용 의무화 및 관련 사건사고가, 마지막인 제 3차 시기에는 주로 사건사고 위주로 토픽이 다뤄진 것을 확인 할 수 있었다. 해당 연구를 통해 마스크 공급, 확보, 착용 외 다른 보건정보에는 상대적으로 소홀했을 가능성을 확인할 수 있었으며, 제2,3차 시기 보도가 사건사고에 치우친 부분은 향후 언론보도의 접근성 및 태도에 대한 개선점이 있음을 시사한다. 따라서 코로나19에 보다 효과적으로 대응하기 위해서는 보다 거시적이고 사회 전체적인 논의가 진행될 수 있도록 언론보도가 변화해야 할 것이다.

ABSTRACT

This study applied LDA topic modeling analysis to collect and analyze news media big data related to face masks in the three waves of the COVID-19 pandemic in Korea. The results empirically show that media reports focused on mask production and distribution policies in the first wave and the mandatory mask wearing in the second wave. In contrast, more reports on trivial, gossipy events consist of the media coverage in the second and third waves. The findings imply that Korea's governmental interventions to address the shortage of face masks and to regulate mask wearing were successful relatively in a short time. In contrast, the study also reports that there may be relative less number of science-based news reports like the ones on the effectiveness of face masks or different levels of filter types. This study exemplifies how a big data analysis can be applied to evaluate and enhance public health communication.

키워드 : 코로나19, 마스크, 뉴스빅데이터, 토픽 모델링 분석
Keywords : COVID-19, Mask, News Big data, Topic modeling analysis

Received 11 March 2021, **Revised** 16 March 2021, **Accepted** 22 April 2021

*** Corresponding Author** Jaewoo Lee (E-mail: jaewoolee@cau.ac.kr, Tel: +82-2-820-5935)
Assistant Professor, The Department of Industrial Security, Chung-Ang University, Seoul, 06974 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.5.731 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

지난해 중국 우한시에서 처음으로 보고된 코로나 바이러스 감염증-19(Coronavirus disease 19, COVID-19, 이하 코로나19)은 2021년 현재에도 진행 중인 전세계인 감염병 사태다. 우리나라에서는 지난 2020년 1월 첫 코로나19 확진자가 발생한 이후 3회의 전국적 대유행 시기를 거쳤고, 언론매체는 각 시기별로 다양한 관점에서 코로나19의 발병상황과 그에 연관된 소식을 꾸준히 보도해왔다[1]. 코로나19와 같은 위기에 적시, 적절하게 대응하기 위해서는 사회 내 효과적인 커뮤니케이션이 중요하다[1]. 최근에는 사회적 의제 설정 및 의사 소통이 소셜 네트워크 서비스(Social Network Service, SNS)를 통해 이루어지고 있으나[2], 아직도 기존 언론은 사회적 담론을 주도하고 형성하는데 상당한 영향력을 미치고 있다[3]. 이에 따라 주요 언론 매체가 코로나 19에 대해 어떻게 보도를 하고 사회적 담론을 형성해가는지 분석할 필요가 있다.

이를 위해 본 연구는 국내 주요 언론사가 보도한 코로나 19에 관한 뉴스를 수집하여 사회적 담론이 어떻게 형성되고 변화했는지 파악하고자 한다. 특히 '마스크'를 키워드로 이용하여 관련 뉴스 보도를 살펴보고자 하는데, 이는 코로나19 초기에 마스크와 관련된 많은 사회적 갈등과 논란이 상당 기간 지속되었기 때문이다[4]. 연구 방법은 주요 언론매체의 뉴스 데이터를 각 대유행 시기에 따라 수집한 뒤, 토픽 모델링 분석을 통해 보도에 사용된 주요 단어 간의 연결성을 파악하여 각 시기 별 변화된 언론의 보도 내용과 태도, 그리고 보도내용을 통해 유추 가능한 시대상을 관찰하려 한다.

본 연구 결과는 데이터 사이언스의 방법을 통해, 향후 발생 가능한 코로나19와 같은 공공 의료 및 재난 상황에서 정부가 공공 의사소통 전략을 만들 때 언론의 보도관행에 어떻게 대응해야 하는지 여러가지 시사점을 제공한다. 또한, 빅데이터를 기반으로 한 연구가 커뮤니케이션 분야에서 어떻게 활용되어 위기 관리 대응 및 공공정책 입안에 도움을 줄 수 있는지 실증적 사례를 제시한다.

Ⅱ. 관련 연구

2.1. LDA 토픽 모델링

토픽 모델링(Topic Modeling)은 텍스트 마이닝 기법 중 최근에 등장한 분석 방법으로, 토픽별 키워드의 분포를 바탕으로 개별 텍스트의 주제를 추정하여 동일한 주제별로 텍스트를 분류한다[5]. 이 방식은 기존 키워드 분석에서는 불가능했던 비정형 데이터 해석 및 이용에 적합하다. 다만 토픽 모델링을 효율적으로 적용하려면 문서에 존재하는 복잡성을 고려해야 한다[6]. 분석한 텍스트 정보가 효율적으로 이용되기 어려운 경우에는 문서와 문서에 나타나는 단어 사이에 토픽 계층이 존재한다고 가정하고, 이 잠재적인 정보(Latent Information)를 추출하여 최종 모형을 만들어낸다[7].

