Research

토픽모델링 기법을 적용한 코로나 관련 언론 키워드 분석

전은수[1], 오승훈[2], 조영목[3]
[1]서울대학교 자연과학대학 통계학과
[2]서울대학교 자연과학대학 생명과학부
[3]서울대학교 공과대학 기계항공공학부 우주항공공학전공

Keyword Analysis in Korean Articles Related to COVID-19 Using Topic Modeling

Eunsu Jeon[1], Seunghoon Oh[2], Yeongmok Cho[3]
[1]Department of Statistics, Seoul National University
[2]School of Biological Sciences, Seoul National University
[3]Department of Mechanical and Aerospace Engineering (Aerospace Engineering Major), Seoul National University

Abstract

Objectives: There was much research that assessed public health policies regarding the COVID-19 pandemic or analyzed media reports on other issues, but statistical analysis on COVID-19 related news reports were scarce. Thus, this study aims to apply LDA (Latent Dirichlet Allocation; topic modeling) method to news reports, and track the effects and interest related to pandemic policies.
Methods: 182,922 news articles from 11 main daily newspapers in Korea during 2020.10.12. ~ 2021.07.19. were used for morpheme analysis and topic classification using Mallet LDA method. 22 topics were decided by coherence score, then visualized by PCA (Principal Component Analysis) and compared with each other. Report statistics were also compared to the timeline of COVID-19 measures, particularly social distancing and vaccination policies.
Results: By comparing confirmed cases with the trend of keywords 'confirmed case status', 'social distancing guidelines', and 'vaccine inoculation', changing patterns in topics over time could be observed, and the cause could be analyzed. Especially, there was a significant difference between the third and fourth wave of the pandemic. At least in the aspect of media focus, the effectiveness of social distancing policies decreased in the fourth wave. Also, there was a significant media interest with the vaccine inoculation started.
Conclusion: Concluding from the media focus level, vaccination policies more influence rather than social distancing policies now.
keywords: *COVID-19, media report, LDA, topic classification*

서론

1. 연구 배경 및 필요성

1) 연구 배경

코로나바이러스감염증-19(코로나19)는 2019년 12월 첫 감염 사례가 발견된 이후 2021년 7월 전 세계에서 1억 9천만 명 이상이 감염될 만큼 세계적인 대유행으로 이어졌다[1]. 메르스와 같은 타 감염병에 비해 치사율은 낮으나, 높은 전염력 때문에 전 세계에서 4백만 명 이상의 사망자가 발생하여 위협을 가하고 있다[2]. 전염력이 높은 코로나19로 인해 국가 간의 이동이 제한되거나 다수의 인원이 모이는 행위가 금지되는 등 인류 사회의 전반적인 모습에 큰 변형이 일어나기도 했다. 일상사회의 변화는 결국 소비 위축과 경제 침체로 이어졌으며 정치, 외교, 경제, 기술 등 사회 전반적인 분야는 물론 사회적 거

*** Corresponding author:** Eunsu Jeon (eunsu0665@snu.ac.kr)
Department of Statistics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea.

리두기에 따른 대인 관계의 문제, 코로나 블루와 같이 개개인에게도 큰 영향을 끼쳤다[3, 4]. 또한 사람 간의 접촉을 최소화하기 위한 비대면, 언택트 관련 주제들이 큰 주목을 받고 있고, 코로나19가 종식된 후의 사회에 대해 다루는 포스트 코로나에 대한 대중들의 관심도 높은 상황이다[5].

한편, 코로나19 감염세를 막기 위해 국가적 차원에서도 적극적인 모습을 보이고 있다. 특히 백신이 개발되지 않았던 전염 초기에 집단 감염을 막기 위해 다수의 사람들이 집합하는 장소를 제한하고 사회적 거리두기를 적극적으로 선전하는 등 정부 차원에서 방역 정책을 시행하였다. 코로나19와 관련된 보건 정책에 있어 가장 중요한 점은 국민들이 상황의 심각성과 사회적 거리두기의 중요성에 대해 충분히 인지하고 최대한 많은 사람들이 동참하고 이행할 수 있도록 유도하는 것이다[6]. 일부 사람들의 사회적 거리두기 참여만으로는 집단 감염을 피할 수 없는 것처럼 모든 국민의 공감대를 얻어야 성공적인 방역 조치를 취할 수 있다. 이처럼 보건 정책은 대중들에게 정확하면서도 신속하게 전달되어야 하는 점이 핵심적이다. 이후에는 백신 개발이 완료되고 접종이 시작되면서 백신 접종 관련 소식과 정보에 대한 관심도 높아진 상황이다. 특히 백신 접종 이전 1.78%에 이르렀던 치명률이 백신 접종 이후 0.68%까지 떨어지는 긍정적인 변화가 이루어지며 주목을 받고 있는 상황이다[7].

이러한 측면에서 국가적 차원의 보건 정책을 모든 국민들에게 효과적으로 전달하는 매체의 중요성 또한 매우 높다. 이 과정에서 정부의 공식적인 보도자료를 전달하는 언론의 영향성이 크고, 그 중에서도 개인의 스마트폰 보급이 이루어진 현재에서는 인터넷 기사를 통한 정보 습득이 높은 비중을 차지하고 있다[8]. 따라서 언론에서 높은 주목을 받는 주제는 매체를 통해 더 적극적으로 전달될 것이고 더 많은 국민들에게 영향을 줄 수 있다. 즉, 언론에서 더 많이 언급되는 키워드들은 국가적인 이슈에 해당한다고 할 수 있다. 따라서 인터넷 기사를 대상으로 하는 토픽모델링 기법은 코로나19와 관련된 핵심적인 토픽들을 도출해내고, 각 토픽들의 비중을 분석해 특정 키워드에 대한 주목도를 수치화할 수 있다. 그리고 특정 보건 정책이 시행되기 전후 관련 토픽에 대한 관심도 추이를 비교해볼 수 있다.

2) 연구의 필요성

코로나19 보건 정책에 대한 평가로 다양한 연구가 진행되었지만 정책의 변화에 따른 확산행태를 분석하는 것에 그쳐 정량적 근거가 부족한 면이 있다[9]. 이러한 면에서 토픽모델링 기법을 이용한 코로나19 관련 키워드 분석을 적용한다면, 특정 키워드에 대한 국민적 관심도를 수치화하여 추이를 정량적으로 확인할 수 있어 정책 평가에 정량적 근거를 더해줄 수 있을 것으로 기대하였다.

