Research Article

KME Review

Korean Medical Education Review 2017; 19(3): 158-168
https://doi.org/10.17496/kmer.2017.19.3.158

pISSN 2092-5603 | eISSN 2093-6370

의학교육에서 기계학습방법 교육: 석면 언론 프레임 연구사례를 중심으로

김준혁[1]* · 허소윤[2]* · 강신익[2] · 김건일[3] · 강동묵[4]

[1]펜실베이니아 대학교 의료윤리 및 건강정책교실, [2]부산대학교 치의학전문대학원 의료인문학교실, [3]부산대학교 의학전문대학원 영상의학교실, [4]부산대학교 석면환경보건센터

Machine Learning Method in Medical Education: Focusing on Research Case of Press Frame on Asbestos

Junhewk Kim[1]* · So-Yun Heo[2]* · Shin-Ik Kang[2] · Geon-Il Kim[3] · Dongmug Kang[4]

[1]Department of Medical Ethics and Health Policy, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA; [2]Department of Medical Humanities, Pusan National University School of Dentistry; [3]Department of Radiology, Pusan National University School of Medicine; [4]Korea Research Center for Asbestos-Related Diseases, Yangsan, Korea

There is a more urgent call for educational methods of machine learning in medical education, and therefore, new approaches of teaching and researching machine learning in medicine are needed. This paper presents a case using machine learning through text analysis. Topic modeling of news articles with the keyword 'asbestos' were examined. Two hypotheses were tested using this method, and the process of machine learning of texts is illustrated through this example. Using an automated text analysis method, all the news articles published from January 1, 1990 to November 15, 2016 in South Korea which included 'asbestos' in the title and the body were collected by web scraping. Differences in topics were analyzed by structured topic modelling (STM) and compared by press companies and periods. More articles were found in liberal media outlets. Differences were found in the number and types of topics in the articles according to the partisanship and period. STM showed that the conservative press views asbestos as a personal problem, while the progressive press views asbestos as a social problem. A divergence in the perspective for emphasizing the issues of asbestos between the conservative press and progressive press was also found. Social perspective influences the main topics of news stories. Thus, the patients' uneasiness and pain are not presented by both sources of media. In addition, topics differ between news media sources based on partisanship, and therefore cause divergence in readers' framing. The method of text analysis and its strengths and weaknesses are explained, and an application for the teaching and researching of machine learning in medical education using the methodology of text analysis is considered. An educational method of machine learning in medical education is urgent for future generations.

Keywords: Asbestos, Structured topic modelling, Automated text analysis, Machine learning, Big data

Corresponding author
Dongmug Kang
Korea Research Center for Asbestos-Related Diseases, Faculty, Research, and Administration Center
3rd floor, 20 Geumo-ro, Mulgeum-eup, Yangsan 50612, Korea
Tel: +82-55-360-1281
Fax: +82-55-360-3779
E-mail: kangdm@pusan.ac.kr
https://orcid.org/0000-0002-0657-0181

Received: December 16, 2016
1st revised: February 24, 2017
2nd revised: May 26, 2017
Accepted: September 4, 2017

*These two authors have contributed equally to this work.

서 론

전문직의 변화에 관한 요청이 점차 거세지고 있다. 그 요청은 최근 미국 대선이 보여주었고 전 세계적 현상으로 나타나고 있는 반이성주의(anti-rationalism)의 물결과 더불어 한층 진보한 정보과학이 이제 전문직의 자리를 넘보며 그 해체 가능성을 제시하는 데에서 기원하는 것으로 보인다[1,2]. 의료인은 변화의 압력을 강하게 받는 전문직 중 하나이다[3]. 인공지능이 의료 전문직을 대체할 수 있는가를 논의하는 사이, 한국에서는 2017년 상반기 6개의 병원이 왓슨 포 온콜로지(Watson for Oncology)를 도입했다. 변화에 대처하기 위해서 요구되는 것 중 하나는 최근의 정보기술의 변화를 이해하고 이를 활용하는 것이다[4]. 따라서 빅 데이터(big data)를 분석하고 활용할 수 있는 기계학습(machine learning) 자료분석(data analytics)방법을 의과대학에서 가르치고 연구현장에서 활용하는 방법에 관한 모색이 활발하게 진행되고 있다[5,6]. 하지만 현장에서 이를 가르치기 위해서 해결해야 할 문제들이 있다.

빅 데이터 교육과 연구를 위해서는 학생들에게 기계학습방법을 가르쳐야 한다. 기계학습이란 자료에서 컴퓨터가 학습하는 방법에

초점을 맞춘 통계학과 컴퓨터과학의 통섭적 학문이다[7]. 이것은 대규모 자료에서 통계적 모형을 구축하는 것에서 출발하며 통계적 기법과 대규모 자료의 취급 양쪽을 다룰 수 있어야 한다. 통계학은 이미 의학 통계학에서 다루고 있으며, 기계학습에서 사용하는 통계적 기법을 이 연장 선상에서 가르치는 것도 가능하다. 하지만 대규모 자료를 수집하고 정리하는 방법을 가르치기는 쉽지 않다. 기존의 통계학 교육에서처럼 같은 자료를 학생들에게 나눠주고 이를 처리하는 방법은 수업 진행상 수월하지만, 학생들의 학업의욕을 고취하기가 어렵다는 난점이 존재한다. 따라서 학생 각각이 자료를 직접 수집하고 처리하는 방법을 가르쳐야 할 필요성이 있다.

이때 활용 가능한 자원으로 첫째, 공공기관이 제공하고 있는 공공데이터가 있다[8]. 둘째, 신문기사나 인터넷 블로그(internet blog), 소셜미디어(social media)의 텍스트(text) 자료가 있다. 후자는 이미 사회학에서 여러 방식으로 분석이 시도되고 있으며, 국내에서도 관련 방법론을 활용한 연구가 활발히 이뤄지고 있다[9,10]. 이런 텍스트 자료의 분석을 교육에 활용하는 것은 여러 장점이 있다. 첫째, 웹 크롤링(web crawling), 즉 인터넷에서 텍스트를 수집하는 방법론을 교육하여 학생들에게 기본적인 컴퓨터 프로그래밍(computer programming), 자료수집 및 정리방법을 가르칠 수 있다. 둘째, 학생들이 직접 관심주제를 선택하고, 현실의 자료를 직접 분석할 수 있어 학생들에게 학습동기를 부여하고 의욕을 고취할 수 있다. 셋째, 과거 분석이 어려웠던 비구조화 자료(unstructured data)인 텍스트 분석방법의 교육을 통해 학생들이 자료분석에 관해 가지는 관점을 넓힐 수 있다.

텍스트 분석방법에는 여러 가지가 있으며, 텍스트를 통해 상품이나 영화 등의 평가점수를 예측하는 모형을 구축하는 감정분석(sentiment analysis), 텍스트에 태그(tag)를 붙여 텍스트의 다면성을 분석하고, 이것이 대상의 평가와는 어떻게 연결되는지 확인하는 양상분석(aspect analysis), 미분류 텍스트를 분류하고, 분류된 결과의 주제를 탐색하는 토픽 모델링(topic modeling) 등이 대표적이다[11]. 본 논문에서는 '석면' 키워드로 검색한 신문기사를 수집하여 수집한 기사를 대상으로 하여 고급 토픽 모델링(advanced topic modeling) 분석을 진행한 결과를 텍스트 분석의 사례로 제시하고자 하였다. 이 사례를 통하여 텍스트 수집과 분석과정을 살피고, 더불어서 의과대학에서 기계학습방법을 교육할 방법을 고찰하는 것이 논문의 목적이다.