본 연구는 잠재 디리클레 할당 알고리즘(Latent Dirichlet Allocation, LDA)을 사용하였다. LDA 토픽 모델링 알고리즘은 각 단어나 문서의 숨겨진 주제를 찾아, 문서와 키워드별로 묶어 주제를 형성하는 비지도 학습 알고리즘(unsupervised learning algorihm)이다[8]. 그림 1은 해당 알고리즘을 설명한다. 원하는 토픽의 개수와 디리클레 분포(Dirichlet Distribution) 조절에 대한 매개변수 값을 조절, 문서와 단어를 하나씩 관측하면, 문서/단어마다 적절한 토픽을 부여되고 Z값이 결정된다. 디리클레 분포는 이항분포가 아닌 연속확률 분포 중 하나로, k 차원의 실수 벡터 중 벡터의 요소가 양수이며 모든 요소를 더한 값을 1로 하여 확률 값이 정의되는 분포를 뜻한다. 따라서 LDA 알고리즘은 디리클레 분포를 가정하고, 토픽이라는 잠재된 변수를 활용해, 각 문서와 단어가 토픽에 할당되는 확률 분포를 그린 것이다[9]. 이런 특징을 바탕으로 LDA 토픽 모델링을 이용한 기존 연구는 대량의 뉴스 텍스트 데이터에서 주요 주제를 추출, 사회적으로 형성된 최신 트렌드나 주요 의제를 인지하고 그 변화상을 확인하였다[10],[11]. 4차 산업 연구 동향을 파악하기 위한 연구[10]에서는 수집한 논문 데이터를 바탕으로 토픽모델링을 사용하였다. 그 결과 현재 4차 산업과 관련한 다양한 분야에서 주요 기술에 대한 연구를 진행 중에 있음을 확인하였다. 마찬가지로 신문 자료의 오피니언 마이닝에 대한 연구[11]에서는 토픽모델링을 통해 매체의 정치적 성향에 따라 주요 단어 분포나 주요 의제가 다름을 확인할 수 있었다. 이와 같이 나타난 토픽모델링의 특성은 코로나19로 인해 발생한 뉴스 매체의 주제적 특징을 확인할 수 있을 것이라 판단된다.

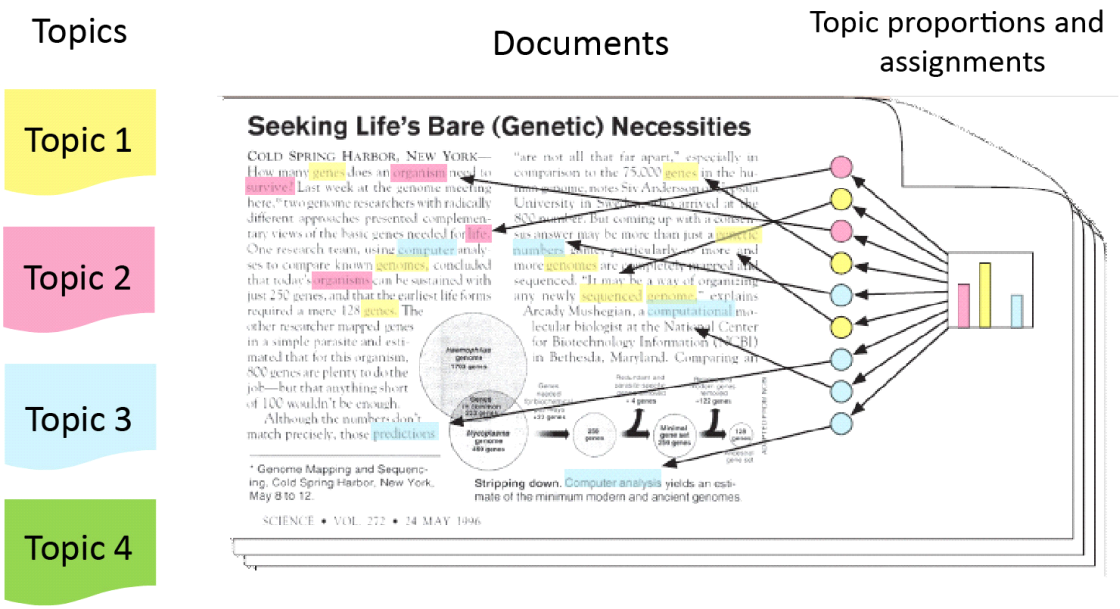

Fig. 1 The Overview of Topic Modeling Process (modified from [19])

2.2. 코로나 관련 연구

코로나19 발생 이후 국내외 많은 관련 연구들이 의료, 경제, 사회, 문화 등 다방면에서 진행돼왔다. 본 논문은 이 가운데 언론 보도와 관련된 연구에 초점을 맞춰 문헌 조사를 진행하였다.

국내 연구의 경우 크게 언론 보도 동향에 대한 연구와 코로나19관련 토픽 연구로 분류할 수 있다. 먼저 코로나19에 관련된 뉴스 프레임에 대한 연구[1]는 공영 방송사 KBS의 주요 뉴스 프로그램 <뉴스 9>을 통해 보도된 뉴스를 분석, 코로나19 이후 변화한 보도의 형식, 유형, 전문성 등 언론 보도 동향에 관해 분석하였다. 해당 연구는 전염병의 '발각/확산/대유행(회복)'이라는 시기적 구분을 기반으로 데이터를 수집하여 분석하였다. 다만 연구가 코로나19 발생 초기인 2020년 1월 20일부터 5월 2일까지로 한정되어 총 473건의 적은 보도 건수만 분석했고, 정치, 경제적 사안과 연관된 보도를 구분하지 않았다는 한계가 있다.

