

03

버트 토픽 (BERTopic) 방법으로 분석한 코로나19 백신 (COVID19 vaccination) 언론보도의 특성과 백신 접종자 수에 미치는 영향 연구: 기계학습과 모델링기법의 결합

●박영은*

숙명여자대학교 홍보광고학과 조교수

본 연구는 최신 전환학습(Transfer learning) 텍스트 모델링 기법인 BERTopic을 통해 총 38,312개의 코로나19 백신 관련 기사를 분석함으로써 22개의 주요 토픽을 생성하였고, 그중에 가장 많이 언급된 토픽들은 코로나 확진자 추세, 백신 접종, 치료제 개발, 변이와 집단감염, 등등의 순이었다. 시계열 요소를 도입한 BERTopic dynamic topic model을 통해 살펴본 토픽의 변화는 상대적으로 예방접종이 잘 진행되고 있던 시기에는 백신 접종에 대한 보도가 줄어들고 델타 변이가 출현한 시점부터 오히려 백신 보도의 비율이 늘어나는 양상을 보였다. 시계열 특성을 회귀분석에 포함한 방법으로 모델을 추정하였을 때, 단순히 코로나 확진자 수를 보도하거나 백신 개발을 강조하는 보도는 오히려 백신 접종자 수를 감소시켰고, 반면에 해외여행 등 백신을 접종함으로써 받을 수 있는 이익을 강조하는 보도가 백신 접종자 수의 증가와 관련이 있었다. 본 연구는 언론보도와 공중의 의견 간의 관계를 밝히는 기존 연구들의 의의를 넘어 특정 언론보도의 주제가 공중의 행동(백신 접종)에 영향을 미친다는 것을 밝혀 향후 연구자, 광고/홍보 전문가 그리고 정책 결정자가 보건 위기 상황에 대응할 방안을 마련했다는 데 그 의의가 있다.

▶투 고 일 : 2024년 02월 13일
▶심 사 일 : 2024년 02월 29일
▶게재확정일 : 2024년 06월 16일

주제어 : BERTopic, 토픽 모델링, 코로나19, 백신, 코로나19백신, 시계열분석, 뉴스, 뉴스 기사

*yepark@sookmyung.ac.kr

1. 연구 배경 및 목적

2020년 3월 11일, 국제 보건 기구(World Health Organization, WHO)는 신종 감염증(이하 코로나19)을 범유행 팬데믹(Pandemic), 즉 사람들이 면역력을 갖고 있지 않은 질병이 전 세계로 전염 및 확산하는 현상으로 선언하였다(권지원, 2023). 이러한 코로나-19는 공중 보건, 경제, 생활 습관 등 수많은 측면에서 국민의 생활에 영향을 미쳤다. 예를 들어 2024년 1월 7일 현재 코로나19 감염자는 총 774,075,242명으로 (한국의 경우 34,571,873 확진자) 7,012,896명의 사망자가 생겼다 (한국의 경우 35,934 사망자) (World Health Organization, 2024). 경제적으로 대한민국은 2020년 0.7%의 경제성장률을 기록한 이후(황정주, 2024) 생산성 하락, 고용률 하락 등 전방위적인 어려움을 겪었다(정선영, 장동산, 2023).

이와 같은 신종 감염병의 발생을 공중보건의 위기로 간주하고 위기를 극복하기 위해 연구자들은 여러 방면에서 많은 노력을 기울여 왔다 (예, 김영옥, 김혜정, 소담이, 2022; 안지수, 구윤희, 노기영, 2021; 오아름, 이준영, 김수진, 김활빈, 2021; 유우현, 오상화, 2023; 정원준 2022). 특히 코로나19의 위험성을 감소시키기 위한 가장 중요한 방안 중 하나로 전문가들은 코로나19 백신 접종을 손꼽아 왔고(Kim, Marks, & Clemens, 2021), 또한 일상생활로 돌아가는데 역시나 코로나19 백신 접종이 가장 중요한 요소로 논의됐다. (NPR, 2021).

이러한 보건 전문가들의 노력에 힘입어 대한민국의 백신 접종률은 최종적으로 미국과 같은 여타 국가들에 비해 높은 양상을 보여왔다. 구

체적으로 2023년 11월 26일 기준으로 2차 백신 포함 접종 완료를 마친 인구는 아래와 같다: (전 세계:67%, 대한민국: 87%, 미국:70%). 하지만 코로나19 위기 상황을 겪으면서 익히 경험한바, 정치적 지향성(예, 김영옥 외, 2022; 김효정, 홍혜현, 2021; Moon et al. 2022), 정부/ 사회에 대한 신뢰(예, 안지수 외, 2021, 오미영, 2023), 정보과부하(예, 김효정, 정우진, 2023) 그리고 문화적 차이(예, 전종우, 2021) 등의 여러 가지 요인들로 인하여 코로나19 관련 정보 추구 행동, 특히 기사를 통한 정보습득에 영향을 미치고 궁극적으로 백신 접종 의도와 행동에 영향을 미쳤다(Steinert et al., 2022; Tahamtan et al., 2022; Troiano, & Nardi., 2021).

이와 같은 연구 추세에 발맞추어 커뮤니케이션, 및 광고/홍보 연구의 측면에서 본 연구는 미디어, 그중에서도 특히 언론보도의 역할에 주목하여 코로나19 백신과 관련하여 각기 다른 주제로 보도된 기사가 실제 백신 접종자 수에 미치는 영향을 실증적으로 분석한다. 맥콤스와 쇼(McCombs, & Shaw 1978)가 정치적 이슈에 관한 언론 기사의 공중에 대한 의제 설정 기능을 처음으로 분석한 이후 40여 년에 걸쳐 미디어 연구자들은 정치 커뮤니케이션 분야뿐만 아니라 보건 커뮤니케이션 (예, 이완수, 정영희, 홍경수, 2022; Kim et al., 2011), 기업 커뮤니케이션(예, 강윤지, 2022; Park, Son, Yang, & Lee, 2019) 분야에서 의제 설정 현상을 연구해 왔다.

따라서 전대미문의 국제적 보건 위기 상황에서 언론의 역할을 고찰해 보기 위해 본 연구는 크게 세 가지를 밝히고자 한다: 1) 코로나19 백신과

관련하여 언론은 어떤 주제로 보도했을까? 2) 코로나19 백신 관련 언론보도의 주제는 시간이 지남에 따라 어떻게 변화했을까? 3) 궁극적으로 코로나19 백신 관련 언론보도의 주제 중 어떤 것이 백신 접종에 얼마나 영향을 미쳤을까?

2. 이론적 고찰

1) 코로나19를 극복하고자 하는 노력

코로나19는 상술한 대로 단순히 사망자와 후유증 등 보건적 측면과 아울러 전 세계적으로 사회, 경제, 문화에 엄청난 영향을 미쳤고, 연구자들은 다양한 분야에서 코로나19를 이해하고 이를 극복하려는 노력을 기울였다(예, 김영옥 외, 2022; 김효정, 정우진, 2023; 안지수 외, 2021; 오미영, 2023; 이병관, 노환호, 이태준, 2021; 전종우, 2021, 2024; 정원준, 2021; 2022; 정재선, 2021). 이러한 노력은 1) 개인적 측면에서 코로나19 관련 정보 추구 행위 분석, 2) 코로나19 위험을 감소시키기 위한 행위(예, 코로나19 백신 접종) 의도 등 크게 두 가지 측면으로 생각해 볼 수 있다. 코로나19 관련 정보 추구 행위와 관련하여 코로나19 위험 인식이 온라인 정보 행동을 통해 궁극적으로 사회 불신에 영향을 미친다는 것(정재선, 2021), 뉴미디어 리터러시에 따라 뉴스 진실검증 행동이 높아진다는 점(이병관 외, 2021)이 연구됐다.

본 연구와 직접적으로 관련이 있는 위험 감소 행위와 위험 지각에 관하여 전종우(2021)는 코로나19의 위험 지각에 있어서 국가 간의 차이와

문화 차원의 차이가 존재한다는 것을 밝혀 향후 커뮤니케이션 전략 수립에 의의를 밝혔다. 정원준 (2022)은 코로나 위험 인식이 코로나 관련 문제해결 동기와 어떤 관계가 있는지 연구하였고, 김영옥과 동료들(2022)은 코로나19 백신 접종 인식에서 정당의 메시지가 정당의 기조와 일치하지 않더라도 공중은 정당의 메시지를 일정 부분 수용하는 경향을 밝혀내었다.

2) 코로나19와 언론의 역할

본 연구는 코로나19 위기와 코로나19 백신 접종 상황에서 언론과 미디어의 역할에 관한 연구에 기반한다 (박주현, 2020; 오준혁, 이보운, 최미연, 최윤희, 2023; 이규호, 이준환, 2021; 이서현, 진명지, 최낙진, 2020; 이주현, 이신행, 2021; 홍주나, 안순태, 2022; Greene, & Murphy, 2021; Mousavi, & Gu, 2023; Ogbodo et al., 2020). 주로 보도의 주제(topic)와 관련하여 이서현과 동료들은 (2020) 코로나19와 관련하여 제주도가 제공한 보도자료가 언론보도에 활용되는 비율을 연구하여 대부분의 보도자료가 언론보도에 활용되는 것을 발견하였고, 오준혁 외(2023)는 코로나19 유행 시기별 언론 기사에 나타난 낙인양상을 분석하여 공포 유발감 (53.6%), 라벨링 (28.9%), 책임 귀인(16.2%)의 순서로 낙인 표현이 사용됐음을 밝혔다. 또한 오지현과 김영옥(2021)은 특히 코로나19 확진자 보도에 있어 위험 수치를 제시하는 형식과 단위에 따라 공중의 예방 행동 의도가 달라질 수 있음을 밝혀내었다. 구체적으로 일 단위의 확진자 보도는 심리적 거리감에 부적절한 영향을 미쳤고, 이러한 부정적 감정이 궁극적으로

예방 행위 의도에 긍정적인 영향을 미쳤다.

또한 코로나19와 코로나19 백신, 그리고 코로나19 백신 접종에 대한 언론 보도는 주로 내용 분석의 측면에서 연구되어 왔다. 박주현(2020)은 보수성향과 진보성향의 신문의 코로나19 보도 프레임을 비교하여 보수신문들은 갈등과 위협적 요소에 초점을 둔 보도가 많았고 진보 신문들은 도덕적 평가 및 예방, 대처 차원의 보도가 많은 것을 찾아내었다. 이주연과 이신형(2021)의 트위터 뉴스 분석은 언론보도에서 심각성과 취약성의 지각된 위협 프레임이 이득과 장애의 행동적 평가 프레임에 비해 부각되고 있다는 것을 발견하였고, 나아가 홍주나와 안순태(2022)는 실험 연구 방법을 통해 코로나19 감염 취약성을 강조한 기사가 코로나19에 대한 지각된 취약성이 높아지고, 궁극적으로 코로나19 백신 접종 의도가 높아지는 것을 밝혀냈다.

한편, 코로나19와 코로나19 백신 맥락에서 언론에 나타나는 주요 주제를 살펴보는 연구는 많이 진행됐으나, 각각 다른 주제로 보도된 기사 공중의 행동을 결정하는 역할을 실증한 연구는 드물었다. 비록 의제설정이론에서 이야기하는 매체 간 의제 설정 기능과 의제 구축 활동이 여론 형성에 미치는 영향은 다양한 맥락에서 실증되었지만, 여론 형성을 뛰어넘어 실제 공중과 소비자의 행동에 미치는 영향을 연구한 연구는 드물다. 예를 들어 코로나19 관련 WHO의 소셜 미디어 포스팅과 트위터 사용자들 사이의 포스팅 주제와는 상관관계가 발견되었고(Tahamtan et al., 2022), 코로나19 관련 기사들의 주요 토픽은 공포/유언비어였다는 것을 밝혀냈지만(Ogbodo et al., 2020), 주지사의 코로나19 관

련한 회복성 관련 메시지(resilience message)가 주민들의 코로나19 관련 행동에 영향을 미친다는 연구 이외에(Mousavi, & Gu, 2023) 이와 같은 다르게 표현된 주요 토픽들이 실제 행위에 어떤 영향을 미치는지는 경험적으로 연구되지 못했다. 구체적으로 주지사의 공식 소셜미디어 계정에 “뉴욕 사람들은 그들이 얼마나 강한지 모른다”(“The people of New York really don't know how strong they are.”)와 같은 회복성 관련 메시지를 게시하였을 때, 외출 제한 명령(stay-at-home order)에 순응하여 집 안에 머무는 시간은 12.5%, 불필요한 이동(unnecessary trips)을 피할 확률은 11% 증가한다는 것을 밝혀냈다(Mousavi, & Gu, 2023). 따라서 이와 같은 논의를 바탕으로 코로나19 백신과 관련하여 언론이 어떤 주제로, 얼마만큼, 얼마 동안 보도하였고, 궁극적으로 이러한 보도 주제(topic)들이 백신 접종률에 얼마나 영향을 미쳤는지 알아보고자 아래와 같은 연구 문제들을 제시하였다.

연구 문제1: 코로나19 백신과 관련하여 언론은 어떤 주제로 보도했을까?

연구 문제2: 코로나19 백신 관련 언론보도의 주제는 어떻게 변하였을까?

연구 문제3: 코로나19 백신 관련 언론보도의 주제 중 어떤 것이 백신 접종에 얼마나 영향을 미쳤을까?