강한 전염력을 지닌 감염병의 특성상 토픽모델링 분석을 이용해 이슈를 분석하려는 시도는 코로나19에만 한정되어 있지 않다. 조재희와 조인호[10]는 2015년 메르스 사태와 2018년 메르스 해외 재유입 시기 각각에 토픽모델링 분석을 적용하여 두 토픽 사이의 유사점과 차이점을 탐구하였다. 강찬희[11]는 소셜미디어를 대상으로 코로나19 관련 토픽모델링 분석을 실시하여 코로나19에 대한 사회적 인식과 이슈에 대해 분석하였다. 그러나 이 연구들은 관련 이슈를 분석하는 것에만 중점을 두었고 토픽 서로 간의 인과관계를 분석하거나 세밀하고 다양한 관점에서 관찰하는 것이 필요하였다. 따라서 토픽모델링 분석 결과를 객관적 근거로 활용해 다른 지표와 비교하며 능동적인 해석을 시행한 본 연구와는 차별점이 있다. 이 외에도 감염병 관련 언론보도 내용을 토픽모델링 기법 등을 이용해 분석한 연구들이 다양하다. 김태종[12]은 코로나19 뉴스 빅데이터를 토대로 토픽모델링 분석을 실시하여 언론을 통해 형성된 사회적 의제에 관해 연구하고 언론보도의 방향성을 제시하였다. 김상미[13]는 코로나19와 관련해 온라인 교육에 관한 언론 기사에 토픽모델링 분석을 실시해 주요 이슈와 시기별 변화 동향을 분석하고, 주목도가 높은 토픽을 분석하는 한편 언론에서 소외되는 의제를 제시하는 방향의 연구를 진행하였다. 서예령 등[14]은 국내 코로나19 대유행 시기별로 마스크 관련 기사들에 토픽모델링 분석을 적용하였으며, 언론보도가 사건·사고에 치우쳐져 있고 마스크에 관한 보건 정보에는 소홀하다는 점을 들어 언론보도의 변화 방향성을 제언하였다. 허용강 등[15]은 에볼라 바이러스와 관련한 언론보도 내용을 분석하여 언론사들이 감염병 보도 준칙을 준수하고 있는지 분석하며 감염병과 언론의 관계에 집중하였다. 그러나 이와 같은 언론보도 내용을 분석한 연구들은 언론을 평가하는 것에 초점을 두었다는 점에서 본 연구와 차별점이 있음을 확인하였다.

이처럼 코로나19에 대한 연구와 감염병 관련 언론보도 내용 분석, 토픽모델링을 이용한 이슈 분석 연구는 다양하였으나 코로나19 관련 기사에 토픽모

델링 분석을 실시해 통계적 분석을 시도한 경우는 없다.

2. 연구 목적

본 연구에서는 토픽모델링 기법으로 언론에서 다루는 기사를 분류하고, 토픽별 기사 수 추이를 기반으로 다양한 주제에 대한 사람들의 관심도와 이슈화 정도를 간접적으로 파악하고자 하였다. 이를 통해 확진자 수 추이와 코로나19 보건 정책의 두 축이라고 할 수 있는 거리두기 정책과 백신 접종 정책에 따른 각 주제의 추이를 확인하고, 이로부터 시간에 따른 주목도 변화를 분석하고 정책적으로 중요한 사건을 파악할 수 있을 것으로 기대하였다.

연구 방법

1. 인터넷 기사 자료 수집

신문기사 데이터베이스시스템 BIGKINDS(https://www.kinds.or.kr/)를 활용하여 2020.10.12. ~ 2021.07.19. 기간 동안에 '코로나'가 제목이나 본문에 포함되는 중앙지(경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보) 기사를 검색하였고, 총 182,922건의 인터넷 기사 자료를 수집하였다. 기사들은 일자와 키워드, 본문 내용의 일부를 포함하는 엑셀 형태의 파일로 저장하였다. 바이그램 검색을 통해서 혹시나 '코로나19' 또는 '코로나바이러스'가 수집되지 않는 경우를 배제하였다. 형태소 검색의 경우에는 '코로나19'와 같은 단어를 하나의 형태소로 인식하여 검색이 되지 않을 수 있지만, 바이그램 검색은 기사의 글을 2글자씩 떼어내서 순차적으로 비교하기 때문에 앞의 경우에도 검색이 된다. 또한, 분석 제외 필터를 적용하여 중복되거나 유사도가 너무 높은 기사들을 제외시켰고, 중앙지 기사들만을 선별하여 경제신문, 디지털신문과 같은 특정 분야의 기사들의 쏠림 현상을 방지하였다.

인터넷 기사의 수집 기간을 코로나19가 대두되기 시작한 2020년 초반이 아닌 2020.10.12.로 설정한 이유는 2차 대유행이 마무리되고 사회적 거리두기 단계가 전국적으로 1단계로 완화된 날짜가 10월 12일이었기 때문이다. 1, 2차 대유행의 경우 특정 집단에서 집단 감염이 이루어지는 형태로 전파되었고 감염 초기이므로 이후의 유행과는 차이가 있다고 볼 수 있어서 3차 대유행을 시작으로 하는 기간을 조사 기간으로 설정하였다. 아울러 비교적 코로나19가 국민들의 일상생활에 녹아들게 된 시점부터 분석을 진행하는 것이 좋을 것이라고 판단하였다. 특히나 코로나19가 처음 대두되기 시작하는 기간에 코로나19관련 기사들이 가장 많이 쏟아졌는데 비해 대부분의 기사들이 거의 유사하고, 분석에는 도움이 되지 않는 기사들이 많아 이 기간의 데이터들은 분석에 부적절하다고 판단하여 제외하였다.

2. 형태소 분석을 통한 키워드 추출

토픽모델링을 하기 위해서는 각 신문 기사 데이터를 형태소 분석을 통해서 형태소들의 집합으로 치환한 뒤, 그 중에서 명사이고 분석에 필요 없는 단어는 제외한, 이른바 키워드들만 남기는 작업이 필요하다. 연구에서는 BIGKINDS에서 제공하는 신문기사의 키워드 추출 데이터를 사용하였다. BIGKINDS에서는 제공하는 키워드에 대해서 "'키워드' 항목은 본문 내에서 추출된 키워드 중 단순 숫자(1, 2, 2018, 2019 등), 이메일 주소, 시간을 뜻하는 단어(밤, 낮, 새벽 등)를 제외한 결과가 표시됩니다." 라고 설명하고 있다. 실제로 분석을 진행하고 난 뒤에 키워드들을 살펴본 결과 '올해', '이번', '불과', '마찬가지' 등의 무의미한 명사들은 보이지 않은 것으로 보아 제공된 키워드 데이터는 신뢰할 수 있을 것으로 생각한다.