한편 토픽 분석 연구의 경우 주로 LDA를 사용한 경우가 많았는데, 코로나19 전체를 분석한 연구와 코로나19와 관련된 특정 키워드를 분석한 연구로 나눌 수 있다. 코로나19 전체를 분석한 연구는 주로 뉴스 빅데이터를 활용한 연구가 대다수였다[12],[13]. 코로나19의 언론보도를 분석한 연구[12]에서는 각 토픽의 주제를 분석하여 과거 신종플루·메르스 시기의 언론태도를 분석 및 비판하였다. 또한 CNN의 코로나19 뉴스 영상 분석에 대한 연구[13]에서는 뉴스의 구조, 시각적 주제, 표현 방법에 대한 연구를 통해 코로나19에 대한 해외 언론의 태도 변화에 대해 분석하였다. 두 연구 결과 모두 코로나19에 대해 장기적으로 분석하기보다 2020년 1월~3월로 한정한 좁은 시점의 데이터를 분석했다는 점에서 한계를 보인다. 특정 주제만을 분석한 경우 코로나19로 인해 수요가 급증한 온라인 교육에 대한 토픽모델링을 진행한 연구[14]가 있었는데, 해당 연구의 경우 키워드의 중요도를 알아보는 TF-IDF (Term Frequency - Inverse Document Frequency)와 LDA를 통한 토픽모델링을 실시하였다. 다만 또한 수집된 데이터의 기간과 양이 2020년 5월 15일부터 5월 30일까지이며, 총 602건을 분석했다는 점에서 시계열적인 특성을 살펴보지 않았다는 한계가 있다.

국외 연구의 경우에는 언론매체 보다는 주로 SNS 데이터를 대상으로 토픽모델링이 사용하였다[15],[16], [17]. 먼저 해외에서 많이 사용하는 트위터를 분석한 연구[15]는 토픽 모델링으로 단어 빈도분석을 실시하여 코로나19 발생 이후 트위터 이용자들의 감정 변화를 확인하였다. 또한, 트위터 사용률과 신뢰도가 높아 보건당국이 올바른 정보 전달을 하는 창구로 사용될 수 있음을

시사하였다. 마찬가지로 트위터를 통해 코로나19에 대한 대중의 인식과 주제형성에 대한 연구[16]에서는 트위터에 대한 토픽모델링을 통해 코로나19에 대한 대중의 관심과 인식을 이해하는데 효과적으로 사용될 수 있음을 밝혔다.

한편 국외에서 많이 사용되는 Facebook에서 수집한 데이터를 기반으로 한 연구[17]의 경우, 가짜뉴스를 토픽모델링으로 분석하여 Facebook에서 유포되는 정보로 인해 사회적 분열이 생길 수 있음을 제시하였다. 다만 마찬가지로 짧은 기간에 수집된 데이터에 한정하였고, Facebook의 특성상 기존 언론 매체보다는 대체 언론을 중심으로 분석되었다는 한계가 있다.

본 연구의 경우 선행연구와는 다르게 1월 1일부터 11월 30일까지로 대상 기간을 선정하여 오랜 기간 주류 언론을 중심으로 데이터를 수집하였다는 점에서 의미가 있다.

Ⅲ. 데이터 및 연구 방법

3.1. 연구 방법의 개요

본 연구 데이터는 검색포털 네이버의 뉴스 크롤링을 통해 수집된 2020년 1월 1일부터 2020년 11월 30일까지 총 11개월간의 뉴스데이터이다. 매체의 유료부수, 열독률, 신뢰도 조사 등을 참조하여 조선일보, 중앙일보, 동아일보, 한겨레, 경향신문, 오마이뉴스 등 6개 언론사와 KBS, MBC, SBS 등 3개의 방송사로 연구 대상을 한정하였다. 코로나19 유행 기간 마스크 구입과 착용에 관한 논란이 많았던 것을 감안하여 분석 뉴스 데이터는

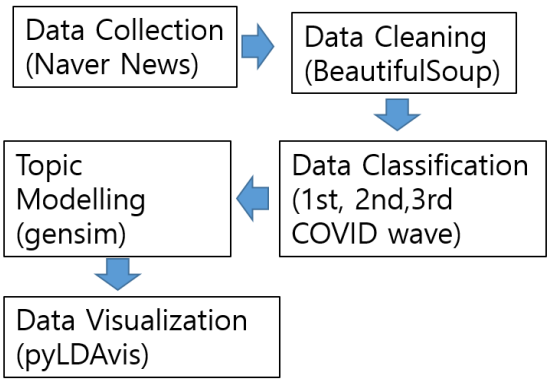

Fig. 2 The overview of research method

'마스크'라는 단어가 사용된 기사에 국한하였다. 연구 진행의 개요는 그림2와 같다: 데이터 수집한 후, 데이터를 정제하고, 기간에 따라 데이터를 분류한 후, 토픽 모델링을 설계하여 분석하고, 데이터를 시각화하였다. 각 단계에 대한 자세한 설명은 아래에서 소개하도록 한다.

3.2. 데이터 수집 및 데이터 정제

본 연구에서는 HTML과 XML파싱 파이썬 소프트웨어인 BeautifulSoup[1)]을 이용하여 개별 기사의 날짜, 제목, 매체, 링크를 직접 수집해 최종 9,566개의 뉴스데이터를 확보하였다. 한국어 자연어 처리 파이썬 라이브러리인 KoNLPy[2)]을 이용하여 수집한 모든 텍스트로부터 0부터 9까지 숫자, 대문자 A부터 소문자 z까지 모든 영문자, ' @ ', ' ? ', ' ", ' ! ' 등과 같은 기호와 특수문자를 제거하였고, 일부 영어 약어 및 단어는 한글로 바꾸어 데이터 누락 가능성을 최소화하였다. 예를 들어 WHO는 세계보건기구로, KF94는 케이에프구사로, KF80은 케이에프팔공으로 변환하였다. 마찬가지로 KoNLPy 모듈의 okt.nouns()를 이용한 한글 형태소 분석을 통해 명사 형태소를 뽑아내는 과정을 수행하였다. 이를 통해 '은/는', '을/를', '이/가', '와/과' 등의 조사나 '이것', 저것' 등의 대명사, 형용사와 같은 형태소를 제거하였다. 마지막으로 데이터 수집시에 사용한 '마스크,' 자주 등장하지만 분석에 의미가 없는 '코로나', '코로나19'와 같은 단어를 삭제하였다.

