조사연구(2023. 2) 24권 1호: 113-154
DOI http://dx.doi.org/10.20997/SR.24.1.4

연구논문

이론기반 텍스트마이닝의 방법과 절차:
비지도 및 준지도 토픽모델링을 이용한 감염병보도에 대한 컴퓨터보조내용분석*

안도현**

텍스트마이닝은 컴퓨터보조내용분석 혹은 전산 내용분석으로서 내용분석 과정을 자동화함으로써 대량의 자료를 비교적 저렴한 비용으로 신뢰성있게 처리할 수 있는 조사방법이다. 기존의 텍스트마이닝을 이용한 연구는 비지도학습 방식의 토픽모델링에만 의존하는 경우가 많아 자료기반의 탐색적 연구에 그쳤다. 이 연구는 준지도 기계학습을 이용하여 자료기반과 이론기반 텍스트마이닝을 통합적으로 적용할 수 있는 방법을 제시했다. 이를 위해 이론기반의 씨앗사전을 두 가지 방식(비지도 토픽모델링과 연결망분석과 해석적 탐색 및 기존에 개발된 사전 활용)을 통해 구성하여 가설검정에 필요한 변수를 구성하는 방법론적 틀을 제시했다. 이 방법을 감염병보도에 대한 이론기반 텍스트마이닝에 적용하여, (1) 해결지향보도, 의제설정, 틀짓기 등 3종의 이론을 토대로 가설을 설정하고, (2) 준지도학습으로 변수를 구성하여 (3) 결과를 분석한 사례를 제시했다. 준지도 토픽모델링의 군집성능이 비지도 토픽모델링의 성능보다 우수함도 확인하고 이론기반 텍스트마이닝이 지니는 사회조사의 함의에 대해 논의했다.

주제어: 토픽모델링, 연결망분석, 의제설정, 프레임, 해결지향보도

* 이 논문은 2022학년도 제주대학교 교원성과지원사업에 의하여 연구되었음.
** 제주대학교 언론홍보학과 교수(dohyun@socialbrain.kr).

Ⅰ. 연구 배경

텍스트마이닝(textmining)은 컴퓨터보조내용분석(computer-assisted content analysis)으로서 사회조사의 핵심 도구 중 하나다. 사람의 수작업에 의존하던 내용분석 과정의 상당부분을 기계로 대체함으로써 대량의 자료를 저렴한 비용으로 처리할 수 있을 뿐 아니라 분석결과의 신뢰성을 크게 향상시킬 수 있는 장점이 있기 때문이다. 이런 장점을 살려 텍스트마이닝은 대량의 텍스트(청와대의 국민청원 게시판에 탑재된 6천여 건의 게시글 혹은 한 학술지에 게재된 20년치 논문의 내용)를 분석하는 데 활용되고 있다(설동훈 등 2020; 양혜진·안정민·이태헌 2020). 텍스트마이닝은 비용문제로 설문조사에서는 채택하기 어려웠던 개방형 문항의 활용가능성도 크게 높였다. 미투운동에 대한 응답자의 생각을 자유롭게 서술하도록 하거나(송준모, 2021), 신고리 5·6호기 공론화 과정에 만족하거나 불만족하는 이유를 자유롭게 기술하도록 한 개방형 문항을 설문조사에 활용하기도 했다(정우연 2022). 동일한 내용에 대하여 폐쇄형 문항과 개방형 문항을 함께 사용하여 개방형과 폐쇄형 문항의 성능을 평가한 연구에서는 개방형 문항의 성능이 더 우수한 것으로 나타났다(Baburajan et al. 2021).

텍스트마이닝 방법으로는 단어빈도의 계산, 연결망분석, 기계학습(machine learning) 이용 등 다양하다. 사회과학과 디지털인문학 연구의 내용분석 도구로 널리 사용되는 텍스트마이닝 기법은 토픽모델링(topic modeling)이다. 토픽모델링의 종류는 다양하나 가장 널리 사용되는 토픽모델링이 비지도학습 방식의 LDA토픽모델링이다. 단어와 문서의 확률분포로 군집한 결과를 통해 문서에 포함된 다양한 주제(topic)를 추론하는 접근이다(Blei 2012).

텍스트마이닝이 내용분석 자동화 도구로 사용되고 있지만, 사회조사 방법론으로서의 내용분석으로서는 충분한 역할을 하지는 못하고 있다. 비지도학습 방식의 토픽모델링에 대한 편향 때문이다. 비지도 토픽모델링은 자료기반 접근 도구로서 문서에 포함됐을 잠재적 주제를 귀납적으로 탐색하는 도구로는 유용하나 분석과정에 이론적인 틀을 적용할 수 없는 한계가 있다. 비지도학습에 대한 편향은 연구자의

편의가 작용한 면도 있다. 지도학습(supervised learning)을 이용하면 이론을 적용한 내용분석이 가능하지만 비용이 많이 든다. 대량의 학습자료로 알고리듬을 훈련시켜야 하기 때문이다.

지도학습의 대안이 준지도학습(semi-supervised learning)이다. 모형구성 및 분류에 이론을 적용할 수 있을 뿐 아니라 비지도학습의 원리와 지도학습의 원리를 통합하여 학습자료를 훈련하는 데 필요한 비용을 크게 줄일 수 있기 때문이다. 다양한 비지도학습 알고리듬이 소개되어 있는데, 그중 하나가 씨앗사전을 이용하여 학습자료를 비지도방식으로 구성하는 접근이다(예: Lu et al. 2011; Watanabe & Zhou 2022).

이 연구는 준지도학습 방식의 토픽모델링을 소개하고, 이를 비지도학습 및 연결망분석과 연계하여 이론기반 텍스트마이닝의 방법과 절차를 제시한 다음, 이를 감염병보도 내용분석에 적용하고자 한다. 이를 통해 기존의 텍스트마이닝 연구가 탐색적 수준의 기술적 분석에만 그쳤던 한계를 극복하여, 텍스트마이닝을 이론기반과 자료기반을 통합하는 온전한 사회조사 방법론으로서의 가능성을 제시하고자 한다.

Ⅱ. 내용분석과 텍스트마이닝

1. 내용분석

내용분석은 설문조사와 함께 추상적인 개념을 측정가능하도록 양화하는 사회조사의 핵심 방법 중 하나다. 사회학, 심리학, 매체학, 정치학, 인류학 등 다양한 분야에서 활용되고 있다.

내용분석은 크게 이론기반 내용분석과 자료기반 내용분석으로 구분할 수 있다. 이론기반 내용분석은 개념화를 통해 구체적인 연구질문 혹은 가설을 설정하고, 개념을 측정 가능하게 조작화한 다음, 자료를 수집하고 측정하여 분석하는 경험과학 연구를 위한 기법이다. 따라서 이론기반 내용분석은 전형적으로 개념화 → 조작화(분석유목 및 분석단위 설정) → 측정(자료 수집 및 코딩) → 분석 등의 절차를 거친다(Krippendorf 2003; Neuendorf 2002).

이론기반 내용분석을 이용한 연구설계는 변수를 설정하는 방식에 따라 크게 3종

으로 구분할 수 있다(Krippendorf 2004). 첫째, 서로 다른 복수의 텍스트를 내용분석하여 변수를 측정하는 방법이다. 특정 사건의 발생 이전, 발생 시점, 발생 이후 등으로 구분하여 비교할 수 있다. 분석 대상이 되는 텍스트를 구분하여 매체A와 매체B의 내용분석 결과를 비교할 수 있다.

둘째, 복수의 내용분석을 수행하여 변수를 측정하는 방법이다. 텍스트에 등장하는 다양한 주제(예: 환경, 건강, 고용, 긍정성, 부정성 등)를 변수로 구성하여 이 변수들 사이의 관련성에 대한 가설검정을 수행할 수 있다.

셋째, 텍스트의 내용분석과 별도의 관측결과(예: 설문조사) 사이의 관계에 대한 가설검정이다. 언론보도의 내용분석을 통해 구성한 변수와 같은 시기의 여론조사를 통해 구성한 변수 사이의 관련성에 대한 가설검정을 수행하는 경우다.

자료기반 내용분석은 귀납적 연구방법으로서 이론이나 개념에 대한 전제 없이 자료에서 떠오르는 주제를 포착하는 연구기법이다. 질적 내용분석이라고도 한다. 자료기반 내용분석의 절차로서 널리 이용되는 방법이 근거이론(grounded theory)이다(Glaser & Strauss 1967). 자료에 근거(grounded)하여 일정한 패턴을 발견하여 이론을 도출하는 일련의 귀납적 추론 절차다. 자료기반 내용분석은 전형적으로 패턴 발견 - 정교화 - 확인 등의 과정을 여러 차례 반복하는 절차를 거친다(권향원 2016).

내용분석은 측정방법에 따라 인간코딩과 컴퓨터코딩으로 구분할 수 있다. 코딩의 모든 작업이 온전히 사람에 의해 이뤄지는지 혹은 컴퓨터에 의해 이뤄지는지의 차이다. 컴퓨터코딩에 의한 내용분석은 컴퓨터를 이용한 내용분석이기에 기계보조 내용분석(machine-assisted content analysis), 컴퓨터보조내용분석(computer-assisted content cnalysis), 혹은 전산 내용분석(computational content analysis)이라고 한다. 인간코딩과 컴퓨터코딩 모두 사전에 정한 일련의 절차와 방법에 따라 진행된다. 다만, 인간코딩은 코딩과정에 인간의 주관이 개입하기 때문에 이론기반 내용분석의 경우, 2인 이상의 코더가 코드북과 코딩양식을 숙지한 후 시험적으로 코딩하여 코딩결과가 일관되게 산출되는지 확인하는 과정이 필요하다. 자료기반 내용분석에서는 2인 이상의 코더 사이의 일치도를 구하기보다 1인의 코더가 코딩 - 해석 - 재코딩 - 재해석 등의 과정을 반복함으로써 코딩의 신뢰성을 확보한다. 컴퓨터코딩은 인간코딩과 달리 일관된 결과를 산출하므로 신뢰도를 확인하는 절차는 필요없지만 소프트웨어가 분석목적에 맞는 결과를 산출하는지에 대한 확인이 필요하다.

2. 텍스트마이닝

텍스트마이닝은 비구조적인 텍스트자료에서 컴퓨터의 연산과 알고리듬(algorithm)을 동원하여 정보를 추출하는 텍스트분석의 한 방법이다(Hotho · Nürnberger · Paaß 2005). 알고리듬은 '문제를 푸는 방도'로서 '문제해결을 위해 수학적으로 표현된 단계적 절차들의 형식적 과정이나 그 집합'이라 정의할 수 있다. 컴퓨터 연산의 맥락에서는 기계가 수행해야 할 특정 과제를 순서대로 알려주는 구체적인 지시의 집합이라 할 수 있다. 반면 알고리즘(algorism)은 아랍식 기수법이란 의미다(이재현 2019; Striphas 2015).

텍스트마이닝이 컴퓨터보조내용분석이므로 텍스트마이닝도 이론기반 텍스트마이닝과 자료기반 텍스트마이닝으로 구분할 수 있다. 알고리듬의 적용방식에 따라 크게 규칙기반(rule-based) 접근과 기계학습(machine learning) 접근으로 구분할 수 있다. 규칙기반 접근은 특정 요소(예: 단어 빈도 혹은 연결구조)에 대한 연산 알고리듬을 사람이 구체적으로 지정하는 반면, 기계학습 접근은 입력자료와 산출자료를 통해 기계로 하여금 연산 알고리듬을 만들어 문제를 해결하는 방법이다. 두 기법 모두 이론기반과 자료기반 접근으로 구분할 수 있다.

1) 자료기반 텍스트마이닝

(1) 규칙기반

자료기반으로 단어의 빈도를 계산하기 위해서는 사전 없이 단어의 빈도를 계산해야 한다. 사전 없이 문서를 분류하려면 상대빈도를 계산해야 한다. 개별 문서의 상대적인 빈도를 계산하는 데 널리 사용되는 알고리듬이 TF-IDF(Term Frequency-Inverse Document Frequency)다. Term은 텍스트 최소 단위인 단어이고, Document는 텍스트를 구성하는 문서로서 내용분석의 분석단위가 된다. 문서는 분석 목적에 따라 기사 한 꼭지가 될 수도 있고, 소설의 한 장이나 절이 될 수 있다. 예를 들어, 개별 문장을 분석단위로 설정한다면 한 문장이 문서가 된다. 문서의 집합단위인 텍스트는 말뭉치(corpus)라고 한다. 예를 들어, 신문 말뭉치는 여러 신문의 기사 모음집이다. 말뭉치는 상대적인 개념이다. 소설 문집이 말뭉치라면 개별 소설이 문서가 된다. 소설 한편이 말뭉치라면 소설의 각 장 또는 각 페이지가 문서가 된다. TF-IDF는 개별문서에 등장하는 단어의 빈도(TF: Term Frequency)와 말뭉치 전반

에 등장하는 단어의 빈도를 역산(IDF: Inverse Document Frequency)한다. 이를 통해 말뭉치 전반에 두루 등장하는 단어의 가중치는 줄이고, 개별 문서에 고유하게 등장하는 단어의 가중치를 높일 수 있어 주제어를 식별할 수 있다(예: 유은순·최건희·김승훈 2015; Qaiser & Ali 2018).