3. 토픽모델링을 통한 분석

문서별로 추출한 키워드를 바탕으로 182,922건의 기사 자료에 대해 토픽모델링을 실시하였다. 토픽모델링 방법은 기본 LDA 기법에서 단어의 사후 확률을 효율적으로 계산하는 깁스 샘플링 모델을 최적화시킨 Mallet LDA[16]를 사용하였다. 또한, 효율적인 계산을 위해 50개 이하의 빈도를 나타내는 키워드는 삭제하였다.

얻고자 하는 모델의 목적은 새로운 기사의 토픽을 예측하는 것이 아니므로 토픽 내의 기사들의 유사성이 높은 것이 더 중요하다고 판단하여 최적화된 토픽 개수를 정하기 위해 혼잡도가 아닌 일관성 점수를 계산하였다. 10개부터 49개까지 각 토픽 개수에 대해 토픽모델링을 진행하여 일관성 점수를 계산하였고, 그 결과는 Figure 1과 같았다.

일반적으로 토픽의 개수가 증가하면 한 토픽에

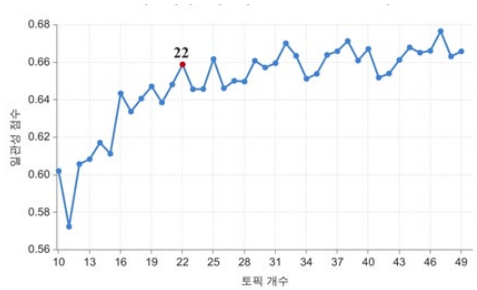

Figure 1. 토픽 개수에 따른 일관성 점수

해당하는 문서의 수는 감소하므로 토픽 내 문서의 유사성이 더 높아지게 되어 일관성 점수가 높아지는 경향을 보이게 된다. 그러나 무조건적으로 토픽의 개수를 늘리게 되면 오히려 많은 데이터를 주제별로 그룹화해서 보고 싶다는 본래의 목적을 잃어버리게 된다. 때문에, 적절한 범위 안에서 극대점에 해당하는 값을 토픽 개수로 설정하는 것이 중요하다. 위의 토픽 개수에 따른 일관성 점수를 살펴보면 토픽 개수가 22개일 때까지는 이후에 비해 큰 증가폭을 보이나, 그 이후로는 일관성 점수가 거의 일정하다시피 이어지고 있다. 따라서 토픽 개수가 22개인 것이 가장 적절하다고 판단하였고, 토픽 개수가 22개인 토픽모델링 결과를 이용해 분석을 진행하였다.

4. 토픽모델링 시각화 분석

토픽모델링을 적용하여 얻어낸 결과를 Python에서 제공하는 LDAvis 패키지를 통해서 시각화할 수 있다. 이때, 중요도에 따라서 가장 상위의 키워드를 뽑아낼 수 있는데, 토픽 내에서 키워드의 중요성을 나타내는 지표로는 크게 2가지를 생각할 수 있다. 첫 번째는 salience(특성) 값으로 $P(\omega|t)$, 즉 토픽 t의 문서 중 단어 ω를 포함하는 기사의 비율을 나타낸다. 한 토픽의 키워드가 되려면, 그 토픽에서 많이 나타나야 하는 단어여야 한다는 것이다. 그러나 이 경우에는 빈도수가 워낙 높아서 다른 문서에서도 많이 나타나는 단어가 키워드가 될 가능성이 크다. 이는 다른 토픽과의 차별성이 떨어짐을 의미한다. 이를 해결하기 위해 제시되는 두 번째 방법은 discriminative power(분별력) 값으로 $P(\omega|t)/P(\omega)$, 즉 salience 값에서 전체 기사에서 단어 ω가 나타나는 비율 $P(\omega)$로 나눠준 값을 계산한 것이다. 해당 토픽에서 많이 나오는 단어이더라도 원래 자주 등장하는 단어라면 그 중요도를 낮추겠다는 것이다. 그러나 이 역시 빈도수가 낮은 드문 단어들이 키워드가 될 수 있다는 단점이 존재한다. 때문에 LDAvis에서는 파라미터 λ 값을 통해 salience값과 discriminative power 값을 적절히 고려하면서 중요도가 가장 높은 키워드를 보여줄 수 있도록 설정할 수 있다. 구체적인 키워드 랭킹 점수는 다음과 같이 계산된다.

$$relevance(\omega|t) = \lambda * P(\omega|t) + (1-\lambda) * P(\omega|t)/P(\omega)$$

파라미터 λ값에 따라 중요도가 다르게 나타나는데, λ가 1에 가까울수록 salience 값에 가까워지는, 즉 토픽별로 자주 등장하는 단어들을 우선적으로 선택한다는 뜻이고, λ가 0에 가까울수록 discriminative power 값에 가까워지는, 즉 토픽 간에 차이가 크게 나는 단어를 우선적으로 선택한다는 뜻이다.

결과분석 및 논의

1. 토픽모델링 기법을 적용한 기사 키워드 분석 결과

1) 토픽모델링 분석 결과 시각화

Figure 2는 토픽모델링을 적용하여 얻어낸 기사 분석 결과를 Python에서 제공하는 LDAvis 패키지를 통해서 html 코드로 분석 결과를 시각화한 모습을 보여주고 있다. 그림에서 왼쪽의 좌표평면은 키워드의 총 개수만큼의 차원을 가지는 토픽 데이터를 차원축소방법(PCA)을 통해 2차원으로 축소시켜 놓은 결과를 보여준다. 좌표평면에서 각 축에 해당하는 PC1과 PC2는 바로 PCA 방법을 통해서 얻어진 2개의 주성분에 해당한다. 좌표평면 위의 토픽 데이터들은 하나의 원으로 표현되어 있는데, 원의 크기가 클수록 해당 토픽이 차지하는 비중이 크다는 것을 의미한다. 정확한 크기는 밑의 축적을 통해서 확인할 수 있다. 또한, 토픽 간의 유사성이 높을수록 해당하는 원들도 서로 가까운 위치에 놓이게 된다. Figure 2의 오른쪽에는 해당하는 토픽에서 가장 중요하게 여겨지는 키워드 30개를 차트 형태로 보여준다. 이때, 오른쪽 위에서 파라미터 λ값을 결정함에 따라 중요도가 달라진다. 차트에서 파란색 막대는 모든 기사를 통틀어서 나타나는 키워드의 빈도수를 나타내고, 빨간색 막대는 해당하는 토픽의 기사에서만 나타나는 키워드의 빈도수를 나타낸다. 파란색 막대에 대한 빨간색 막대의 비율이 높을수록 키워드가 해당 토픽에서만 나타나는 특징적인 키워드라고 볼 수 있다.