3. 방법론적 고찰

1) 토픽모델링, BERT, BERTopic

따라서 본 연구는 언론에서 주로 다루어지는 코로나19 백신 관련 주제를 탐구하고자 전환 학습 (Transfer learning) 패러다임에 속하는 최신 기계 학습 (Machine Learning) 텍스트 마이닝 (Text mining)의 한 기법인 BERTopic 기법을 적용하고자 한다. 토픽 모델링 (Topic modeling)은 수집한 문서들의 맥락 (Context)을 바탕으로 특정 단어 (Word)들이 어떤 주제와 군집으로 모여있는지 함축적인 의미 구조를 제시하는 방법의 총칭으로서 대표적으로 Latent Dirichlet Allocation (LDA, Blei et al., 2003), non-negative matrix factorization (NMF, Shi, Kang, Choo, & Reddy, 2018), Topic to vector (Top2Vec Angelov, 2020), 그리고 Bidirectional Encoder Representations from Transformers Topic (BERTopic, Grootendorst, 2022) 등의 알고리즘들이 자연어 처리 (Natural Language Processing, NLP)의 주요 기법들로 사용되어 왔다 (Egger, & Yu, 2022). 특히 LDA를 활용한 토픽 모델링은 사회 과학 분야에서 최근 큰 관심을 받고 있다 (예, 강운지, 김재영, 강지훈, 김상훈, 2021; 김대욱, 2021; 문안나, 2016; 신성일, 이은순, 2023; 오혜라, 정운재, 2021; 2022; 2023, 정원준, 2018; 2021, 함승경, 최지명, 김영욱, 2019).

대표적으로 강운지와 동료들은 (2021) 2003년부터 2020년까지의 학술지들을 토픽 모델링하여 CSR 분야의 연구 동향을 분석하였고, 김대욱 (2021)은 에이즈에 대한 수용자 인식과 지

식수준, 미디어 시각에 대한 설문조사를 연구하여 한국인들이 에이즈에 대한 인식은 불치병, 면역결핍, 문란한 생활 등의 크게 세 가지 주제로 나타낼 수 있다는 것을 밝혀내었다. 또한 오혜라와 정운재 (2023)는 최근 오프라인과 온라인 스포츠 팬덤의 차이점을 그 활동 동기에 주목하여 오프라인 활동 동기는 경험적 몰입감, 유희성, 공존감, 공간적 경험으로, 온라인 활동 동기는 편의성, 사회적 상호작용, 사회적 소속감, 그리고 여가 활용으로 차이가 있다는 것을 발견하였다. 특히 정원준 (2021)의 연구는 위험 커뮤니케이션을 이론적 패러다임으로 코로나19와 관련한 정부와 일반인 유튜브 콘텐츠와 댓글 내 토픽들을 상호 비교함으로써 정부의 효율적이고 전략적인 양방향 커뮤니케이션 역할을 고찰해 보았다는데 그 의미가 본 연구와의 직접적인 관련성이 있다.

본 연구에서 사용될 방법인 BERTopic은 BERT 모델에 기반한다. 구체적으로 BERT는 Transformer (Vaswani et al., 2017) 기반의 언어 모델로서 현재까지 가장 정확한 텍스트 분석 방법으로 알려져 있고 (Egger & Yu, 2022; Grootendorst, 2022), 다중층양방향 트랜스포머 인코더 (Multi-layer bidirectional Transformer encoder)를 사용한 사전 훈련 모델 (Pre-trained model)을 사용하여 자연어 처리 분야의 신기원을 연 기술로 알려져 있다. (Egger & Yu, 2022, Grootendorst, 2022). 즉, 사전 학습된 (pre-trained) 언어 모델을 바탕으로 각 단어가 문장 내에서 갖는 의미를 판단한다.

이러한 기본적인 BERT 모델을 기반으로 그루텐도스트 (Grootendorst, 2022)는 BERT를 토픽 모델

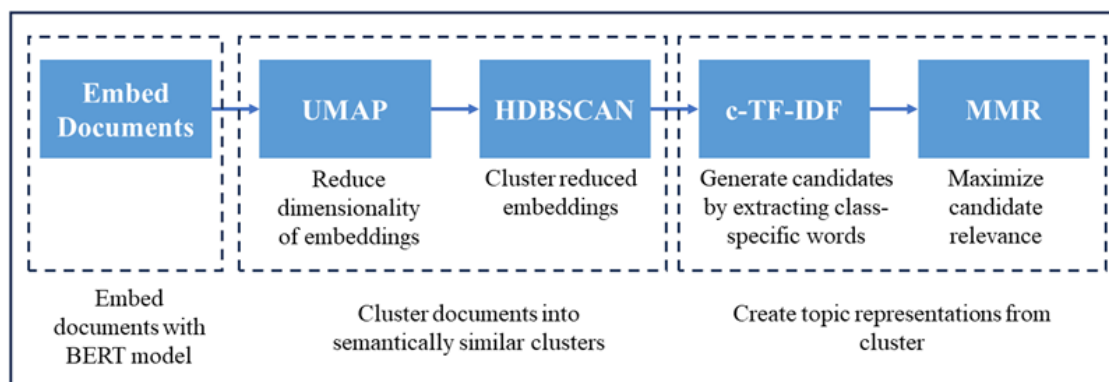
링에 사용하는 방법을 제안하고 BERTopic이라 명명하였다. 이는 토픽 표현(topic representation)을 군집 기법(clustering technique), 특히 단어 빈도-역 문서 빈도(term frequency-inverse document frequency, TF-IDF)를 단어나 문서 기반이 아닌 클래스 기반(class-based) 변화의 측면에서 살펴보고 이것을 특징 생성(feature generation) 과정에서 사용하는 데 그 근원적인 의의가 있다. 따라서 BERTopic은 아래와 같이 세 가지 단계를 거쳐 토픽을 생성해 낸다: 1) 문서 연결(document embedding), 2) 차원 감소와 군집화(dimension reducing and clustering), 그리고 3) 토픽 생성(topic representations) (Goel & Sharma, 2021; Joachims, 1997). BERTopic의 분석 과정에 대한 요약은 <그림 1>을 참조.

2) BERTopic의 장점

이러한 BERT의 기본적인 정의와 모델 구성에 기반하여 BERTopic은 특별히 Sentence-BERT (SBERT) 모델을 응용하여 사용한다(Reimers & Gurevych, 2020). 이 방법을 통하면 기존의 LDA

토픽 모델링이 갖는 몇 가지 한계를 극복하고, 더 효율적으로 대규모의 데이터를 분석할 수 있다 (Grootendorst, 2022; Egger & Yu, 2022). 구체적으로 BERTopic이 갖는 장점을 설명하자면, 첫째, BERTopic은 데이터의 전처리와 형태소 분석 같은 과정이 필요치 않다. 텍스트 전처리를 하는 이유는 전체 문서에서 의미를 제한적으로 전달하는 불용어들(stopwords)이나 같은 단어의 변화 형태들(예, 좋다, 좋았다, 좋았었다, 좋을걸, 등등)을 일정 규칙을 통해 분석이 용이한 대표적인 형식으로 (대부분 원형으로) 어간 추출(stemming)이나 표제어 추출(lemmatization) 등을 통해 변환하는 데 있다. 하지만 BERTopic은 이미 대규모 언어 모델로 훈련된(pre-trained) 데이터 세트를 바탕으로 그 이름에서 알 수 있듯이, 선 훈련된 모델을 연구자들의 데이터로 전환(transfer)하여 분석하는 BERT(Bidirectional Encoder Representations from Transformers, Devlin et al., 2018) 모델을 기반으로 토픽모델링 분석을 수행한다.

특히 BERT는 33억 개 이상의 말뭉치를 바탕으로 1백만 단계 이상의 모델 훈련을 거친 모델로 현존하는 최고의 언어 분석 모델로 손꼽히

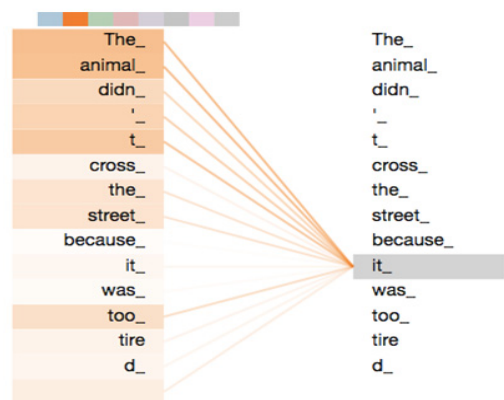


<그림 1> BERTopic 방법론 (Kim, Kim, & Lee, 2022)

며, 이 BERT를 바탕으로 많은 변종 모델이 연구되어 오고 있다 (예, SBERT, Reimers & Gurevych, 2020). BERTopic에서 사용되는 BERT 모델은 특히 영어뿐만 아니라 다국어 분석 모델에 특화된 형식으로, 한국어 포함 104개 언어, 12개 층의 레이어, 768개의 숨겨진 레이어(hidden layer), 그리고 1억 1천만 개의 모수(parameter)를 사용하는 모델이다. 따라서 BERTopic에서 사용되는 BERT를 사용할 경우 self-attention mechanism을 사용하여 단어의 의미를 양방향으로(Bidirectional) 가중치를 계산하여 어떤 단어의 의미를 문장 내의 맥락에서 해석하게 된다. 예를 들어 “The animal didn't cross the street because it was too wide”라는 문장이 있을 때, 일반적으로는 it과 같은 단어는 의미를 제한적으로 나타내기 때문에 분석에서 제외되는 경우가 많다. 하지만 BERT를 사용하였을 경우 <그림 2>에서 볼 수 있듯이, 일정 기준에 따라 각 문장의 단어들을 가리고(mask) 다시 분석하여 특정 단어의 의미가 문장이 복잡해질수록 갖게 되는 다른 의미들을 종합적으로 고려할 수 있게 된다. 즉, 문장의 다른 모든 단어가 초점 단어(masked word)에 미치는 영향을 설명할 수 있다. 즉, <그림 2>의 “it”은 “it”과 인접한 단어들의 의미를 attention mechanism을 사용하여 어떤 단어와 가장 연관이 있는지 계산하고, “it”이 “he”를 나타내는 것이 아니고 “animal”을 나타내는 것임을 수치상으로 표현할 수 있다. (Lutkevich, 2020). 따라서 BERTopic은 문장에 사용되는 단어 그대로를 통째로 분석할 수 있기 때문에 전처리 과정이 필요치 않다.

둘째, LDA 토픽 모델링에서 발생할 수 있는

연구자의 주관적 판단을 배제하고 더 정확한 토픽을 생성할 수 있다. 비록 LDA가 대규모 텍스트 데이터를 분석하는 데 많은 장점이 있지만, Latent Dirichlet 분포에 의존하는 통계기법의 특성상 내재적으로 몇 가지 방법론적인 한계를 갖는다. 그중 가장 큰 문제는 토픽의 개수를 결정하는 방법이다. 토픽 개수를 설정하는 방법은 토픽 모델링의 가장 중요한 이슈 중에 하나로서 통계적으로 perplexity score나 coherence score (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012)를 사용하여 그 값들이 급격히 줄어든거나 늘어나는 토픽의 개수를 선택해 왔다 (Ahn, Son, & Chung, 2021; Son, Ahn, Chung, & Drumwright, 2023). 하지만 BERTopic에서는 대규모 언어 분석을 바탕으로 토픽의 유사성(similarity)을 계산하여 토픽을 생성하기 때문에 통계적으로, 그리고 좀 더 과학적으로 토픽의 개수를 정할 수 있다.



<그림 2> Self-attention mechanism을 사용한 BERT 모델 예시(Alammar, 2018)

마지막으로 가장 중요하게, 기존 토픽모델링 방법들과 다르게 BERTopic은 더 정확한 결과를 얻을 수 있다. 이는 BERTopic의 첫 번째 장점으로 설명되었던 데이터 전처리 과정이 필요 없다는 것과도 연관되어 있는데, BERTopic은 사전 학습된 모델을 사용함으로써 단어의 뜻과 그 단어들의 뜻이 맥락(context)과 같은 복잡한 관계를 좀 더 정확하게 설명할 수 있다(Jung, Lee, & Kim, 2023). 또한 토픽 생성 과정에서 거치는 UMAP (Uniform Manifold Approximation and Projection), HDBSCAN(Hierarchical Density-based spatial clustering of applications with noise) 과정은 비선형적인(non-linear) 차원 감소(dimensionality reduction) 과정의 효율성과 정확성을 얻을 수 있다. (Grootendorst, 2022; Jung et al., 2023).

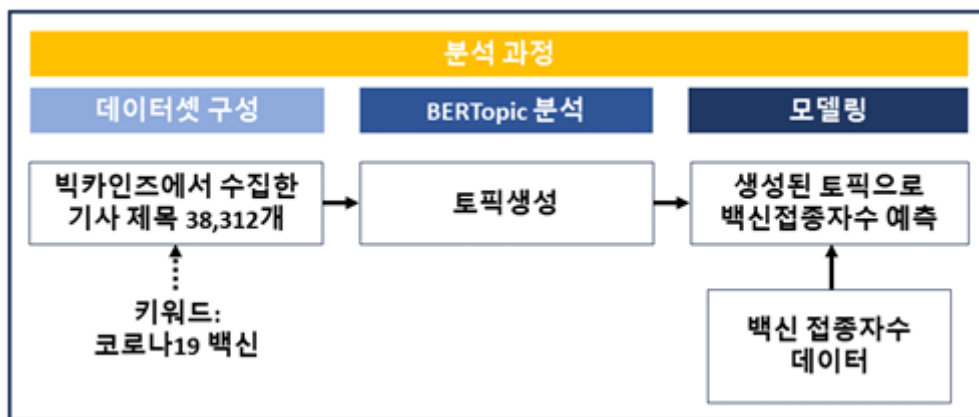
4. 연구 방법

본 연구는 코로나19 백신에 대한 보도 주제의

차이를 살펴보기 위해 총 38,312개의 기사 제목을 수집하여 분석하였다. 먼저 시계열적인 측면을 고려하기 위해 코로나19와 코로나19 백신 관련 타임라인을 정리하자면 <부록 I>과 같다. 이와 같은 코로나19와 코로나19 백신 관련 주요 현황들을 고려하여, 본 연구의 방법론을 요약하자면 <그림 3>과 같다. 즉 먼저 “코로나19 백신”을 키워드로 기사들을 수집하고, BERTopic을 이용하여 주요 토픽들을 생성한 후, 주요 토픽들의 시계열적인 변화 양상을 살펴본 후 시계열적인 특징을 포함한 회귀분석 식으로 각 토픽과 코로나19 백신 접종자 수의 관계를 규명한다.