3.3. 기간 분류

효율적인 데이터 분석을 위하여 전체 코로나19 기간을 분석하기보다, 세 번의 대유행 기간을 반영하여 일부 기간을 연구에 이용하였다. 대한민국 통계청 코로나19 일간 확진자 발생현황데이터를 기반으로 그림 3과 같은 코로나19 확진자 발생 추세 그래프를 그릴 수 있었다. 그림 3을 분석하면 제1차 대유행은 2020년 2월 22일부터 3월 7일, 제2차 대유행은 2020년 8월 20일부터 9월 3일, 제3차 대유행은 2020년 11월 19일부터 2020년 11월 30일으로 나타났다. 본 연구는 해당 대유행 기간 최고점을 기준으로 전후 7일을 가져와 각 시기당 15일의 데이터를 분석하였고, 연구를 위한 마지막 데이터 분석 기간

1) https://www.crummy.com/software/BeautifulSoup/
2) https://konlpy.org/

중 최고점이 11월 27일이기 때문에 해당 날짜 이후 7일이 아닌 데이터 수집 마지막 기간인 11월 30일까지로 설정하였다.

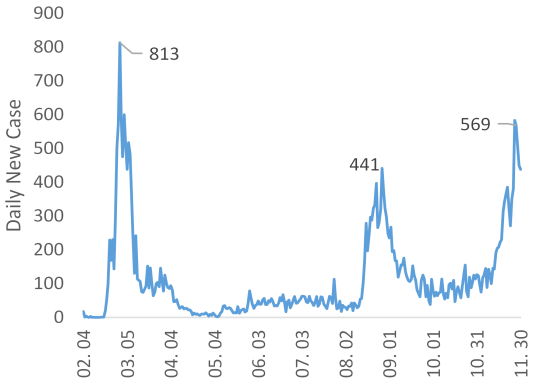

Fig. 3 Daily Number of Confirmed Cases

3.4. LDA 토픽 모델링 설계

본 연구의 LDA 토픽 모델링은 토픽 모델링 연구에서 일반적으로 사용하는 파이썬 소프트웨어인 gensim[3]을 사용하였다. 또한, gensim 모듈의 CoherenceModel의 연산과정을 이용해 토픽 개수에 따른 일관성 점수(coherence score)를 산출하여 그림 4과 같은 그래프를 도출하였다. 일관성 점수란 토픽의 개수를 정할 때 사용하는 척도로, 주어진 토픽 개수(x축)에 대해 각 토픽 내의 단어가 의미론적으로 일치하는지를 수치로 계산한다. 따라서 일관성 점수가 가장 높은 토픽의 수 값이 토픽모델링을 돌리고자 하는 데이터의 적정 토픽 수이다.

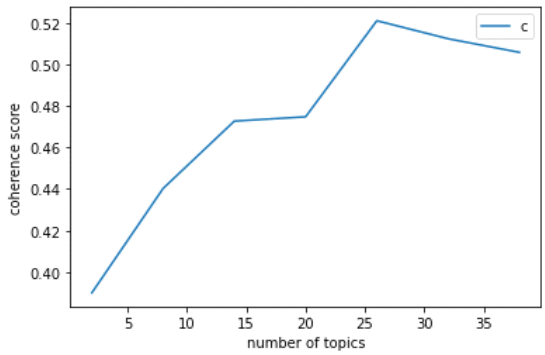

Fig. 4 Estimation of Coherence Score

LDA 모델링을 수행하려면, 토픽의 수 외에도 데이터 샘플링의 반복횟수와 토픽별 키워드의 수를 연구자가 지정해야 한다. 본 연구에서 사용한 샘플링 횟수는 토픽 모델링 연구에서 일반적으로 사용하는 500회로 설정하였다. 이러한 설정하에서, 가장 높은 일관성 점수를 가지는 토픽 수를 선택하였다. 그 결과, 1차 시기에는 14개, 2차 시기에는 26개, 3차 시기에는 20개의 토픽이 선택되었다. 마지막으로 도출된 토픽 중 분석에 의미없는 토픽을 제외하여 1차시기 8개, 2차와 3차시기에 각 10개씩의 토픽을 도출하였다.

파이썬용 LDA 시각화 라이브러리인 pyLDAvis[4]를 통해 그림 5와 같이 학습된 모델을 시각화하였다. 원의 위치는 토픽 간의 거리가 멀수록 판별 타당도가 높고 주제가 뚜렷하게 구분된다. 토픽 간의 거리가 가깝거나 겹쳐져 있다면 판별 타당도가 낮음으로 비슷한 주제를 나타낸다. 크기는 해당 토픽이 해당 데이터 내에서 얼마만큼의 비중을 차지하는 가를 의미한다. 이를통해 각 토픽별 분포를 파악할 수 있고, 토픽 내의 키워드가 다른 토픽에는 어느 정도의 비율로 차지하고 있는지 쉽게 파악할 수 있다.

3.5. 실험 환경

데이터 분석을 위한 실험환경은 Google 사의 클라우드 컴퓨팅 환경인 CoLab서비스[5]를 이용하였다. 본 논문의 데이터 분석을 위하여 Colab에서 Intel Xeon 2.2GHz 만큼의 CPU 성능과 13GB 크기의 RAM을 할당 받아 사용하였다.