연결망분석도 규칙기반으로 텍스트의 주제를 식별하는 방법으로 널리 사용된다. 연결망 분석을 통해 텍스트에 포함된 단어(node)의 등장빈도와 단어와 단어 사이의 연결(edge) 정도 및 구조를 파악할 수 있다. 각 노드의 중심성을 계산하여 군집을 찾을 수 있어 텍스트에 잠재된 다양한 주제를 도출할 수 있다(Doerfel 2018).

(2) 기계학습: 비지도학습

기계학습(machine learning)에는 다양한 방식이 있으나 크게 지도학습과 비지도학습으로 구분할 수 있다. 비지도학습은 학습자료 없이 기계 스스로 자료에서 규칙성을 찾아 유사한 내용끼리 군집하는 분석방법이다. 학습자료의 투입 없이 분류하는 비지도학습이 자료기반 텍스트마이닝에 해당한다. 널리 사용되는 비지도텍스트마이닝이 LDA(Latent Dirichlet Allocation) 토픽모델링이다. 말뭉치의 문서와 단어의 확률분포를 계산한 단어의 군집을 통해 문서 안에 잠재된 주제를 추론하는 방법이다. 문서의 주제(topic)를 추론하기에 토픽모델(topic models) 혹은 토픽모델링(topic modeling)이라고 한다. 주제모형 혹은 주제모형분석이라고도 한다(Blei 2012; Blei et al. 2003).

LDA토픽모델링은 문서를 문법 등에 의해 규정되는 짜임새 있는 구조가 아니라 주머니에 섞여 있는 혼합물로 본다(bag of words). 혼합물이지만 무작위 혼합이 아니라 유사한 단어들끼리 확률적으로 함께 모여 있는 군집의 합이다. 즉, 토픽모델링은 문서를 잠재된 다양한 주제의 혼합물로 파악한다. 문서는 여러 주제가 섞여 있는 혼합물이고, 문서마다 주제(예: 예술, 교육, 예산 등)의 분포 비율이 상이하며, 주제(예: 예술)마다 단어(예: 오페라, 교향악단)의 분포가 상이하다. 따라서 LDA토픽모델링은 각 문서가 해당 주제에 속할 확률값과 각 단어가 해당 주제에 속할 확률값을 산출한다(Blei 2012).

비지도 토픽모델링은 텍스트의 내용에 대한 전제 없이 주제를 귀납적으로 추론하기에 질적 내용분석에 해당한다(Nelson 2020). Nelson(2020)은 근거이론의 절차(패턴감지 - 패턴정제 - 패턴확인)를 비지도 토픽모델링에 적용해 전산근거이론(computational grounded theory)이라는 자료기반 텍스트마이닝의 틀을 제시했다. 1단계인 패턴감

지 단계는 귀납적인 텍스트의 탐색 단계로서 비지도 기계학습을 이용해 텍스트자료에서 다수의 주제를 가능한 많이(예: 20~100개) 추출한다. 2단계인 패턴 정제는 해석적 탐색 단계로서 추출한 주제에 해당하는 개별 문서에 대하여 인간의 전문지식을 적용해 질적으로 탐색한다. 이 과정에서 주제의 수를 줄여나간다. 3단계인 패턴 확인 단계는 확보한 주제에 대한 확정 단계로서 지도학습 혹은 사전을 이용하여 2단계에서 해석적으로 탐색한 주제를 확정한다.

사회과학과 디지털인문학 연구자들의 토픽모델링을 이용한 연구는 전형적으로 LDA분석을 수행하여 주제의 수를 선정하고, 각 주제별로 출현 확률이 높은 단어와 대표적인 문서를 선정한 다음 대표문서에 대하여 심층적인 분석을 진행하여 해석가능한 주제를 귀납적으로 선정하는 절차를 거친다(예: 양혜진·안정민·이태언 2001). 이는 전산근거이론의 2단계까지 적용한 질적 내용분석에 해당한다. 결국 넬슨(2020)의 관점에서 보면 토픽모델링을 이용한 다수의 내용분석 연구는 미완의 질적 내용분석인 셈이다.

2) 이론기반 텍스트마이닝

(1) 규칙기반

사전을 이용한 단어빈도 계산은 규칙기반 접근으로 오래전부터 널리 사용되는 이론기반 텍스트마이닝 기법이다(Stone et al. 1966). 사전은 미리 지정한 범주에 해당하는 주제어의 집합이다. 이론 혹은 전문지식에 의해 설정한 개념에 대하여 사전에 포함된 단어를 통해 조작함으로써 텍스트에 등장하는 사전 단어의 빈도를 계산하여 문서의 주제를 추론할 수 있다. 예를 들어, 감정이론은 크게 감정을 차원감정과 개별감정으로 구분한다(Barrett 1998). 감정분석에 이용하는 감정사전은 차원감정 혹은 개별감정이론에 따라 감정어의 분류를 달리한다. 차원감정이론(dimensional emotion theory)을 적용하는 경우 감정을 긍정과 부정의 축에 놓여진 연속선상의 속성으로 파악하므로 감정어를 긍정과 부정의 정도에 대하여 점수로 부여한다. 반면 개별감정이론(discrete emotion theory)을 적용하는 경우, 다양한 개별감정(분노, 공포, 기쁨 등)을 나타내는 단어를 선별하여 각 감정에 해당하는 라벨을 부여한다.

사전방식이 제대로 작동하기 위해서는 사전에 포함된 단어의 분류가 특정 맥락에서 사용되는 단어의 용법과 잘 맞아야 한다. 즉, 사전을 이용한 단어빈도 계산 접근방법은 맥락의존적이어서 사전이 제작된 바깥 영역에 적용할 경우 심각한 오류 가능성이 있다. 같은 단어라도 영역에 따라 긍정적 의미로 사용될 수도 있고 부정

적으로 사용될 수도 있기 때문이다. 예를 들어, '세금' '비용' '암' 등과 같은 단어는 통상적으로 부정적 함의를 지닌 것으로 여겨지나 회계관련 혹은 건강관리 기업의 실적 보고서에서는 긍정적인 함의를 지닌 단어가 될 수 있다(Grimmer & Steward 2013).

(2) 기계학습: 지도학습

지도학습은 미리 범주에 따라 분류한 학습자료를 이용해 기계가 학습자료를 통해 생성한 알고리듬으로 분류하거나 예측하는 분석방법이다. 범주에 따라 미리 설정한 기준에 따라 텍스트를 분류하는 지도학습이 이론기반 텍스트마이닝에 해당한다. 지도학습을 위한 다양한 알고리듬이 소개돼 있는데 크게 기호주의, 연결주의, 진화주의, 베이즈주의, 유추주의 등 5종의 접근으로 나눌 수 있다(Domingos 2015).

지도학습 방식의 이론기반 텍스트마이닝으로 수행한 사례 중 하나가 8천건의 공공문서를 분석해 러시아의 정계와 군부 지도층이 취한 외교정책의 차이를 비교한 연구다(Stewart & Zhukov 2009). 이 연구는 정계와 군부 지도층 사이의 쟁점 현저성, 무력사용선호도 및 견해수렴 정도에 대한 연구가설을 제시한 뒤, 8천 건의 문서 중 300건을 무선추출하여 능동, 보수, 중립 등 3개 분석유목을 설정하고 내용분석의 절차에 따라 문서를 분류하여 학습자료를 구축했다. 이 300건의 학습자료로 4종의 지도학습 알고리듬(랜덤포리스트, SVM 등)을 혼합하여 훈련시킨 뒤 나머지 7,500건의 문서에 대한 분류작업을 수행했다.

결국, 지도학습 방식의 텍스트마이닝은 내용분석의 과정을 온전하게 컴퓨터로 자동화했다기보다, 인간코딩(학습자료 분류)에 컴퓨터코딩(예측자료 분류)을 추가한 것이라 할 수 있다. 분류대상의 양이 많으면 그만큼 학습자료의 양도 함께 늘려야 한다. 지난 20년간 국내에서 보도된 한미동맹 관련 기사 5만여 건을 자주와 동맹의 틀로 분석한 연구에서는 사설 1,534건을 외교-국방분야 전문가가 수작업으로 내용분석하여 자주와 동맹으로 분류하여 학습자료를 구성한뒤, 기사 5만여 건에 대한 분류작업을 수행했다(정재철·이종혁 2020).

이론기반 텍스트마이닝에서 규칙기반과 기계학습 방식의 선택 기준은 분석대상의 양에 있다. 기계학습은 분석대상의 양이 충분히 많아야 분류 정확성을 담보할 수 있다. 즉, 지도학습 방식의 텍스트마이닝은 분석대상의 양이 많은 경우에만 특정 단어의 빈도를 계산하는 사전방식에 비해 정확성이 더 높고 맥락에 덜 민감하다(Grimmer & Steward 2013). 따라서 분석대상의 양이 많지 않다면 기계학습 방식보다는 규칙기반의 텍스트마이닝이 더 적절하다.

(3) 기계학습: 준지도학습

준지도학습은 분류된 자료와 분류되지 않은 자료를 혼합하여 모형을 훈련시키는 기계학습이다. 지도학습은 모든 예제를 분류하여 학습자료로 구축한 뒤 알고리듬을 훈련시키는 반면 비지도학습은 분류되지 않은 자료로 알고리듬을 훈련시킨다. 준지도학습 알고리듬은 미리 분류한 학습자료를 제한적으로 투입하여 알고리듬을 훈련시킨다. 제한된 양의 예제를 학습자료로 투입하지만 지도학습에 버금가는 정확도를 달성할 수 있다. 다만, 준지도학습이 지도학습만큼의 성능을 내기 위해서는 근사치(smoothness)와 군집(cluster)의 전제가 충족돼야 한다. 즉, 예제 자료에 일정한 구조가 있어 이 구조를 활용해 예측을 할 수 있어야 한다. 예를 들어, 두 표본 x1과 x2가 인접한 입력공간에 있다면, 두 표본의 라벨 y1과 y2에도 동일하게 적용할 수 있어야 한다(근사치 전제). 또한 한 자료의 특정 값을 중심으로 동일한 군집을 이룬다면, 그 특정 값들은 서로 유사한 특성이 있어야 한다(군집 전제). 텍스트의 문서, 문장, 단어는 일정한 구조를 이루기 때문에 텍스트분석은 준지도학습의 두 전제를 충족한다(Chapelle et al. 2006).

LDA토픽모델링에 적용한 준지도학습으로는 씨뿌린 LDA(seeded LDA)가 있다. 모형에 투입할 주제어(씨앗단어: seed words)를 미리 설정하여 씨앗단어를 중심으로 군집하는 방법이다. 단어에 대한 주제의 사전분포(prior distribution)에 가중치를 부여하여 특정 주제를 언급하는 문장을 탐지한다. 즉, 이론기반으로 선정한 단어를 씨앗으로 삼아 각 문장에 대하여 비지도학습기법으로 레이블링하는 셈이다. 예를 들어, 음식 리뷰 텍스트의 주제로서 음식과 서비스가 있다고 할 때 음식 주제에 대한 씨앗단어는 '음식, 치킨, 소고기, 스테이크', 서비스 주제의 씨앗단어는 '서비스, 직원, 웨이터, 예약' 등으로 설정된 씨앗사전을 구성할 수 있다. '돼지갈비는 맛있는 음식' 같은 문장은 음식 주제의 씨앗단어를 통해 음식 주제로 레이블되고, "직원들이 친철하고 서비스도 좋아"와 같은 문장은 서비스 주제로 레이블된다. 이렇게 각 주제별로 레이블된 문장이 학습자료로서 문서와 단어의 주제분포를 계산하는 데 이용된다(Lu et al. 2011). 뉴스텍스트를 대상으로 한 주제분류 과제에서 비지도 LDA에 비해 씨앗사전을 적용한 LDA의 분류정확도가 높았다 (Jagarlamund et al. 2012).

씨앗사전을 구성할 때는 이론을 적용하여 선정한 단어를 이용하는데 크게 세가지 방식으로 나눠 접근할 수 있다. 첫째, 이론과 지식을 이용하여 새롭게 구성하는 방식이다. 설문조사에서 척도를 새롭게 구성하는 것에 해당한다. UN총회의 어록을 분석한 연구에서는 LDA기반 분류에서는 말뭉치에서 추출한 고빈도 단어를 이론기

반으로 재구성할 때 씨앗단어 사전 성능이 더 우수했다고 보고했다(Watanabe & Zhou 2022). 둘째, 비지도 토픽모델링으로 씨앗사전에 포함될 후보군을 산출하여 이론과 지식으로 재분류한 다음 씨앗사전을 구성하는 방식이다. 비지도학습으로 다수의 주제(예: 40개)를 설정하여 확보한 후 이론기반으로 설정한 개념의 범주에 포함되는 단어를 선별하는 방식이다. 비지도 토픽모델링으로 구성한 단어를 이용하기에 씨앗사전 구성이 첫째 방식에 비해 용이하다는 장점이 있다. 셋째, 기존에 개발된 사전을 이용하는 방식이다. 설문조사에서 기존에 개발된 척도를 사용하는 것에 해당한다. 예를 들어, 전망이론에 따라 문서의 프레임을 긍정과 부정으로 분류한다면(Kahneman & Tversky 1979), 감정을 공포, 분노, 기쁨 등으로 범주화한 개별감정이론과 달리 감정을 긍정-부정 등의 복수의 벡터공간에 배열하는 차원감정이론을 적용한 감정사전에서 감정어를 추출하여 씨앗사전을 구성할 수 있다.