먼저 기계학습 적용의 근거를 개괄하기 위해 대상과 방법 설정의 이유를 설명하고 그 이론적 바탕을 살펴볼 것이다. 연구자는 과거 석면공장이 있었던 지역에 거주하던 시민 중 석면으로 인한 질병인 석면폐증에 걸린 환자 몇 명을 면담하였다. 면담과정에서 환자들이 자신의 불편과 고통이 바깥에 잘 전달되지 않는다고 불만을 토로하고 있음을 알게 되었다. 국내에서 석면을 다룬 서적은 한 권이 발간된 바 있으며, 논문은 주로 중피종이나 폐증 등 관련 질환과 증상을 다룬 의학 계열의 논문, 석면의 물리적, 화학적, 환경적 특성과 그 처리법을 다룬 공학 계열 논문, 그리고 분쟁을 다룬 사회, 법학 계열 논문이 주를 이룬다[12-15]. 학술논문은 전문적인 독자를 대상으로 하므로 환자가 불만을 표한 것은 일반 독자를 대상으로 하며 사회적 현실을 다루고 있는 뉴스매체이며, 따라서 환자는 뉴스기사가 자신의 견해를 대변하지 않는다고 인식하고 있다고 가정하였다[16].

이 가정을 검증하기 위해서 지금까지 발간된 모든 신문자료 중에서 석면을 다룬 기사를 찾아 그 중에서 환자들의 불편과 고통을 다룬 기사가 있는지를 확인하는 방법을 상정할 수 있다. 기존에는 석면을 다룬 기사를 모두 찾는다는 것이 불가능했고, 다 찾는다고 해도 기사 전체를 확인하여 주제를 파악하는 것은 오랜 시간과 많은 노동력이 필요했기에 해당 방법의 접근 가능성이 낮았다. 그러나 컴퓨터를 통해 인터넷에서 기사를 자동으로 수집하고, 이것을 텍스트 분석하는 방법은 비교적 간단하게 위에서 제기된 질문을 검증할 수 있는 방법론을 제시한다[17,18].

따라서 처음에는 인터넷에서 관련된 신문기사 전체를 수집하고자 하였으나, 인터넷에서 기사를 검색할 수 있는 것은 1990년 이후의 기사뿐이다. 따라서 검색대상을 1990년 이후의 기사로 한정하였다. 또한 단순하게 신문기사 전체를 모아 그 주제를 파악하는 것을 넘어 신문사의 정파성과 시기에 따라 신문기사의 주제에 차이가 발생하는지를 확인해보고자 했다. 신문이 세상을 보는 하나의 창이라면 그 창은 세상을 보는 특정한 틀을 제시하고 있으며, 그 틀에 따라서 같은 사실이 다르게 해석될 수 있다는 것, 그리고 그 틀은 정파성, 즉 정치적 입장에 따라 크게 결정된다는 프레이밍 이론(framing theory)에 따르면, 석면이라는 현실을 신문이 제시하는 특정한 방식이 존재할 것이며, 그것은 신문의 정파성에 따라 차이가 날 것이라는 가설을 세울 수 있기 때문이다[19-21]. 따라서 정파성이 크게 드러나며 시민들이 주로 구독하는 네 개의 신문에서 기사를 수집하기로 하였다[22].

정리하자면, 본 논문은 석면 관련한 신문기사를 대상으로 한 텍스트 분석방법을 통해 의학교육에서 텍스트 분석 및 기계학습방법을 교육하기 위한 사례를 제시하고자 했다. 사례에서 검증하고자 하는 가설은 두 가지이다. 첫째, 석면 키워드의 신문기사에는 환자의 불편과 고통 주제가 잘 다뤄지지 않는다. 둘째, 신문의 정파성에 따라 석면 키워드의 신문기사 주제는 차이가 날 것이며, 이것은 석면 관련 현실을 다르게 조망하고 있을 것이다.

연구대상 및 방법

1. 데이터 수집 및 처리

1) 데이터 수집

1990년 1월 1일부터 2016년 11월 15일까지 제목 및 내용에

석면이 포함된 신문기사로 보수적인 성향으로 분류되는 조선일보와 중앙일보, 진보적인 성향으로 분류되는 한겨레와 경향신문의 기사를 추출하여 연구대상으로 삼았다. 조선일보와 중앙일보의 경우 외부 사이트에서 검색을 지원하지 않아 각 신문사 홈페이지에서 웹 크롤링 기법을 통하여 수집하였다. 한겨레와 경향신문의 기사는 한국언론진흥재단의 뉴스 빅데이터 분석서비스 페이지(http://www.bigkinds.or.kr)에서 마찬가지로 웹 크롤링 기법을 통해 수집하였다. 웹 크롤링 기법이란 인터넷 페이지(web page)를 수집하여 그 중에서 관심 있는 내용을 추출하는 것을 말하며, 본 논문에서는 Python의 Scrapy 패키지를 통해 자동으로 '석면'을 포함하고 있는 신문기사를 수집하여 신문사, 제목, 보도 일자, 본문을 추출하였다[23,24]. 이 방식을 통해 수집한 기사의 총수는 2,019개였다. 이 중에서 검색되었으나 실제로 석면과 관련된 내용이 아닌 기사 50개, 기간에서 벗어나거나 검색어를 통해서는 수집되었으나 실제로 본문을 담고 있지 않은 기사 319개를 제외하였다. 결과적으로 분석에 사용한 신문기사의 총수는 1,650개였다.

2) 데이터 전처리

텍스트 분석은 단어의 출현빈도에 따른 통계적 모형을 구축한다. 한글은 어미에 결합하는 조사가 다양하며 동사의 활용형이 불규칙적이다. 따라서 텍스트를 그대로 활용하지 않고, Python KoNLPy 패키지(http://konlpy.org/ko/latest/)의 Twitter 품사 태거(speech tagger)를 활용하여 명사를 추출하였다[25]. 또한 의미가 명확하지 않은 단어는 불용어(stop word)로 처리하여 분석에서 제외하였다. 불용어로 처리한 단어는 '석면,' '기자' 등 모든 기사에 등장하는 단어, '년,' '월,' '일' 등 주제를 파악하는 데 도움이 되지 않는다고 판단되는 단어, 의밋값이 명확히 드러나지 않는 '이,' '등,' '것,' '및' 등의 단어이다.

2. 데이터 분석

1) Latent dirichlet allocation 모델을 통한 주제 추출

본 연구에서 활용한 토픽 모델링 알고리즘(topic modeling algorithm)이란 여러 텍스트에서 반복적으로 나타나는 단어에 기초하여 여러 문서에서 공통으로 나타나는 잠재적인 패턴(latent pattern)을 추정하는 것을 가리킨다[26]. 비슷한 주제를 가지고 있는 텍스트는 비슷한 단어를 활용하여 기술하고 있을 것으로 생각할 수 있다. 예컨대 폐렴, 질병, 항생제, 외과적 술식 등의 표현이 자주 나타나는 텍스트는 의료 텍스트로, 축구, 골키퍼, 관중, 심판 등의 표현이 자주 나타나는 텍스트는 스포츠 텍스트라고 볼 수 있다. 그리고 예컨대, 심장내과 논문에는 myocardial infarction (MI)이 많이 등장하지만, 신경외과 논문에는 잘 등장하지 않을 것이라고 예상할 수 있으며, MI가 많이 등장하는 논문의 주제가 MI와 연결되어 있을 것이라고 가정하는 것은 타당할 수 있다. 이렇게 텍스트의 집합에서 자주 출현하는 단어를 그 집합의 특성을 반영하고 있는 중심단어라고 본다면, 그런 단어를 모아 텍스트 집합의 주제를 구성하는 것이 가능하리라는 가정에서 출발한 것이 토픽 모델링이다.

