한국정보전자통신기술학회논문지(jkiiect)'22-10, Vol.15 No.5
논문 22-15-05-423 http://dx.doi.org/10.17661/jkiiect.2022.15.5.423

LDA를 사용한 COVID-19 관련 국내 논문의 연구 토픽 분석

김은회*, 서유화**

Research Topic Analysis of the Domestic Papers Related to COVID-19 Using LDA

Eun-Hoe Kim*, Yu-Hwa Suh*

요 약 본 논문은 학술연구자들이 COVID-19 관련 논문의 전체적인 연구 동향을 파악할 수 있도록 한다. KCI 사이트에서 수집한 2020년 1월부터 2022년 7월까지 총 10,599편의 COVID-19 관련 논문 정보를 LDA 토픽 모델링으로 분석한 결과를 제시한다. 또한 학술연구자들이 자신의 관심 연구분야의 토픽을 쉽게 파악할 수 있도록 LDA 토픽 모델링의 결과를 주요 연구 카테고리별로 분석하고, 토픽별로 연구가 많이 이루어지는 세부 연구 카테고리 정보를 분석한다. 학술연구자들이 시간의 흐름에 따른 연구 토픽의 추세(trend)를 파악하는 것은 연구 동향을 파악하는데 매우 중요하다. 따라서 이를 위해 본 논문에서는 시계열 분해를 사용하여 토픽들의 추세(trend)를 분석하여 제시한다.

Abstract This paper analyzes a total of 10,599 papers related to COVID-19 from January 2020 to July 2022 collected from the KCI site using LDA topic modeling so that academic researchers can understand the overall research trend. The results of LDA topic modeling are analyzed by major research categories so that academic researchers can easily figure out topics in their research fields. Then, the detailed research category information in which a lot of research is done by topic is analyzed. It is very important for academic researchers to understand the trend of research topics over time. Therefore, in this paper, the trend of topics is analyzed and presented using time series decomposition.

Key Words : COVID-19, LDA, Research Topic, Topic Modeling, Time Series Decompose

1. 서론

COVID-19(Coronavirus Disease-2019) 팬데믹으로 인하여 전 세계가 사회, 경제, 교육 등 많은 분야에서 큰 충격을 받았고, 또 많은 것이 변하였다. 학술연구자들도 COVID-19 관련 연구를 수행하며 각 분야에서 다양한 논문을 발표하고 있다. 학술연구자들은 보통 연구를 시작하기 전에 관련 연구분야의 전체적인 연구 토픽을 살펴본다. 또 연구자의 관심분야에 해당하는 토픽들을 집중적으로 파악하며 연구의 방향과 주제를 찾는다. 연구 토픽을 파악하기 위해서는 관련된 많은 논문들을 검토하고 토픽을 분류해야 한다. 그러나 연구자가 직접 많은 양의 논문을 검토하는 것은 매우 어려운 일이므로 텍스트 마이닝 기법을 활용하고 있다. 현재 대용량의 비정형 데이터의 주제 분류를 위해 많이 사용하는 텍스트 마이닝 기법은 LDA(Latent Dirichlet Allocation)[1] 토픽 모델링이다.

본 논문은 COVID-19 관련 논문들을 대상으로 LDA 토픽 모델링을 사용하여 연구 토픽을 찾아 분석한다. LDA 토픽 모델링으로 도출된 토픽들을 변별성 있게 분석하기 위해서 토픽에 기여하는 단어들에 대하여 연관성(relevance)을 계산하여 각 토픽에 기여하는 상위 단어들을 선정한다. 또한 학술연구자들이 자신의

*Department of Software Engineering, Seoil University
**Corresponding Author : Baird University College, Soongsil University(yhsuh@ssu.ac.kr)
Received September 30, 2022 Revised October 24, 2022 Accepted October 26, 2022

관심 연구분야의 토픽을 쉽게 파악할 수 있도록 토픽 모델링의 결과를 주요 연구 카테고리별로 분석하고, 토픽별로 연구가 많이 이루어지는 세부 연구 카테고리 정보를 분석하여 제시한다. 또한 시계열 분해를 사용하여 토픽들의 추세(trend)를 분석한다.