1) 데이터소스 및 데이터 수집

본 연구는 최신 NLP 기법을 동원한 신문 기사 제목을 바탕으로 연구 문제 1, 2, 그리고 연구 문제 3을 위한 분석에 사용될 변수를 생성하였다. 구체적으로 한국언론진흥재단이 국내 54개 주요 언론사의 기사를 데이터베이스화한 빅카인즈(BIG KINDS, <https://www.bigkinds>).



<그림 3> 연구 모델

or.kr/)에서 2021년 1월 21일부터 8월 24일까지 “코로나 백신”을 키워드로 기사들의 제목을 수집하였다. 이는 정부가 코로나19 백신을 전국민을 대상으로 접종할 계획을 발표한 날부터 1주일 전, 그리고 델타 변이가 출현한 날을 기준으로 7일 후를 선정하였다. 이를 통해 정부의 백신 접종 계획 발표, 실제로 국민 대상 접종 시행, 그리고 새로운 변이바이러스의 출현까지 7개월간의 언론보도를 분석에 포함하기 위함이다. 이와 같은 과정을 통해 총 38,312개의 기사의 제목을 추출하였다. 기사의 본문이 아닌 제목을 분석한 이유는 빅카인즈를 통해 추출한 데이터에는 본문의 전문이 실려있지 않고, 기계학습 기법을 통한 뉴스 기사 분석 연구에 의하면 제목에 집중하여 분석하는 것이 분석의 방법론 측면에서 전문을 분석하는 것과 큰 차이가 없고, 오히려 분석의 효율성을 제공해 주기 때문이다(Ghasiya & Okamura, 2021; Qiang et al., 2020; Thakur, Saritha, & Jain, 2022; Xia et al., 2015, Yang, Li, & Zhao, 2019). 구체적으로 본 연구의 초점은 언론사별 보도 프레임링과 같은 언론사 특징이 아니라 전체 언론사의 보도 주제 중 특정 주제가 코로나19 백신 접종에 효과가 있는지에 있기 때문에 특정 언론사의 홈페이지에 접속해 기사 전문을 크롤링하는 것보다는 비록 전문이 실려 있지 않더라도 전체 보도 기사가 체계적으로 정리되어 있는 빅카인즈 데이터베이스를 활용하는 것이 더 객관적인 데이터베이스를 구성할 수 있다고 가정하여 빅카인즈를 활용하여 데이터셋을 구성하였다. 특히 기사 전문을 사용하여 자연어처리를 한 결과가 기사 제목이나 기사 요약본을 사용하

여 자연어 처리를 한 결과가 크게 다르지 않기 때문에(Yang et al., 2019) 이러한 방법을 사용하였다.

따라서 총 38,312개의 기사 제목이 수집되었고 향후 분석에 사용되었다. 또한 본 연구는 일자별로 변하는 토픽의 변화를 효과적으로 분석하기 위해, 그리고 종속변수로 사용될 날짜별 코로나19 백신 접종자 수에 맞추어 토픽의 변화를 일별로 살펴보았다.

종속변수와 관련해서 본 연구는 날짜별 코로나19 백신 접종자 수를 원자료로 전국, 시도별 코로나19 예방접종 실적 통계정보를 사용하였다(<https://kdx.kr/data/view/30239>). 이 데이터는 공공데이터의 일종으로 질병관리청에서 전일 0시 기준으로 집계된 자료를 매일 9시 35분에 갱신하고 있다. 원자료는 2021년 3월 11일부터 집계를 종료한 2022년 11월 9일까지 1차 백신 접종자, 2차 접종자 그리고 3차 추가 접종 접종자 수를 포함하고 있다.

2) BERTopic 모델링

구체적으로 BERTopic 방법을 살펴보자면, 첫 번째, Sentence-BERT 모델을 기반으로(Reimers, & Gurevych, 2019) 문서를 벡터 공간으로 치환하여 각 단어의 의미를 비교하는(semanticly compared) 과정을 거친다(document embedding). 그 이후 두 번째로 UMAP 과정을 통해 이전 과정에서 생성된 문서들의 벡터 차원이 감소하고(McInnes et al., 2018), 따라서 HDBSCAN이 군집을 형성하는데 사용된다. (McImms et al., 2017).

마지막으로 이전 토픽 모형화에서 사용되던 TF-IDF 대신 클래스 기반의 TF-IDF, 즉 c-TF-IDF가 문서 간의 벡터 유사도를 계산하여 토픽들의 그룹을 결정하는 데 사용된다. 그다음 토픽들을 대표하는 단어들을 가능한 최대로 다양하게 늘리기 위해 Maximize Candidate Rebalance Algorithm(MMR) 방법이 사용된다. 이 과정을 통해 단어 간의 코사인 유사도가 계산되고, 이것들을 바탕으로 c-TF-IDF가 계산된다.

이를 요약하자면, TF-IDF는 문서 d 의 단어 빈도가 t 일 때 아래와 같이 정의된다.

$$W_{t,d} = tf_{t,d} \cdot idf_t \quad (1)$$

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (2)$$

방정식(2)에서 N 은 전체 말뭉치에서의 문서의 개수이고, 말뭉치에서 t 를 포함한 문서의 개수로 정의된다. 따라서 IDF는 특정 단어가 포함하고 있는 정보의 양을 나타내고, BERTopic에서 이 과정은 TF-IDF 가중치로 계산되어 문서의 군집을 만드는 데 사용된다. 즉, 문서들을 연결함으로써 군집 내의 모든 문서를 하나의 문서로 취급하고(Grootendorst, 2022), 단어들은 문서 수준이 아닌 군집 수준에서 단어들의 기여도가 계산된다(Hristova & Netov, 2022). 따라서 IT-IDF가 문서 군집 수준에서 클래스 수준 재현(class-level representation)을 설명할 수 있도록 변환된다.

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right) \quad (3)$$

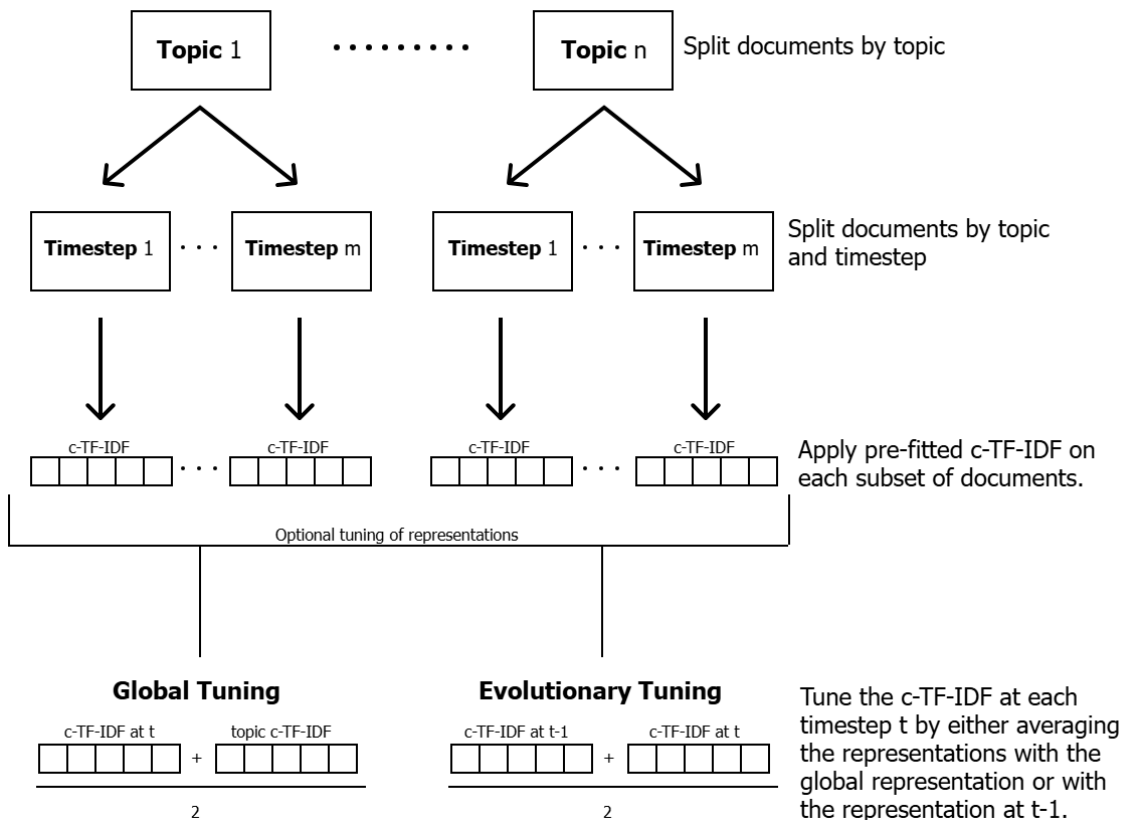
따라서 방정식(3)에서 c 는 각각의 군집에서 하나로 연결된 문서들의 모음을 나타내고, 클래스 A 의 평균 단어 술에 로그를 취한 값을 모든 클래스에 단어 빈도 t 를 나누어서 역문서 빈도(inverse document frequency)는 클래스 수준에서 각 단어가 클래스의 정보 값에 얼마나 기여하는지를 나타내는 역 클래스 빈도(inverse class frequency)로 변환된다(Grootendorst, 2022).

이런 과정을 거쳐 문서의 군집을 선 훈련된 BERT 모델을 통해 분석할 수 있고, 따라서 연구자들은 단어와 맥락 간에 좀 더 복잡하고(complex) 정확한(accurate) 관계를 포함하여 단어 재현(word representation)을 통해 실행할 수 있다. 이는 BERT를 사용한 분석의 장점으로 논의되는 점으로, 본 연구의 맥락인 코로나19와 코로나19 백신 보도를 고려해 보면, 본 연구에서 사용된 텍스트들이 선훈련된 BERT 모델을 바탕으로 Context-Aware Embedding(Ethayarajh, 2019; Li et al., 2020)과정을 거쳐 문맥 정보를 가진 단어들이 코로나19와 코로나19 백신에 관련된 보도의 문맥에 맞는 정보를 갖게 된다. 즉, 기존 토픽 모델링 과정에서 발생할 수 있는 문맥을 고려하지 않는 토픽 생성 과정을 효과적으로 방지할 수 있다. 구체적으로 BERTopic을 사용한 본 연구에서는 비록 “코로나 백신”이라는 키워드가 포함되었지만, “北, 코로나 백신 치료제 기술탈취 시도 화이자 해킹”과 “일본, 코로나 백신 접종 시작 하루 사망

자 다시 100명 넘어”와 같은 텍스트는 맥락상 대한민국의 코로나 백신접종이라는 연구 맥락에서 벗어난 것으로 취급되어 아웃라이어 토픽 (-1)로 구분된다.

또한 연구 문제 2(코로나19 관련 언론보도의 주요 토픽은 어떻게 변화하였을까?)와 관련하여 BERTopic 기법 중 특히 Dynamic Topic Modeling (DTM) 기법을 사용하여 시간이 지남에 따라 변화하는 토픽의 변화를 분석하였다. 즉, BERTopic은 각각 설정된 시간 변곡점에 따른 토픽 분포를 여러 번의 임의적인 변곡점 구분 없이 계산할

수 있다. 구체적으로 각각의 토픽과 변곡점에 따라 c-TF-IDF를 계산하고 이미 생성된 embedding에서 cluster를 생성할 필요 없이 각 시간 변곡점에서 특정 주제에 관한 재현(representation)이 만들어진다. 이어서 특정 시간 t 에 대한 토픽 재현(representation)은 전역적(globally), 그리고 진화적(evolutionary) 미세 조정(fine-tuning) 과정 등 두 가지 방법에 따라 이루어진다. (그림 4).



〈그림 4〉 Dynamic Topic Modeling의 미세조정(fine-tuning) 과정

(https://maartengr.github.io/BERTopic/getting_started/topicsvertime/topicsvertime.html)

5. 연구 결과

1) 연구 문제1: 코로나19 백신과 관련하여 언론은 어떤 주제로 보도했을까?

코로나19 백신 관련 언론보도의 주요 토픽들을 알아내기 위해 BERTopic을 사용한 토픽 모델링을 수행하였다. 구체적으로 <표 1>은 “all-MiniLM-L6-v2”를 사용하여 문장과 문단을 384 차원의 벡터로 변환한 transformer를 사용하고 (Koupaee & Wang, 2018), n-그램 범위를 1에서 3으로 고정한 초매개변수(hyperparameter, Houvardas & Stamatatos, 2006)를 사용한 BERTopic 결과를 나타낸다. 해석의 용이성을 위해 100개 이하의 기사가 모이는 토픽은 제외하였다 (Grootendorst, 2022).

BERTopic은 HDBSCAN 알고리즘을 통해 생성된 토픽에 소속되지 못하는 가외 치(outlier)를 토픽-1에 포함한다. 즉 코로나19 백신과 관련하여 보도된 기사 중 군집으로 모이지 못하는 기사들이 이 그룹에 포함된다. 예를 들어 “北, 코로나 백신 치료제 기술 탈취 시도 화이자 해킹”과 같은 기사는 백신 치료제라는 키워드로 검색되는 기사지만 사실 기사의 핵심은 북한의 기술 탈취 시도와 해킹으로서 토픽 0(코로나 확진)과 토픽1(코로나 백신 접종) 등과는 무관한 기사로 분류된다.