Ⅳ. 연구 결과

4.1. 1차시기 분석 결과

2020년 2월 22일부터 3월 7일까지의 뉴스 기사를 통해 도출된 토픽 결과는 표 1과 같다. 분석 결과 14개의 토픽 중 의미를 가진 토픽은 그림 6과 같이 8개가 도출되었다. 이러한 토픽 개수의 차이는 데이터 처리 과정 중 데이터의 양이 많아 토픽이 하나로 겹치기 때문에 발생하는데[18], 이런 경우 대부분 비중이 낮은 내용으로 구성되므로 분석의 정확성을 위해 해당 토픽들은 삭제

3) https://radimrehurek.com/gensim/

4) https://github.com/bmabey/pyLDAvis

5) https://colab.research.google.com

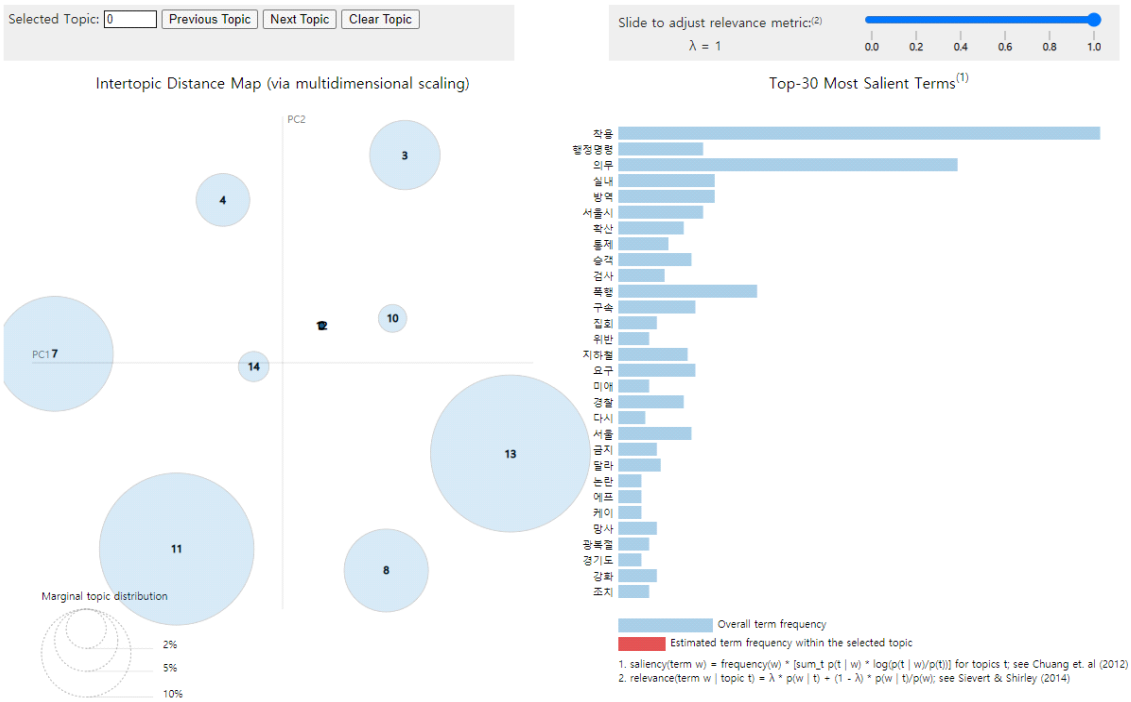

Fig. 5 Visualization of Topic Modeling

하였다. 또한, 표1에서 토픽별 5개의 단어를 뽑은 이유는 6순위 단어부터는 비중이 급격하게 떨어지기 때문에, 5개의 단어를 선택해서 표시하였다.

도출된 토픽의 주제를 확인해본 결과 제1차 대유행 시기에는 대부분 마스크 구입과 공급에 대한 주제가 대부분이었다. 먼저 가장 비중이 높은 토픽인 [10번 토픽]은 '판매', '수출', '공급', '공적', '제한'과 같은 단어가 주요 단어로 구성되어 있다.

Table. 1 Top 5 Word in each Topic During the 1st Covid Wave

Period	Topic num	Top 5 words in each topic
1st period (02.22. ~03.07.)	1	공급, 제한, 정부, 수출, 지원
	2	공급, 판매, 오늘, 정부, 우체국
	4	대통령, 정부, 생산, 판매, 업체
	5	하나로마트, 약국, 판매, 생산, 지원
	6	공급, 진자, 정부, 생산, 부족
	8	판매, 공급, 정부, 대구, 오늘
	10	판매, 수출, 공급, 공적, 제한
	11	대구, 공급, 경북, 이마트, 지역

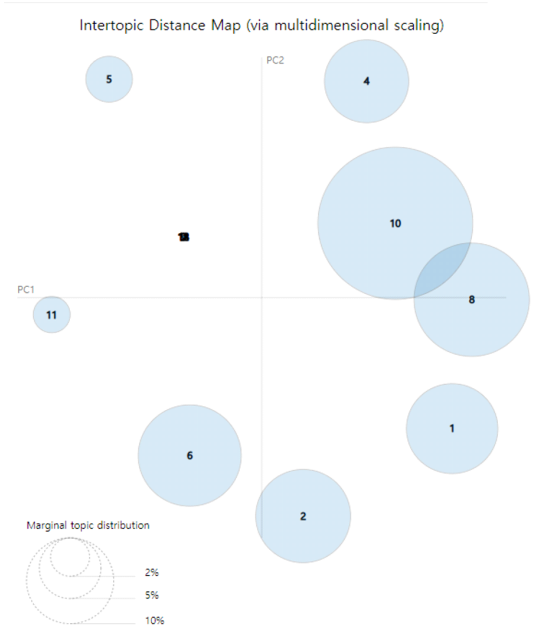

Fig. 6 The Intertopic Distance Map and Top 5 Words in Each Topic During the 1st Covid Wave