그 중 최근 가장 많이 활용되고 있는 latent dirichlet allocation (LDA)은 각 문서는 여러 개의 토픽을 가지고 있다고 가정하는 것에서 출발한다. 이때 각 토픽은 여러 문서에서 반복적으로 나타나는 단어의 집합이므로 일관성을 지닌다. 단, 단어의 묶음인 토픽이 가지는 일관성과는 별개로 선정된 단어를 어떻게 해석할 것인지는 최종적으로 연구자에게 달려 있다[27]. 본 연구에서는 의료인문학과 산업의학을 전공한 세 연구자가 최종적으로 선정된 토픽 단어집합을 해석하고 이를 비교하여 최종적으로 토픽의 제목을 결정하였다.

LDA는 텍스트에서 나타나는 토픽의 수가 선험적으로 주어진 것으로 가정한다. 즉, 예를 들면 현재 텍스트의 주제가 10개라고 주어지면, 알고리즘은 열 개의 주제를 탐색하여 결과를 제출한다. 기계학습 일반에서는 모형의 조건(parameter)을 결정할 때, 예측모형을 통한 예측값과 실제값의 차이를 나타내는 손실함수(loss function)로 평균제곱오차(mean squared error) 등의 값을 활용하여 최적 조건을 결정한다. 하지만 텍스트 자료에서는 평균제곱오차를 사용할 수 없어 문장 길이에 따른 복잡성 비교를 위해 계산된 엔트로피 값을 단어 개수로 나눈 단어당 비트(bit per words)를 통해 구한 혼란도(perplexity)를 활용한다. 최근 이 값을 통해 적절한 토픽 수를 결정할 수 있다는 연구가 제시되고 있으며, 본 연구에서도 Zhao 등[28]의 연구를 따라 혼란도 변화율(rate of perplexity change)을 지표로 사용하였다.

혼란도는 정보이론(information theory)에서 통계적 모형이 자료를 잘 설명하는지의 정도를 평가하는 데에 흔히 사용되는 수치로 낮을수록 더 좋다(Figure 1). 이 값에 기반을 두어 토픽의 개수 결정을 위해서는 혼란도 변화율을 구간마다 계산하여 혼란도 변화율이 최초로 최소화되는 값을 토픽의 수로 삼았다. 그 결과 일간지 기사에 있어 토픽 개수에 따른 혼란도 변화율을 구하여 토픽 수를 6개로

$$perplexity(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}$$

M: No. of documents
N_d: No. of words in specific document
W_d: Specific word in document d

$$RPC(i) = \left| \frac{P_i - P_{i-1}}{t_i - t_{i-1}} \right|$$

P_i: Perplexity of topic number i
t_i: Topic number i

Figure 1. Formula of perplexity and rate of perplexity change.

결정하고 모형을 구축하였다(Figure 2). 이때 변화율 곡선의 진행 방향에 명확한 차이가 나타나는 지점, 즉 최소 변화율을 나타내는 구간을 선택하였다.

2) 구조화 토픽 모델링을 통한 정파성 및 시계열에 따른 토픽분석

LDA 자체는 문서의 군집을 위하여 개발되었으며, 점차 대규모 문서를 분석하기 위한 유용한 도구로 자리 잡고 있다[29]. 그러나 토픽 모델링은 단순히 전체 문서의 토픽을 나열할 뿐이므로 결과에서 도출할 수 있는 내용은 많지 않다. 따라서 토픽 모델링에 공변량(covariate)을 결합하여 분석하고자 하는 시도가 있었다[30-32]. 국내에서도 이를 통해 신문자료의 오피니언 마이닝을 시도한 연구가 있었다[9]. 하지만 선행연구는 토픽비율에 대한 시계열적 분석일 뿐 토픽구성과 변화에 공변량이 어떻게 영향을 미치는지는 보여주지 못한다는 한계가 있었다. 최근 발표된 구조화 토픽 모델링(structured topic modeling, STM)은 여기에서 한 발짝 더 나아가 공변량을 직접 모델링에 결합하여 텍스트의 공변량이 토픽비율과 내용에 영향을 미치는 정도를 분석하는 방법론이다[33]. 여기에서 공변량이란 본 연구에서의 신문사나 보도일자와 같이 텍스트 자료 각 항의 매개, 수치 변수를 가리킨다. 예컨대 신문사에 따른 토픽비율의 차이를 확인하여 신문사가 해당 사건을 바라보는 관점의 차이를 확인할 수 있다.

앞서 혼란도 변화율을 통해 결정한 토픽 수 6개를 사용하여 매체와 보도일자를 공변량으로 활용하여 전체 기사를 대상으로 R 환경에서 STM 패키지(http://cran.r-project.org/web/packages/stm/index.html)를 통해 분석을 실시하였다[34]. 환자의 불편과 관련한 토픽이 나타나는지, 그리고 각 매체의 토픽구성의 차이를 통해 프레이밍의 차이를 확인할 수 있는지의 가설을 확인하기 위해 세 단계로 분석하였다. 우선, 매체별 기사의 빈도수 차이를 확인하였다. 다음, 매체와 시계열에 따라 토픽의 변화를 관찰하였다. 매체의 프레이밍에 따라 토픽의 구성은 어떻게 변하는지, 그리고 시간에 따라 어떤 차이를 보이는지 확인하였다. 이어서 이를 통해 진보매체와 보수매체에 발표된 석면 관련 기사의 토픽구성을 비교하였다. 논문에서의 분석과정을 요약하여 Figure 3에 제시하였다.

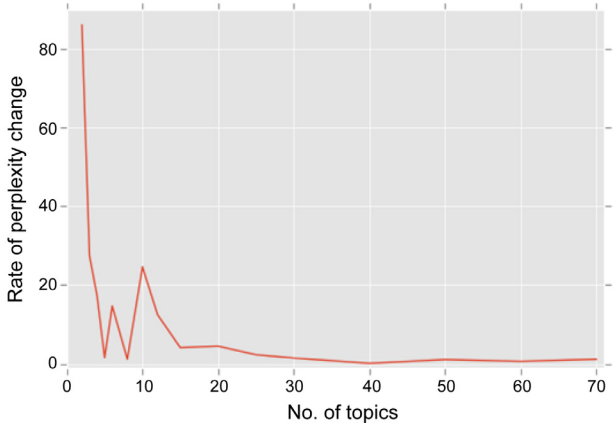


Figure 2. Rate of perplexity change for topic numbers in gathered news articles.