본격적으로 군집 된 토픽을 살펴보자면, 토픽 0은 가장 많은 17,688개의 기사를 포함하고 있고, 코로나(.018), 대통령 (.009), 확진(.009), 신규(.008), 확진자(.008), 방역(.008), 거리 두기 (.007), 4차 (.007), 신규 확진(.006) 등의 주

요 단어들을 포함한다. 즉 코로나19 확산 추세에 대한 보도가 토픽 0에 주로 포함된다고 볼 수 있다. 반면에 그다음으로 많은 기사를 포함하는 토픽 1 (n = 9,406) 은 코로나 백신 접종과 관련한 기사들이 아래와 같은 단어들을 포함한다: 백신(0.057), 접종(0.038), 코로나 백신 (0.033), 코로나(0.029), 백신 접종(0.026), 코로나19(0.015), 백신 접종(0.012), 코로나19 백신(0.011), 화이자(0.011), 시작(0.010). 기타 토픽들의 세부 단어들과 C-TF-IDF 값에 기반한 토픽에서 단어가 나타날 빈도에 관한 기술은 표 1에 요약하였다.

본 분석의 신뢰도를 확인하기 위해 토픽 유사도 분석을 실행하였다. 토픽 유사도는 토픽 간의 코사인 유사도(cosine similarity)를 A와 B 두 벡터 사이의 거리를 아래와 같은 방식을 통해 계산한다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

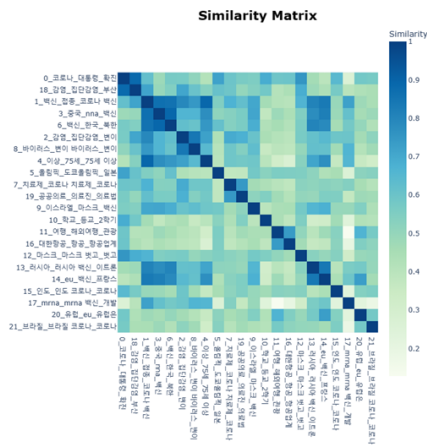
그림 5 와 표 2는 토픽 간의 상관관계와 유사도를 나타낸다. 토픽 유사도가, 8 이상인 토픽을 살펴보자면 토픽5(중국_nna_백신_중국산)와 토픽 7(백신_한국_북한_대통령), 토픽 0(코로나_대통령_확진_신규)과 토픽 18(집단감염_감염_부산_확진)등이 있었으나, 비록 토픽 5~토픽7이 모두 백신을 다루고 있으나, 세부 주제에서는 기사의 대상이 다르다고 판단하였기 때문에 향후 분석에서 그대로 사용하기로 하였다.

〈표 1〉 BERTopic 결과

토픽	토픽 내 기사 수 (Document Distribution)	자동 생성된 토픽 이름 (Topic Theme)	단어들 (c-TF-IDF 기반 확률)	대표적인 문장 (Representative Documents)
-1	6,381	백신_코로나_접종_코로나 백신	백신(.025), 코로나 (.023),접종(.018), 코로나 백신(.014), 백신 접종(.011), az(.009), 코로나19(.009), 화이자 (.008), 사망 (.008), 변이(.005)	[속보] “北, 코로나 백신 치료제 기술탈취 시도 화이자 해킹”. [나이트포커스] AZ 백신 첫 접종...내 일부터 화이자 백신 접종, 일본, 코로나 백신 접종 시작 하루 사망자 다시 100명 넘어]
0	17,688	코로나_대통령_확진_신규	코로나(.018), 대통령 (.009), 확진 (.009), 신규(.008), 확진자(.008), 방역 (.008), 거리 두기 (.007), 4차 (.007), 신규 확진(.006)	코로나 신규 확진 494명 서울째 400명대 코로나 신규 확진 668명...4차 대유행 한실화되나, 코로나 신규 확진 671명 4차 유행 우려 지속
1	9,406	백신_접종_코로나 백신_코로나	백신(.0057), 접종(.0038), 코로나 백신 (.0033), 코로나(.0029), 백신 접종 (.0026), 코로나19(.0015), 백신접종 (.0012), 코로나19 백신(.0011), 화이자 (.0011), 시작(.0010)	코로나 백신 접종 안착하려면, 문 대통령, 코로나 백신 2차 접종, 진천군, 코로나 백신 접종 시작
2	564	치료제_코로나 치료제_코로나_셀트리 온	치료제(.062), 코로나 치료제(.051), 코 로나(.032), 셀트리온(.018), 먹는 (.016),제약(.014), 임상(.014), 개발 (.014), 먹는 코로나(.012), 식약처(.012)	오송첨단의료산업진흥재단, 코로나 치료제 개발 속도, “셀트리온 코로나 치료제 임상 3상 전체로 허가 가능”, 日 시노노키, 먹는 코로나 치료제 임상 시작
3	425	감염_집단감염_코로나 _변이	감염(.113), 집단감염 (.040),코로나 (.024), 변이(.023), 확진 (.023), 코로나 감염(.019), 확산(.019), 감염자(.018), 감 염 확산 (.015),돌파감염(.015)	대구시장 ‘백신 논란’ 사과, 일상 감염 확산, 소규모 일상 감염 지속...신규 확진 666명,美 백악 관 돌파 감염 델타 변이 급속 확산
4	370	올림픽_도쿄올림픽_일 본_도쿄	올림픽(.145), 도쿄올림픽(.104), 일본 (.065), 도쿄(.050), 긴급사태 (.033),IOC(.033), 스가(.028), 도쿄 올림 픽(.023), 취소(.023), 코로나(.022)	[올림픽 선수단 일본 입국 후 코로나 첫 확진 미, 일본 여행금지 권고... 올림픽 영향 있다, 美 일본 여행금지 권고 올림픽 영향 있다]
5	365	중국_nna_백신_중국산	중국(.081), nna(.062), 백신(.060), 중국산(.039), 코로나 백신 (.362), 중국 백신(.032), 대만(.030), 접 종 (.029),홍콩(.028), 싱가포르(.028)	WHO, 중국 백신 시노팜 긴급사용 승인, 홍콩, 중국 시노팜 코로나 백신 긴급사용 승인, WHO, 중국 시노팜 코로나 백신 긴급사용 승인
6	292	바이러스_변이 바이러스_변이_변이바 이러스	바이러스(.207),변이바이러스(.139), 변이 (.119),변이바이러스(.034), 확산 (.030), 코로나바이러스 (.027),코로나 (.023), 변이바이러스(.022), 바이러 스에(.022), 델타(.019)	변이 바이러스 너마저... (악수터) 변이 바이러스, [여백] 변이 바이러스
7	286	백신_한국_북한_대통령	백신(.062), 한국(.051), 북한(.035), 대 통령(.022), 코로나 백신(.022), 공급 (.022),북한에(.019), 한국군(.018), 문 재인 대통령 (.018),한국은(.018)	한미 백신 파트너십 구축...美, 한국군 55만 명 백신 접종, 문 대통령 “북한 동의하면 백신 공급 협력 적극 추진”, 문 대통령 “북한이 동의하면 백신 공급 협력 적극 추진”
8	273	학교_등교_2학기_교육	학교(.073), 등교(.057), 2학기(.050), 교 육(.028), 확대 (.024),전면등교(.023), 수업(.023), 교사(.022), 대학 (.019),학 부모(.018)	등교 확대 추진...새 학기에는 학교 일상 회복할 까?, 2학기 전면등교 계획대로 학교 학원 방역 강화, [인터뷰] 유은혜 “학교, 상대적으로 안전 2학기 등교 확대”
9	243	마스크_착용_마스크	마스크(.247), 착용 (.036), 마스크 벗고	‘마스크 프리’ 두 달 만에 미국, 다시 마스크

토픽	토픽 내 기사 수 (Document Distribution)	자동 생성된 토픽 이름 (Topic Theme)	단어들 (c-TF-IDF 기반 확률)	대표적인 문장 (Representative Documents)
		벋고_마스크 착용	(.033), 마스크 착용(.033), 벋고(.033), 미국(.027), 의무화(.025), 마스크 의무	쓴다, 미 “실내 마스크 완화해야” 영, 학교서 노 마스크, 마스크 마스크 마스크 코로나 1년 당산이 화(.022), 다시 마스크(.019), 자침(.018) 영웅입니다 [+사진]]
10	232	러시아_러시아 백신_이트론_이아이디	러시아(.219), 러시아 백신(.130), 이트론(.064), 이아이디(.062), 이트론 이아이디 (.059),백신(.057), 이화전기(.057), 이아이디 이화전기(.054), 이수엠피지스(.054), 이화전기 이수엠피지스(.050)	“러시아 백신 기술이전 원할”...관련주 이트론, 이아이디, 이화전기, 이수엠피지스 등은?, “휴온스글로벌, 러시아 백신 스푸트니크 라이트 만든다”...관련주 이트론, 이아이디, 이화전기, 이수엠피지스 등은?, “러시아 백신 델타 델타변이에 효과”...관련주 이트론, 이아이디, 이화전기, 이수엠피지스 등은?]
11	227	해외여행_여행_관광_여행업계	해외여행(.061), 여행(.052), 관광(.043), 여행업계(.030), 해외(.029), 상품(.026),관광산업(.019), 사이판(.019), 트래블(.017), 트래블 버블(.015)	트래블 버블 기대감 해외여행 상품 불티, 접종자 해외여행, 시민 여행업계 기대감 고조, 접종자 해외여행, 시민 여행업계 기대감 고조
12	202	eu_백신_유럽_프랑스	eu(.107),백신(.057), 유럽 (.055),프랑스(.054), 코로나 백신(.035), 접종(.031), EU 백신(.030), 코로나(.027), 이탈리아 (.024), EU 코로나(.023)	EU, 백신 접종 관광객 입국 허용 이르면 내일부터 EU 집행위, 백신 수출 통제 규정 강화 EU 영국 신경전, [이 시각 세계] EU, “코로나 백신 접종 관광객 입국 허용”
13	189	인도_인도 코로나_코로나_생지옥	인도(.229), 인도 코로나 (.062),코로나(.042), 생지옥(.038), 하루(.037), 생지옥 인도(.033), 코로나 생지옥(.032), 코로나 지옥(.031), 지옥 인도(.030), 지옥(.030)	인도, 코로나 신규 확진 31만명 넘어 세계 최다, 코로나 지옥 인도 하루 확진 31만명 세계 최다, 코로나 생지옥 된 인도 미 유럽 “긴급 지원”
14	187	대한항공_항공_항공업계_화물	대한항공(.060), 항공(.046), 항공업계(.030), 화물(.029), 항공기(.026), 항공사(.025), 여행(.025), 공항(.024),현대글로벌(.023), 아시아나(.022)	현대글로벌, 유럽에 전락 거점 구축 항공 물류 강화, 현대글로벌, 글로벌 ‘항공 물류’ 강화 유럽에 전락 거점 구축, 현대글로벌, 글로벌 항공 물류 사업 확대 獨 공항 거점 구축
15	148	브라질_브라질 코로나_코로나_사망	브라질(.249), 브라질 코로나(.048), 코로나(.043), 사망(.036), 시위(.035), 사망자(.034), 브라질 대통령 (.032),넘어(.030), 코로나 사망(.028), 코로나 사망자 (.026)	美 코로나 사망 51만명...브라질 변이 확산, 미 이어 브라질 코로나 사망 50만명 민심 부글부글, 브라질, 코로나 신규 확진 10만 명대로 급증
16	147	mrna_mrna 백신_개발_에스티팜	mrna(.249), mRNA 백신 (.147),개발(.068), 에스티팜(.048), 백신 개발(.045), 백신(.045), 추가 (.044), mRNA 코로나(.035), 아이진(.036), mRNA 백신 (.030)	에스티팜, 국내 최초 LNP 방식 코로나 mRNA 백신 개발 나간다, 에스티팜, 코로나19 mRNA 백신 개발 본격화한다, 에스티팜, 국내 최초 LNP 방식 코로나 mRNA 백신 개발 착수
17	145	이스라엘_백신_1위 이스라엘_1위	이스라엘(.237), 백신 (.050), 1위 이스라엘 (.044),1위(.042), 접종 (.033), 접종 1위(.032), 이스라엘 백신(.030), 이스라엘서(.030), 이스라엘 코로나(.025), 접종률(.025)	백신 접종 1위 이스라엘 어제부터 ‘마스크 벗었다’ , 백신 접종 1위 이스라엘 어제부터 ‘마스크 벗었다’ , 백신 접종 1위 이스라엘 어제부터 ‘마스크 벗었다’
18	145	집단감염_감염_부산_확진	집단감염(.080), 감염 (.077),부산(.029), 확진 (.028),확산(.028), 확진자(.026),돌파 감염(.026), 감염 확산(.024), 발생(.023), 변이. 023).	제주 사회적 거리두기 2단계로 격상 ...돌파 감염 도 처음 발생, 제주 사회적 거리두기 2단계로 격상 ...돌파 감염 도 처음 발생, 제주 사회적 거리두기 2단계로 격상 ...돌파 감염 도 처음 발생

토픽	토픽 내 기사 수 (Document Distribution)	자동 생성된 토픽 이름 (Topic Theme)	단어들 (c-TF-IDF 기반 확률)	대표적인 문장 (Representative Documents)
19	142	유럽_eu_유럽은_봉쇄	유럽(.117),EU(.066),유럽은(.031), 봉쇄 [.코로나19 국제뉴스] 첫 봉쇄 후 1년, 독일서 방역 (.031), 독일 (.028),코로나(.024), 프랑 반대 대규모 시위 [코로나19 국제뉴스] 봉쇄 1년, 스(.021), 시위(.020), 다시 (.019),3차 못 막은 3차 유행 유럽 정치지형 '흔들' , 유럽, (.017).	코로나 3차 확산 우려..미국도 확진자 다시 증가
20	132	삼성_삼성바이오_삼성 바이오로직스_모더나	삼성(.183), 삼성바이오 (.076), 삼성바 이오로직스 (.059),모더나(.050), 투자 삼성바이오 "3분기부터 모더나 코로나 백신 생 (.046),LG(.038), 삼성바이오 모더나 산" , 삼성 3년간 240조원 투자한다, 삼성 3년간 (.037), 위탁생산 (.036),생산(.035), 삼 240조원 통 큰 투자 4만명 고용한다 성 LG(.031)	
21	123	이스라엘_마스크_해제 _국제뉴스 이스라엘	이스라엘(.227), 마스크, (.071), 해제 [코로나19 국제뉴스] 이스라엘, 방역 조치 전면 (.044), 국제뉴스 이스라엘(.038), 실내 해제 실내 마스크 출입국 제한은 유지 [코로나19 (.031), 압사(.030), 교환 (.028), 코로나 국제뉴스] 이스라엘 "실내 마스크 이외 모든 방역 19 국제뉴스 (.026), 국제뉴스(.026), 이 조치 해제" , [코로나19 국제뉴스] 이스라엘 마지막 방역 조치 '실내 마스크 착용 의무' 도 해제	



〈그림 5〉 토픽 유사도 행렬표

〈표 2〉 토픽 유사도

토픽 A	토픽 B	A, B 토픽 간 거리
Topic5	Topic7	0.8901
Topic0	Topic18	0.8828
Topic1	Topic5	0.8689
Topic1	Topic7	0.8548
Topic3	Topic6	0.8487
Topic10	Topic12	0.8403
Topic12	Topic1	0.8351
Topic12	Topic5	0.8182
Topic10	Topic1	0.8110
Topic17	Topic5	0.8075
Topic10	Topic7	0.8056

2) 연구 문제2: 코로나19 관련 언론보도의 주요 토픽은 어떻게 변하였을까?