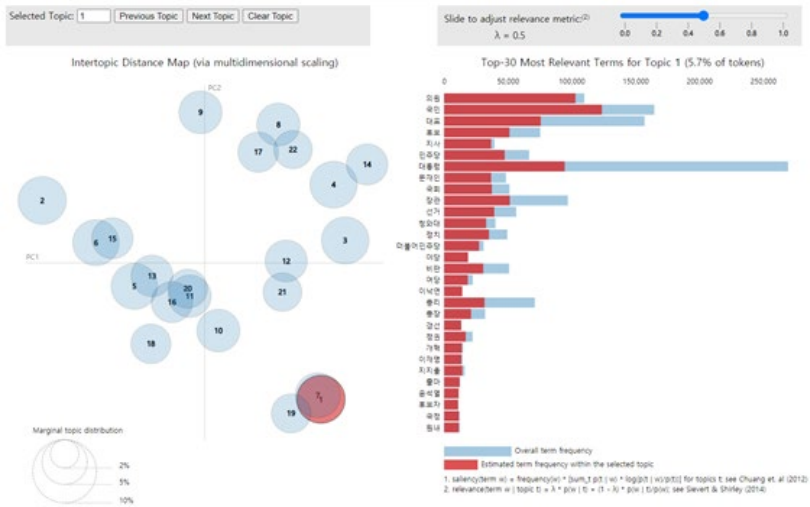

Figure 2. 토픽모델링 시각화 (*그림에 보이는 토픽의 번호는 이후 결과에 제시되는 토픽 번호와 무관하다)

2) 토픽의 이름 결정

토픽모델링을 통해서 총 182,922건의 기사들을 22개의 토픽으로 분류하였으나, 분류된 토픽들이 각각 어떤 주제를 담고 있는 지까지는 알려주지 못한다. 따라서 각각의 토픽들을 자세히 관찰하여 토픽의 이름을 붙이는 작업이 필요한데, 이를 위해서 이전의 시각화과정에서 보았던 각 토픽의 주요 키워드를 바탕으로 결정하였다. 아래의 Table 1은 파라미터 λ가 각각 0일 때와 1일 때의 상위 5개 키워드를 토픽별로 나타낸 뒤 이를 통해 결정한 토픽 이름을 함께 나타낸 것이다.

Table 1에서 볼 수 있듯이 각 토픽을 대표하는 키워드들을 확인해보면 해당 토픽이 어떤 주제들을 중점적으로 다루고 있는지를 쉽게 예측할 수 있다. 다만, 파라미터 값이 0인 경우에는 빈도수가 낮은 특이한 키워드들이 몇몇 발견되고 있다. 15번 토픽인 확진 현황의 키워드 중에는 코로나19 확산세가 나타났던 대전 BTJ 열방센터, 대전 IEM 국제학교와 관련된 키워드들을 볼 수 있었다. 또한, 6번 통계자료의 키워드 46%, 15번 확진 현황의 키워드 58명과 같이 수치와 관련된 키워드들도 볼 수 있는데, 숫자 자체는 키워드 대상에서 제외되지만, 퍼센트나 명과 같이 뒤에 단위가 붙게 되면 키워드 대상에 포함되기 때문에 나타나게 된 결과라고 보인다. 하지만 이 키워드 역시도 적어도 50번 넘게 나타난 단어이기 때문에(50번 이하의 단어들은 모두 제외되었기 때문) 그만큼의 빈도수를 가지는 단어라면 자체적으로도 의미를 가질 가능성이 있을 수도 있다.

Table 1. 토픽모델링 기법으로 분류한 22 개의 토픽과 이에 대한 상위 5 개 키워드

#	Topic Name	λ	Top-5 Most Relevant Terms
1	노동 문제	0	노조, 항공사, 아시아나항공, 파업, 비행
		1	직원, 노동자, 업무, 상황, 근무
2	백신 접종	0	접종, 아스트라제네카, 화이자, 접종자, 모더*
		1	백신, 접종, 아스트라제네카, 예방, 정부
3	국내 정치	0	야당, 후보자, 당내, 야권, 안철수
		1	국민, 의원, 대통령, 대표, 장관
4	문화예술	0	공연, 영화, 음악, 넷플릭스, 예능
		1	공연, 영상, 영화, 작품, 온라인
5	변이	0	변이, 영국발, 남아공, 거브러여수스, 팔레스타인
		1	영국, 코로나, 바이러스, 세계, 국가

6	통계 자료	0	감소, 취업자, 서비스업, 감소폭, 46%
		1	증가, 감소, 대비, 조사, 기록
7	인간 관계	0	아버지, 결혼, 어머니, 엄마, 저자
		1	사람, 생각, 사람들, 자신, 여성
8	소비	0	쇼핑몰, 샐러드, 온라인몰, 홍삼, 쿠팡이츠
		1	서비스, 제품, 온라인, 판매, 사용
9	미국 정치	0	부통령, 미국인, 인종, 흑인, 취임식
		1	미국, 대통령, 트럼프, 바이든, 대선
10	지역 문제	0	먼지, 관광지, 주차장, 해수욕장, 폭염
		1	지역, 서울, 운영, 주민, 도시
11	올림픽	0	선수, 도쿄, 프로, 출전, IOC
		1	일본, 경기, 대회, 올림픽, 선수
12	국제 관계	0	외교, 대북, 이란, 태평양, 공산당
		1	중국, 한국, 미국, 북한, 정부
13	기술	0	탄소, ESG, 선도, 패러다임, 탄소중립
		1	사업, 기업, 지원, 분야, 산업
14	거시경제	0	주식, 금리, 증시, 급등, 한국은행
		1	경제, 상승, 금융, 전망, 투자
15	확진 현황	0	위중증, BTJ, 동일집단, IEM, 58명
		1	확진자, 확진, 감염, 검사, 발생
16	방역 수칙	0	수칙, 격상, 5인, 방역수칙, 중대본
		1	방역, 거리, 마스크, 조치, 사회
17	기업	0	삼성전자, 전기차, 현대차, 부품, 애플
		1	기업, 시장, 생산, 회장, 반도체
18	교육	0	학교, 학생, 대학, 수업, 학생들
		1	교육, 학교, 학생, 대학, 수업
19	의료	0	병상, 질환, 연구팀, 응급실, 응급
		1	병원, 환자, 치료, 코로나, 의료
20	사회봉사	0	기부, 장애인, 목사, 기부금, 이웃들
		1	사회, 행사, 교회, 진행, 활동
21	범죄	0	경찰, 혐의, 부대, 선고, 고발
		1	경찰, 사건, 조사, A씨, 혐의
22	정부 지원	0	지급, 지원금, 소상공인, 추경, 재원
		1	지원, 정부, 지급, 지원금, 재난