결 과

1. 제목 및 내용에 석면이 포함되어 있는 기사의 신문사별 분포

진보적인 매체로 분류된 주요 언론은 보수적인 매체에 비해 더

Table 1. Distribution of news articles containing 'asbestos' in the title or the body by press companies

Variable	Press company	No. of articles
Conservative (N=552)	Chosunilbo	183
	Chungang Daily	369
Progressive (N=1,098)	Hankyoreh	663
	The Kyunghyang Shinmun	435

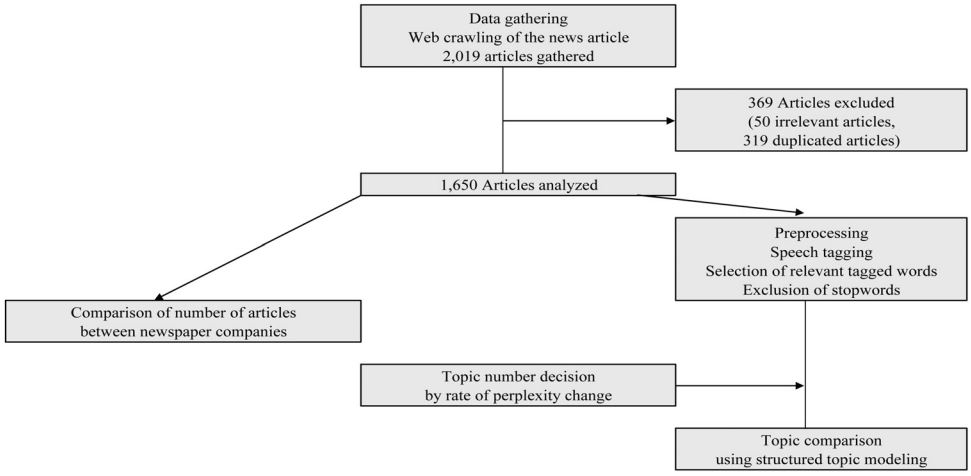


Figure 3. Flow chart of the text analysis method applied in this study.

빈번하게 석면을 다루고 있었다(Table 1). 또한 정파성에 따라 기사 수에 큰 차이가 있었다. 조선일보는 관련 기사 183건으로 석면 문제를 가장 적게 다루고 있었지만 한겨레는 관련 기사 663건으로 석면 문제를 가장 많이 다루고 있었다.

2. 자동화된 텍스트 분석을 위한 기사의 토픽 수 선정

6개의 토픽에 대해 토픽 브라우저(topic browser)를 통해 각각을 조사, 개별 토픽을 결정하고 토픽에 제목을 부여하였다[26]. 토픽 브라우저란 해당 토픽의 중심단어와 가장 관련성이 높은 텍스트를 제시하여 토픽의 제목 설정을 용이하게 하는 제시법을 가리킨다. 이를 통해 확인한 전체 기사의 토픽은 다음과 같다(Appendix 1).

각 토픽의 전체 문서에서의 비율을 표시하였다. 다음, 예시로 토픽 1, 2와 최대 연관성을 보이는 문서를 도표로 표시하였다(Appendix 2). 건물, 현장 석면 검출 토픽의 비율이 상대적으로 높고, 나머지 토픽의 비율은 유사하였다. 사고 토픽의 경우 1994년 발생한 삼풍백화점 사고 등에서 사고환경을 묘사할 때 '석면'과 '석면 가루'가 언급되어 이를 다룬 다수의 기사가 포함되어 있었다.

3. 구조화 토픽 모델링 분석을 통한 언론사별, 기간별 기사 비교

다음, STM기법을 통해 언론사의 정파성 및 시기가 토픽의 비율에 미치는 영향을 분석하였다.

1) 정파성이 토픽 비율에 미치는 영향

정파성이 토픽 비율에 미치는 영향을 보면, 진보매체는 토픽 1(환경 문제), 토픽 2(건물, 환경 석면 검출), 토픽 3(석면 피해자)을 더 많이 다루고 있었다. 반면 보수매체는 토픽 4(발암물질), 토픽 5(생활제품 석면 검출), 토픽 6(사고)을 더 많이 다루고 있었다. 토픽 1은 전반적인 환경 문제를 다루고 있으며, 토픽 4는 석면을 포함한 발암물질 모두를 주제로 했다. 토픽 6은 사고를 주제로 삼고 있었는데, 이것은 사고현장 등에서 석면이 유출되는 때도 있기 때문이다. 이 토픽들은 직접 석면과 관련되어 있지 않는다는 점에서 진보매체에서 석면을 더 많이 다루고 있다고 볼 수 있다(Appendix 3).

이 중에서 석면과 직접 관련된 토픽, 즉 건물, 현장 석면 검출, 석면 피해자, 생활제품 석면 검출 토픽을 좀 더 집중적으로 살펴보았다(Appendix 4). 토픽 2(건물, 현장 석면 검출)의 경우 진보언론에서 더 많이 다뤄지고 있으며 건물, 철거, 슬레이트, 공사 등의 주제어가 대두되고 있지만, 보수언론은 많이 다루고 있지 않으며 주로 교육, 마을 등이 핵심어로 나타났다(Appendix 4A). 이에 기반을 둘 때 진보언론은 공사현장의 문제를, 보수언론은 지역 교육시설의 문제에 초점을 맞추고 있다. 또한 토픽 3(석면 피해자)의 경우 보수언론과 진보언론의 어휘에서 두드러진 차이를 보였다(Appendix 4B). 보수언론은 주민, 광산 등 환경 문제를, 진보언론은 암, 폐암, 피해자 등 질병 문제를 중점적으로 다루고 있었다. 다음 토픽 5(생활제품 석면 검출)를 살펴보면, 보수언론은 주로 화장품에서 석면 성분이 검출된 것을 문제 삼고 있지만, 진보언론은 약품에서 석면 성분이 검출된 것을 문제로 제시하고 있었다(Appendix 4C).

위에서 확인할 수 있는 것은 다음과 같다. 첫째, 진보언론이 석면을 더 비중 있게 다루고 있었다. 둘째, 보수언론은 실생활과 관련된 토픽을 주로 내세웠지만, 진보언론은 공사현장, 질병 등 좀 더 사회와 관련된 토픽들을 주로 다루고 있었다. 셋째, 보수언론과 진보언론이 석면 문제를 부각하는 관점에 차이가 있음을 확인할 수 있다.

2) 시기 및 정파성이 토픽 비율에 미치는 영향

이어서 시기에 따른 토픽비율의 변화를 살펴보았다(Figure 4). 토픽 1, 3, 5는 시기에 따라 유의미한 변화가 관찰되지 않았다. 반면, 토픽 2(건물, 현장 석면 검출)는 1990년도에 거의 나타나지 않다가 2015년에 가까워질수록 그 빈도가 증가하여 2015년에는 30%를 넘어서고 있었다(Figure 4A). 반면, 토픽 4(발암물질)와 토픽 6(사

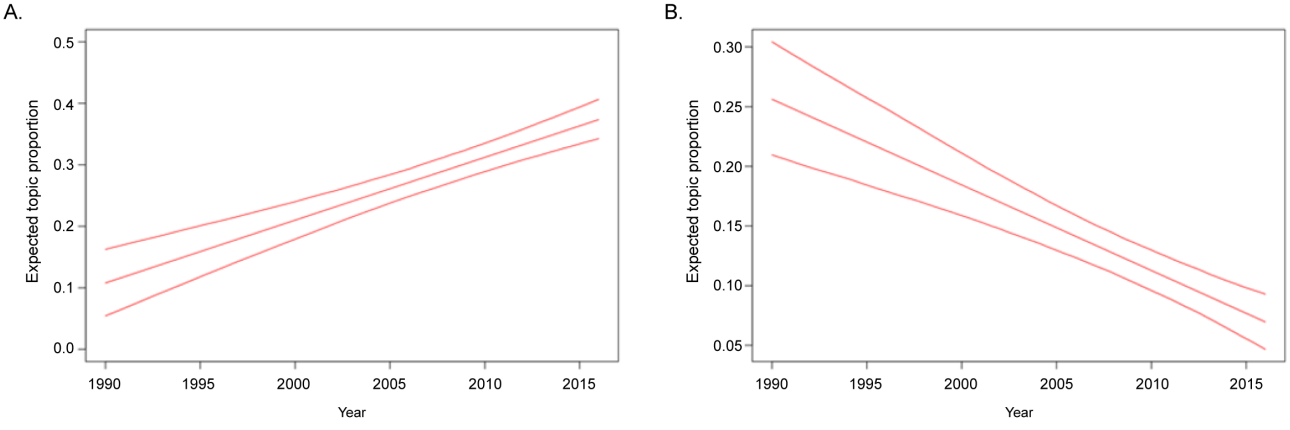


Figure 4. Topic proportion change by year. (A) Topic 2: topic proportion of topic 2 by year (asbestos detection in buildings and construction sites). (B) Topic 4: topic proportion of topic 4 by year (carcinogen).