이는 코로나 초기 마스크 공급과 관련하여 물량 확보를 위해 정부가 마스크 수출 규제를 진행하고, 동시에 마스크 생산량을 공적 판매로 전환한 것과 관련된 단어다. 둘째, 2번, 5번, 8번, 10번, 11번 토픽에서는 마스크 공적 판매에 내용이 도출되는데, 2번과 5번의 경우 마스크 판매 장소에 대한 내용으로 2번은 우체국에서, 5번은 하나로마트에서 마스크 공적 판매를 진행한다는 주제였다. 8번과 11번의 경우 마스크 공급에 있어서 1차 시기에 주로 코로나19 확진자가 발생되었던 대구경북지역에 우선하여 마스크 공적 공급을 진행한다는 내용이었다. 셋째, 1번과 10번에서는 마스크 수출 및 매점매석 방지를 통해 국내 물량을 확보하기 위한 노력을 확인할 수 있다. 마지막으로 4, 6번에서는 국내 마스크 물량 부족을 극복하기 위해 생산량을 늘려야 한다는 내용이 나타난다. 이와 같이 제1차 대유행 시기에는 마스크의 확보와 관련된 내용이 주요 언론을 통해 다뤄지고 있음을 알 수 있다.

4.2. 2차시기 분석 결과

Table. 2 Top 5 Word in each Topic During the 2nd COVID Wave

Period	Topic num	Top 5 words in each topic
2nd period (08.20. ~09.03.)	1	착용, 미애, 폭행, 망사, 행동
	5	착용, 의무, 서울시, 서울, 집회
	8	착용, 의무, 시설, 미착, 기사
	9	의무, 착용, 실내, 감염, 구속
	10	착용, 의무, 폭행, 행정명령, 확산
	16	이슈, 가로, 확대, 업체, 북적
	21	착용, 의무, 서울, 요구, 사업
	24	지하철, 승객, 과태료, 파티, 부과
	25	실내, 금지, 다시, 행사, 행정고시
	26	시술, 완료, 철썩, 철통, 전파

코로나19 제2차 대유행 시기인 2020년 8월 20일부터 9월 3일까지의 뉴스 기사를 통해 도출된 토픽 결과는 표 2와 같고, 이를 그림 7에서 시각화하였다. 제2차 시기에는 모두 10개의 토픽이 도출되는데, 제1차시기의 주요 이슈였던 마스크 공급과 관련된 내용은 없고, 대부분 마스크 착용과 관련된 내용이 주를 이룬다. 이는 마스크 공급및 분배가 제1차 시기에 이미 안정이 되어 국민 대부분이 마스크를 원활히 확보할 수 있는 시대상을 반영한다.

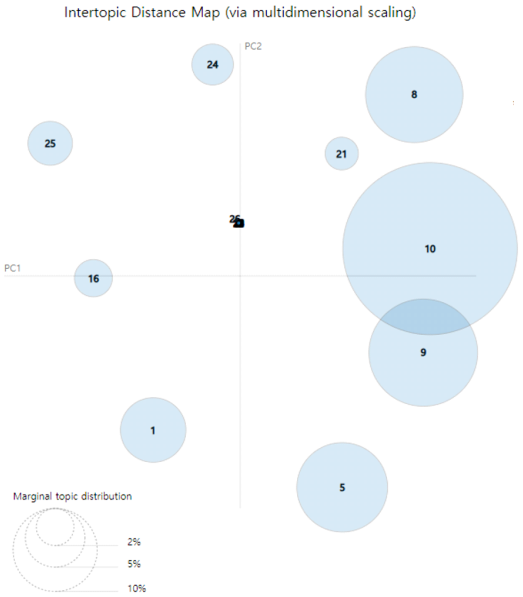

Fig. 7 The Intertopic Distance Map and Top Five Words in Each Topic During the 2nd Covid Wave

가장 비중이 높은 10번 토픽에서는 '착용', '의무', '폭행', '행정명령', '확산'과 같은 단어를 볼 수 있는데 이는 해당 시기 마스크 착용과 그에 관련된 사건사고 등이 발생함을 확인할 수 있다. 실제로 다른 토픽의 주제들 역시 마스크 착용 의무화와 사건사고, 두 가지로 크게 나눌 수 있다. 5, 8, 9, 21번 토픽의 경우 대체로 실내외 마스크 착용 의무화에 대한 정보 전달에 초점이 맞춰져 있다. 1, 16, 24, 26번의 경우 마스크 착용과 관련된 사건사고형 토픽이 주로 나타나는데 특히 1번의 경우 해당 시기가 더운 여름철이라 망사 마스크, 비말 마스크 등 KF94, KF80이 아닌 마스크의 사용이 늘면서 발생하였던 사회적 논란이 반영된 것으로 보인다. 또한 마스크 미착용자들로 인한 사건사고가 관련된 단어를 중심으로 구성되었음을 표2에서 확인할 수 있다.