* '나'를 조사로 인식하여 모더나를 모더로 인식한 것으로 보임

2. 토픽모델링 결과와 일일 신규 확진자 수, 방역 정책의 대조

1) 토픽모델링 결과와 포털사이트 검색 데이터의 관계

토픽모델링의 결과를 해석하기에 앞서 다음의 결과를 살펴볼 필요가 있다. 아래의 Figure 3는 포털사이트 네이버에서 이루어진 '확진자' 단어의 검색 데이터량과 확진 현황 토픽 관련 일일 평균 기사의 수를 주별로 나타내어 비교한 그래프이다. 네이버 데이터랩에서 제공하는 네이버 검색 데이터와 토픽모델링의 결과를 통해서 확진 현황 토픽의 기사 수 데이터를 활용하였다. 그 결과 7주차, 10주차, 19주차, 26주차, 39주차 등에서 공통적인 피크가 나타나고 그 상대적인 크기가 유사하다는 점에서 두 곡선은 흡사한 추이를 보이는 것을 확인할 수 있다. 또한, 두 데이터가 유사한 관계를 가지고 있음을 수치적으로 확인하기 위해서 코사인 유사도를 계산한 결과 약 94.20%의 유사도를 보임을 확인할 수 있었다. 따라

서 포털사이트의 확진자 단어의 검색량과 확진 현황 토픽의 기사 수는 서로 상관관계가 있다고 결론지을 수 있을 것이다.

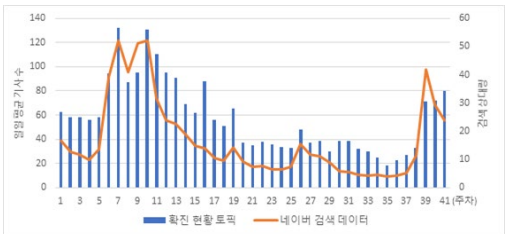

Figure 3. '확진 현황' 토픽 관련 기사의 수와 '확진자' 네이버 검색 데이터의 비교 그래프

포털사이트 검색은 대중의 능동적 행위다. 포털사이트의 검색량이 높다는 것은 더 많은 대중들이 관심을 가지고 능동적으로 검색했다는 의미이므로 대중의 관심도를 반영하고 있다고 볼 수 있다. 이때, 포털사이트의 검색량과 관련 기사 수가 상관관계를 가졌던 앞선 결과를 보았을 때, 토픽모델링을 통해 얻어낸 결과를 국민적 관심도와 연결지어 살펴볼 수 있을 것이라는 아이디어를 착안하였다. 또한, 언론은 국민들에게 정보를 전달해주는 역할을 하므로 언론에서 큰 관심을 가지는 토픽은 대중들의 큰 관심으로도 이어질 수 있는만큼, 언론의 주목도가 가지는 의미는 더 클 것이다.

따라서, 각 토픽의 추이를 일일 신규 확진자 수, 방역 수칙의 변화 시점과 함께 대조하여 분석한다면 주목도가 큰 중요한 이슈에 대해 파악할 수 있고 나타나는 양상의 변화를 분석해 코로나 이슈에 대한 주목도의 변화를 해석할 수 있을 것이다.

대조할 일일 신규 확진자 수는 KDX 한국데이터거래소와 보건복지부에서 제공하는 자료를 사용하였다. 발행된 기사의 수와 일일 신규 확진자 수는 모두 요일의 영향을 받기에 2020년 10월 12일부터 2021년 7월 19일까지의 데이터를 주별로 평균 내어 총 41주간의 추이를 비교하였다.

2) 확진 현황, 방역 수칙 토픽의 추이

확진 현황 토픽에 대한 추이는 Figure 4의 그래프와 같이 나타난다. 3차 대유행, 4차 대유행과 함께 피크가 나타나는 것을 확인할 수 있다.

방역 수칙 토픽에 대한 추이는 Figure 5의 그래프와 같이 나타난다. 확진자 수의 증가와 더불어 방역 수칙 변화가 일어난 시점에 증가하는 것을 확인할 수 있다.

'확진 현황', '방역 수칙' 두 가지 토픽의 추이에 큰 변동이 일어나는 시점을 구체적으로 명시하면 다음과 같다.

(1) 6~7주차 (2020년 11월 16일 ~ 2020년 11월 29일)

5주차 평균 일일 확진자 수가 159.9명에 불과했으나 6주차에 312.4명, 7주차에 441.6명으로 확진자 수가 급증하며 3차 대유행이 시작되었다. 이와 같은 시점에 확진 현황 토픽 관련 기사의 수가 5주차 일일 평균 58.1개에서 94.3개(6주차), 132.1개(7주차)로 역시 증가한 것을 확인할 수 있다.

6주차 2020년 11월 17일에 수도권 지역의 거리두기를 1단계에서 1.5단계로 격상한다는 사실이 발표되었고 6주차에 방역 수칙 토픽에 대한 관심도 역시 크게 증가한 것을 확인할 수 있다. 5주차 방역 수칙 토픽 관련 기사의 수는 31.6개였으나, 6주차에는 66.1개로 증가하였다.

(2) 9~11주차 (2020년 12월 7일 ~ 2020년 12월 27일)

10주차 평균 일일 확진자 수가 985.6명으로 매우 빠른 속도로 증가하였으며, 일일 신규 확진자가 1,000명을 넘는 등 역대 최다 기록을 보였다. 이와 함께 확진 현황 토픽 관련 기사의 수 역시 9주차 95.1개에서 10주차에 130.9개까지 증가하였다.