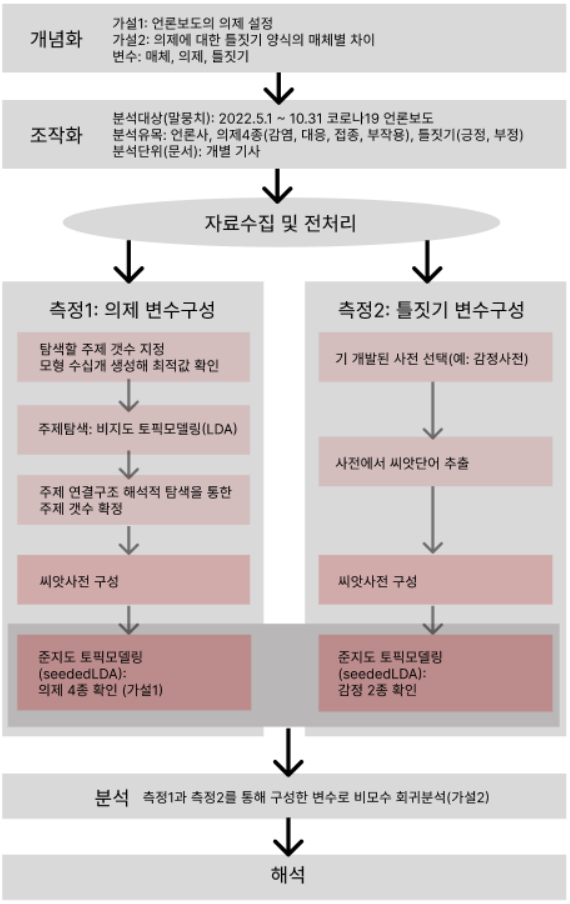
LDA 외에도 NMF(Non-negative Matrix Factorization), Top2Vec, BERTopic 등 다양한 토픽모델링 기법이 소개돼 있다. LDA가 생성적 확률모형을 사용하는 반면, NMF는 선형대수접근을 이용한다. Top2Vec과 BERTopic은 단어를 벡터공간에 배치하는 워드임베딩(word embedding)을 이용한다. 각 기법별로 장·단점이 있다. LDA는 워드임베딩에 비해 적은 수의 주제를 산출해 해석이 용이하다. 한 문서에 복수의 주제를 담고 있다고 전제하기 때문에 긴 문서로 이뤄진 텍스트 분석에 적당하다. 문서와 단어에 대한 각 주제별 확률분포 값을 제공하기 때문에 산출된 주제를 연속형 변수로 사용할 수 있다. NMF는 LDA에 비해 연산효율이 좋아 대규모 자료 처리에 유리하다. Top2Vec과 BERTopic은 워드임베딩이기 때문에 형태소 분석과 같은 전처리작업이 불필요하고 다국어분석이 가능하다. 문서당 하나의 주제만을 할당하기 때문에 짧은 문서로 이뤄진 텍스트 분석에 적당하다. 긴 문서로 이뤄진 텍스트를 분석하기 위해서는 문장 단위로 분리해야 한다(Egger & Yu 2022).

Ⅲ. 이론기반 텍스트마이닝의 방법과 절차

이론기반 텍스트마이닝은 내용분석의 방법과 절차와 유사하다. 즉, 개념화 → 조작화 → 측정 → 분석 → 해석 등의 절차를 거친다. 연구설계 역시 내용분석의 연구

설계와 다르지 않다. 이 연구에서는 서로 다른 복수의 텍스트를 내용분석하여 유사한 현상을 비교하는 방법과 텍스트에 대하여 복수의 내용분석을 수행한 결과 사이의 관계에 대한 가설을 검정하는 두 방법을 혼용하고자 한다.

이를 위해 텍스트를 4종의 언론매체(복수의 텍스트)에서 수집하여 감염병 창궐(유사한 현상)에 대해 어떻게 다뤘는지 비교하고, 해결지향보도의 이론을 적용하여 의제(측정 1) 및 틀짓기 양식(측정 2)에 대한 내용분석(복수의 변수 구성)을 수행하여 언론매체, 보도의제 및 틀짓기 3종의 변수의 관계에 대한 가설을 검정하고자 한다. 따라서 변수는 언론매체(복수의 텍스트), 의제(측정 1), 틀짓기(측정 2) 등 3종이 된다. <그림 1>에 절차를 요약했다.



<그림 1> 이론기반 텍스트마이닝의 절차

1. 문제의식과 이론적 틀

코로나19 바이러스 창궐로 인한 위기 상황에서 언론매체는 문제해결에 제대로 기여했는지에 대한 판단근거를 마련하고자 한다. 이를 위해 해결지향보도(solutions journalism), 의제설정이론(agenda setting theory), 및 틀짓기이론(framing theory)을 적용한다. 해결지향보도는 '문제에 대한 엄격한 대응'으로서 문제 자체를 지적하는 데 그치지 않고 그 문제에 대하여 '어떻게 대응하는지'까지 다루는 보도다(McIntyre & Lough 2021). 즉, 언론이 공중으로 하여금 해결책을 찾아 집행할 수 있는 공론을 형성함으로써 사회문제에 대한 가능한 해법을 탐색하도록 하는 보도양식이다(Lough & McIntyer 2019).

언론보도는 크게 '무엇을' 보도하는가와 '어떻게' 보도하는가로 구분하여 분석할 수 있다. 전자에 대한 이론으로 의제설정이론(agenda setting theory)이 있고, 후자에 대한 이론으로는 틀짓기이론(framing theory)이 있다. 의제설정이론에 따르면 언론은 공간과 시간의 제약 그리고 매체의 성향에 의해 특정 주제에 대하여 차별적으로 다룰 수밖에 없으며 언론이 차별적으로 강조하는 내용이 공중의 의제로 전이된다(반현·McComb 2007). 틀짓기이론에 따르면, 동일한 내용에 대해서도 뉴스의 구성과 해석의 틀(frame)에 따라 수용자는 동일한 의제에 대해서도 달리 지각하게 된다. 긍정 혹은 부정적 측면의 강조, 의혹제기, 책임지우기 등 함축된 가치를 달리하여 뉴스 해석의 틀을 구성한다(이준웅 2000).

2. 가설 설정

의제란 한 사회가 주의를 기울여야 하는 광범위한 주제로서 쟁점이 될 수 있는 공적 사안이라 할 수 있다. 그 주제의 광범위성은 층위에 따라 규정되는 상대적인 특성이 있다. 생존을 상위 범주로 설정하면 그 하위 범주인 위기 혹은 기회를 의제로 볼 수 있고, 혹은 위기를 상위 범주로 설정한다면 전염병위기, 경제위기, 전쟁위기 등이 의제가 될 수 있다. 이 연구에서는 전염병창궐에 주목하여 코로나19를 상위 범주로 설정하여 감염, 백신 등의 하위 주제를 의제로서 탐구하고자 한다. 전염병창궐 상황에서 하위 주제를 분류하기 위해서는 다양한 접근을 취할 수 있는데, 그중 하나가 해결지향 접근이다.

해결지향보도는 문제 자체의 지적뿐 아니라 그 문제에 대한 대응을 다루는 보도 양식이다. 따라서, 전염병창궐 상황에서 언론이 설정하게 되는 의제는 크게 문제지적과 문제대응으로 구분하여 접근할 수 있다. 보다 구체적으로는 문제의 지적에 해당하는 의제는 병원균의 감염 및 감염에 대한 다양한 대응의 부작용(백신부작용) 등이 있다. 문제의 대응에 해당하는 의제는 환자를 수용하고 치료해야 하는 의료자원의 할당과 감염에 대응하기 위한 백신접종 등이 있다. 즉, 감염, 대응, 백신접종, 백신부작용은 코로나19창궐이라는 상위 범주에서 언론매체가 선택적으로 강조할 수 있는 의제가 된다. 언론매체가 감염을 집중적으로 보도한다면 공중은 감염을 사회적 의제로 지각할 가능성이 증대하고, 백신부작용에 대한 보도빈도를 늘린다면 백신부작용이 사회적 의제로 부각하게 될 가능성이 증대한다.

가설 1: 코로나19 창궐상황에서 언론보도의 주요 의제는 감염, 대응, 백신부작용, 백신접종 등의 주제로 구분될 것이다.

틀짓기이론에 따르면 유사한 내용에 대하여 보도하는 방식에 따라 여론형성이 달리 작동한다. 틀을 짓는 방식에는 전망이론, 커뮤니케이션효과론적 접근 등 다양한 이론이 소개돼 있다(이준웅 2000). 커뮤니케이션효과론적 연구에서는 메시지가 조직되는 방식으로 인간적 흥미, 갈등, 도덕가치와 원칙, 혹은 경제적 이해관계 등의 프레임을 제시했다(예: Price & Tewksbury 1997). 전망이론(prospect theory)에 따르면 프레임은 예상되는 이득 또는 손실을 강조하는 틀로서 긍정 프레임과 부정 프레임이 제시돼 있다(Kahneman & Tversky 1979). 동일한 의제에 대하여 긍정 혹은 부정 프레임을 차별적으로 적용하는 기제는 다양하나 개인 혹은 집단의 부정편향 혹은 낙관편향의 차이에서 비롯된다. 언론매체는 정치적, 사회적, 혹은 경제적 성향별로 고유한 특색을 지니기 때문에 언론매체의 성향에 따라 동일한 주제에 대하여 긍정적으로 혹은 부정적으로 틀지어 보도할 수 있다.

가설 2: 언론매체의 성향에 따라 동일한 주제에 대하여 긍정 혹은 부정의 틀 적용이 다를 것이다.

3. 연구방법

국내 주요 언론에서 코로나19 상황에 대한 위험과 기회를 어느 정도로 실재적인

측면을 반영하는지 가늠하기 위해, 코로나19 상황의 문제와 문제의 대응을 반영하는 주제어를 구성하고, 이 주제어를 통해 준지도학습 토픽모델링을 수행한다. 또한 기 개발된 감정사전에서 추출한 긍정어와 부정어를 사용하여 준지도학습 토픽모델링으로 문서의 감정을 분류한 뒤, 2종의 변수(의제와 틀짓기) 및 매체 사이의 관계에 대하여 분석한다.

1) 사용 도구

모든 분석은 R(4.22)을 이용했다(R Core Team 2022). R 기본함수 및 tidyverse (자료 전처리), quanteda(텍스트자료 전처리), seededlda(비지도 및 준지도 토픽모델링), igraph(연결망분석), ggplot2(시각화) 등의 패키지를 이용하여 자료 수집 및 전처리 →측정(비지도 토픽모델링 및 연결망분석) →준지도 토픽모델링 및 비모수회귀분석 등을 수행했다.

2) 자료 수집 및 전처리

코로나19 백신이 상용보급되던 초기인 2021년 5월1일부터 10월31일 사이의 코로나19 관련기사를 한국언론재단의 빅카인즈를 통해 수집했다. 분석방법과 절차에 대한 사례연구이기에 자료 수집 범위를 특정 시기로 제약했다. 검색어는 빅카인즈에서 코로나19 관련 어휘로 제시한 “((코로나19) OR (코로나) OR (코로나 바이러스) OR (신종 코로나바이러스) OR (COVID-19) OR (코비드19))”를 이용했다. 기사 분류는 사회분류에서 의료건강으로 국한했다. 인사, 부고, 동정 등으로 중복 분류된 기사는 제외했다. 언론사는 성향이 뚜렷하게 대비되는 경향신문, 조선일보, 중앙일보 및 한겨레신문을 선정했다. 중복기사 및 영문기사를 제외한 4,180건의 기사를 분석대상으로 삼았다.

빅카인즈는 저작권 문제로 기사 본문은 200자만 제공하지만, 키워드는 기사 전문에서 추출하여 제공한다. 이 키워드는 형태소 분석을 통해 추출한 명사에 해당한다. 자료분석을 위해 키워드 열에서 먼저 특수문자, 불용어, 및 문자와 숫자를 제외한 모든 기호를 제거했다. 또한 다양하게 사용되는 용어를 통일했다(예: 부스터샷, 부스터샷 접종 등→추가 접종). ‘코로나’ 키워드로 검색한 기사이므로 키워드에서 ‘코로나’도 제거했다. 최종적으로 4,180건의 기사에서 796,668개의 키워드를 분석대상으로 삼았다. 각 기사에 포함된 키워드의 개수는 평균 246.6개(SD=146.25)다. 중간값은 218개다. 오른쪽 꼬리가 왼쪽 꼬리보다 긴 분포다(skewness=2.6). 첨도는 10.96

으로 중심값에 대한 집중도가 높다. 가장 짧은 기사는 11개의 키워드로 이뤄졌고, 가장 긴 기사는 1,295개의 키워드로 구성됐다. 유니그램(단일 키워드) 최빈도 키워드는 접종(37,152), 백신(29,470), 확진자(7,302) 순이었다. 바이그램(연속하는 두 키워드) 최빈도 키워드는 백신 접종(10,385), 예방 접종(2,620), 접종 백신(1,938) 순이었다. 유니그램과 바이그램 키워드의 빈도는 <그림2>에 단어구름으로 시각화했다.