4.3. 3차시기 분석 결과

대유행 제3차 시기 중 본 연구의 데이터 수집시간에 해당하는 2020년 11월 19일부터 11월 30일까지의 뉴스를 통해 도출된 토픽 결과는 표 3과 같다. 도출된 총 8개의 토픽 가운데 가장 눈에 띄는 부분은 과거 제1, 2차시기에 비해 공통적인 주제가 적고, 정보전달이나 마스크

생산, 확보, 착용의무 등과 같은 거시적 담론 보다는 사건사고 전달 위주로 바뀌었다는 점이다. 이같은 제3차 시기 감염자 수가 급증하는 상황에서도 더이상 마스크 공급 및 착용은 주요의제가 아님을 확인해 준다.

Table. 3 Top 5 Word in each Topic During the 3rd COVID Wave

Period	Topic num	Top 5 words in each topic
3rd period (11.19. ~11.30.)	1	확진, 과태료, 해외, 외교, 도박
	2	폭행, 택시, 기사, 역무, 격분
	3	미국, 등장, 의무, 시위, 봉쇄
	4	활용, 관리, 수능, 구매, 확산
	5	폭행, 도주, 지하철, 경찰, 피부
	6	폭행, 기사, 구매, 달라, 택시
	7	감염, 폭행, 택시, 기사, 착용
	8	공무원, 당진시, 행패, 직위해제, 달라

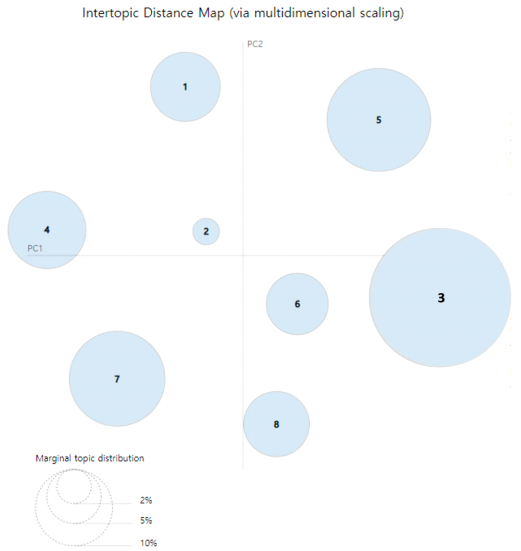

Fig. 8 The Intertopic Distance Map and Top Five Words in Each Topic During the 3rd Covid Wave

그림 8을 보면 토픽 1,2,3,7,8번의 크기가 균일한 것을 알 수 있는데, 이는 곧 해당 시기 마스크와 관련한 주도적인 이슈가 적음을 의미한다. 토픽 3, 4번을 제외한 토픽 대부분은 마스크 착용거부로 생긴 사건사고였다. 토픽 3번은 미국 내 코로나19 재유행에 따른 주정부들의 봉쇄조치와 이에 맞서는 방역거부 및 시위를, 토픽4번의 경우 해당 시기 중 국내에서 치뤄진 대학수학능력시험과 관련된 뉴스에서 생성된 토픽으로 분석된다.

Ⅴ. 결 론

본 연구는 LDA 토픽모델링 분석을 적용, 한국 내 세 번의 코로나19 대유행 시기를 기준으로 마스크와 관련된 뉴스 빅데이터를 수집, 분석하였다. 분석 결과 각 시기별 마스크라는 단어를 중심으로 언론보도가 마스크 정책과 관련된 주제에서 사건사고 위주로 바뀌어가는 것을 실증적으로 살펴볼 수 있었다.

즉 제1차 시기의 경우 마스크 생산과 공급이, 제2차 시기에서는 마스크 착용 의무화 및 관련 사건사고가, 마지막인 제3차 시기에는 주로 사건사고 위주로 토픽이 다뤄졌다. 전례없는 전세계적인 감염병 위기 상황에서 빠른 시간 , 세계 다른 나라에 비해 마스크 확보와 착용이 단시간 내에 해결된 한국의 상황을 생각해보면, 주요 매체의 집중적인 언론보도가 정부의 마스크 공급정책 수립 및 실행에 적절한 자극을 제공했을 가능성이 있다. 또한 언론 보도가 마스크 착용율 제고, 마스크 5부제 협력 등 정부 보건정책에 대한 공공의 협력을 이끄는 데도 긍정적인 영향을 미친 가능성이 있다. 이에 반해, 마스크 공급, 확보, 착용 외 다른 보건 정보에는 상대적으로 소홀했을 가능성 역시 이 연구를 통해 확인할 수 있다. 특히 마스크의 효과, 올바른 착용법, 필터 성능차 등 보다 심층적이고 과학적인 정보가 보다 많은 횟수로 제공되지 않았다는 점을 본 연구를 통해 실증적으로 알 수 있었다. 또한 제2차 시기의 데이터에서 망사, 비말 마스크가 언급되었으나 그 효과성과 관련된 토픽들이 추출되지 않았던 점, 제2,3차 시기 보도가 개별적 사건사고에 보다 치우쳤던 부분은 향후 언론보도의 접근성 및 태도에 대한 개선점이 있음을 시사한다[6]. 향후 코로나19와 같은 위기 상황에서는 보다 거시적인 담론 형성과 사회 통합적인 대응에 언론이 보다 건설적으로 기여할 수 있도록 사회 구성원이 모니터링을 해야 할 것이며, 언론 역시 변화해야 할 것이다.

본 연구는 '마스크'라는 단어를 기준으로 뉴스 데이

6) 뉴스 보도의 특성상 심층적이고 과학적인 정보보다는 사건/사고에 보도가 치우칠 수 밖에 없다는 점도 고려해야 한다.