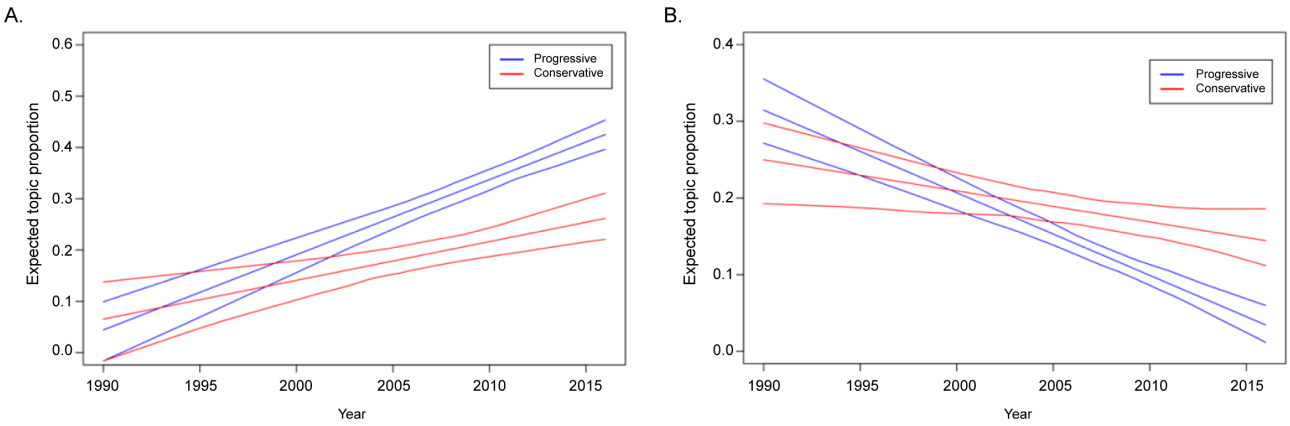


Figure 5. Topic proportion change by year and partisanship. (A) Topic 2: topic proportion of topic 2 with interaction of time and politics (asbestos detection in buildings and construction sites). (B) Topic 4: topic proportion of topic 4 with interaction of time and politics (carcinogen).

고)의 경우 1990년도에는 그 비율이 높다가 시간이 지나면서 그 비율이 점차 감소하여 2015년에는 거의 비중이 사라졌다(Figure 4B). 즉 1990년대 석면은 주로 발암물질의 하나로 사고현장에서 노출되는 것으로 인식되고 있었으나 최근 들어서면서 점차 건축물과 공사현장에서의 석면 검출이 중요한 문제로 대두되고 있음을 알 수 있었다. 반면 석면 피해자나 생활환경에서의 석면 문제에 대한 문제 인식은 시간에 따라 크게 변화가 없다는 점도 확인할 수 있었다.

정파성과 시기가 동시에 토픽비율에 미치는 영향을 분석한 결과, 토픽 1, 3, 5, 6에서는 경향성에 큰 차이가 없었다(Figure 5). 반면, 토픽 2(건물, 현장 석면 검출)의 경우 위의 시계열분석에서 본 것처럼 1990년대에는 모든 언론이 별로 다루고 있지 않다가 점차 증가하였다(Figure 5A). 이때 진보언론이 이 토픽을 더 많이 다루고 있었다. 또한 토픽 4(발암물질)의 경우 1990년대에 진보언론이 대대적으로 다루었지만, 2015년에 가까워지면서 진보언론은 이 주제를 거의 다루지 않고 있었다(Figure 5B). 반면, 보수언론은 적은 양이나마 해당 토픽을 계속하여 다루고 있음이 관찰되었다. 위의 시계열분석에서 확인한 것처럼 전체에서 이 토픽이 차지하는 비율은 감소한 것으로 나타나는데, 그것은 보수언론의 기사량이 기본적으로 적기 때문이다.

고 찰

본 논문에서 활용한 텍스트 분석방법은 표본자료를 통한 원자료의 수치 추정이나 회귀모형 구축을 통한 변수의 설명력을 확인하는 방식이 아닌 모집단 또는 그에 가까운 큰 규모의 자료를 수집하여 자료를 가장 잘 설명할 수 있는 모형을 제시하는 것을 목표로 한다. 물론 자료수집이 어렵고 측정기준 설정의 문제가 있는 의학자료 전반을 대상으로 기계학습방법을 일반화하는 것은 어려우며, 오히려 기존의 통계학적 방법이 더 타당한 결과를 제시할 수 있는 영역이 많다는 주장이 제기되고 있다[35]. 하지만 빅 데이터를 통한 접근이 의학에 가져올 이점에는 여러 가지가 있으며, Murdoch과 Detsky [36]는 빅 데이터가 새로운 지식의 생산, 지식의 보급, 시스템 생물학과 병원자료의 통합을 바탕으로 한 정밀의학의 활용, 환자 측면에서의 정보 확대가 의료전달을 혁신할 것이라고 보았다.

먼저 연구사례를 통해 두 가지 가설을 검증할 수 있는지를 검토해 보자. 첫 번째 가설인 환자의 불편, 고통에 관한 주제가 나타나는가에 있어 토픽 3이 석면 피해자를 다루고 있으나 세부적으로 볼 때 질병과 소송, 환경을 중심으로 구성되어 환자의 고통이나 불편에 관한 주제는 잘 드러나고 있지 않음을 확인할 수 있었다(Appendix 4). 다음, 정파성에 따라 토픽의 차이가 나타나는가에 관해서는 결과 3에서 확인한 것처럼 명확한 차이를 관찰할 수 있었으며, 이것은 정파성에 따라 같은 현실이 다르게 제시된다는 프레이밍 이론의 주장을 시각화하여 보여주는 하나의 사례가 된다. 결과 2와 3에서 확인할 수 있는 것처럼 정파성에 따라 신문은 석면이라는 같은 현실을 서로 다른 모습으로 형상화하고 있다는 점이 나타난다.

두 번째 가설을 구체적으로 살펴보자면 우선, 언론의 기사 수 차이에서 보수적인 언론은 석면을 덜 언급하는 방향으로, 진보적인 언론은 석면을 많이 언급하는 방향으로 석면 문제에 접근하고 있었다. Kim과 Cheong [22]의 국정원 민간 사찰 의혹과 민간인 사찰 주장에 대한 언론보도 연구에 따르면, 매체는 정파성에 따라 의도적으로 사건을 배제한다. 다시 말하면 정파성에 따라 신문매체가 특정 주제를 다루는 빈도가 달라진다는 것이며 같은 결과가 석면 기사에서도 관찰됨을 확인할 수 있었다.