<그림 2> 유니그램(좌) 및 바이그램(우) 키워드 빈도 단어구름

3) 측정 변수

(1) 언론매체

성향이 대체로 분명하게 나뉘는 언론사 4곳을 선정했다. 경향신문과 한겨레신문을 하나로 묶고, 조선일보와 중앙일보를 하나로 묶었다. 일반적으로 언론매체를 보수와 진보로 구분하지만, 이 연구에서는 보수 혹은 진보라는 용어를 사용하지 않았다. 정치적으로는 진보지와 보수지라기보다 여당지와 야당지의 측면이 더 강하고, 사회적 쟁점과 경제적 쟁점에 대해서는 진보적 성격과 보수적 성격이 혼재돼 있기 때문이다.

(2) 의제

해결지향보도의 논리를 적용해 언론보도의 의제로서 문제지적과 문제대응을 나눠, 문제지적은 감염과 백신부작용으로 설정했고, 문제대응은 병상확보나 거리두기와 같은 대응 및 백신접종으로 설정했다. 먼저 비지도 토픽모델링으로 주제 수십개를 탐색적으로 산출한 다음, 연결망분석과 해석적 탐색을 통해 가설에서 설정한 주제로 축소될 수 있는지 탐색했다. 축소된 주제에 따라 씨앗사전을 구성해 준지도 토픽모델링을 수행했다. 준지도 토픽모델링을 통해 산출되는 각 문서별 해당 주제에 속할 확률값이 각 변수의 측정값이 된다.

(3) 틀짓기

틀짓기 양식은 전망이론에 따라 긍정과 부정의 틀로 구분했다. 이를 위해 기개발된 한국어감정사전(온병원 등 2018)에서 긍정어와 부정어를 추출하여 긍정어 및 부정어 씨앗사전을 구성했다. 이 사전으로 준지도 토픽모델링을 수행해 긍정적인 내용과 부정적인 내용의 기사를 분류했다. 준지도 토픽모델링을 통해 산출되는 각 문서별 해당 주제에 속할 확률값이 각 변수의 측정값이 된다.

4. 측정 및 분석

1) 의제

가설 1. '코로나19 창궐상황에서 언론보도의 주요 의제는 감염, 대응, 백신부작용, 백신접종 등의 주제로 구분될 것이다'를 검정하기 위해 먼저 비지도 토픽모델링으로 주제를 탐색적으로 선별하고 해석적 분석을 수행한 후 준지도 토픽모델링을 통해 주제를 확정했다.

(1) 탐색할 주제의 개수 지정

단어-문서 행렬을 구성한 다음, 자료에서 추출할 수 있는 적절한 주제가 몇 개일 때 최적의 개수가 될지 탐색했다. 단어-문서 행렬은 단어를 행에 배치하고, 문서를 열에 배치할 행렬자료인데, quanteda패키지를 이용해 구성했다. 주제의 최적 개수는 ldatuning 패키지를 이용하여, 가능한 주제의 개수를 1개부터 25개까지 모두 25개의 LDA모형을 Gibbs샘플링으로 계산했다. ldatuning패키지는 4종의 지표를 제공하는데, 2종은 문서의 일관성과 관련된 지표로서 값이 클수록 최적치를 의미하고, 다른 2종은 복잡성과 관련된 지표로서 값이 작을수록 최적치를 의미한다. 분석결과 21개가 복잡성과 일관성을 함께 고려한 최적의 값이었다.

(2) 비지도 토픽모델링으로 주제 탐색

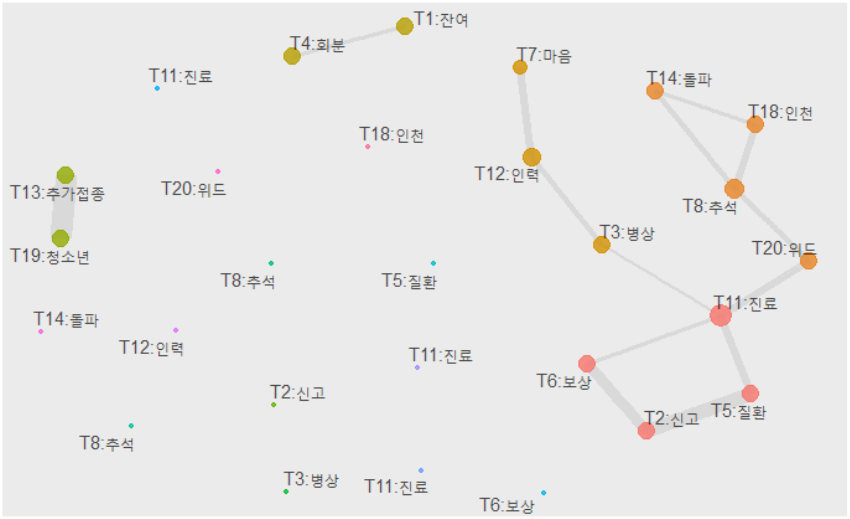
최적값으로 산출된 21개를 주제산출 개수로 설정하여 비지도 토픽모델링을 수행했다. seededlda패키지를 이용했다. seededlda패키지는 각 키워드가 해당 주제에 포함될 확률값을 *phi*값으로 제공한다. 각 주제에 포함된 키워드의 *phi*값을 모두 더하면 1이 된다. topicmodels나 stm패키지의 *beta*에 해당한다. 각 주제별로 *phi*값을 기준으로 내림차 정렬하면 각 주제별로 속할 확률이 높은 대표단어를 선별할 수 있다.

각 문서가 해당 주제별로 포함될 확률값은 *theta*값으로 제공한다. topicmodels나 stm패키지의 *gamma*에 해당한다. 각 주제에 포함된 기사(문서)의 *theta*값을 모두 더

하면 1이 된다. *theta*값을 기준으로 내림차 정렬하면 각 주제별로 속할 확률이 높은 대표문서를 선별할 수 있다.

(3) 주제 연결 구조 탐색

탐색적으로 산출한 주제를 가설을 통해 설정한 변수(감염, 대응, 백신접종, 백신 부작용)에 따라 분류하는 단계다. 이를 위해 먼저 tidygraph와 igraph패키지를 이용해 주제에 대한 연결망분석을 했다. 각 주제별로 개별 단어에 *phi*값이 부여돼 있어 주제 사이의 상관분석을 할 수 있다. 피어슨 상관계수가 0.08 이상인 관계에 대하여 연결성을 부여하여 주제의 연결구조를 파악했다(<그림 3>).



<그림 3> 21개 주제의 연결구조

주: 노드의 색이 동일 커뮤니티.
노드의 색은 페이지랭크로 계산한 중심성.
엣지의 굵기는 연결된 관계가 많은 정도.
즉, 단어 사이의 상관계수가 0.08 이상이 많은 정도.

분석결과 5개 커뮤니티를 찾을 수 있다. 이 5개의 커뮤니티에 속한 주제를 임시로 묶어, 각 주제에 속한 기사의 제목과 본문을 확인하여 가설에서 설정한 범주를 적용해 재분류했다. 예를 들어, 연결망분석에서는 T3, T7, T12를 하나의 묶음으로 제시했지만, 기사 제목과 본문을 확인한 결과, T12는 '인력 간호사 공공 노조 근무 확충 파업(보건의료노조 총파업 D-1, 최후 교섭 시작 결렬시 내일 오전 파업 돌입)'

등에 대한 내용으로 가설에서 설정한 변수(감염, 대응, 접종, 부작용)와는 관련이 없는 내용이었다. T3은 '병상 재택 전담 재택치료 경증 생활 생활치료센터(재택치료 때, 접종완료 동거인은 같은 집 격리가 가능하다)', T7은 '마음 아이 자신 남편 간호사 걱정 아들(외로운 치매 할머니 간호사는 치료 위해 화투 들었다)' 코로나19 대응에 대한 내용이었다. 따라서 T3과 T7은 대응변수에 할당하고, T12는 기타변수로 할당했다.

T9는 연결망분석에서 다른 주제와 관련성이 나타나지 않았지만, 내용을 확인한 결과 '직원 의무 휴가 의무화 거부 기업 직원들(LG그룹 이틀, 삼성전자 SK하이닉스는 하루 '백신휴가')'로서 대응에 해당하는 내용이기에 대응변수에 할당했다.

T20은 연결망 분석에서 '위드 내과 전환 달성 정책 의대 집단면역(오명돈 중앙예방접종센터장이 '집단면역 달성 어렵다'고 말한 이유는)'으로서 감염과 관련된 주제와 관련성이 있는 것으로 나왔지만 의미적인 내용으로는 대응에 해당하여 대응변수에 할당했다.

이와 같은 식으로 연결망분석 결과를 기초로 21개 주제를 가설에서 설정한 감염, 대응, 백신접종, 백신부작용 등 4개 변수에 할당했다. 4개 변수에 할당되지 않은 T5, T6, T11, T12는 기타변수에 할당했다. 5개 변수로 구분한 각 주제별로 대표단어와 대표기사의 제목을 <표 1>에 정리했다. 토픽번호와 함께 대표단어 7개를 표시하고, 같은 열에 대표기사 제목 3건을 제시했다. 각 주제에 속할 확률이 높은 대표기사의 본문은 <첨부 1>에 정리했다. 각 주제에 속한 문서가 완벽하게 해당 주제에 속하지 않는 경우도 있지만, 대체로 기사의 내용과 해당 주제의 내용이 부합한다.

<표 1> 5개 변수별 21개 주제에 속할 확률이 높은 키워드와 기사의 제목

변수	제목	*Theta*
감염	T8: 추석 연휴 반장 중중 확산세 중대본 수습	
감염	코로나 이틀째 2000명대...정부 '정점인지 아직 알 수 없어'	0.83
감염	잇단 연휴에도 확진자 20% 떨어져...정부 '백신 효과 나타나'	0.81
감염	주말효과에도 확진자 2000명대 10월초 황금연휴 또 확산 우려	0.8
감염	T14: 돌파 감염자 돌파감염 요양병원 요양 방대본 추정	
감염	국내 돌파감염 0.03% '접종률 80% 돼도 요양시설 감염 위험'	0.84
감염	접종 후 확진 '돌파감염' 4명 '접종 14일 뒤 감염된 2명 무증상'	0.82
감염	한달만에 부산 휩쓴 '델타 변이'... 요양병원도 돌파감염에 뚫렸다	0.81

감염	T15: 청해부대 국방부 장병 부대 키트 현지 음성	
감염	[그렇군] 서욱 국방장관 대국민 사과에 합참의장 국방차관 '병풍' 선 까닭은	0.9
감염	나라 지키는 그들을 나라는 지켜주지 않았다	0.9
감염	'아덴만 회군' 청해부대 전원 귀국 군 병원, 생활치료센터로	0.9
감염	T18: 인천 외국인 부산 대구 경남 대전 경북	
감염	인천 코로나19 123명 확진 10명 중 2명 외국인	0.85
감염	경기도 역대 2번째 최다 367명 확진자 나왔다	0.83
감염	코로나 신규 확진 681명 이틀 연속 600명대	0.8
대응	T3: 병상 재택 전담 재택치료 경증 생활 생활치료센터	
대응	재택치료 때, 접종완료 동거인은 같은 집 격리 가능하다	0.87
대응	동작구, 단계적 일상회복 전환을 위한 재택치료 체계 본격 추진	0.86
대응	재택치료 대상자 확대에 지자체 방역당국도 '분주'	0.85
대응	T7: 마음 아이 자신 남편 간호사 걱정 아들	
대응	외로운 치매 할머니 간호사는 치료 위해 화투 들었다	0.89
대응	93세 할머니와 29세 간호사 '방호복 화투' 끝내 코로나 이겼다	0.89
대응	"할머니 기운 내셔야죠"... 방호복 화투, 간호사가 제안했다	0.88
대응	T9: 직원 의무 휴가 의무화 거부 기업 직원들	
대응	LG그룹 이틀, 삼성전자 SK하이닉스는 하루 '백신휴가'	0.89
대응	최장 사흘 쉰다 삼성 이어 LG 하이닉스 네이버도 '백신휴가'	0.87
대응	美선 백신 안맞으면 잘리기도 CNN, 미접종 출근 3명 해고	0.85
대응	T20: 위드 내과 전환 달성 정책 의대 집단면역	
대응	오명돈 중앙예방접종센터장이 "집단면역 달성 어렵다"고 말한 이유는	0.79
대응	오명돈 감염병 중앙임상위원장 "집단면역 도달 어려워"	0.78
대응	독감처럼 위드 코로나? '9월말 시행은 희망고문'	0.77
백신 접종	T1: 잔여 잔여백신 보건소 사전예약 안내 의료기관 당일	
백신 접종	27일부터 네이버 카카오서 '잔여백신' 검색시 당일 접종 된다	0.95