본 연구의 가장 큰 특징 중 하나는 토픽을 분석하는 데 시계열적인 요소를 포함하였다는 것이다. 토픽 모델링의 방법으로는 모든 텍스트 데이터를 한꺼번에 큰 말뭉치(corpus)로 가정하고 그 말뭉치 안에 문서(document)와 단어(word)가 어떻게 군집화될 수 있는지 살펴보는 기본적인 토픽 모델링 방법이 있지만, 말뭉치가 시대별로, 시기별로 어떻게 변화하며, 주요 토픽들이 나타나는 횟수가 어떻게 변하는지 살펴보는 방법이 있다. LDA 기반의 모델링에서도 특정 시기별 토픽이 나타나는 횟수를 이전 시기의 모델링 결과와 상관관계를 계산한 후 추론식에 더하는 Dynamic Topic Modeling(Blei & Lafferty, 2006)의 방법이 사용되어 왔다. 마찬가지로 BERTopic에서도 각각의 토픽이 주어진 시간 간에 어떻게 변하는지 살펴보는 접근법이 제안되었다. 구체적으로 언론보도가 매일 어떻

게 변하는지를 시계열적으로 분석하여 본 결과, <부록 II>에서 볼 수 있듯이 분석 기간 내에 토픽 0(코로나 확진자 보도)이 가장 많이 보도되었다. 하지만 특이하게도 2021년 2월 말 코로나 접종을 시행한 시기는 토픽1(코로나 백신 접종)이 단순 코로나 확진자를 보도하는 토픽 0보다 많이 관찰되었고, 7월에는 거의 보도되지 않고 확진자 수 보도가 백신 관련 보도의 수를 압도하였고, 8월을 지나 다시 두 보도 간의 차이가 좁혀지는 패턴을 보였으나 8월 말 델타 변이가 시작되면서 다시 두 보도 간의 격차가 커지면서 확진자 수 위주의 보도가 주를 이루었다. 즉 상대적으로 예방접종이 잘 진행되고 있던 시기(2021년 6월, 7월)에는 오히려 백신 접종에 대한 보도가 줄어들고 델타 변이가 출현한 시점부터 오히려 백신 보도의 비율이 줄어드는 양상을 보였다. <부록 III>은 가장 많은 보도 양상을 보였던 토픽 0과 토픽1을 제외한 나머지 토픽들을 시각화하였다.

3) 연구 문제3: 코로나19 백신 관련 언론 보도의 주제 중 어떤 것이 백신 접종에 영향을 미쳤을까?

(1) 분석단위와 변수의 조작화

이와 같은 코로나19 백신 관련 언론보도의 양상을 분석한 결과를 바탕으로 어떤 토픽의 보도가 궁극적으로 코로나19 백신 접종자 수와 얼마나 어떻게 관련이 있는지 탐구하기 위하여 본 연구는 시계열 회귀분석을 하였다. 먼저 분석의 단위로는 종속변수로 사용될 1차 백신 접종자 수가 국가통계자료로 열람할 수 있는 2021년 3

월 11일부터 언론보도 데이터 수집이 끝난 8월 24일까지 168일 중 각각의 날을 분석단위로 설정하였다. 즉, 시계열 회귀분석 식에는 3월 11일 보도된 토픽 0에서 토픽 21까지 각각의 기사 개수가 독립변인으로, 그 해당 날짜의 1차 접종자 수가 종속변인으로 모델을 추정하였고, 궁극적으로 168일간의 시계열 변화를 분석하였다. 따라서 독립변수로는 BERTopic을 통해 생성된 해당일에 100개 이상의 기사가 관측된 토픽에서 관측된 각 주제의 언론보도 개수로 정의되었고, 종속변수로는 해당일에 1차 접종자 수로 정의하였다.

(2) 기술 통계

본 연구의 마지막 단계는 생성된 토픽들을 바탕으로 어떠한 토픽이 백신 접종률과 관련이 있는지 시계열분석을 통한 계량 경제학적 추정 방식을 통해 알아보는 것이다. 본격적인 모델 추정에 앞서 각 변수의 기술 통계와 상관관계, 분산 팽창 요인(Variance Inflation Factor, VIF), 그리고 종속변인의 분포에 대해 살펴보았다 <부록 IV>. 먼저 시계열 데이터인 백신 접종자 수를 효과적으로 분석하기 위해 백신 접종자 기록이 존재하기 시작하는 날(2021년 3월 11일)부터 8월 24일까지 168일의 1차 백신 접종자 수(종속변인)의 특징을 살펴보았다. 이 기간에, 하루에 평균 151,493.59명의 국민이(SD=208,541.85) 1차 백신을 접종하였고, 가장 많이 보도된 토픽 0은 분석 기간 평균 하루에 86.57번(SD=37.93), 토픽 1은 38.42번(SD=18.37) 보도되었다. 각 토픽의 보도 횟수와 백신 접종자 수의 상관관계를 비교해 보면 0.13에서 (토픽 7, 토픽 13) .66(토픽-1, 토픽1)으로 큰

상관관계를 보이는 독립변수들은 존재하지 않았다. 또한 독립변수 간의 다중 공산성을 판단하기 위해 VIF 값을 살펴본 결과 가장 높은 VIF 값이 2.47에 불과하여 다중 공산성이 없는 것으로 판명되었다(Hair, Sarstedt, Ringle, & Mena, 2012; Hair, Sarstedt, Hopkins, & Kuppelwieser, 2014; Sarstedt, Hair, Ringle, Thiele, & Gudergan 2016).

4) 모델 선택과 추정(Model Selection and Estimation)

먼저 시계열 종속변수인 코로나19 백신 접종자 수가 시점에 상관없이 시계열 특성이 일정하지 또는 일정하지 않은지 확인하기 위해 Augmented Dickey Fuller Test (ADF test, Dickey, & Fuller, 1981)을 통해 정상성(stationary)을 확인하였다. ADF 분석 결과 ADF Statistic이 2.69로 5% 결정 값(Critical value)인 2.88보다 크기 때문에 5% 신뢰구간으로 ADF 분석의 귀무가설을 기각하지 못하고 데이터의 구조가 정상성을 갖추지 못했다는 것이 밝혀졌다 ($p = .075$).

따라서 백신 접종자 수는 크게 아래와 같은 세 가지 측면에서 일반 최소제곱법(Ordinary Least Squares, OLS)으로 추론된 회귀모형이 1) 추세 요인(Trend), 2) 계절 요인(Seasonality), 3) 이분산(heteroscedasticity, Clements, & Hendry, 1999) 이렇게 세 가지 측면에서 오류를 발생시킬 수 있다. <부록 V>는 종속변수의 자기상관(Autocorrelation)을 나타낸다.

종합하여 ADF 값과 자기상관 값을 나타낸 그래프에 의하여 본 연구는 1) 일반 OLS 회귀, 해석의

용이성을 위한 종속변수를 로그 변환한 OLS, 그리고 3) 상술한 세 가지 요소를 통제하기 위한 robust Newey-West variance estimator를 사용한 OLS(Kim, 2022; Newey & West, 1987; Smith & McAleer, 1994) 이렇게 세 가지 모델을 추정하였다 <표 3>.

모델 추정 결과 공통으로 백신 접종자 수에 영향을 미치는 토픽은 토픽 0(코로나 확진), 토픽 11(해외관광), 토픽 20(삼성바이오 모더나 백신)이었다. 구체적으로 모델 1(수정된 결정계수: .12, $F(23, 168) = 2.01$)에서 토픽 0(코로나 확진, $\beta = -.1452, 11$, $SE = 682.67$, $t = -2.13$, $p < .05$)과 토픽 8(학교등교, $\beta = -38,250$, $SE = 14,900$, $t = -2.57$, $p < .05$), 토픽 10(러시아 백신, $\beta = -17,320$, $SE = 9020.98$, $t = -1.91$, $p < .05$), 토픽 20(삼성바이오 모더나백신, $\beta = -46,960$, $SE = 17,600$, $t = -2.66$, $p < .01$)은 백신 접종자 수를 통계적으로 유의미하게 감소시키고 아웃라이어 토픽-1을 제외하고 토픽 11(해외여행, $\beta = 30,950$, $SE = 9398.96$, $t = 3.29$, $p < .001$)은 백신 접종자 수와 유의미한 정의 관계가 있었다. 백신 접종자 수를 로그 변환한 OLS 모델(모델 2, 수정된 결정계수: .11, $F(23, 168) = 1.88$)과 거기에 Newey-West variance estimator를 사용한 모델(모델 3, 수정된 결정계수: .11, $F(23, 168) = 2.21$) 모두에서 공통으로 토픽 0(코로나 확진, $\beta_{\text{모델2}} = -.016$, $SE_{\text{모델2}} = .009$, $t_{\text{모델2}} = -1.91$, $p_{\text{모델2}} < .05$; $\beta_{\text{모델3}} = -.016$, $SE_{\text{모델3}} = .010$, $t_{\text{모델3}} = -1.64$, $p_{\text{모델3}} < .05$)과 토픽 20(삼성바이오 모더나백신, $\beta_{\text{모델2}} = -.838$, $SE_{\text{모델2}} = .221$, $t_{\text{모델2}} = -3.79$, $p_{\text{모델2}} < .001$; $\beta_{\text{모델3}} = -.837$, $SE_{\text{모델3}} =$

.253, $t_{\text{모델3}} = -3.32$, $p_{\text{모델3}} < .001$)은 백신 접종
자 수와 유의미하게 음의 관계가 있었고, 반면
에 토픽 11(해외여행)은 양 모델 모두 유의미하
게 백신 접종자 수를 증가시켰다($\beta_{\text{모델2}} = .283$,
 $SE_{\text{모델2}} = .118$, $t_{\text{모델2}} = 2.41$, $p_{\text{모델2}} < .05$; $\beta_{\text{모델3}}$

$= .283$, $SE_{\text{모델3}} = .075$, $t_{\text{모델3}} = 3.76$, $p_{\text{모델3}} < .001$). (모델추론 결과는 <표 3>, 모델추론 결과
의 요약본은 <표 4>를 참조).