다음, Lakoff [19,20,37]는 인지이론과 신경과학에 기반을 두어 사건을 제시하는 틀, 즉 '프레이밍'에 따라 사람들이 다르게 반응하며, 그것은 도덕적 기반에 호소한다고 주장하였다. 한국에서는 보편복지와 선별복지에 호소하는 '무상급식'과 '선별급식'의 프레이밍이 논란이 된 적이 있다. Lem [38]은 보수, 진보신문의 무상급식

논쟁 프레이밍을 분석하였으며, 담론유형에 기초하여 보수와 진보 언론의 대항 담론 6가지를 파악하였다. 여기서 담론이란 믿음과 견해의 언어적 재현일 뿐만 아니라 권력의 반영이자 그 구현이다[39]. 그러나 Lem [38]의 연구는 담론유형을 미리 선정하고 선정된 담론을 대표할 수 있는 기사제목이나 표현, 은유를 찾아보는 방식으로 접근하였으며 이것이 적절한 분석방법이라고 보기에는 의문의 여지가 있다. 담론유형의 선정 이유, 그리고 선정된 담론이 과연 신문기사에 얼마만큼이나 드러날 것인지에 관한 분석이 제시되지 않았기 때문이다.

본 논문의 사례는 진보언론이 사회적 담론을, 보수언론이 실생활과 관련한 담론을 제시하고 있음을 관찰하였으며, 자료를 검토하기 전 미리 선정한 담론유형이나 프레이밍에 맞춰 기사를 재단하는 것이 아니라 자료 내에서 담론을 찾아내며, 그것이 진보와 보수의 견해 차이에 따라 어떤 식으로 차이를 나타내는지 시각화하여 제시하였다는 데에서 그 가치가 있다. 또한 기존 연구에서 제시한 견해가 석면 기사에서도 같이 나타나는 것도 확인할 수 있었다. Feinberg와 Willer [40]는 환경에 관한 태도가 정파성에 따라 차이를 보이며 진보는 보수보다 피해와 돌봄(harm and care)의 도덕적 축에 영향을 더 많이 받는다고 보고하였다. 본 사례에서도 진보언론에서 현장과 질병의 담론이 더 두드러지게 나타나고 있음을 확인하였다. 반면 보수언론은 개인적 측면에 더 가까운 교육, 환경, 화장품 등을 담론의 축으로 삼아 석면을 제시하고 있었다.

이 연구방법론으로 검색 가능한 모든 신문기사를 모아 분석할 수 있다는 것은 결과의 타당성을 보강하기는 하나 신뢰도를 검증할 필요가 있다. 즉 사례의 연구결과의 보강을 위해서는 두 분류로 수집한 기사를 독자에게 제시하는 방식에 따라서 그들이 석면에 관해 가지는 의견이 어떻게 변화하는지 확인할 필요가 있다. 프레이밍이 환경에 관한 의견이 미치는 영향을 검토한 기존 심리학적 연구에서 활용한 평가방법을 도입하여 추가적인 연구를 진행하는 것이 필요할 것으로 보인다[40,41].

이렇게 텍스트 분석, 특히 토픽 모델링은 관심 주제의 텍스트를 대량으로 수집한 경우 이를 효과적이고 빠르게 분석하는 방법을 제시하여 기존 연구방법에서 분석하기 어렵다고 느껴졌던 문제를 탐구해볼 수 있는 길을 제시한다. 이것은 특히 의료와 관련된 사회적 문제나 수치로는 표현하기 어려웠던 의료적 현실을 연구하고 이를 시각화하여 제시할 수 있는 방법론을 제시한다는 점에서 강점을 지닌다. 또한 텍스트 관련 연구를 수행하면서 기존의 질적 연구가 지니는 난점, 즉 의료 관련 연구를 진행하기 위한 재원의 한계, 기획 및 수행에 필요한 기술, 시간, 노력의 양, 능숙한 연구자의 부족과 같은 문제 제기에 대한 대안을 제시할 수 있을 것으로 보인다[42,43].

처음 질문으로 돌아가서 이 방법을 의학교육현장의 기계학습교육에 활용할 수 있는지를 살펴보자. 서두에서 제시한 것과 같이 의학교육에서 기계학습을 가르쳐야 할 필요성은 시급하나, 학습에 필요한 자료와 기계학습모형을 찾기 어려우며, 학생들이 흥미를 느끼고 자신의 연구를 수행할 수 있는 연구방법을 제시하기 어렵다는 난점이 존재한다. 본 논문에서 제시한 연구사례를 의과대학에서의 데이터 과학교육에 응용하여 상기의 문제를 해결할 수 있을 것으로 보인다. 텍스트 분석은 학생들에게 기계학습이 전통적인 통계학과는 다른 비정형 자료나 대규모 자료를 다룰 수 있음을 보여준다. 또한 본 논문의 사례에서 제시한 방법론은 학생들이 사회 의학적 가설을 설정하고, 데이터 분석을 통해 연구를 시도할 가능성을 제시한다. 더불어 학생들이 자신의 문제를 설정하고, 직접 자료를 구축하며, 직접 모형을 만들어볼 수 있으므로 학생들의 참여도와 이해도가 높아질 것으로 기대해볼 수 있다. 또한 단순한 통계 패키지 활용법 교육뿐만 아니라 어느 정도의 프로그래밍 기법에 관한 이해가 필요하므로 학생들에게 최근 요구되고 있는 코딩 및 알고리즘(coding and algorithm) 교육을 위한 접근법으로 활용할 수 있다는 장점을 지닌다.

이 논의가 가지는 함의는 의학교육에서 학생들이 의료적 문제를 바라보는 방식에 관한 심도 있는 논의를 촉진할 수 있다는 것이다. 앞서 언급한 것처럼 본 연구의 사례에서 활용한 토픽 모델링은 그동안 데이터 과학에서 다루기 어려웠던 비구조화된 자료의 하나인 텍스트를 처리하는 방법을 제시한다. 이 방법론은 의학교육의 실제에서 적용 가능성을 지닌다. 의학교육현장에서 최근 글쓰기와 성찰기록이 강조되면서 학생들의 서술자료를 분석할 방법이 요청되고 있다. 진료현장에선 의료진이 기록한 전자의무기록(electronic medical record) 등의 전자문서를 분류, 분석, 처리하는 방법에 관한 연구가 필요하다. 또한 환자와 의료진 각각이 인터넷 공간에 남긴 진료현장과 질병의 현장에서 일어난 일, 겪은 일에 관한 기록을 분석하여 의료경험을 확인할 수 있다. 신문과 방송매체는 의학을 둘러싼 여러 사건을 전달하며, 여기에서 생성되는 수많은 문서는 의학과 의료를 이해하며, 환자와 의사를 둘러싼 질환과 사회를 해석하는 데에 있어서 필수적인 자료가 된다.

그러나 지금까지는 이런 텍스트 자료를 분석하기 위한 적절한 도구가 없었기 때문에 이를 다루는 데에 어려움이 있었다. 질적 연구를 통해서 분석할 수 있지만, 숙련된 연구자만이 수행할 수 있다는 접근성의 이슈는 의학과 질적 연구 모두에 관한 심도 있는 지식을 갖춘 연구자가 드물다는 문제를 제기하고 있다. 또한 현 의학교육환경에서 질적 연구와 같이 수행에 많은 시간이 드는 연구방법을 학생들에게 가르치는 것은 현실적으로 한계가 있다는 문제도 존재한다[44]. 본 연구사례에서 활용한 연구방법론은 시간과 처리방법의 문제로 그동안 접근하기 어려웠던 문서를 빠르고 간략하게 정리하며, 그 결과를 다른 수치와 결합하여 분석의 자료로써 제시한다. 이 방법론을 통해 텍스트 자료를 더 폭넓게 이해하여 의료적 문제의 사회적, 문화적 차원을 검토할 수 있는 토대를 제시할 수 있다.