백신 접종	‘남는 AZ 백신’ 당일 접종 예약, 27일부터 가능	0.9
백신 접종	남은 백신 ‘당일 접종’ 27일부터 네이버 카카오 앱으로 예약	0.88
백신 접종	T4: 회분 간격 수급 차질 대통령 정은경 상반기	
백신 접종	8월 온다던 모더나 절반도 안온다 접종간격 4주→6주로	0.84
백신 접종	정부, 내년 접종할 화이자 백신 3,000만 회분 구매계약 체결	0.83
백신 접종	5일 화이자 43만 여회분 온다지만	0.78
백신 접종	T10: 임상 항체 교차 시험 허가 치료제 식약처	
백신 접종	SK 바이오, 국산 백신 최초 3상 승인 ‘AZ와 비교 임상 시작’	0.93
백신 접종	식약처, 국내 개발 코로나19 백신 첫 3상 임상시험 승인	0.92
백신 접종	SK바이오 코로나19 백신 ‘3상 허가’ ‘백신 자급 첫걸음’	0.89
백신 접종	T13: 추가접종 이스라엘 cdc fda 승인 성인 식품	
백신 접종	美도 면역 취약층에 부스터 샷 접종 ‘FDA, 48시간 내 승인’	0.88
백신 접종	미 FDA, 고령층 고위험군 국한 부스터샷 사용 승인	0.87
백신 접종	글로벌 백신부족 우려에도 미 FDA, 면역취약층 ‘부스터샷’ 전격 결정	0.85
백신 접종	T16: 식당 인센티브 허용 카페 실외 패스 실내	
백신 접종	사적모임 10~12명, 유흥시설은 자정까지	0.93
백신 접종	미접종자 헬스장 환불? 위드 코로나 달라지는 것들 Q&A	0.89
백신 접종	[Q&A]야구장 접종자 전용구역에 미접종자 자녀가 함께 갈 수 있나요?	0.89
백신 접종	T17: 학생 학교 교사 어린이집 교직원 교육부 수능	

백신 접종	‘화이자 티켓’ 9월 모평, 온라인응시 허용 ‘시험장 추가 확보’	0.83
백신 접종	교육부 ‘9월 모의수능 응시 30 40대도 화이자 접종’	0.83
백신 접종	화이자 ‘약발’? 9월 모의평가 졸업생 신청자 전년대비 3만명 ‘껑충’	0.83
백신 접종	T21: 면제 인도 입국 일본 중국 세계 산소	
백신 접종	격리면제, 백신 맞고 온 브라질 CEO는 ○, 바이어는 ×	0.88
백신 접종	거래처 브라질 직원은 백신 맞아도 격리...알쏭달쏭 격리 면제[Q&A]	0.86
백신 접종	백신 맞았다면 괌 사이판 여름휴가 하와이 몰디브는 ‘음성 확인서’면 통과	0.86
백신 부작용	T2: 신고 혈전증 청원 호소 두통 의심 여성	
백신 부작용	‘AZ 희귀 혈전’ 국내 두번째 발생 접종 9일 후 두통 구토	0.84
백신 부작용	‘백신 이상신고’ 접종자대비 0.44% 대부분 근육통 두통 발열	0.84
백신 부작용	백신 이상 반응 신고 사흘간 6556건 ‘사망 13명’ 인과성 미확인	0.82
백신 부작용	T19: 청소년 임신부 심근염 추가접종 소아 심낭염 어린이	
백신 부작용	초6~고2, 임신부 백신 접종, 60세 이상 부스터샷 내달 시작	0.79
백신 부작용	다음달부터 부스터샷 청소년 임신부 접종 ‘단계적 일상회복’으로 간다	0.79
백신 부작용	스웨덴, 30세 이하 모더나 백신 접종 일시중단키로 덴마크도 12세 이상 청소년 중단	0.77
기타	T5: 질환 원인 운동 도움 음식 피부 수술	
기타	마스크로 건조한 피부 외출 1시간 전 보습제 바르세요	0.93
기타	[건강한 가족] 만사 귀찮고 다리 가늘어졌나요? 노화 문제 아닐 수 있어요	0.92
기타	열사병 주의보! 몸은 뜨거운데 땀 안 나면 위험	0.92
기타	T6: 보상 인정 보험 부담 의료비 인과성 보장	
기타	[김용하의 이코노믹스] 10년간 38% OECD 국가 중 가장 빠르게 늘었다	0.86

기타	다음달부터 저소득층 ‘재난적 의료비’ 80%까지 지원	0.86
기타	“백신 피해, 정부가 인과성 입증하라” 헌소	0.84
기타	T11: 진료 서비스 세계 활동 분야 의학 팬데믹	
기타	정창현 한국한의약진흥원장 “국민과 함께하는 한의약의 가치 만들겠습니다”	0.88
기타	한의약진흥원, ‘글로벌 전통의약 협력’ 국제컨퍼런스 연다	0.87
기타	파키스탄 결핵 봉사단체, 제16회 고촌상 수상	0.86
기타	T12: 인력 간호사 공공 노조 근무 확충 파업	
기타	보건의료노조 총파업 D-1, 최후 교섭 시작 결렬시 내일 오전 파업 돌입	0.93
기타	‘이달 중 코로나19 전담병원 간호사 배치 기준 마련 내년부터 생명안전수당’ 노정 합의	0.93
기타	보건의료 노정 협상 결렬로 모레 총파업 가능성 정부 ‘파업 자제하고 대화로 해결하길’	0.92

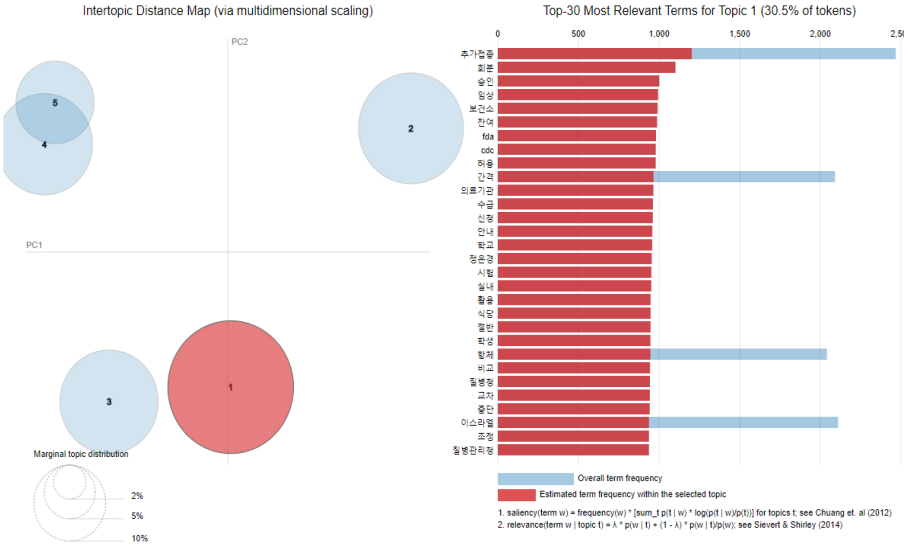
주: 제목열의 T로 시작하는 행은 각 주제에 속할 확률이 높은 키워드.
T 다음의 숫자는 비지도학습을 통해 기계가 부여한 번호.

(4) 변수 확정

2단계를 통해 구성한 5개 변수는 준지도 토픽모델링을 이용해 확인했다. 이를 위해 quanteda패키지를 이용해 각 주제에 속할 확률이 높은 대표단어 40개를 선별하여 5개 변수로 묶어 씨앗사전을 구성했다(감염=돌파, 중증, 방역당국, 급증, 접촉 등; 대응=위드, 전환, 내과, 의견, 의무, 제시 등; 접종=추가접종, 회분, 승인, 임상, 보건소 등; 부작용= 소년, 추가접종, 신고, 혈전증, 임신부; 기타=진료, 인력, 간호사, 논의, 추진 등). seededlda패키지를 이용해 준지도 토픽모델링을 수행하여 5개의 주제로 구성된 모형을 만들었다. 5개 주제 사이의 잰슨-새넌(Jensen-Shannon)의 발산도(divergence)를 비지도 토픽모델링 결과와 비교했다. 잰슨-새넌의 발산도(divergence)는 토픽모델링에서 주제의 숫자를 정할 때 주제의 일관성과 관련된 지표로서 값이 클수록 최적에 근접한다(Watanabe & Xuan-Hieu 2022). 비지도 토픽모델링의 21개 주제로 구분한 모형의 발산도가 0.5이고, 5개 주제로 구분한 준지도모형의 발산도가 0.49다. 주제의 개수가 크게 줄었음에도 불구하고 발산도가 나빠지지 않았다. 5개 주제로 비지도 토픽모델링을 수행하면 발산도 0.37로 크게 떨어진다. 따라서 준지도 토픽모델링을 이용한 5개 주제로의 분류는 타당하다고 판단할 수 있다.

각 주제가 의미별로 제대로 분류됐는지 확인하기 위해서 각 주제 사이의 관련성을 파악했다. 이를 위해 ldaViz패키지를 이용해 시각화했다(<그림 4>). ldaViz는 주성분분석(PCA, Principal Component Analysis)를 이용하여 다수의 토픽 사이의 거리 및 각 주제에 속한 대표단어 30개에 대하여 시각화한다.

분석결과 크게 4개의 군집(접종, 감염, 대응, 기타+부작용)이 확인됐다. 접종과 대응은 거리가 가까웠지만 중복되지 않았다. 부작용과 기타로 분류된 주제가 일정 부분 중복됐다. 감염은 모든 주제와 거리가 멀었다. 즉, 가설에서 설정한 4개의 주제가 준지도 토픽모델링에서 확인됐다고 할 수 있다. 기타로 분류한 주제는 가설에서 설정한 주제와 의미적으로 일치하지 않았어도 일정 부분 중복되는 이유는 두 주제 사이에 공유하는 내용이 많기 때문이다. 따라서 가설을 통해 예측한 코로나19에 대한 언론보도의 의제가 감염, 대응, 접종, 부작용 등 4개 주제로 이뤄질 것이란 가설은 지지됐다고 할 수 있다.



<그림4> 5개 주제 사이의 관련성과 1번 주제(백신접종)에 속할 확률이 높은 단어

주: 1=백신접종, 2=감염, 3=대응, 4=기타, 5=부작용

추가로 주제별 기사의 분포를 파악하기 위해서 각 주제별로 기사의 *theta* 평균을

비교했다. *theta*의 평균을 구하면 해당 주제에 대해 다룬 기사의 양을 가늠할 수 있다. 분석결과 접종 주제의 기사가 가장 많았고, 감염과 대응 관련 기사가 그 다음으로 많았다. 즉, 가설에서 설정한 주제의 기사가 말뭉치에서 대부분을 차지했다. 부작용 주제의 기사가 가장 적었다. 각 주제별 단어-문서의 분포를 <그림 4>로 시각화했다.

<그림 4> 주제별 단어-문서 분포

2) 매체 성향과 주제에 대한 프레임

가설 2. “언론매체의 성향에 따라 동일한 주제에 대하여 긍정 혹은 부정의 틀을 달리 적용할 것이다”를 검정하기 위해 앞서 확인한 4종의 주제에 대한 긍정성과 부정성 정도를 준지도 토픽모델링으로 감정분석한 뒤, 의제 4종과 언론매체를 긍정에 대하여 비모수 회귀분석을 수행했다.

(1) 감정분석

각 주제에 대한 긍정 및 부정으로의 틀짓기 양식을 측정하기 위해서 기존에 만들어져 있는 감정사전의 긍정 및 부정어를 씨앗사전으로 활용하여 준지도 토픽모델링을 수행했다. 감정사전은 KNU한국어감성사전(온병원 등 2018.5.11)을 이용했다. KNU한국어감성사전에 포함된 단어 중 말뭉치에 포함된 단어를 추려 씨앗사전을 만들었다. 긍정 주제의 상위 단어는 인정, 개선, 도움, 최고, 필수, 희망, 이득, 혜택, 감사, 정상 등이었다. 부정 주제의 상위 단어는 질환, 의심, 독감, 피해, 부담, 통증, 걱정, 두통, 불안, 발열 등이었다. 대표적인 긍정 및 부정 기사는 <표 3>에 정리했다.