<표 3> 모델추론 결과

변수	종속변인					
	백신 접종자 수			Log (백신 접종자 수)		
	모델 1. OLS		모델 2. OLS		모델 3. OLS with Newey–West variance estimator	
	계수	표준오차 (SE)	계수	표준오차 (SE)	계수	표준오차 (SE)
절편	169,000***	46,900	11,626***	.589	11,626***	.532
토픽-1	4907.18*	2030.79	.130	.025	.013	.026
토픽0	-1452.11*	682.67	-.016*	.009	-.016*	.010
토픽1	302.43	1237.37	-.019	.016	-.018	.015
토픽2	34.19	8675.02	.051	.109	.050	.111
토픽3	3699.00	10,100	.037	.127	.037	.108
토픽4	-5230.52	7210.61	-.035	.089	-.035	.089
토픽5	2109.47	6472.37	.103	.081	.103	.068
토픽6	11,970	11,600	.193	.145	.193	.129
토픽7	1087.26	11,000	.174	.139	.174	.132
토픽8	-38,250**	14,900	-.041	.187	-.041	.156
토픽9	-6305.09	9264.17	-.126	.116	-.126	.121
토픽10	-17,320*	9020.98	.068	.113	.068	.085
토픽11	30,950***	9398.96	.283*	.118	.283***	.075
토픽12	-19,850	13,800	-.030	.173	-.030	.145
토픽13	-6061.98	5272.33	-.006	.066	-.006	.052
토픽14	12,530	17,400	-.190	.219	-.190	.271
토픽15	-23,810*	12,500	-.282	.157	-.282	.143
토픽16	34,120*	17,800	.422	.223	.422	.213
토픽17	-3696.11	11,700	-.077	.146	-.077	.103
토픽18	15,230	17,600	.247	.221	.247	.169
토픽19	13,800	17,300	.273	.217	.273	.169
토픽20	-46,960**	17,600	-.838***	.221	-.837***	.253
토픽21	-13,970	13,600	-.105	.170	-.105	.137
관측수(Observations)		168		168		168
수정된 결정계수(Adjusted R ²)		.12		.11		.11
로그 우도(Log-Likelihood)		-2272.1		-376.06		-376.06
AIC(Akaike Information Criterion)		4,592		800.1		800.1
BIC(Bayesian Information Criterion)		4,667		875.1		875.1

〈표 4〉 모델추론 결과 요약

	백신 접종자 수 증가	백신 접종자 수 감소
모델 1. OLS	토픽1(백신접종 아웃라이어 토픽), 토픽11(해외여행)	토픽0(코로나 확진), 토픽8(학교등교), 토픽10(러시아 백신), 토픽20(삼성바이오 모더나 백신)
모델 2. OLS (log백신접종자수)	토픽11(해외여행)	토픽0(코로나 확진), 토픽20(삼성바이오 모더나백신)
모델 3. OLS Newey–West variance estimator (log백신접종자수)	토픽11(해외여행)	토픽0(코로나 확진), 토픽20(삼성바이오 모더나백신)

5) 확인 분석(Robustness Check)

추가로 본 회귀분석의 결과가 믿을만한지 확인하고자 두 가지 확인 분석을 시행하였다: 1) 역인과(reverse causality)의 가능성 확인, 2) 민감도 검사(sensitivity analysis). 역인과 가능성은 독립변수와 종속변수를 반대로 넣어서 회귀분석이 유의미한지 확인하는 방법과 그레인저 인과관계(Granger causality, Granger, 1969)의 두 가지 방법으로 확인하였다. 먼저 종속변수인 백신 접종자 수를 독립변수로, 독립변수인 토픽 0, 토픽 8, 토픽 11, 토픽 20을 각각 종속변수로 선정하여 회귀분석을 실행한 결과, 모든 β 값의 유의도 .05 이상으로 나타나 회귀분석 식이 통계적으로 유의미하지 않았다. 마찬가지로 독립변수들(토픽 0, 토픽 8, 토픽 11, 토픽 20)을 종속변수로 선정하고 백신 접종자 수를 종속변수로 놓은 뒤 짝지어(pairwise) 그레인저 인과관계 분석을 실행한 결과 어떤 F-test의 값들도 유의미하지 않았고 ($p > .05$), 따라서 역 인과의 가능성이 희박하다고 결론내릴 수 있다.

비록 본 연구의 초점이 언론보도의 주제와 코로나19 백신 접종자 수의 관계를 밝히는 데 있지만, 우리가 추출한 22개의 주제 이외에도 설명하지 못한 교락 요인(unobserved confounder)이 존재

할 가능성이 있기 때문에 회귀계수를 명목 가치(face value) 그대로 해석하는 것은 위험할 수 있다. 따라서 확인 분석의 일종으로 민감도 분석(sensitivity analysis, Cinelli, & Hazlett, 2020)을 실시하였다. 분석 결과 각각의 독립변인들과 직교(orthogonal)한 unobserved confounder는 독립변인과 종속변인의 관계에 있어 26.9% 이상의 잔차 분산(residual variance)을 설명한다. 다시 말하자면, 회귀분석 식에 포함되지 않은 변수들이 독립변수인 백신 접종자 수를 최대 73.1%까지 설명할 수 있고, 회귀분석 식에 포함된 독립변수가 설명하는 비중과 미포함된 독립변수가 설명할지도 모르는 비중의 차이가 1:4가 넘지 않기 때문에 독립변수 효과 추정이 효과적이었다고 결론 내릴 수 있다(Altonji, Elder, & Taber, 2005; Oster, 2019).

6. 결론 및 논의

본 연구는 최신 텍스트 분석기법인 BERTopic과 계량경제학 모델링을 결합하여 대규모의 코로나19 백신 관련 기사들이 주로 어떤 주제를 나타내는지, 이러한 주제들이 시간이 감에 따라 어떻게 변화하는지 분석하고 궁극적으로 각기 다른 토픽

들이 백신 접종률에 미치는 영향을 분석하였다.

1) 결과 요약

구체적으로 본 연구는 38,312개의 코로나19 백신 관련 기사가 주로 어떤 주제로 보도되고 있는지, 이 주제들은 시간이 지남에 따라 어떻게 그 중요도가 변하는지, 그리고 궁극적으로 어떤 토픽이 백신 접종자 수를 유의미하게 예측할 수 있는지 분석하였고, 그 결과 몇 가지 주요 결과를 도출해 내었다. BERTopic을 통해 본 연구는 가장 먼저 22개의 주요 토픽을 생성해 내었고, 코로나 확진자 수와 확진자 증가 추세에 집중된 주제가 가장 많이 보도되었고 (토픽 0, $n = 17,688$), 뒤를 이어 코로나 백신 접종 추세와 접종 현황을 이야기하는 토픽이 많았다 (토픽 1, $n=9,406$) (연구 문제1). 이는 코로나19 시기에 시기별 정부 기관의 유튜브 콘텐츠에서 코로나 발생 현황과 증상 및 예방책, 그리고 전파 상황을 알려주는 콘텐츠가 가장 많았다는 정원준(2021)의 연구에서 알려진 것과 일치한다. 하지만 글로벌 미디어에서 나타나는 코로나19 관련 기사들의 주요 토픽 역시 공포/유언비어 (OgBodo et al., 2020)이고, 코로나19 유행 시기별 언론 기사에 나타난 낙인 양상 중 공포 유발감이 가장 낙인 표현이었음을 고려하면(오준혁 외, 2023), 단순히 확진자 수를 강조하여 공포감을 조성하는 기사는 코로나19 위기 극복에 도움이 되지 않을 수도 있다.

연구 문제2와 관련해 시계열적인 추세를 가장 많이 보도된 두 토픽(토픽 0 = 확진자 보도, 토픽 1 = 백신 접종)을 살펴본 결과 백신 접종이

본격적으로 실시된 2021년 2월 말부터 토픽1(백신 접종) 이 토픽 0(확진자 보도)보다 많이 보도되었으며, 백신 접종이 비교적 성공적으로 진행되던 7월, 8월에는 오히려 백신 관련 보도가 감소하고, 8월 말 새로운 변종이 유행함에 따라 다시 확진자 추세 보도가 많아졌다. 이는 언론의 역할에 대해 다시 한번 생각해 볼 기회를 제공한다. 의제설정이론에서 알 수 있듯이 언론이 무엇을 보도하냐에 따라 공중의 이슈 중요도 판단은 크게 영향을 받는다 (McCombs & Shaw, 1972; McCombs, Shaw, & Weaver, 2014). 따라서 새로운 변이가 출현하였을 때, 오히려 백신 관련 보도를 늘려서 전파를 효과적으로 방지해야 함에도 오히려 백신 관련 보도가 줄어드는 것은 코로나19 위기를 극복하는데 바람직하지 않을 수 있다.

시계열적인 요소를 고려한 회귀모델 추정해 결과에 따르면 코로나19 백신 접종자 수와 유의미하게 정의 관계가 있는 토픽은 가장 많이 보도되던 두 토픽이 아니라 백신接种의 부가적 이익을 강조하는 토픽 11(해외여행)이 유일하였다. 반면에 단순히 코로나 확진자 수를 보도하는 것(토픽 0)과 새로운 백신 개발과 관련한 뉴스(토픽 20)는 접종자 수를 오히려 감소시키는 경향이 있었다 (연구 문제 3). 앞서 논의한 연구 문제1과도 관련 있는 논의로서 단순히 코로나 확진자 수를 보도하는 것은 공포 관련 낙인 표현으로 간주하여(오준혁 외, 2023) 메시지 회피를 발생시킬 수 있다(Tannenbaum et al., 2015).

반면에 자주 보도되지는 않았지만, 백신의 직접적인 장점(건강 이슈) 이외에 부가적인 장점

을 강조하는 보도(토픽 11, 해외여행 가능)가 매우 강력하게, 그리고 유일하게 백신 접종자 수를 증가시키는 것을 고려하였을 때, 향후 보건 관련 설득 전략을 수립할 때 본 연구의 결과가 사용될 수 있다.

2) 이론적 함의

따라서 본 연구는 크게 네 가지 분야의 이론적 함의를 갖는다. 첫 번째로, 본 연구는 지금까지 널리 수행되어 온 언론의 역할에 관한 연구에서 주로 다루어지던 언론이 공중의 의견에 어떻게 영향을 미치는지를 넘어 실제 공중의 행동에 어떠한 영향을 미치는지 살펴보았다. 그동안 미시적인 개인의 심리학적 요인과 설득 메시지 구성이 예방 행동 의도에 미치는 영향은 널리 연구되어 왔다(예, 김효정, 정우진, 2023; 오미영, 2023; 오지현, 김영옥, 2021). 본 연구는 이러한 심리학적 그리고 개인 간의 차이(individual difference) 요소뿐만 아니라 실제 언론 기사의 내용(토픽)이 행위 의도를 넘어 실제 행위까지 영향을 미칠 수 있다는 점을 밝혔다.

둘째, 본 연구는 궁극적으로 설득 커뮤니케이션 분야에서, 특히 보건 커뮤니케이션 맥락에서 널리 연구됐던 획득-상실 메시지 프레이밍(Gain vs. loss message framing, Kahneman, & Tversky, 1979; O'keefe, & Jensen, 2007; O'keefe, & Nan, 2012; Park, Son, Lee, & Go, 2020; Rothman, Martino, Bedell, Detweiler, & Salovey, 1999)의 효과를 대규모의 기사 데이터로 실제 코로나19 백신 접종 상황에서 적용해 보았다는 데 그 의의가 있다. 예방 행동

(prevention behavior) 관련하여 그동안의 메시지 프레이밍 연구는 획득(gain) 프레이밍 메시지가 예방 행동을 장려하는데 더 설득력이 있는 것으로 알려져 왔지만(O'keefe & Jensen, 2007) 특별히 백신을 접종하는 예방 행동에 있어서는 백신 접종을 통해 얻을 수 있는 이익을 강조하는 전략과 백신을 접종하지 않았을 때 경험할 수 있는 상실을 강조하는 전략 두 가지의 효과 차이가 통계적으로 유의미하지 않았다(O'keefe, & Nan, 2012; Penta, & Baban, 2018). 특히 코로나19 백신 접종을 격려하는 메시지 프레이밍과 관련한 연구 결과 역시 국가/연령 그리고 연구 참여자의 심리학적 특성 별로 그 효과가 일정하지 않았다는 점을 생각해 보면(Borah, Hwang, & Hsu, 2021; Chen, Dai, Xia, & Zhou, 2022; Gursoy, Ekinci, Can, & Murray, 2022), 본 연구의 결과는 코로나19 백신接种의 이득을 강조하는 기사가 실제 접종률을 높였다는 점을 고려하여 메시지 전략을 수립할 수 있게 돕는다.

다음으로는 위기관리(crisis management) 측면에 대한 시사점을 몇 가지 찾아볼 수 있다. 코로나19, 메르스와 같은 국가적 보건 위기 상황에서 보도 프레이밍이 어떻게 변하였는가를 살펴보는 연구는 널리 행해져 왔으나(예, 박주현, 2020; 이주연, 이신행, 2021), 코로나19 백신 예약 절차, 백신 접종 결과 및 부작용에 대한 공중 보건 커뮤니케이션 관점에서의 함의는 제한적으로 연구되어 왔다(김수진, 차희원, 2022). 따라서 본 연구 결과를 바탕으로 위기 소통 메시지, 특히 예방적 행동(예, 코로나19 백신 접종)을 강조하고 전달하는 데 효과적인 주제를

선정하는데 도움을 줄 수 있다.

마지막으로, 방법론적으로 최신 기계학습법을 도입하여 대용량의 텍스트(본 연구에서는 기사 제목)를 효율적으로 연구하는 방법의 가이드라인을 제시하였다. 이는 광고/홍보 분야에서 앞으로 널리 사용될 기계학습과 빅데이터를 이용한 분석을 한 단계 나아간 방법론으로 실행할 수 있게 했다는 데 그 의의가 있다. 기존 사회과학 분야에서는 LDA 방법론에 기반한 토픽 모델링이 방법론적으로 주를 이루었다(강윤지 외, 2021; 강윤지, 2022; 김대욱, 2021; 문안나, 2016; 신성일 · 이은순, 2023; 오혜라 · 정윤재, 2021; 2022; 2023, 정원준, 2018; 2021, 함승경 외, 2019). LDA기반의 모델링 방법은 그 나름의 장점이 있어서 어떤 방법이 더 좋다고 말할 수는 없을 것이다. 하지만 BERTopic은 대용량의 텍스트 데이터를 효율적이고 엄밀한 통계적 가정(assumption) 없이 분석을 수행할 수 있는 장점이 있기 때문에 향후 연구자들은 BERTopic도 토픽 모델링의 방법으로 고려해 볼 수 있을 것이다.

3) 실용적 함의, 전략적 제언 및 후속 연구로의 초대

대한민국과 전 세계는 코로나19라는 유례 없는 보건 위기 상황을 경험하였다. 이러한 미증유의 위기 상황을 맞이하여 본 연구를 통해 향후 보건 위기 상황에서의 언론보도와 설득 메시지 구성에 관하여 실무적인 의미를 제한할 수 있다. 본 연구에서 밝혔듯이 단순히 확진자 수를 보도하는 것만으로는 백신 접종이라는 예방

행동에 긍정적인 영향을 미치지 못했다. 오히려 이러한 보도들은 백신 접종자 수와 부의 관계가 있었고, 해외여행과 같은 백신 접종으로 얻을 수 있는 이익을 강조하는 것이 접종자 수를 유의미하게 증가시킴을 생각해 볼 때, 향후 코로나19와 같은 신종 감염병이 출현하였을 때 위기 관리(crisis management)의 측면에서 이를 반영해야 할 것이다. 구체적으로 정부에서 신종 감염병이 출현하였을 때 그 여파를 최소화 하기 위한 공익 캠페인 디자인과 PR 활동, 더 나아가 광고 메시지 디자인에 본 연구에서 밝혀낸 점을 적극적으로 활용할 수 있을 것이다.