방역 수칙 토픽의 추이가 증가한 9주차에는 사회적 거리두기 강화(수도권 2.5단계, 비수도권 2단계, 12월 6일 발표)가 이루어졌다. 10주차에는 확진자 수의 증가에도 불구하고 새로운 방역 수칙이 발표되지 않았는데, 이 시기 방역 수칙 토픽은 확진 현황 토픽과는 대조적으로 감소했다. 11주차 12월 23일부터는 수도권 5인 이상 모임 금지라는 새로운 거리두기 정책이 발표되었고, 방역 수칙 토픽 관련 기사는 일일 평균 63.4개까지 증가하였다.

(3) 12~15주차 (2020년 12월 28일 ~ 2021년 1월 24일)

9주차 사회적 거리두기가 강화되고 11주차에 5인 이상 모임 금지 정책이 새롭게 시행된 이후 3차 대유행이 마무리되는 시점으로 신규 확진자 수가 꾸준히 감소하는 구간이다. 확진 현황 토픽 관련 기사의 수도 꾸준히 감소하는 것을 확인할 수 있다.

(4) 16주차 (2021년 1월 25일 ~ 2021년 1월 31일)

15주차 평균 일일 확진자 수가 392.0명이었으나 16주차에는 445.9명으로 53.9명(13.8%) 증가하였으나 확진 현황 토픽 관련 기사의 수는 15주차 일일 평균 61.6개에서 16주차 88.0개로 26.4명(42.9%) 증가하였다. 확진자 수의 증가가 크지 않았음에도 불구하고 확진 현황 토픽 관련 기사의 수는 크게 증가하였다. 17주차 평균 일일 확진자 수는 384.4명으로 다시 감소하게 되었다.

(5) 19주차 (2021년 2월 15일 ~ 2021년 2월 21일)
18주차 평균 일일 확진자 수가 375.6명이었으나 19주차에는 495.3명으로 119.7명(31.9%) 증가하였고 확진 현황 토픽 관련 기사의 수는 18주차 일일 평균 51.3개에서 19주차 65.6개로 14.3개(27.9%) 증가하였다. 이후 20주차 평균 일일 확진자 수는 383.4명으로 다시 감소하였다.

(6) 26~34주차 (2021년 4월 5일 ~ 2021년 6월 6일)
대유행은 아니지만, 이 기간에 일일 확진자 수는 600명 이상을 유지하며 확산세를 보였다. 특히 다시 600명 이상의 일일 확진자가 발생한 26주차의 경우 방역 수칙이 변화하지 않고 기존의 정책을 연장하는 방향으로 결정되었다. 이 시기 확진 현황 토픽 관련 기사의 수가 47.7개로 전주에 비해 46.3% 증가, 방역 수칙 토픽 관련 기사의 수가 42.3개로 전주에 비해 77.0% 증가하였다. 하지만 확진 현황 토픽 관련 기사의 양이 3차 대유행, 16주차, 19주차에 비해 절대적으로 작았다. 또한 이전과 다르게 확진자 수가 바로 감소하지 않고 8주간 일일 확진자 수가 평균 600명 이상 유지하였다.

(7) 39~41주차 (2021년 7월 5일 ~ 2021년 7월 19일)
39주차 평균 일일 확진자 수가 1137.4명으로 1,000명을 돌파하며 4차 대유행이 시작되었고, 7월 9일에는 오후 6시 이후 3인 이상 모임이 금지되는 등 수도권 거리두기 4단계 정책이 새롭게 발표되었다. 39주차 방역 수칙 토픽 관련 기사의 수도 84.9개까지 증가하며 3차 대유행보다도 높은 최고 수치를 기록하였다. 한편 확진 현황 토픽 관련 기사의 수는 39주차 71.1개, 41주차 80.0개로 증가는 하였지만 3차 대유행 시기인 7주차 132.1개에 비해서는 절대적으로 줄어들었다. 그리고 1,000명 이상의 일일 확진자 수가 계속 유지되면서 확진자의 감소세도 보이지 않았다.

정리하면, 확진 현황과 방역 수칙 토픽 관련 기사의 수는 확진자 수, 그리고 방역수칙의 변화에 따라 함께 증가하고 감소하는 추이를 보였다. 다만 시기