<표 3> 대표적인 긍정 및 부정 기사

긍정	부정
청와대 “내년도 백신 9천만회분 신규 구매” 청와대가 내년에 쓰기 위해 새로 구매하는 코로나19 백신 물량이 애초 5천만회분보다 1.8배 늘어난 9천만회분이고, 올해 다 쓰지 못하는 백신까지 포함하면 내년에 모두 1억7천만회분을 활용할 수 있다고 밝혔다. 정은경 질병관리청장은 코로나19 방역 체계를 ‘위드 코로나’(코로나19와의 공존)로 전환하는 시점에 대해 “9월 말이나 10월 초에 준비 작업을..	열 끓는 아이들 50명 돌연 숨졌다. 印 덮친 '미스터리 열병' 신종 코로나바이러스 감염증(코로나19) 델타 변이의 확산으로 몸살을 앓는 인도에서 원인 모를 열병이 퍼져 일주일만에 어린이 최소 50명이 사망했다. 현재 같은 증상으로 입원한 어린이 환자는 수백명으로, 사망자는 더 늘어날 것으로 보인다. ― 일주일만에 어린이 50명 사망 1일(현지시간) 영국 BBC 방송은 최근 인도의 북부 우타르프라데시..
‘국외 접종 격리면제자’ 10명 입국 뒤 확진 5명 시노팜 접종자 이달부터 국외 코로나19 예방접종 완료자에 대한 격리면제 제도가 시행 중인 가운데, 최근 입국한 격리면제자 가운데 확진자가 10명 발생한 것으로 나타났다. 특히 이 가운데 5명이 시노팜 백신 접종자인 것으로 나타나 정부가 격리면제 제도 적용 대상을 재검토해야 하는 것 아니냐는 목소리가 나온다. 중앙사고수습본부(중수본)는 지난 14일 기준 국내에 입국한 ..	코로나19 재택치료 60대 환자 병원 이송 중 심정지로 숨져 코로나19 확진 판정 후 재택치료를 받던 환자가 병원 이송 중 심정지로 숨지는 일이 발생했다. 22일 서울 서대문구 중앙사고수습본부(중수본) 등에 따르면 서대문구에서 재택치료 중이던 코로나19 환자 A씨(68)가 21일 오전 갑자기 상태가 악화해 병원으로 이송되던 중 심정지가 발생해 끝내 숨졌다. 서대문구 관계자는 “A씨는 전날인 20일 오후 코로나1..
오늘부터 ‘잔여백신’ 당일 예약 서비스...어떻게 조회・접종 하나 오늘(27일)부터 네이버・카카오 앱 검색을 통해 주변 병원의 신종 코로나바이러스 감염증(코로나19) 백신 잔여량을 검색하고 당일 접종을 위한 예약을 할 수 있게 된다. 앱을 통해 집이나 직장 가까운 곳의 접종 의료기관을 미리 지정해 두면 잔여 백신이 발생했을 때 알림을 받을 수도 있다. 27일 코로나19 예방접종대응추진단에 따르면 네이버, 카카오..	[단독] 부작용 무서운 다이어트약 자살 충동까지 일으킨다 “아침부터 밤까지 심장이 너무 빨리 뛰어요. 누가 가슴을 막 짓누르는 느낌도 나고” “무기력과 우울감이 심해지고, 몸이 다 피폐해지는 느낌까지 듭니다.” 블로거들이 인터넷에 올린 다이어트약 후기엔 이 같은 부작용 후기가 적잖게 씌여있다. 코로나 ‘집콕’ 생활에 살찌는 사람들이 자꾸 불어지면서 다이어트약을 찾는 사람도 느는 가운데, 다이어트약을 석 ..

각 주제별 긍정 및 부정 단어는 <표 4>에 정리했다.

<표 4> 주제별 긍정 및 부정 단어

감염		대응		접종		부작용		기타	
부정	긍정	부정	긍정	부정	긍정	부정	긍정	부정	긍정
재난	안심	걱정	감사	독감	최고	질환	인정	부담	개선
비판	적극	불안	친구	실수	혜택	두통	이득	질환	도움
의심	여유	피해	희망	어려움	필수	발열	정상	위기	인정
감기	안정	불편	신뢰	불만	희망	의심	이익	피해	소득
함정	상승세	독감	성공	제한적	효과적	통증	안심	장애인	필수
폐렴	정상	고통	신중	결석	성공	뇌출혈	휴식	우울증	발전
한계	조화	죽음	효과적	분통	능력	독감	적극	결핵	능력
기침	자신감	불신	사랑	가짜	신중	염증	효과적	통증	휴식
어려움	최상	전염병	정상	불균형	안정	마비	안정	스트레스	향상

(2) 기술통계

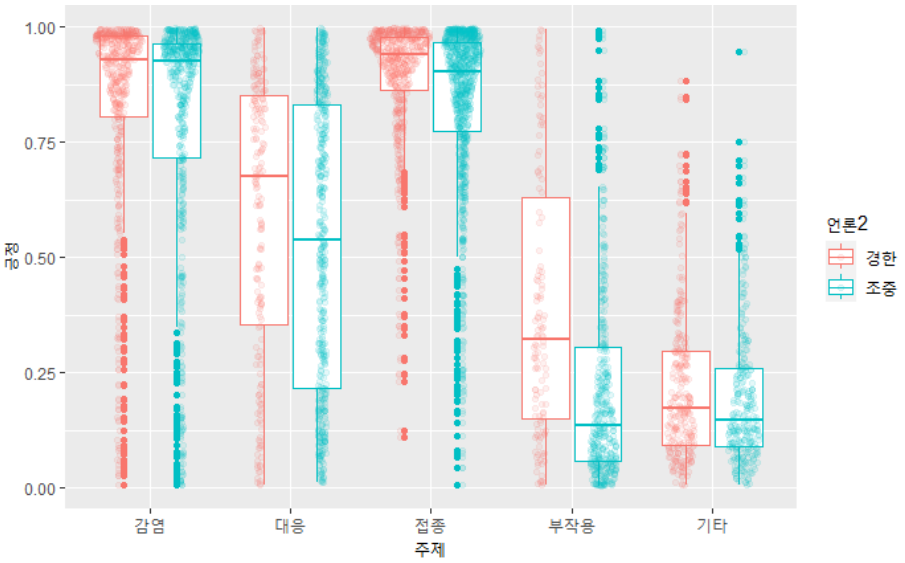
보도된 기사 4,180건 중 언론사별 기사의 건수는 중앙일보(N=1,674), 경향신문(N=927), 조선일보(N=837), 한겨레신문(N=742) 순이었다. 연구변수(감염, 대응, 접종, 부작용, 기타, 긍정, 부정)에 대한 기술통계는 <표 5>에 정리했다.

<표 5> 연구변수의 산술평균, 중간값, 표준편차, 분포

변수	산술평균	중간값	표준편차	히스토그램
감염	0.24	0.11	0.27	
대응	0.18	0.11	0.18	
접종	0.31	0.21	0.26	
부작용	0.12	0.03	0.19	
기타	0.15	0.07	0.2	
긍정	0.63	0.79	0.35	
부정	0.37	0.21	0.35	

(3) 비모수 회귀분석

언론매체의 성향에 따라 코로나19 보도에 대한 4개 주제에 대한 긍정 혹은 부정 틀 적용의 차이를 파악하기 위해 긍정에 대한 시각화 및 회귀분석을 수행했다. 이를 위해 성향이 유사한 경향신문과 한겨레를 하나로 묶고(경한), 조선일보와 중앙일보를 하나로 묶었다(조중). 시각화는 상자도와 시나플롯을 이용했다. 시각화 결과 대체로 경한이 조중에 비해 긍정적인 내용의 보도가 많았다. 특히 부작용에 대한 보도는 경한과 조중의 차이가 컸다(<그림 5>).



<그림 5> 언론매체별 보도주제에 따른 긍정적 틀짓기의 관계에 대한 상자도와 시나플롯

주: N=4,180.

경한과 조중의 차이가 통계적으로 유의한지 확인하기 위해 언론매체와 4개 주제(감염, 대응, 접종, 부작용)를 독립변수로 투입하여 긍정에 대해 회귀분석했다. 4개 주제와 언론매체에 대해서는 모두 상호작용항을 만들어 독립변수로 투입했다. 언론매체는 경한을 0으로 더미코딩했다. 회귀모형은 상호작용항을 넣지 않은 모형 1과 상호작용항을 넣은 모형 2 두 개를 구성했다. 모든 변수의 분포가 과도하게 편향돼 있어 1,000회 복원 재표집하는 부트스트랩하여 비모수 회귀분석를 수행했다. 따라서 모든 계수는 1,000개가 생성됐으며, 1,000개에 대한 평균치를 보고했다. 재표집할

때는 정규분포함수를 이용하지 않고 무작위로 복원추출했다(Mooney et al. 1993).

상호작용항이 없는 모형 1에서 조중은 경향에 비해 긍정적인 기사의 양이 통계적으로 유의하게 적었다(β=-.04, $p < .001$). 언론매체와 4개 주제(감염, 대응, 접종, 상호작용)의 상호작용항을 투입한 모형 2에서 언론매체*부작용의 상호작용항이 통계적으로 유의했다(β=-.16, $p = .003$). 자세한 내용은 <표 6>에 정리했다. 따라서 의제에 대한 긍정 및 부정틀 적용에 대한 가설 2는 부작용 의제에 대해서만 지지됐다.

<표 6> 긍정에 대한 연구변수의 비모수 회귀분석 결과(1,000회 재표집)

모형 1							
변수	*B*	*CI*: .025	*CI*: .5	*CI*: .975	*SE*	*t*	*p*
(Intercept)	-0.10	-0.11	-0.10	-0.08	0.01	-7.53	< .001
언론(조중=1)	-0.04	-0.05	-0.04	-0.03	0.01	-6.42	< .001
감염	1.06	1.03	1.06	1.08	0.02	65.26	< .001
대응	0.6	0.56	0.6	0.65	0.02	28.74	< .001
접종	1.24	1.22	1.24	1.26	0.02	77.63	< .001
부작용	0.13	0.09	0.13	0.16	0.02	6.15	< .001
모형 2							
(Intercept)	-0.10	-0.12	-0.10	-0.07	0.02	-5.46	< .001
언론(조중=1)	-0.04	-0.07	-0.04	-0.01	0.03	-1.54	0.182
감염	1.05	1.01	1.05	1.08	0.02	45.77	< .001
대응	0.63	0.56	0.63	0.7	0.03	18.15	< .001
접종	1.2	1.17	1.2	1.23	0.02	52.46	< .001
부작용	0.25	0.18	0.25	0.32	0.03	7.25	< .001
언론:감염	0.02	-0.03	0.02	0.07	0.03	0.58	0.51
언론:대응	-0.03	-0.13	-0.04	0.05	0.04	-0.80	0.43
언론:접종	0.06	0.02	0.06	0.11	0.03	1.85	0.116
언론:부작용	-0.16	-0.25	-0.16	-0.08	0.04	-3.78	0.003

주: 경한(경향신문+한겨레)은 0으로 더미코딩. 조중(조선일보+중앙일보).
모든 계수는 표본 1,000개의 평균값. *CI*: 신뢰구간.
모형1: 결정계수=0.7; 수정결정계수=0.7.
모형2: 결정계수=0.71; 수정결정계수=0.71.

Ⅳ. 결어

이 연구는 준지도학습 방식의 토픽모델링을 활용하여 이론기반 텍스트마이닝을 수행했다. 이를 위해 개념화, 조작화, 자료수집과 정제, 측정, 해석의 다섯 단계를 거쳐 이론기반 텍스트마이닝 절차와 방법을 제시했다. 이를 바탕으로 해결지향보도, 의제설정, 그리고 틀짓기 이론을 적용하여 가설 2종을 설정하고, 제시한 이론을 기반으로 만든 씨앗사전을 구축했다. 이 씨앗사전은 준지도 토픽모델링으로 변수를 측정하는 데 활용됐다. 매체별 코로나19 감염병 보도의 의제와 프레임의 관계는 이론기반 텍스트마이닝의 분석사례로서 제시됐다.

이 연구에서는 해결지향보도의 틀에서 언론보도의 의제를 크게 문제지적과 문제대응으로 구분하고, 코로나19 관련 의제가 감염, 대응, 백신접종, 백신부작용 4가지로 이뤄질 것으로 가설을 설정했다. 또한, 감염, 대응, 백신접종, 부작용 등 4가지 의제가 언론매체의 성향에 따라 긍정과 부정의 틀이 달리 적용될 것이라는 두 번째 가설을 설정했다.

이에 따라, 백신공급이 시작되던 코로나창궐 초기의 언론보도를 분석 대상으로 삼아, 언론매체, 의제 4가지, 그리고 긍정과 부정의 틀을 분석의 유목으로 설정하고, 개별 기사를 분석 단위로 삼아 빅카인즈의 기사를 수집하여 전처리하여 가설에서 설정한 변수를 측정했다. 이를 통해, 설정한 가설들이 얼마나 타당한지 검증했다.

변수의 측정은 이번 연구에서 적용한 새로운 방법으로 이뤄졌다. 먼저, 씨앗사전을 두 가지 방식으로 구성했다. 의제변수의 측정은 비지도 토픽모델링을 통해 탐색적으로 추출한 주제에 이론을 적용하여 학습용 씨앗사전을 구축하는 방법을 이용했다. 비지도 토픽모델링으로 주제를 수십개 생성한 다음, 해결지향보도의 논리를 적용하여 가설에서 설정한 주제 4개와 기타 주제 1개로 축소했다. 5개의 주제로 구성한 씨앗사전을 사용하여 준지도 토픽모델링을 수행하고 가설에서 설정한 주제 4종을 확인했다. 발산도를 계산한 결과, 준지도 토픽모델링으로 구성한 주제분류 성능이 비지도 토픽모델링으로 구성하는 주제분류 성능보다 우수한 것으로 나타났다.

또한 도출한 주제에 대하여 주성분분석으로 군집한 결과 도출한 주제 4개가 서

로 중복되지 않아 코로나19 보도는 해결지향보도의 논리로 구성한 의제 4종으로 나눌 수 있음을 확인할 수 있었다. 이러한 결과는 가설 1과 가설 2를 확인하는 데에 매우 유용한 정보를 제공한다. 씨앗사전을 이용한 준지도 토픽모델링은 이번 연구에서 매우 유용한 분석 방법임이 증명됐다.