터를 검색했고, 모든 언론사를 포함시키지 못했다는 점, 연구기간이 지난 2020년 11월 30일까지만 이뤄진 점이 한계다. 예를 들어 2020년 하반기 부터는 자영업자나 소상공인 지원 및 보조금 등에 대한 논의들이 사회적으로 진행되었으나 본 연구 데이터에서는 보조금 관련 뉴스는 3건, 소상공인 관련 뉴스는 8건, 자영업은 0건으로 나타나 유의미한 분석이 가능하지는 않았다. 하지만 이는 본 연구가 '마스크'와 관련된 뉴스 데이터 분석에 목적을 둔 때문이고 연구방법론에 문제가 있다는 점을 의미하는 것은 아니다.

향후 연구에서는, 다른 단어들－예를 들어, '백신', '보조금' 등－을 중심으로 코로나19에 대한 언론보도 분석을 시행하거나, 국내와 해외 뉴스의 교차 비교, SNS과 기존 언론간 비교 분석, 시간 추이에 따른 언론보도 키워드 변화 등을 분석한다면 보다 다양한 방향에서 코로나19와 관련된 국내외 여러 사회상과 의사 소통 과정 및 행태를 분석할 수 있을 것이다.

ACKNOWLEDGEMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2018-0-01799, IITP-2021-2020-0-01655) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

[1] S. Y. Pyo, "Frame Analysis of Corona-19 News on Korean Public Broadcasting System: Focused on KBS," *The Journal of the Korea Contents Association*, vol. 20, no. 12, pp. 112-122, Dec. 2020.

[2] S. T. Ha, "Impact of Using SNS on Individuals Social Capital," *Society for Journalism and Communication Studies*, vol. 21, no. 2, pp. 73-99, May. 2017.

[3] S. R. Lee and E. J. Choi, "Comparison of responses to issues in SNS and Traditional Media using Text Mining-Focusing on the Termination of Korea-Japan General Security of Military Information Agreement (GSOMIA)-," *Journal of Digital Convergence*, vol. 18, no. 2, pp. 277-284, Feb. 2020.

[4] H. L. Kim, "COVID-19 : Our Memory: A Digital Archive for Social Changes caused by SARS-CoV-2," *Journal of Korean Society of Archives and Records Management*, vol. 20, no. 4, pp. 229-237, Nov. 2020.

[5] I. Vulić, W. D. Smet, J. Tang, and M. Moens, "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications," *Information Processing & Management*, vol. 51, no. 1, Jan. 2015.

[6] C. H. Nahm, "An Illustrative Application of Topic Modeling Method to a Farmer's Diary," *Cross-Cultural Studies*, vol. 22, no. 1, pp. 89-135, Jan. 2016.

[7] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, Apr. 2012.

[8] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169-15211, Nov. 2018.

[9] D. M, Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993-1022. Jan. 2003.

[10] K. W. Cho, K. Won, and Y. W. Woo, "Topic Modeling on Research Trends of Industry 4.0 Using Text Mining," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 23, no. 7, pp. 764-770, Jul. 2019.

[11] B. I. Kang, M. Song, and W. S. Jho, "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling," *Journal of the Korean Society for Library and Information Science*, vol. 47, no. 4, pp. 315-334, Nov. 2013.

[12] T. J. Kim, "COVID-19 News Analysis Using News Big Data : Focusing on Topic Modeling Analysis," *Journal of the Korea Contents Association*, vol. 20, no. 5, pp. 457-466, May. 2020.

[13] J. S. Lee, "Visual Analysis of CNN's COVID-19 News Coverage : From the Disease of the 'Others' to the 'Us'," *Media, Gender & Culture*, vol. 35, no. 2, pp. 245-298, Jun. 2020.

[14] S. M. Kim, "Analysis of Press Articles in Korean Media on Online Education related to COVID-19," *Journal of Digital Contents Society*, vol. 21, no. 6, pp. 1091-1100, Jun. 2020.

[15] D. P. Kaila and D. A. Prasad, "Informational flow on Twitter – Corona virus outbreak – topic modelling approach," *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 3, pp. 128-134, Mar. 2020.

[16] S. Boon-Itt and Y. Skunkan, "Public Perception of the

COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study," *JMIR Public Health Surveill*, vol. 6, no. 4, Nov. 2020.

[17] S. Boberg, T. Quandt, T. Schatto-Eckrodt, and L. Frischlich. (2020, April). Pandemic populism: Facebook pages of alternative news media and the corona crisis--A computational content analysis [Internet]. Available: https://arxiv.org/abs/2004.02566.

[18] Media Big Data Research Center, "Visualization of text mining, analysis and utilization of topic modeling," *BFly News Letter*, Jun. 2020.

[19] D. M. Blei. "Probabilistic topic models," *Communications of ACM*, vol. 55, no. 4, pp. 77-84, 2012.

서예령(Ye-Ryoung Suh)

2019 중앙대학교 산업보안학과 졸업(학사)
2019~ 중앙대학교 융합보안학과 석사과정
※ 관심분야 : 블록체인, 사이버 물리 시스템 보안

고금석(Keumseok Peter Koh)

2002년 서울대학교 농경제사회학부(학사)
2004년 서울대학교 농경제사회학부(석사)
2004-2006년 서울신문 기자
2016년 Michigan State University, PhD in Geography
2016-2018년 The Ohio State University, Postdoctoral Researcher (Public Health)
2018년-현재 University of Hong Kong, Assistant Professor (Geography)
※ 관심분야: Health/medical geography, Computational social science, Urban planning

이재우(Jaewoo Lee)

2006년 서울대학교 컴퓨터공학부 (학사)
2008년 서울대학교 컴퓨터공학부 (석사)
2017년 University of Pennsylvania, Ph.D in Computer and Information Science
2018년~현재 중앙대학교 산업보안학과 조교수
※ 관심분야 : 실시간 시스템, 사이버 물리 시스템 보안