본 논문은 2장에서 관련연구를 논하고, 3장에서는 토픽 모델링 및 토픽 분석 방법을 설명한다. 4장에서는 LDA 토픽 모델링 결과를 분석하고 5장에서 결론과 향후 연구를 제시한다.

2. 관련연구

토픽 모델링을 사용하여 대용량의 COVID-19 관련 비정형 데이터에서 토픽을 분석하는 연구는 국내 언론보도기사, 인터넷 포털, 소셜미디어 등을 대상으로 많이 이루어져 왔다. 김[2]은 국내 언론보도기사를 토픽 모델링과 키워드 네트워크 분석기법을 사용하여 분석하여 COVID-19 관련 온라인 교육에 대한 사회적인 이슈들을 제시했다. 윤[3]은 인터넷 포털과 소셜미디어에서 간호사 관련기사를 수집하고, COVID-19 발생 전후 토픽을 비교 분석하여 간호사에 대한 대중들이 요구 변화를 제시하였다. 김[4]은 소셜미디어에서 수집한 데이터를 TF-IDF 키워드 추출과 LDA 토픽 모델링을 사용하여 정부 정책 변화에 따른 사회적 이슈와 대중의 의견을 분석하였다. 김[2], 윤[3], 김[4]와 같은 연구들은 특정 분야의 COVID-19 관련 연구 상황을 파악할 수 있게 하지만, 전체적인 COVID-19 관련 논문들의 연구의 동향은 제시하지 못한다.

국내 COVID-19 관련 논문을 대상으로 연구 동향을 파악하는 논문으로는 허[5],[6]이 있으며, 두 논문은 같은 저자들이 쓴 논문으로 모두 LDA 토픽 모델링을 통해 연구 토픽을 탐색한다. 허[5]는 추가적으로 다범주 로짓모형을 사용하여 연구 토픽의 추세를 분석한다. 허[6]은 토픽별로 연구 논문의 어조를 감성분석하여, COVID-19 관련 연구의 동향을 파악하는데 기여를 하고 있다. 그러나 허[5]는 2020년 10월 10일까지의 290편의 논문, 허[6]는 2020년 12월 31일까지의 571편의 논문들을 대상으로 하므로 최신 논문들을 추가한 토픽 분석이 필요한 실정이다.

따라서 본 논문은 학술연구자들이 COVID-19 관련 논문의 전체적인 연구 동향을 파악할 수 있도록 논문 정보를 분석 대상으로 한다. 2022년 7월까지의 최신 논문을 추가하여 총 10,599편의 COVID-19 관련 논문을 LDA로 토픽 모델링하고 연구 토픽을 분석한다. 시계열 분해를 사용하여 토픽의 추세를 분석하여 제시하는 특성이 있다. 또한 주요 연구 카테고리와 세부 연구 카테고리 정보를 사용하여 토픽을 분석함으로써 학술연구자들이 자신의 관심분야별로 연구 토픽을 쉽게 파악할 수 있도록 제공하는 차별성이 있다.