비록 최신 기계학습 방법론과 계량경제모델을 통해 본 연구는 많은 함의를 전달하였지만, 어떤 연구도 완벽하지는 않다. 따라서 연구의 한계점을 바탕으로 아래와 같은 후속 연구를 제안한다. 먼저 종속변수와 관련하여 본 연구는 전국 단위의 1차 접종자 수를 선정하였다. 이를 1차 접종자 및 2차 접종자, 그리고 추가 접종 접종자를 시/도 단위로 분석하여 본다면 좀 더 풍부한 함의를 답을 수 있을 것이다.

또한 본 연구는 백신 접종자들의 사회/경제적 구분과 같은 개인적인 요소(individual difference)를 고려하지 않았다. 우유현과 오상화(2023)의 연구에서와 같이 성별, 학력, 소득수준과 같은 사회적 결정요인을 모델에 고려한다면 이러한 요소들이 언론보도와 백신 접종자 수의 관계에 미치는 조절 효과를 알아볼 수 있을 것이다. 데이터 수집의 측면에서도 비록 본 연구는 코로나 백신 접종과 그에 따른 여러 주제를(예, 백신 접종에 대한 부정적인 감정, 부작용, 백신 개발과정 등등) 통합하여 데이터를 수집하기 위해 “코로나19 백

신”이라는 키워드로 언론 기사들을 수집하였으나 백신과 백신접종은 다른 개념이자 범위임을 감안하여 키워드를 다변화하여 데이터를 수집하고 살펴볼 필요가 있다. 비슷한 맥락으로 본 연구는 전체 언론 기사를 기관별 특징에 따라 분류하지 않았다(박주현, 2020). 따라서 박주현(2020), 이주연, 이신행(2021)의 연구처럼 코로나19 백신에 관련된 보도를 정치적 지향성, 보도 의도(현상 보도 vs. 의도적 기사)와 같은 언론사별 특징이 백신 접종자에 미치는 조절 효과 역시 연구되기를 기대한다.

또한 본 연구는 비록 변수 간의 인과관계를 담보할 수 있는 여러 가지 통계적 기법을 활용하였으나 더욱더 엄밀하게 인과관계를 추론할 수 있는 실험연구 방법을 활용한 후속 연구 역시 필요하다(예, 오지현, 김영옥, 2021; 홍주아, 안순태, 2022).

마지막으로 본 연구에서 연구된 언론의 효과 이외에도 다루지 못한 공중의 의견(public opinion)과 광고/홍보를 통한 의제 구축적(agenda building) 측면을 고려한 후속 연구가 필요하다. 기계학습 기법인 Word2Vec를 이용한 기업의 사회적 공헌 보도자료의 의제 구축 효과를 연구한 선행연구(Part et al., 2019)처럼 본 연구에서 소개된 BERTopic과 같은 기계학습 기법을 이용한 언론과 보도자료가 공중의 의견에 얼마나 어떻게 영향을 미치는지, 그리고 궁극적으로 공중의 행위에 어떻게 영향을 미치는지에 대한 연구가 활발히 이루어지기를 기대한다.

참고문헌

- 강윤지, 김재영, 강지훈, 김상훈. (2021). 의미연결망과 토픽모델링을 활용한 광고홍보학 분야의 CSR 연구동향 분석(2003~2020). *광고학연구*, 32(8), 7-40.
- 강윤지. (2022). 공유가치창출 (Creating Shared Value)에 대한 뉴스기사와 블로그의 프레임 및 네트워크 의제 설정 연구: 토픽모델링과 의미연결망을 중심으로. *경영학연구*, 51(5), 1165-1220.
- 권지원. (2023, 5. 13). 확진자부터 엔데믹까지...3년 4개월의 기록[코로나 종식①]. *뉴시스*, https://mobile.newsis.com/view.html?ar_id=NISX20230512_0002301434
- 김대옥. (2021). 북한이탈주민에 대한 한국 사람들의 인식 형성 과정 연구: 미디어 보도와 수용자 인식에 대한 의미 연결망 분석을 중심으로. *PR연구*, 25(1), 185-207.
- 김수진, 차희원. (2022). 코로나19 위기 소통 대국민 평가 지표 개발 연구. *광고학연구*, 33(2), 103-131.
- 김영옥, 김혜정, 소담이. (2022). 코로나19 백신의 정치화: 정치적 성향과 음모론에 대한 신념이 백신 접종에 미치는 영향. *광고연구*, 132, 97-132. doi: 10.16914/ar.2022.132.97
- 김효정, 정우진. (2023). 코로나 19 백신 기사에 대한 선택적 노출 연구: 정보과부하와 개별 감정의 영향을 중심으로. *PR연구*, 27(3), 95-122.
- 김효정, 홍혜현. (2021). 코로나 19 예방행동과 정보추구 의도 연구: 보호동기이론의 확장과 지지정당의 영향력 검증을 중심으로. *한국광고홍보학보*, 23(2), 42-76.
- 문안나. (2016). 한국 사회의 IR (Investor Relations) 개념: 1994~2014년 신문기사 내용분석 및 의미연결망 분석. *홍보학연구*, 20(1), 50-78.
- 신성일, 이은순. (2023). 충주시 유튜브 콘텐츠 댓글을 통해 본 지자체 행정 PR의 함의: 텍스트마이닝의 의미연결망 분석을 중심으로. *한국콘텐츠학회 논문지*, 23(12), 277-287.
- 안지수, 구윤희, 노기영. (2021). 코로나 19 초기에 정부 신뢰가 다채널 정보 추구에 미치는 영향: 정보 불충분성과 감정 반응의 매개 역할. *한국광고홍보학보*, 23(2), 113-139.
- 오미영. (2023). 사회신뢰와 집단효능감이 지각된 위험, 자기효능감 및 코로나19 예방행동에 미치는 영향. *광고학연구*, 34(1), 49-83.
- 오아름, 이준영, 김수진, 김활빈. (2021). 코로나19 펜데믹에서 흡연자의 관여인식이 금연행동의도에 미치는 효과에 관한 연구: 공포감과 자기효능감의 매개 및 조절된 매개효과. *광고연구*, 128, 5-34. doi: 10.16914/ar.2021.128.5
- 오준혁, 이보윤, 최미연, 최윤희. (2023). 확진보다 무서운 낙인? 유행시기별 코로나 19 보도 헤드라인에 나타난 낙인 양상 분석. *한국광고홍보학보*, 25(1), 45-85.
- 오지현, 김영옥. (2021). 위험 수치 제시 형식과 단위가 코로나 19 예방 행동 의도에 미치는 영향: 심리적 거리감과 부정적 감정의 매개 효과 및 구성원 효능감의 조절효과 분석. *광고학연구*, 32(2), 7-39.
- 오혜라, 정윤재. (2021). 상관토픽모델을 활용한 유튜브 크리에이터 소비자의 온라인 구전 특성에 관한 연구. *한국광고홍보학보*, 23(3), 37-72.
- 오혜라, 정윤재. (2022). LDA (Latent Dirichlet Allocation) 토픽 모델링을 활용한 숏폼 (Short-Form) 광고 특성에 관한 탐색적 연구. *광고연구*, 135, 51-86.
- 오혜라, 정윤재. (2023). 토픽 모델링을 활용한 오프라인 vs 온라인 스포츠 팬덤 활동 동기에 관한 탐색적 연구. *광고학연구*, 34(5), 1-29.
- 유우현, 오상화. (2023). 건강의 사회적 결정요인이 코로나 19 예방행동 의도에 미치는 영향: 정보 처리와 정보 탐색의 매개 역할. *광고연구*, 137, 47-74. doi: 10.16914/ar.2023.137.47
- 이병관, 노환호, 이태준. (2021) 뉴미디어 리터러시가 뉴스 팩트체크 행동에 미치는 영향: 불확실성에 대한 민감성의 조절효과와 가짜뉴스 태도 및 사실확인 동기의 매개효과를 중심으로. *PR연구*, 25(4), 31-56.
- 이서현, 진명지, 최낙진. (2020). 지방정부의 보도 자료는 언론 보도에 어떻게 활용되었나? 제주 지역 '코로나 19' 관련 주요 이슈를 중심으로. *PR연구*, 24(6), 1-32.
- 이완수, 정영희, 홍경수. (2022). 코로나19 백신 접종 뉴스 보도 형식과 의제 구성: 국가기간뉴스통신사

- 〈연합뉴스〉 사례를 통해. 헬스커뮤니케이션연구, 21(2), 1-49.
- 전중우. (2021). 코로나19 위험지각의 국가 차이와 문화 차원의 조절효과. PR연구, 25(4), 133-149.
- 정선영, 장동산. (2023). 코로나19 이후 생산성 변화의 주요 특징 및 시사점. BOK 이슈노트(한국은행 보고서), <https://www.bok.or.kr/portal/bbs/P0002353/view.do?nttId=10075270&menuNo=200433&pageIndex=1>
- 정원준. (2018). 사드 (THAAD) 이슈를 둘러싼 한국과 중국 간 갈등 쟁점의 변화 추이 연구: 빅데이터를 이용한 시계열 토픽 모델링과 언어 네트워크 분석 기법으로. 한국광고홍보학보, 20(3), 143-196.
- 정원준. (2021). 코로나 19 위험 커뮤니케이션 탐색 연구: 소셜미디어 의존 이론을 응용한 정부와 일반인 유튜브 콘텐츠와 댓글 내 담론 분석. PR연구, 25(2), 59-101.
- 정원준. (2022). 위험정보 프로세싱 (Processing) 다양화에 의한 코로나 예방 실천 행동 연구: 문제해결상황 이론을 확장하여. 한국광고홍보학보, 24(3), 69-103.
- 정재선. (2021). 코로나 19 위험인식이 온라인 정보행동과 사회불신에 미치는 영향: 가짜뉴스 신뢰와 내집단신뢰의 조절효과를 중심으로. 홍보학연구, 25(4), 83-104.
- 함승경, 최지명, 김영옥. (2019). 언론 보도의 여성 혐오 그리고 남성 혐오 분석: 언어 네트워크와 비판적 담론 분석의 결합. 홍보학연구, 23(6), 24-51.
- 황경주. (2024, 1, 25.). 지난해 경제성장률 1.4% ...코로나19 이후 최저. KBS. <https://news.kbs.co.kr/news/pc/view/view.do?ncd=7874779>
- Ahn, J., Son, H., & Chung, A. D. (2021). Understanding public engagement on twitter using topic modeling: The 2019 Ridgecrest earthquake case. *International Journal of Information Management Data Insights*, 1(2), 100033. <https://doi.org/10.1016/j.ijime.2021.100033>
- Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113(1), 151-184.
- Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. Paper presented at *the Proceedings of the 23rd International Conference on Machine learning*, 113-120.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.
- Borah, P., Hwang, J., & Hsu, Y. C. (2021). COVID-19 vaccination attitudes and intention: Message framing and the moderating role of perceived vaccine benefits. *Journal of Health communication*, 26(8), 523-533.
- Chen, T., Dai, M., Xia, S., & Zhou, Y. (2022). Do messages matter? Investigating the combined effects of framing, outcome uncertainty, and number format on COVID-19 vaccination attitudes and intention. *Health Communication*, 37(8), 944-951.
- Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1), 39-67.
- Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. MIT Press.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4), 1057-1072.
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7, 886498.
- Ethayarajah, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

- Ghasiya, P., & Okamura, K. (2021). Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach. *IEEE Access*, 9, 36645–36656.
- Goel, R., & Sharma, R. (2021). Studying leaders & their concerns using online social media during the times of crisis—A COVID case study. *Social Network Analysis and Mining*, 11, 1–12.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438. doi: 10.2307/1912791
- Greene, C. M., & Murphy, G. (2021). Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. *Journal of Experimental Psychology: Applied*, 27(4), 773–784. <https://doi.org/10.1037/xap0000371>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Gursoy, D., Ekinici, Y., Can, A. S., & Murray, J. C. (2022). Effectiveness of message framing in changing COVID-19 vaccination intentions: Moderating role of travel desire. *Tourism Management*, 90, 104468.
- Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM). *European Business Review*, 26(2), 106–121.
- Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40, 414–433.
- Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications* (pp. 77–86). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *ICML*, <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c52eb66e23b201cb44f567cbb270feadca532c9a>
- Jung, H. S., Lee, H., & Kim, J. H. (2023). Unveiling cryptocurrency conversations: Insights from data mining and unsupervised learning across multiple platforms. *IEEE Access*, 11, 130573–130583. <https://doi.org/10.1109/ACCESS.2023.3334617>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.
- Kim, J. H., Marks, F., & Clemens, J. D. (2021). Looking beyond COVID-19 vaccine phase 3 trials. *Nature Medicine*, 27(2), 205–211.
- Kim, J. W. (2022). How other minorities gained access: The war on poverty and Asian American and Latino community organizing. *Political Research Quarterly*, 75(1), 89–102.
- Kim, J., Ragas, M., Son, H., Park, K.-G., Chung, Y. J., & Park, Y. E. (2011). Examining influence during a public health crisis: An analysis of the H1N1 outbreak from an agenda-building and agenda-setting perspective. *Journal of Health & Mass Communication* 3, 116–137.
- Koupaei, M., & Wang, W. Y. (2018). Wikihov: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.
- Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., ... & Xiang, T. (2020). Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access*, 8, 46868–46876.
- McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176–187.
- McCombs, M. E., Shaw, D. L., & Weaver, D. H. (2014). New directions in agenda-setting theory and research. *Mass Communication and Society*, 17(6), 781–802.
- McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN:

- Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205.
- McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Moon, W.-K., Atkinson, L., Kahlor, L. A., Yun, C., & Son, H. (2022). US political partisanship and COVID-19: Risk information seeking and prevention behaviors. *Health Communication*, 37(13), 1671-1681.
- Mousavi, R., & Gu, B. (2023). Resilience messaging: The effect of governors' social media communications on community compliance during a public health crisis. *Information Systems Research*, <https://doi.org/10.1287/isre.2021.0599>
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *Econometrica*, 55(3), 703-708.
- NPR. (2021). *Vaccination is key to winning the battle against COVID, former CDC official says*. Retrieved October 14 from <https://www.npr.org/2021/10/12/1045182576/vaccination-is-key-to-winning-the-battle-against-covid-former-cdc-official-says>
- Ogbodo, J. N., Onwe, E. C., Chukwu, J., Nwasum, C. J., Nwakpu, E. S., Nwankwo, S. U., Nwamini, S., Elem, S., & Iroabuchi Ogbaeja, N. (2020). Communicating health crisis: A content analysis of global media framing of COVID-19. *Health Promotion Perspect*, 10(3), 257-269. <https://doi.org/10.34172/hpp.2020.40>
- O'Keefe, D. J., & Jensen, J. D. (2007). The relative persuasiveness of gain-framed loss-framed messages for encouraging disease prevention behaviors: A meta-analytic review. *Journal of Health Communication*, 12(7), 623-644. doi: 10.1080/10810730701615198
- O'Keefe, D. J., & Nan, X. (2012). The relative persuasiveness of gain- and loss-framed messages for promoting vaccination: A meta-analytic review. *Health Communication*, 27(8), 776-783. doi: 10.1080/10410236.2011.640974
- Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187-204.
- Park, S.-Y., Son, H., Lee, J., & Go, E. (2020). Moderating effects of social norms and alcohol consumption on message framing in responsible drinking campaigns: Value from deviance regulation theory. *Health Communication*, 35(7), 793-803. doi: 10.1080/10410236.2019.1593077
- Park, Y. E., Son, H., Yang, S.-U., & Lee, J. K. (2019). A good company gone bad: An examination of corporate social responsibility agenda-building in times of corporate crisis using machine learning techniques. *Journal of Communication Management*, 23(1), 31-51. <https://doi.org/10.1108/JCOM-11-2017-0132>
- Pența, M. A., & Băban, A. (2018). Message framing in vaccine communication: A systematic review of published literature. *Health Communication*, 33(3), 299-314.
- Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 1427-1445.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Rothman, A. J., Martino, S. C., Bedell, B. T., Detweiler, J. B., & Salovey, P. (1999). The systematic influence of gain-and loss-framed messages on interest in and use of different types of health behavior. *Personality and Social Psychology Bulletin*, 25(11), 1355-1369.
- Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, 69(10), 3998-4010.

- Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018, April). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1105–1114).
- Smith, J., & McAleer, M. (1994). Newey-West covariance matrix estimates for models with generated regressors. *Applied Economics*, 26(6), 635–640. <https://doi.org/10.1080/00036849400000034>
- Son, H., Ahn, J., Chung, A. D., & Drumwright, M. E. (2023). From the black box to the glass box: Using unsupervised and supervised learning processes to predict user engagement for the airline companies. *International Journal of Information Management Data Insights*, 3(2), 100181.
- Steinert, J. I., Sternberg, H., Prince, H., Fasolo, B., Galizzi, M. M., Bütte, T., & Veltri, G. A. (2022). COVID-19 vaccine hesitancy in eight European countries: Prevalence, determinants, and heterogeneity. *Science Advances*, 8(17), eabm9825.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Tahamtan, I., Potnis, D., Mohammadi, E., Singh, V., & Miller, L. E. (2022). The mutual influence of the World Health Organization (WHO) and twitter users during COVID-19: network agenda-setting analysis. *Journal of Medical Internet Research*, 24(4), e34321.
- Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, 141(6), 1178.
- Thakur, O., Saritha, S. K., & Jain, S. (2022). Topic modeling, sentiment analysis and text summarization for analyzing news headlines and articles. *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*.
- Troiano, G., & Nardi, A. (2021). Vaccine hesitancy in the era of COVID-19. *Public Health*, 194, 245–251. <https://doi.org/https://doi.org/10.1016/j.puhe.2021.02.025>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- World Health Organization (2024). *Number of COVID-19 cases reported to WHO*. Retrieved from <https://data.who.int/dashboards/covid19/cases?n=c>
- Yang, P., Li, W., & Zhao, G. (2019). Language model-driven topic clustering and summarization for news articles. *IEEE Access*, 7, 185506–185519.

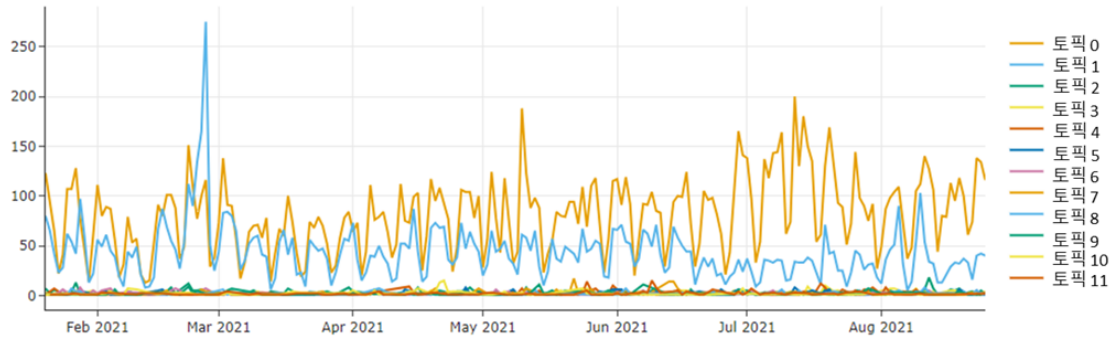
부록(Appendix)

부록 1. 코로나19 관련 주요 타임라인

날짜	주요 사건	데이터 수집
2020-01-20	국내서 코로나19 확진자 1명 발생 위기경보 주의로 상향	뉴스기사 수집기간
2020-01-27	감염병 위기경보 경계로 격상	
2020-02-12	정부, 신종코로나 공식 명칭 코로나19로 확정	
2020-02-23	코로나19 위기경보 심각단계 진입	
2020-03-11	세계보건기구(WHO) 팬데믹 선언	
2020-03-22	1차 강화된 사회적 거리두기 추진	
2020-05-26	마스크 없이 대중교통 이용 제한 실시	
2020-08-23	사회적 거리두기 2단계 전국 확대	
2020-10-13	마스크 착용 의무화	
2021-01-21	코로나19 예방접종 실시 방안 발표 1주일 전	
2021-01-28	코로나19 전 국민 무료 예방접종 실시 방안 발표	
2021-02-26	국내 첫 백신접종 시작	
2021-05-05	예방접종완료자 자가격리조치 완화	
2021-06-01	예방접종완료자 일상회복 지원 발표	
2021-06-20	사회적 거리두기 개편안 5단계에서 4단계로 간소화	
2021-07-09	수도권 사회적거리두기 4단계 시작	
2021-08-17	델타변이 출현	
2021-08-24	델타변이 출현 1주일 후	

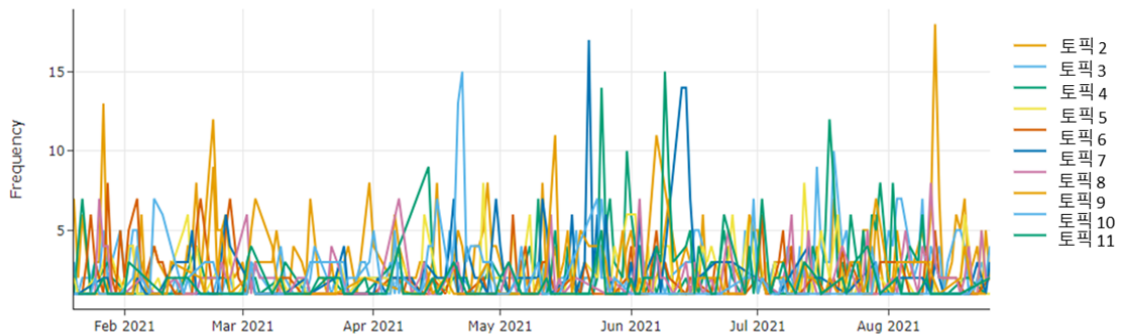
부록 II. 시간에 따른 토픽의 개수 변화

Topics over Time



부록 III. 시간에 따른 토픽의 개수 변화(토픽0, 토픽1 제외)

Topics over Time

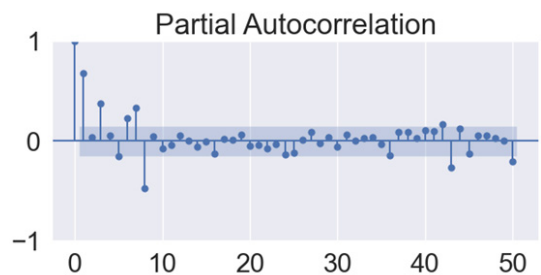
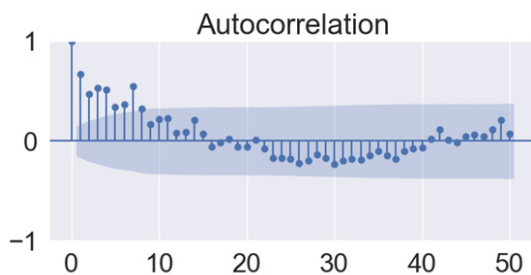


부록 IV. 기술 통계

변수	평균	SD	최소값	중앙값	최대값	VIF
백신접종자수	151,493.59	208,541.85	0	39,075	855,642	1.27
토픽1(백신_코로나_접종_코로나 백신)	28.30	12.97	4.00	29.50	60.00	1.00
토픽0(코로나_대통령_확진_신규)	86.57	37.93	13.00	87.00	200.00	2.47
토픽1(백신_접종_코로나 백신_코로나)	38.42	18.37	4.00	37.00	99.00	1.97
토픽2(치료제_코로나 치료제_코로나 셀트리온)	2.18	1.20	0.00	2.00	11.00	1.29
토픽3(감염_집단감염_코로나 변이)	1.80	1.70	0	1.00	8.00	1.26
토픽4(올림픽_도쿄올림픽_일본_도쿄)	1.78	2.33	0	1.00	14.00	1.20
토픽5(중국_nna_백신_중국산)	1.55	2.54	0	1.00	17.00	1.18
토픽6(바이러스_변이 바이러스_변이 변이바이러스)	1.32	1.57	0	1	10	1.40
토픽7(백신_한국_북한_대통령)	1.27	1.55	0	1	8	1.29
토픽8(학교_등교_2학기_교육)	0.94	1.22	0	1	6	1.50
토픽9(마스크_착용_마스크 벗고_마스크 착용)	1.09	1.77	0	1	12	1.19
토픽10(러시아_러시아 백신_이트론_이아이디)	1.07	2.03	0	0	16	1.46
토픽11(해외여행_여행_관광_여행업계)	1.17	1.81	0	1	15	1.37
토픽12(eu_백신_유럽_프랑스)	0.87	1.23	0	0.5	7	1.21
토픽13(인도_인도 코로나_코로나 생지목)	1.13	3.17	0	0	31	1.22
토픽14(대한항공_항공_항공업계_화물)	0.67	1.00	0	0	6	1.33
토픽15(브라질_브라질 코로나_코로나 사망)	0.74	1.30	0	0	8	1.19
토픽16(mrna_mrna 백신_개발_에스티팜)	0.69	0.93	0	0	4	1.21
토픽17(이스라엘_백신_1위 이스라엘_1위)	0.80	1.54	0	0	13	1.41
토픽18(집단감염_감염_부산_확진)	0.55	1.00	0	0	5	1.37
토픽19(유럽_eu_유럽은 봉쇄)	0.61	1.00	0	0	5	1.30
토픽20(삼성_삼성바이오_삼성바이오로직스_모더나)	0.66	0.93	0	0	4	1.30
토픽21(이스라엘_마스크_해제_국제뉴스_이스라엘)	0.59	1.24	0	0	8	1.25

관측치 (Observations) 168일

부록 V. 종속변수의 자기상관





BERTopic Analysis of The Impact of Media Coverage of COVID-19 Vaccinations on Vaccination Rates: Combination of Machine Learning and Modeling Techniques

•Park, Young Eun*

Professor, Sookmyung Women's University, Department of Public Relations & Advertising

Utilizing cutting-edge text modeling techniques such as BERTopic employing a transfer learning approach, the current study investigates a total of 38,312 news articles regarding COVID-19 vaccination. We found that COVID-19 confirmed cases, vaccination efforts, the development of COVID-19 treatments, variations, and mass infections are the prominent topics among the 22 topics that we extracted. Considering the time series dynamics, we observed a decrease in topics related to vaccination as the actual vaccination rate steadily increased, whereas it increased after the appearance of the Delta variant. Regarding the relationship between topics and actual vaccination cases, we identified a negative association when newspapers reported the number of confirmed cases. However, emphasizing the benefits of vaccination, such as the availability of overseas travel, led to an increase in vaccination rates. These findings offer novel insights for theory developers, advertising/public relations experts, and policymakers dealing with public health crises.

Keywords: BERTopic, Topic Modeling, COVID19, Vaccination, COVID19 Vaccination, Times Series Analysis, News, News Media

*yepark@sookmyung.ac.kr