데이터 과학의 물결이 사회 각 영역과 전문직의 변화를 요구하는 현실에서 의학교육은 데이터 분석방법론을 가르쳐야 하고, 의료 전문직은 의료현장에서 활용할 수 있는 데이터 분석기법에 관한 연구를 수행해야 한다[45]. 의료인이 빅 데이터와 보건정보학(health informatics)에 관한 지식을 갖추는 것은 단지 의료인의 생존뿐만 아니라 환자와 보호자, 더 나아가 사회 모두를 위한 일이 될 것이다[46]. 기계학습을 통한 예후 예측의 향상, 일부 의료작업의 대체, 진단 정확성의 개선이 의료를 크게 변화시킬 것이라는 예측이 제시되고 있다. 물론 진단의 모호함, 의료자료 처리의 한계, 기계학습을 통해 구축한 모형의 검증 필요성은 이 변화의 속도를 늦출 것이다. 의료에 직접적인 영향을 미칠 때까지 아직 시간이 남아있는 지금, 미래의 주역이 될 의과대학 학생들에게 기계학습 관련 교육을 위한 적절한 방법론을 논의해야 한다.

감사의 글

이 논문은 2015년도 부산대학교 의생명융합연구소의 지원을 받아 연구되었다(30-2015-033).

REFERENCES

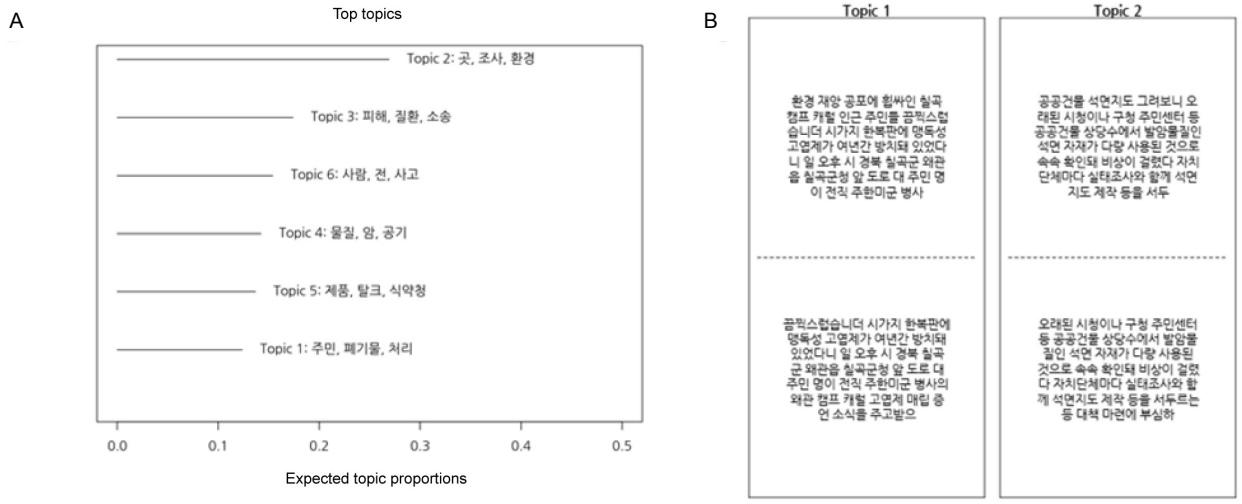
1. Nichols T. The death of expertise: the campaign against established knowledge and why it matters. New York (NY): Oxford University Press; 2017.
2. Susskind RE, Susskind D. The future of the professions: how technology will transform the work of human experts. Oxford: Oxford University Press; 2017.
3. Choi YS. Artificial intelligence: will it replace human medical doctors? Korean Med Educ Rev. 2016;18(2):47-50.
4. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. Developing the role of big data and analytics in health professional education. Med Teach. 2014;36(3):216-22.
5. Tighe PJ, Harle CA, Hurley RW, Aytug H, Boezaart AP, Fillingim RB. Teaching a machine to feel postoperative pain: combining high-dimensional clinical data with machine learning algorithms to forecast acute postoperative pain. Pain Med. 2015;16(7):1386-401.
6. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature. 2008;455(7209):47-50.
7. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-30.
8. Lee M. Big data and utilization of public data. Internet Inf Secur. 2011;2(2):47-64.
9. Kang B, Song M, Jho W. A study on opinion mining of newspaper texts based on topic modeling. J Korean Soc Libr Inf Sci. 2013;47(4):315-34.
10. Park JH, Park E, Jo DJ. Automated text analysis of North Korean new year addresses, 1946-2015. Korean Polit Sci Assoc. 2015;49(2):27-61.
11. Jockers ML. Text analysis with R for students of literatures. New York (NY): Springer; 2014.
12. An JJ. Asbestos, a silent killer. Paju: Hanul; 2008.
13. Park EK. Environmental molecular epidemiological mechanism studies on asbestos-related diseases. Environ Health Toxic. 2012;10:87-9.
14. Jung JS, Jung HS, Lee JY, Lee WS, Kwon OS, Kim SM. A study of asbestos characteristics and correlation of environmental factors in naturally occurring asbestos areas. Korean Soc Living Environ Syst. 2015;22(5):639-46.
15. Ham T, Jeong M. A review of legal issues over relief of damages resulting from asbestos. Environ Law Policy. 2011;6:179-216.
16. Coleman S. New mediation and direct representation: reconceptualizing representation in the digital age. New Media Soc. 2005;7(2):177-98.
17. Wilkerson JD, Casas A. Large-scale computerized text analysis in political science: opportunities and challenges. Annu Rev Polit Sci. 2017;20:529-44.
18. Jacobi C, van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit Journal. 2016;4(1):89-106.
19. Lakoff G. Don't think of an elephant!: know your values and frame the debate. White River Junction (VT): Chelsea Green Publishing; 2014.
20. Lakoff G. Moral politics: how liberals and conservatives think. 2nd ed. Chicago (IL): University of Chicago Press; 2002.
21. Scheufele DA. Framing as a theory of media effects. J Commun. 1999;49(1):103-22.
22. Kim SJ, Cheong YG. Non-reporting, media ethics and ideological conflicts in South Korea: focus on media coverage relating to surveillance of civilians by the National Intelligence Service and the Defense Security Command. Korean J Commun Inf. 2011;53:5-28.
23. Castillo C. Effective web crawling. ACM SIGIR Forum. 2005;39(1):55-6.
24. Myers D, McGuffee JW. Choosing Scrapy. J Comput Sci Coll. 2015;31(1):83-9.
25. Park EL, Cho S. KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology; 2014 Oct 10; Chuncheon, Korea. Seoul: Korean Society of Speech Sciences; 2014.
26. Jacobi C, van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit Journal. 2016;4(1):89-106.
27. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. Reading tea leaves: how humans interpret topic models. In: Bengio Y, editor. Advances in neural information processing systems 22. New York (NY): Curran Associates Inc.; 2009. p. 288-96.
28. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics. 2015;16 Suppl 13:S8.
29. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77-84.
30. Ahmed A, Xing EP. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In: Li H, Marquez L, editors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010 Oct 9-11; Cambridge, USA. Stroudsburg (PA): Association for Computational Linguistics; 2010. p. 1140-50.
31. Blei DM, Lafferty JD. Dynamic topic models. In: Cohen W, Moore A, editors. Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25-29; Pittsburgh, USA. New York (NY): Association