의제변수 4종 외에도 틀짓기 변수를 추가로 측정했다. 차원감정이론을 기반으로 만들어진 사전을 이용하여 학습용 씨앗사전을 구성하고 준지도 토픽모델링을 수행하여 틀짓기 변수를 구성했다. 분석결과, 부작용 주제의 경우, 전형적인 부정어로는 질환, 두통, 발열, 의심, 통증 등이 사용되었고, 전형적인 긍정어로는 인정, 이득, 정상, 이익, 안심 등이 사용됐다.

가설 2(언론매체별 긍정과 부정의 틀의 차별적 적용)의 검정을 위해 회귀분석을 수행했다. 이를 위해 언론매체와 의제변수 4종을 독립변수로 사용하고, 틀짓기변수(긍정틀)를 종속변수로 사용했다. 분석결과, 부작용 주제에 대해서만 언론매체의 틀짓기에 차이가 있었다. 경향신문과 한겨레신문(경한)이 조선일보와 중앙일보(조중)에 비해 백신부작용을 긍정의 틀로 보도하는 경향이 있었다. 다시 말해, 조중은 경한에 비해 백신부작용의 부정적인 측면을 강조하여 질환, 두통, 발열, 의심, 통증 등을 중심으로 한 기사를 더 많이 보도했다. 이러한 차이가 나타난 이유에 대해서는 다양한 해석이 가능하다. 예를 들어, 코로나19 창궐 당시 야당지였던 조중이 백신부작용을 정부에 대한 견제도구로 활용한 것일 수 있다.

이 연구의 가장 큰 기여는, 기존의 토픽모델링을 이용한 연구의 자료 기반의 한계를 극복할 수 있는 방안을 제시한 것이다. 기존의 텍스트마이닝은 대량의 텍스트 자료를 비교적 저렴하게 분석할 수 있는 장점이 있었지만, 가설 검정이 어렵다는 한계가 있었다. 그러나 이 연구는 해결지향보도, 의제 설정, 그리고 틀짓기 이론을 통해 가설을 설정하고, 준지도 토픽모델링을 통해 개념을 양화하여 복수의 변수를 구성하였으며, 비모수 회귀분석을 통해 변수들 사이의 가설을 검정함으로써 이러한 한계를 극복했다. 이는 텍스트마이닝에 이론기반의 접근을 제시하여 사회조사 연구의 선택폭을 넓힌 것으로, 탐색적이거나 기술적인 분석에 더해 이론기반의 추론까지 할 수 있게 됐다는 의의가 있다.

또 다른 기여는 토픽모델링에서 가장 중요한 문제인 주제 지정 방식에 있다. 주제연결망분석을 활용해 탐색적인 해석을 수행하고 주성분분석을 통해 주제의 개수

지정과 그 정당화를 해결할 수 있는 방법을 제시한 것이다. 기존의 토픽모델링에서는 주제의 개수를 어떻게 지정하고 그 지정된 개수가 얼마나 타당한지에 대한 문제가 항상 있었는데, 이번 연구에서 제안한 방법은 주제의 개수를 명확하게 지정하고 그 근거를 제시할 수 있다는 장점이 있다.

이 연구는 감정분석에 대한 새로운 방법인 준지도학습을 도입하여, 감정분석의 활용도를 크게 높일 수 있었다. 이 방법은 확률적인 접근을 취하기 때문에, 분석대상 문서에 감정사전에 등재된 단어가 포함돼 있지 않아도 문서 분류가 가능하다는 큰 장점을 가지고 있다. 반면에 사전기반 감정분석은 사전에 등재된 단어만을 이용하여 문서를 분류할 수 있다는 제약이 있다. 또한 각 문서에 대한 개별적인 감정값을 계산할 수 있으므로, 감정변수를 회귀분석에 변수로 활용할 수 있는 장점이 있다. 이러한 방법론적인 기여는 감정분석 분야에서 큰 의의가 있다.

이 연구는 토픽모델링을 활용하여 추출한 토픽들이 어떤 프레임과 연결되는지 파악하는 방법을 제시했다. 프레임은 문서에서 주제가 표현되는 방식을 나타내는 구조이므로, 토픽모델링으로 추출한 주제를 그대로 프레임으로 해석하는 것은 무리가 있다. 따라서, 토픽모델링을 이용하여 프레임을 식별하려면, 우선 주제를 식별한 다음, 해당 주제에 대한 추가적인 분석이 필요하다. 이러한 이유로, 보통 주제를 추출하는 자료기반 토픽모델링을 수행한 다음, 추출한 주제에 대한 프레임 분석은 인간코더들이 직접 수행했다(예: 문안나·신행 2020). 이 연구는 추가적인 프레임 분석도 기계를 통해 자동화할 수 있는 방법을 제시했다.

이 연구에 제한점이 없지 않다. 토픽모델링에 다양한 알고리듬이 소개돼 있는데 LDA에 국한했기 때문이다. 향후 워드임베딩, NMF 등 다양한 방식의 토픽모델링을 적용한 연구가 필요하다.

내용분석은 설문조사와 실험과 함께 사회조사방법의 3대 축이라 할 수 있을 만큼 핵심적인 연구방법임에도 불구하고, 비용과 시간의 문제로 대부분의 내용분석은 작은 규모에 그칠 수밖에 없는 한계가 있었다. 텍스트마이닝을 통해 이전에는 불가능했던 대규모의 텍스트분석이 가능해졌다. 이번 연구는 씨앗사전을 활용한 준지도 토픽모델링 방법과 절차를 소개하여, 사회조사방법론의 새로운 가능성을 열어주었다.

참고문헌

권향원. 2016. “근거이론의 수행방법에 대한 이해: 실천적 가이드라인과 이론적 쟁점을 중심으로.” ≪한국정책과학학회보≫ 20(2): 181-216.

문안나·이신행. 2020. “사회서비스원 정책 보도의 프레임 분석: 구조적 주제모형 (structural topic modeling) 과 내용분석 (content analysis) 의 보완적 적용.” ≪한국광고홍보학보≫ 22(4): 100-134.

반현. 2007. “의제설정 이론의 재고찰: 5 단계 진화 모델을 중심으로.” ≪커뮤니케이션 이론≫ 3(2): 7-53.

설동훈·고재훈·유승환·이기재. 2020. “한국조사연구학회 발표 논문 내용분석, 1999~2019 년: 학문분야·방법론·연구대상·데이터분석기법의 지속과 변동.” ≪조사연구≫ 21(1): 1-32.

송준모. 2021. “세대, 성차별, 권위주의 그리고 미투 (# MeToo): 개방형 문항을 통한 미투 운동에 대한 태도 분석.” ≪한국사회학≫ 55(3): 113-158

양혜진·안정민·이태헌. 2021. “텍스트 빅데이터 분석을 통한 한국인의 불공정 경험 분석: 국민청원 게시판 데이터 분석 결과를 중심으로.” ≪조사연구≫ 22(1): 25-59.

온병원·박상민·나철원 2018.5.11. KNU한국어감성사전. https://github.com/park1200656/KnuSentiLex.

유은순·최건희·김승훈 2015. “TF-IDF와 소설텍스트의 구조를 이용한 주제어 추출 연구.” ≪한국컴퓨터정보학회논문지≫ 20(2): 121-129.

이준웅. 2000. “프레임, 해석 그리고 커뮤니케이션 효과.” ≪언론과 사회≫ 29: 85-153.

이재현. 2019. ≪인공지능 기술비평≫ 커뮤니케이션북스

정우연. 2022. “신고리 5· 6 호기 공론화 과정에 대한 시민참여단의 평가: 정치적 이념에 따른 차이를 중심으로.” ≪조사연구≫ 23(3): 65-95.

정재철·이종혁. 2022. “한미동맹 보도에 대한 의제 도출과 ‘동맹-자주’관점의 비교 분석: BERT 모델 기반 딥러닝 모형의 활용.” ≪사이버커뮤니케이션학보≫ 39(4): 205-263.

Albugh, Q., J. Sevenans, and Soroka, S. 2013. “Lexicoder Topic Dictionaries.” June 2013 Versions. McGill University. http://www.lexicoder.com/docs/LTDjun2013.zip

Baburajan, V., J. de Abreu e Silva, and F.C. Pereira. 2021. “Open-Ended Versus

Closed-Ended Responses: A Comparison Study Using Topic Modeling and Factor Analysis." *IEEE Transactions on Intelligent Transportation Systems* 22(4): 2123-2132. https://doi.org/10.1109/TITS.2020.3040904.

Barrett, L.F. 1998. "Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus." *Cognition & Emotion* 12(4): 579-599. https://doi.org/10.1080/026999398379574.

Blei, D.M. 2012. "Probabilistic Topic Models." Communications of the ACM 55(4): 77-84. https://doi.org/10.1145/2133806.2133826.

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

Chapelle, O., B. Scholkopf, and A. Zien. 2006. *Semi-supervised Learning*. MIT Press.

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.

Doerfel, M.L. 1998. "What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies." *Connections* 21(2): 16-26.

Egger, R. and J. Yu. 2022. "A Topic Modeling Comparison between lda, nmf, top2vec, and Bertopic to Demystify Twitter Posts." *Frontiers in Sociology* 7. https://doi.org/10.3389/fsoc.2022.886498.

Hotho, A., A. Nürnberger, and G. Paaß. 2005. "A Brief Survey of Text Mining." *Journal for Language Technology and Computational Linguistics* 20(1): 19-62. https://doi.org/10.21248/jlcl.20.2005.68.

Glaser, B.G. and A.L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter.

Jagarlamudi, J., H. Daumé III, and R. Udupa. 2012, April. "Incorporating Lexical Priors into Topic Models." in P*roceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204-213).

Kahneman, D. and A. Tversky. 1979. "Prospect Theory: An Analysis of Decisions Under Risk." *Econometrica* 47(2): 263-292.

Krippendorff, K. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage. https://dx.doi.org/10.4135/9781071878781.

Lu, B., M. Ott, C. Cardie, and B.K. Tsou. 2011. Multi-aspect Sentiment Analysis with Topic Models. in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 81-88). IEEE.

Lough, K. and K. McIntyre. 2019. “Visualizing the Solution: An Analysis of the Images That Accompany Solutions-oriented News Stories.” *Journalism* 20(4): 583-599. https://doi.org/10.1177/1464884918770553

McIntyre, K.E. and K. Lough. 2021. “Toward a Clearer Conceptualization and Operationalization of Solutions Journalism.” *Journalism* 22(6): 1558-1573. https://doi.org/10.1177/1464884918820756.

Mooney, C.Z., C.F. Mooney, R.D. Duval, C.L. Mooney, and R. Duvall. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference* (No. 95). Sage.

Nelson, L.K. 2020. “Computational Grounded Theory: A Methodological Framework.” *Sociological Methods & Research* 49(1): 3-42. https://doi.org/10.1177/0049124117729703.

Neuendorf, K.A. 2017. *The Content Analysis Guidebook*. Sage.

Qaiser, S. and R. Ali. 2018. “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents.” *International Journal of Computer Applications* 181(1): 25-29.

Price, V., D. Tewksbury, and E. Powers. 1997. “Switching Trains of Thought: The Impact of News Frames on Readers’ Cognitive Responses.” *Communication Research* 24(5): 481-506. https://doi.org/10.1177/00936509702400500.

R Core Team. 2022. “R: A Language and Environment for Statistical Computing.” *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Scheufele, D.A. and D. Tewksbury. 2007. “Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models.” *Journal of Communication* 57(1): 9-20. https://doi.org/10.1111/j.1460-2466.2006.00326_5.x.

Stewart, B.M. and Y.M. Zhukov 2009. “Use of Force and Civil–military Relations in Russia: An Automated Content Analysis.” *Small Wars & Insurgencies* 20(2): 319-343. https://doi.org/10.1080/09592310902975455.

Stone, P.J., D.C. Dunphy, and M.S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Striphas, T. 2015. “Algorithmic Culture.” *European Journal of Cultural Studies* 18(4-5): 395-412.

https://doi.org/10.1177/1367549415577392.

Watanabe, K. and Y. Zhou. 2022. “Theory-driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches.” *Social Science Computer Review* 40(2): 346-366.
https://doi.org/10.1177/0894439320907027.

Watanabe, K. and P. Xuan-Hieu. 2022. *Seededlda: Seeded-LDA for Topic Modeling*. R Package Version 0.8.2,
https://CRAN.R-project.org/package=seededlda.

<접수 2023.02.06; 수정 2023.02.22; 게재확정 2023.02.23>

Theory-driven Textmining Methodological Frameworks: Computer-assisted Content Analysis of Infectious Disease News Report Using Unsupervised and Semisupervised Topic Modeling

Dohyun Ahn
(Jeju National University)

Text mining, also known as computer-assisted content analysis or computational content analysis, is a research method that automates the content analysis process, enabling the reliable processing of large amounts of data at a relatively low cost. However, existing studies on text mining often rely on unsupervised topic modeling and are limited to data-based exploratory studies. This study proposes a method to integrate data-based and theory-based text mining using semi-supervised machine learning. To this end, this study presents a methodological framework for constructing variables required for hypothesis testing by creating theory-based seed dictionaries through unsupervised topic modeling and network analysis, interpretive exploration, and utilization of previously developed dictionaries. Applying this framework to theory-based text mining of infectious disease news reports, this study presented a case study (1) setting hypotheses based on three theories(solutions journalism, agenda setting, and framing), (2) measuring variables using semi-supervised learning, and (3) analyzing results. This study also confirmed that the clustering performance of semi-supervised topic modeling was better than that of unsupervised topic modeling. The implications of theory-based text mining for social research were discussed.