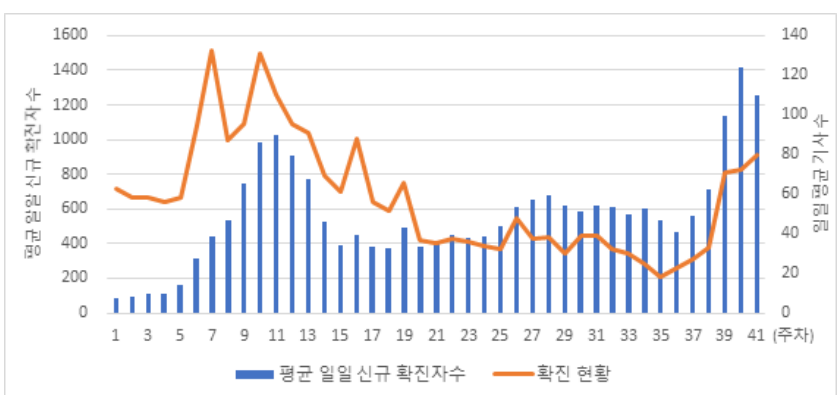

Figure 4. '확진 현황' 토픽과 일일 신규 확진자 수의 추이 비교 그래프

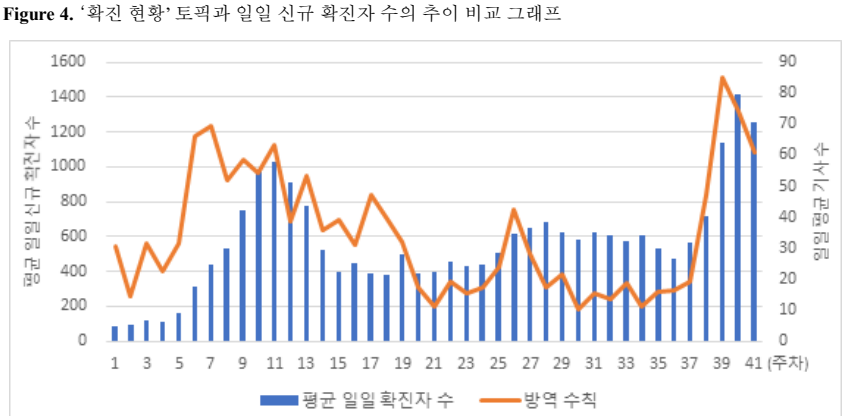

Figure 5. '방역 수칙' 토픽과 일일 신규 확진자 수의 추이 비교 그래프

에 따라 그 형태가 조금 다르게 나타났는데, 3차 대유행과 16주차, 19주차의 확산세에서는 작은 확산세에서도 방역 수칙 토픽 관련 기사의 수와 확진 현황 토픽 관련 기사의 수가 모두 크게 증가했고 이후 확산세가 빠르게 감소하였다. 그러나 그 이후 26주차의 확산세, 4차 대유행에서는 방역 수칙 토픽 관련 기사의 수의 증가에 비해 확진 현황 토픽 관련 기사의 수의 증가량이 상대적으로 적었고, 확진자 수의 명확한 감소로 이어지지도 않았다.

3) 백신 접종 토픽의 추이

백신 접종 토픽에 대한 추이는 Figure 6의 그래프와 같다. 백신 공급에 대한 정부 발표가 이루어진 9~12주차에서 높은 양상을 보이고, 국내 첫 백신 접종이 시작된 20주차에서 백신 접종 토픽 관련 기사의 수가 일일 평균 112.4개로 가장 높은 관심을 받은 것을 확인할 수 있다. 이후 백신 접종은 꾸준히 높은 관심을 받고 있다.

백신 접종이 시작되면서 언론에서 높은 주목도를 받은 시기에 확진 현황 토픽의 추이에도 변화가 나타났다. 실제로 백신 접종 토픽과 관련하여 가장 높은 주목도를 받은 시기인 20주차 이전에는 작은 확산세에도 강한 증가를 보였던 확진 현황 토픽이 그 이후에는 상대적은 낮은 증가세를 보이기도 하였다.

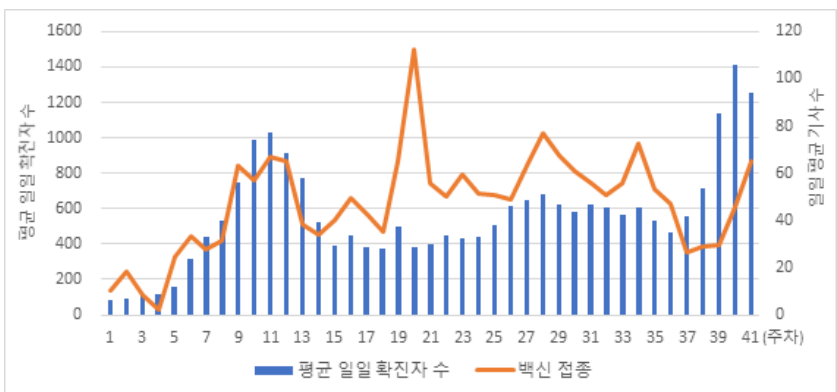

Figure 6. '백신 접종' 토픽과 일일 신규 확진자 수의 추이 비교 그래프

결론 및 제언

지금까지 토픽모델링 기법을 통해서 2020.10.12. ~ 2021.07.19. 기간 동안에 '코로나'로 검색된 182,922건의 인터넷 기사 자료를 분석해보았다. 일관성 점수 계산을 통해 토픽 개수를 22개로 결정한 뒤, 각 토픽에 대해 상위 키워드들을 살펴본 결과 노동 문제, 백신 접종, 국내 정치, 문화예술, 코로나바이러스 변이, 통계자료, 인간관계, 소비, 미국 정치, 지역 문제, 올림픽, 국제 관계, 기술, 거시경제, 확진 현황, 방역 수칙, 기업, 교육, 의료, 사회봉사, 범죄, 정부 지원의 22가지 주제들로 분류되었음을 확인하였다. 이후 기준일로부터 각 주차별로 특정 토픽들의 기사 수가 어떻게 변화하는지를 확인함으로써, 코로나 사태가 진행되어감에 따라 각각의 토픽 관련 기사 수가 어떤 양상을 나타내는지를 살펴보았다.

이때 확진 현황 토픽 관련 기사 수와 포털사이트 네이버에서 '확진자' 단어 검색 데이터량이 높은 상관관계를 보이는 것을 확인하였다. 이는 코로나 확진 현황 관련 키워드에 대해 언론의 주목도와 대중의 관심도가 상관관계를 가지고, 본 주제에 대한 언론 키워드 분석이 가지는 의의를 시사하고 있다고 볼 수 있다.

그리고 확진 현황, 방역 수칙, 백신 접종 토픽과 가장 큰 관련이 있을 것이라 생각되는 주차별 평균 확진자 수 데이터와 비교하여 추이를 설명하고자 하였다. 그 결과 확진 현황 토픽과 방역 수칙 토픽 관련 기사의 수가 3차 대유행과 4차 대유행에서의 추이가 어떻게 변화했는지 관찰할 수 있었고, 또한 정부의 방역 수칙이 변화되는 시점을 관찰하여 새로운 방역 수칙의 발표와 시행 전후 두 토픽 관련 기사의 수가 어떠한 방향으로 변화하였는지 파악하였다. 구체적으로, 3차 대유행에서는 초기에서부터 확진 현황과 방역 수칙 토픽에 대해 많은 관심도를 나타냈고, 이후 확진자 수의 작은 증가폭에도 민감하게 반응하는 모습을 보여주었다. 또한 이 시기는 사회적 거리두기 정책과 5인 이상 모임 금지 정책 등의

시행으로 인해 대유행이 비교적 빠르게 마무리되었다. 그에 비해 4차 대유행에서는 전례 없는 확진자 수의 증가폭에도 불구하고 확진 현황 토픽에 대한 기사 수는 3차 대유행에서만큼 크게 나타나지 않았다. 방역 수칙 토픽에 대한 기사 수는 이전과 같이 높아졌으나 확진 현황에 대한 관심도가 상대적으로 덜 증가하였다. 그러나 같은 시기 백신 접종 토픽에 대한 기사 수는 높은 수를 유지하며 백신 접종 토픽의 비중이 증가한 것을 확인할 수 있었다. 즉, 백신 접종 토픽의 주목도가 높아진 이후 확진 현황 토픽에 대한 주목도가 상대적으로 낮아졌다.