of Computing Machinery; 2006. p. 113-20.
32. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. In: Chickering M, Halpern J, editors. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence; 2004 Jul 7-11; Banff, Canada. Arlington (TX): Association for Uncertainty in Artificial Intelligence; 2004. p. 487-94.
33. Roberts ME, Stewart BM, Tingley D, Airoldi EM. The structural topic model and applied social science. Proceedings of the advances in Neural Information Processing Systems workshop on topic models: computation, application, and evaluation; 2013 Dec 10; Lake Tahoe, USA. La Jolla (CA): Neural Information Processing System Foundation; 2013.
34. R Development Core Team. R: a language and environment for statistical computing [Internet]. Vienna: R Foundation for Statistical Computing; 2008 [cited 2016 Oct 18]. Available from: http://www.R-project.org.
35. Seward JB. Paradigm shift in medical data management: big data and small data. JACC Cardiovasc Imaging. 2017 Jan 12 [Epub]. https://doi.org/10.1016/j.jcmg.2016.10.013.
36. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351-2.
37. Lakoff G. The political mind: why you can't understand 21st-century politics with an 18th-century brain. New York (NY): Penguin Books; 2008.
38. Lem SM. The argument presented by the conservative and progressive seen through controversy over the free elementary school meal project centering on the framing of the newspaper media. Korean Polit Sci Rev. 2011;45(2):251-79.
39. Foucault M. Power/knowledge: selected interviews and other writings 1972-1977. New York (NY): Pantheon; 1980.
40. Feinberg M, Willer R. The moral roots of environmental attitudes. Psychol Sci. 2013;24(1):56-62.
41. Walsko C, Ariceaga H, Seiden J. Red, white, and blue enough to be green: effects of moral framing on climate change attitudes and conservation behaviors. J Exp Soc Psychol. 2016;65:7-19.
42. Chung J, Cho JJ. Use of qualitative research in the field of health. J Korean Acad Fam Med. 2008;29(8):553-62.
43. Hong L, Davison BD. Empirical study of topic modeling in Twitter. Proceedings of the first workshop on social media analytics; 2010 Jul 25-28; Washington DC, USA. New York (NY): Association of Computing Machinery; 2010. p. 80-8.
44. Rosenkranz SK, Wang S, Hu W. Motivating medical students to do research: a mixed methods study using Self-Determination Theory. BMC Med Educ. 2015;15:95.
45. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-9.
46. Kim J. Big data, health informatics, and the future of cardiovascular medicine. J Am Coll Cardiol. 2017;69(7):899-902.

Appendix 1. Vocabulary associated with all topics of articles in selected newspapers (topic number=6)

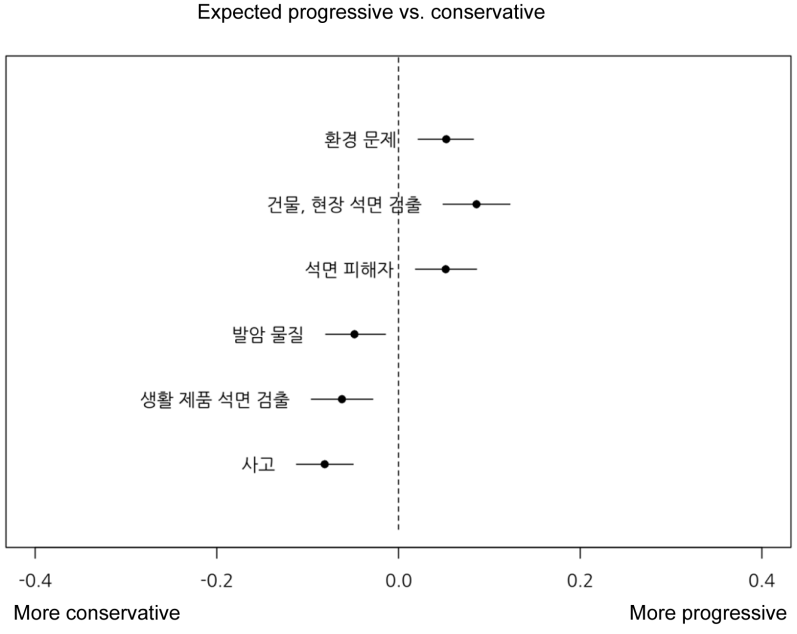
번호	토픽	상위 7개 어휘						
1	환경문제	주민	폐기물	처리	환경	미군	매립	업체
2	건물, 현장 석면 검출	곳	조사	환경	학교	공사	철거	사업
3	석면 피해자	피해	질환	소송	노동자	공장	중피종	피해자
4	발암물질	물질	암	공기	발암	실내	기준	폐암
5	생활제품 석면 검출	제품	탈크	식약청	판매	화장품	약품	금지
6	사고	사람	전	사고	시	뒤	우리	층

Appendix 2. Differences in topic proportions and examples of topics



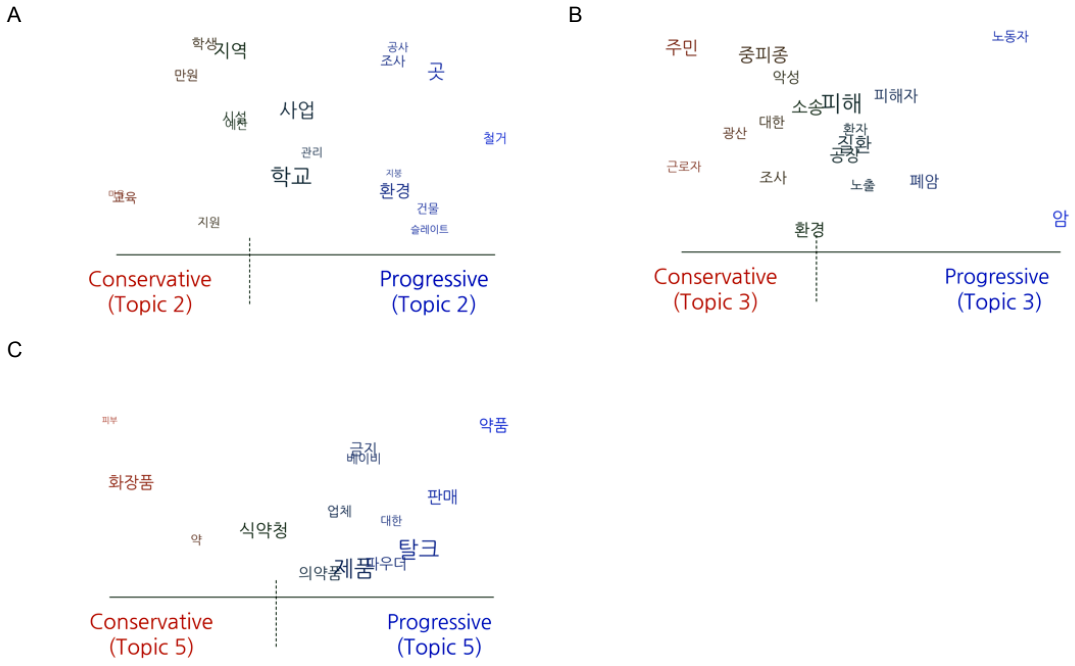
(A) 토픽 비율 추정값. (B) 최대 연관문서의 예시(토픽 1, 2). topic 1 (environmental problem) and topic 2 (asbestos detection in building and construction site).

Appendix 3. Differences in topical coverage by rating, based on the media partisanship



Lines indicate 95% confidence intervals.

Appendix 4. Graphical display of topical perspectives, based on the partisanship



(A) 토픽 2(건물, 현장 석면 검출). (B) 토픽 3(석면 피해자). (C) 토픽 5(생활제품 석면 검출). Topic 2, topic 3, and topic 5 are presented.