3. 토픽 모델링 및 토픽 분석 방법

3.1 논문정보 수집 및 전처리

COVID-19 관련 논문 정보는 국내의 학술지 및 게재 논문에 대한 학술 정보를 제공하는 KCI(Korea Citation Index, 한국학술지인용색인) 사이트[7]에서 2022년 8월 2일에 수집하였다. 검색어로, 코로나-19, 코로나19, 코비드 19, COVID-19, COVID 19를 검색하였고, 기간은 2020년 1월부터 2022년 7월까지로 설정하여 총 11,157개의 논문을 검색하였다. 검색된 논문을 대상으로 Python의 BeautifulSoup, selenium을 사용하여 논문 정보, 즉 영문 초록, 발행일자, 대분류(주요 카테고리), 소분류(세부 카테고리) 정보 등을 크롤링(Crawling)하였다. BeautifulSoup은 HTML(HyperText Markup Language) 문서를 파싱하여 파스 트리를 만들어 구문 분석을 지원하는 패키지다. Selenium은 WebDriver 모듈을 사용하여 브라우저의 자동화를 지원하는 라이브러리이다. Selenium은 KCI 사이트에서 COVID-19 관련 논문들의 아이디를 수집할 때, 자바스크립트를 호출하여 위해 사용하였다. 수집한 논문 정보 중에 영문 초록이 없거나 중복된 것 등 분석에 적합하지 않은 558개의 논문 정보를 삭제하고 10,599개의 논문 정보를 분석에 사용하였다.

크롤링한 10,599개의 논문정보에서 영문 초록만을 추출하여 LDA 토픽 모델링을 위한 전처리 과정을 거친다. 특수문자 제거, 토큰화, 품사 태깅 과정을 거쳐 명사와 대명사, 수사만을 추출한 후 표제어로 변환시킨

다. 불용어(Stop word)를 삭제하기 위해 Python NLTK(Natural Language Tool Kit) 패지키에서 제공하는 불용어를 기본으로 사용한다. 특히 논문에서 많이 사용하는 명사, 예를 들어 subject, purpose, aim, article, conclusion, objective 등을 불용어에 추가하여 삭제한다. 또한 'COVID-19'를 불용어에 포함시켜 토픽 모델링의 입력으로 사용되지 않게 한다. 문장에서 동시에 출현하는 연속적인 단어들은 하나의 분석 단위로 두는 것이 단어 단위로 토픽 모델을 만드는 것보다 좀 더 맥락적인 정보를 얻을 수 있는 방법이다[8]. 따라서 본 논문에서는 bigram, trigram 모형을 사용하여 2개, 3개의 단어가 연속적으로 많이 등장할 경우, 이를 하나의 단위(unit)로 취급하여 LDA 토픽 모델의 입력으로 사용한다.

3.2 LDA 토픽 모델링

LDA는 잠재 디리클레(Latent Dirichlet) 확률 모델에 기반을 두고 있으며 대용량의 문서 집합에서 토픽들을 찾아내고, 토픽과 관련 있는 문서들과 단어들을 알아낼 수 있다[1][8]. 본 논문에서는 전처리된 영문 초록들을 LDA 모델의 입력으로 사용하여 토픽들을 도출한다. 도출된 토픽들에 변별력 있는 토픽 제목(label)을 선정하기 위하여 본 논문에서는 [9]에서 제안한 relevance를 사용한다[8]. LDA는 기본적으로 토픽에 출현하는 단어의 빈도수를 기준으로 토픽에 기여한 단어들을 산출한다. 사용자는 토픽에 기여한 상위 단어들을 검토하여 토픽의 특성을 분석하고, 토픽의 제목을 결정한다. 그러나 한 토픽에 출현 빈도가 높은 단어는 다른 토픽에도 출현 빈도가 높을 수 있기 때문에, 출현 단어의 빈도수만 기준으로 결정된 토픽의 제목은 변별성이 부족할 수 있다는 단점이 있다[8]. Relevance란 다른 토픽과는 차별화된 상위 단어들을 찾기 위해 고안된 방법으로 수식 1에 의해 계산된다.

수식 1. $$relevance(t,w) = \lambda \cdot P(w|t) + (1-\lambda) \cdot \frac{P(w|t)}{P(w)}$$

Relevance(t,w)는 토픽 t에서 단어 w의 relevance 값, $\lambda(0 \leq \lambda \leq 1)$는 가중치를 나타내는 파라미터, P(w|t)는 토픽 t에서 단어 w가 발생할 수 있는 확률, P(w)는 말뭉치에서 단어 w가 발생할 확률이다[8][9]. 논문에서는 [9]에서 제안한 최적의 λ 값 0.6을 사용하여 relevance 값을 계산하고 토픽에 기여하는 상위 단어들을 20개씩 선정하여 토픽의 제목을 결정한다.