Key words: topic modeling, network analysis, agenda setting, frame, solutions journalism

첨부

– 5개 범주로 구분한 21개 주제의 대표적인 기사 –

〈감염〉

‘t8 추석_연휴_반장’

“코로나 이틀째 2000명대...정부 “정점인지 아직 알 수 없어” 9일 국내 신종 코로나바이러스 감염증(코로나19) 신규 확진자가 2000명 넘게 나왔다. 이틀 연속이다. 4차 유행 중심지인 서울 경기 인천 수도권의 확산세가 여전하다. 정부는 현재가 4차 유행의 정점인지는 명확하지 않다고 판단했다. 다만 9월 중하순부턴 감소세가 나타날 것으로 기대하고 있다. 중앙방역대책본부에 따르면 이날 0시 기준 코로나19 ..“

‘t14 돌파_감염자_돌파감염’

“국내 돌파감염 0.03% \‘접종률 80% 돼도 요양시설 감염 위험\’ 국내 신종 코로나바이러스 감염증(코로나19) 4차 대유행으로 42일째 1000명대 신규 확진자가 쏟아지는 가운데 돌파감염자가 계속 늘고 있다. 돌파감염은 접종을 완료하고 2주가 지난 뒤 코로나19에 감염되는 사례를 말한다. 방역당국은 접종률이 높아지더라도 요양병원·요양원 등 밀집 시설은 돌파감염에 취약할 수 있다고 밝혔다. 17일 중앙방역대책본부..”

‘t15 청해부대_국방부_장병’

“[서욱 국방장관 대국민 사과에 합참의장 국방차관 '병풍’ 선 까닭은·청해부대 감염 책임자로 나와 사과·4월 고준봉함 사태 겪고도 대책 부재·확진자 발생 후에도 \‘어쩔 수 없었다\’·군 안팎 \‘감염병 무지가 빚은 결과\’ 서욱 국방장관이 20일 청해부대(4400t급 문무대왕함) 코로나19 집단 감염과 관련해 대국민 사과를 하면서 또 고개를 숙였다. 올해 들어 군과 관련해 각종 사건 사고가 이어지면서 장관이 매달..”

‘t18 인천_외국인_부산’

“인천 코로나19 123명 확진 10명 중 2명 외국인 인천에서 코로나19 확진자가

123명 나왔다. 지난달 확진자 10명 중 2명은 국내 거주 외국인이다. 인천시는 1일 코로나19 확진자는 123명이라고 밝혔다. 주요 집단감염 관련자는 19명, 확진자의 접촉자 68명, 감염경로 조사 중 36명이다. 신규 집단감염은 부평구에 있는 철판제조업체서 지난달 30일 1명의 확진자가 나온데 이어 이날 7명이 추가..”

〈대응〉

‘t7 마음_아이_자신’

“외로운 치매 할머니 간호사는 치료 위해 화투 들었다 병실 바닥에 깐 매트리스 위에서 환자복을 입은 할머니가 고심하듯 화투 패를 내려다보고 있다. 전신 방호복을 입은 간호사가 마주 앉아 그런 할머니를 바라본다. 지난 1일 한 장의 사진이 온라인을 달궜다. 코로나 병동에 홀로 격리된 치매 할머니 환자를 위해 간호사가 화투 패를 갖고 그림 맞추기를 하는 모습이었다. 코로나로 지친 많은 이의 마음을 위로한..”

‘t9 직원_의무_휴가’

“LG그룹 이틀, 삼성전자 SK하이닉스는 하루 ‘백신휴가’ 신종 코로나바이러스 감염증(코로나19) 백신 휴가를 도입하는 기업이 늘고 있다. 삼성전자가 국내 주요 대기업 가운데 가장 먼저 코로나19 백신 유급휴가를 발표한 데 이어 LG그룹과 SK하이닉스도 백신을 접종한 직원에게 유급 휴가를 제공하기로 결정했다. 13일 LG그룹과 SK하이닉스는 코로나19 백신을 맞은 임직원에게 유급휴가를 부여한다고 밝혔다...”

‘t20 위드_내과_전환’

“오명돈 중앙예방접종센터장이 \‘집단면역 달성 어렵다\’고 말한 이유는 국민 70% 이상이 코로나19 백신을 접종하더라도 집단면역은 달성하기 어려울 것이며, 사실상 코로나19 종식은 불가능할 것이라는 관측이 제기됐다. 오명돈 신종감염병 중앙임상위원회 위원장(중앙예방접종센터장 서울대 의대 감염내과 교수)는 3일 서울 중구 중앙예방접종센터에서 열린 기자간담회에서 “코로나19 바이러스는 토착화하여 (인류는) 코로나19 바이러스..”

〈백신접종〉

‘t1 잔여_잔여백신_보건소’

“27일부터 네이버 카카오서 '잔여백신' 검색시 당일 접종 된다 27일부터 네이버·카카오 앱 검색을 통해 주변 병원의 신종 코로나바이러스 감염증(코로나19) 백신 잔여량을 검색하고, 당일 접종을 위한 예약을 할 수 있게 된다. 앱을 통해 집이나 직장 가까운 곳 접종 의료기관을 미리 지정해두면 잔여백신이 발생했을때 알림을 받을 수도 있다. 코로나19 예방접종대응추진단은 \‘네이버, 카카오의 지도 플랫폼을 활용해 ..”

‘t4 회분_간격_수급’

“8월 온다던 모더나 절반도 안온다 접종간격 4주→6주로 정부가 미국 모더나사(社)로부터 이달 공급 받기로 한 신종 코로나바이러스 감염증(코로나19) 백신이 예정된 물량의 절반 이하로 줄어들게 됐다. 생산문제 여파다. 이 때문에 긴급하게 화이자 모더나 백신의 1 2차 접종간격을 4주에서 6주로 한시적으로 늘렸다. 두 백신은 3분기 주력이다. 정부는 올해 1억9200만 회분의 백신을 확보했다고 밝혔지만 지금까지..”

‘t10 임상_항체_교차’

“SK 바이오, 국산 백신 최초 3상 승인 \‘AZ와 비교 임상 시작\’ 국내 제약사가 개발 중인 신종 코로나바이러스 감염증(코로나19) 백신이 처음으로 3상 임상시험에 진입한다. 3상 시험은 임상시험 마지막 단계다. 사람 대상으로 실제 접종을 해 효과와 안전성을 평가한다. 식품의약품안전처는 10일 “국내 개발 코로나19 백신 ‘GBP510(㈜에스케이바이오사이언스)’의 3상 임상시험 계획에 대해 안전성과 과학적 타당성..”

‘t 13 추가접종_이스라엘_cdc’

“美도 면역 취약층에 부스터 샷 접종 \‘FDA, 48시간 내 승인\’ 미국 식품의약국(FDA)이 곧 면역 취약층에 대한 코로나19 백신 부스터 샷(3차 접종)을 승인할 예정이라고 미 주요 언론이 11일(현지시간) 보도했다. CNN, NBC 등은 정통한 소식

통들을 인용해 FDA가 화이자와 모더나 백신에 대한 긴급사용 승인 내용을 바꿔 면역 체계가 손상된 사람은 부스터 샷을 맞도록 허용할 예정이라고 전했다. CNN은 ..”

‘t16 식당_인센티브_허용’

“사적모임 10~12명, 유흥시설은 자정까지 달라지는 ‘일상회복 1단계’ 백신 접종 여부 따지지 않고 허용 식당 카페만 미접종자 4명 제한 고위험 시설 이용 ‘방역패스’ 필수 1주간 계도기간 거친 뒤 적용키로 코로나19 백신 접종 여부와 관계없이 수도권 10명, 비수도권 12명까지 다음달 1일부터 모일 수 있다. 다만 식당 카페를 이용할 땐 미접종자가 4명 이하여야 한다. 일상회복은 1차..”

‘t17 학생_학교_교사’

“‘화이자 티켓’ 9월 모평, 온라인응시 허용 \‘시험장 추가 확보\’ 백신 접종을 위해 9월 대학수학능력시험(수능) 모의평가 응시자가 몰려 수험생이 피해를 볼 수 있다는 우려가 커지자 교육부가 조치에 나섰다. 그동안 예외적으로 허용해오던 온라인 응시를 접수 때부터 선택할 수 있도록 하고, 오프라인 시험장을 추가로 확보하기로 했다. 교육 당국은 이번 9월 모의평가 응시자에게 신종 코로나바이러스 감염증(코로나19) 예..”

‘t21 면제_인도_입국’

“격리면제, 백신 맞고 온 브라질 CEO는 ○ 바이어는 × 신종 코로나바이러스 감염증(코로나19) 백신 접종이 이어지는 가운데 다음 달부터 백신 접종 완료자에 대한 입국관리체계가 완화된다. 정부는 외국에서 백신 접종을 마친 사람도 7월부터 입국 시 2주간의 자가 시설격리를 면제하기로 했다. 하지만 변이 바이러스 상황, 입국 목적별 대상자 등을 고려해 기준을 정하다 보니 체계가 다소 복잡하다. 격리 면제 내용을..”

‘t2 신고_혈전증_청원’

“'AZ 희귀 혈전' 국내 두번째 발생 접종 9일 후 두통 구토 아스트라제네카(AZ) 신종 코로나바이러스 감염증(코로나19) 백신 부작용으로 알려진 희귀 혈전증 사례가 국내에서 두 번째로 발생했다. 코로나19예방접종추진단은 “두 번째 혈소판 감소성 혈전증(TTS Thrombosis with thrombocytopenia syndrome) 확정 사례가 발생했

다”고 16일 밝혔다. 추진단에 따르면 이번 사..“

‘t19 청소년_임신부_심근염’

“초6~고2, 임신부 백신 접종, 60세 이상 부스터샷 내달 시작 다음달 18일부터 만 12~17세 소아·청소년의 신종 코로나바이러스 감염증(코로나19) 백신 접종이 시작된다. 그동안 접종 대상에서 빠졌던 임신부에 대한 백신 접종과 만 60세 이상 고령층 등에 대한 부스터샷(추가접종) 접종도 진행된다. 27일 코로나19 예방접종대응추진단 이러한 내용을 담은 ‘코로나19 예방접종 4분기 시행계획’을 발표했다. 정은..”

〈백신부작용〉

‘t5 질환_원인_운동’

“마스크로 건조한 피부 외출 1시간 전 보습제 바르세요 코로나 사태가 길어지면서 얼굴과 손에 아토피 피부염이 심해진 채 진료실을 찾는 아이들을 자주 본다. 코로나는 아토피 피부염과 얼핏 큰 관련이 없어 보이는데 왜 피부 증상이 악화할까. 코로나 감염 예방을 위해 아이들이 손을 수시로 씻는 데다 얼굴 피부의 절반을 마스크로 덮고 다니면서 피부가 가렵고 붉어지고 건조해지는 증상이 더 흔히 나타나기 때문이다. 그..”

‘t6 보상_인정_보험’

“[김용하의 이코노믹스] 10년간 38% OECD 국가 중 가장 빠르게 늘었다 점점 커지는 사회보장 비용 사회보험료 부담이 급격히 증가하고 있다. 이른바 8대 사회보험이 징수한 지난해 보험료 총액은 151조 원에 이른다. 2010년 74조 원에서 두 배 늘어난 것으로 국내총생산(GDP)의 7.8%에 달하는 규모다. 이 기간 사회보험료 연평균 증가율은 7.4%로 GDP 증가율 4.3%의 1.7배에 근접한다. 사회보험료 부담 증가가 소득..”

〈기타〉

‘t11 진료_서비스_세계’

“정창현 한국한의약진흥원장 “국민과 함께하는 한의약의 가치 만들겠습니다.\“

2006년부터 시작된 ‘한의약육성발전종합계획’이 올해로 4단계에 접어들었다. 2025년까지 추진될 ‘제4차 한의약육성발전종합계획’은 한의약 중심의 건강복지 증진, 산업 혁신을 통한 경쟁력 강화와 지속 성장을 위한 인프라 확충에 중점을 두어졌다. 정창현 한국한의약진흥원장(54)은 최근 취임 6개월을 맞아 경향신문과의 인터뷰에서 “한의약 건강돌봄 및 공공..”

‘t12 인력_간호사_공공’

“보건의료노조 총파업 D-1, 최후 교섭 시작 결렬시 내일 오전 파업 돌입 보건복지부와 민주노총 전국보건의료산업노동조합(보건의료노조)이 노조가 예고한 총파업을 하루 앞둔 1일 최후 교섭에 돌입해 막판까지 진통을 겪었다. 정부는 노정간 핵심쟁점 합의가 이뤄지면 필요한 재정 지원을 하겠다는 입장인 반면, 노조측은 재정과 구체적인 시행 시기 방법론을 먼저 약속하라는 쪽이다. 복지부와 노조는 이날 오후 3시쯤 서울 영등포구 의료기..”