그러나 이 연구에서는 몇 가지 내부적인 결함들로 인한 한계점들이 존재한다. 첫 번째로, 분석을 진행한 182,922건의 인터넷 기사가 연구에 적합한 기사를 모두 수집했는가에 대한 한계이다. 분석에 포함되어야 하는 기사가 배제되었다거나, 반대로 제외되어야 할 기사가 분석 대상에 남아있는 경우가 생기지는 않았는지 검토해보아야 한다. 특히나 유행 이후 대부분의 주제의 기사에 '코로나'라는 단어가 포함되어 있을 것이라 예상되므로 과연 '코로나'라는 검색어만으로 충분했는지를 고민해볼 수 있다. 허나 분석 방법이 토픽모델링이기 때문에 분석하기를 원하는 주제들만 선별할 수 있다는 장점이 있기에, false positive에 대해서는 비교적 용인될 수 있다. 두 번째로, 분석 방법에 대한 것이다. LDA 기법을 통해 기사들을 분류하면 기사들이 각 토픽별로 얼마만큼의 비중을 가지는지를 알 수 있는데, 분석을 진행할 때에는 시간별 토픽의 추이를 확인하기 위해 그 중에서 가장 큰 비중을 가지는 토픽 하나에만 기사들을 분류하였다. 하지만 실제로 각 기사들은 하나의 토픽만을 가지기 보다는 확진 현황과 방역 수칙을 함께 다루는 식으로 여러 토픽에 함께 속할 수 있을 것이다. 때문에 이러한 경우에 대한 보정이 필요하지만 그렇지 못한 한계가 존재한다. 또한, LDA 기법의 경우 각 토픽간의 관계가 없다는 가정 하에 이루어지기 때문에 실제로 세밀한 분류가 어려워지는 문제점이 있는데, 이를 해결하기 위해 각 토픽간의 상관관계를 설정한 CTM(Correlated Topic Model)이나 메타데이터를 기반으로 한 STM(Structural Topic Model) 발전된 분석 기법을 활용할 수도 있을 것이다. 마지막으로 분석에 사용한 인터넷 기사의 기간이다. 최근에 계속되고 있는 토로나19의 4차 대유행이 끝나지 않은 시점에서 연구 기간상의 문제로 인해 7월 19일 월요일을 끝으로 더 이상의 기사를 수집할 수 없었다. 때문에 이후의 기사들을 더 수집하여 분석을 진행하지 못한 한계가 존재한다.

본 연구를 통해서 코로나19와 연결되어있는 사회의 많은 토픽들을 알아보고, 그 양상을 관찰할 수 있었다. 특히나 확진 현황과 방역 수칙, 백신 접종의 토픽들이 확진자 수, 정책의 변화에 따라 나타나는 추이를 알아보았다는 점에서 의의가 있다. 이러한 방식의 연구를 지속한다면 사람들의 관심도를 파악하고, 사회의 분위기를 읽는 주요한 역할을 하면서, 동시에 이전의 정책 시행 이후 나타난 효과를 파악하며 새로운 정책을 모색하는 데에 있어서 도움을 줄 수 있을 것이라 생각한다.

참고문헌

1. WHO. WHO Coronavirus (COVID-19) Dashboard, 2020.07.30. Available from: https://covid19.who.int/
2. Lee, Y. The Impact of the COVID-19 Pandemic on Vulnerable Older Adults in the United States. Journal of Gerontological Social Work, 2020; 63(6-7): 559-564.
3. 박희석, 반정화, 정현철, 김수진. 코로나 19 사태가 서울경제에 미치는 영향과 소상공인 및 관광업 대응 방안. 정책리포트, 2020; 297: 1-32.
4. 이동훈, 김예진, 이덕희, 황희훈, 남슬기, 김지윤. 코로나바이러스(COVID-19) 감염에 대한 일반대중의 두려움과 심리, 사회적 경험이 우울, 불안에 미치는 영향. 한국심리학회지: 상담 및 심리치료, 2020; 32(4): 2119-2156.
5. 백선혜, 이정현, 조윤정. 포스트코로나 시대 비대면 공연예술의 전망과 과제. 정책리포트, 2020; 307: 1-30.
6. 김정. 코로나 19 방역 정책의 성공 조건: 한국 사례의 비교연구. 한국과국제정치(KWP), 2021; 37(1): 191-221.
7. 최재규. 백신의 힘... 코로나 치명률 1.78% → 0.68%. 문화일보. 2021.06.16.
8. 서병호, 김춘식. 정부의 대언론 홍보에 대한 연구 – 재정경제부의 보도자료 분석과 평가를 중심으로. 한국언론학보, 2001; 45(2): 216-249.
9. 김동규, 정윤, 이견직. 코로나 19 에 대응하는 세분화된 사회적 거리두기 정책의 효과 분석. 한국시스템다이내믹스연구, 2021; 22(1): 37-57.
10. 조재희, 조인호. 2018 메르스 해외 재유입에 대한 주요 온라인 이슈 탐색: 토픽모델링 분석과 감성 분석을 중심으로. 한국디지털 콘텐츠학회논문지, 2019; 20(5): 1051-1060.
11. 강찬희. 동적토픽모델링과 의미연결망을 통한 코로나 19 이슈분석: 트위터 텍스트데이터를 활용하여. 성균관대학교 일반대학원 석사

학위논문. 2021.
12. 김태종. 뉴스 빅데이터를 활용한 코로나 19 언론보도 분석 :토픽모델링 분석을 중심으로. 한국콘텐츠학회논문지, 2020; 20(5): 457-466.
13. 김상미. 코로나 19 관련 온라인 교육에 관한 국내 언론보도기사 분석. 한국디지털콘텐츠 학회논문지, 2020; 21(6): 1091-1101.
14. 서예령, 고금석, 이재우. 빅데이터 LDA 토픽 모델링을 활용한 국내 코로나 19 대유행 기간 마스크 관련 언론 보도 및 태도 변화 분석. 한국정보통신학회논문지, 2021; 25(5): 731-740.
15. 허용강, 차수연, 서필교, 김소영, 백혜진. 감염병 보도 지침에 따른 에볼라 바이러스 언론보도 내용분석: 국내 주요 일간지를 중심으로. 헬스커뮤니케이션연구, 2015; 12: 75-113.
16. Yao, L., Mimno, D., McCallum, A. Efficient Methods for Topic Model Inference on Streaming Document Collections. KDD, 2009; 937-946.