3.3 연구 분야별 토픽

학술연구자들은 관심 연구분야가 있고, 작성하는 논문의 주제는 이러한 연구분야의 범주를 크게 벗어나지 않는다. 연구를 시작할 때, 학술연구자들은 관심 연구분야의 범주에 속한 논문들의 연구동향을 파악하게 된다. 따라서 학술지 논문들에 대한 학술 정보를 제공하는 사이트들은 학술지에 대한 주제분류를 제공하고, 연구자들 또한 관심 논문을 검색할 때, 이러한 주제분류를 활용한다. 본 논문에서는 LDA 토픽 모델링의 결과와 논문 정보를 크롤링할 때 수집한 논문의 주요 카테고리(대분류)와 세부 카테고리(소분류) 정보를 사용하여 토픽 모델링의 결과를 주요 카테고리별로 분석한다. 그리고 토픽별로 연구가 많이 이루어지는 세부 연구 카테고리 정보를 분석하여 제시한다.

3.4 시계열 분석

시간의 흐름에 따른 연구 토픽의 추세(trend)를 파악하는 것은 연구를 준비하는 연구자들에게 매우 중요한 일이다. KCI 사이트를 검색해 보면, 국내에서 COVID-19 관련 논문이 발행된 시기는 2020년 1월부터이다. 따라서 본 논문에서는 2020년 1월부터 2022년 7월까지 2년 7개월간 발표된 논문 정보를 월별로 수집하였다. 추후 확인한 결과, 논문 정보를 수집한 2022년 8월 2일에는 2022년 7월에 발행된 논문이 많이 누락된 것을 확인하였다. 따라서 2022월 7월 논문정보는 시계열 분석에 적합하지 않다고 판단하여 제외시키고, 2020년 1월부터 2022년 6월까지 30개월의 데이터를 시계열 분석에 사용한다.

시계열 데이터는 다양한 패턴을 나타낼 수 있다. 시계열 분해(time series decompose)란 시계열 데이터에서 추세(trend), 계절성(seasonality), 순환성(cycle) 패턴을 분리해 내어 분석하는 것을 말한다. 추세란 시간의 흐름에 따라 데이터가 증가하거나 감소하는 패턴이다. 계절성은 주, 월, 분기, 년 등의 주기마다 일정한 빈

도로 반복되는 패턴이다. 순환성은 반복적으로 비슷한 형태의 증가 또는 감소하는 패턴을 말하며 일정하지 않는 빈도로 발생한다[10]. 본 논문에서는 시간에 따른 연구 토픽들의 추세, 계절성, 순환성을 파악하기 위해 Python statsmodels 패키지의 tsa[11]에서 제공하는 시계열 분해를 사용한다. Tsa에서 제공하는 시계열 분해는 다양한 상황에서 사용할 수 있는 강력한 STL(Seasonal and Trend decomposition using Loess)[12] 분해 기법을 제공한다. STL은 아래 수식 2와 같이 시계열 데이터를 추세, 계절성, 잔차(residual)로 나눠서 분석한 모델로 덧셈 분해를 제공하며, 추세에 순환성을 포함하고 있다. 잔차(residual)란 STL에서는 원래 데이터에서 추세와 계절성을 뺀 나머지를 말한다.

수식 2. $y_t = S_t + T_t + R_t$

t 는 시점, y_t 는 데이터, S_t 는 계절 성분, T_t는 추세 및 순환 성분, R_t는 잔차이다.

4. 연구 결과

4.1 LDA 토픽 모델

토픽 모델링에는 Python genism 라이브러리의 ldamallet을 사용하였고, 토픽 모델링의 결과를 시각화 하기 위해서 pyLDAvis 모듈을 사용했다. 최적의 토픽 개수를 결정하기 위해 그림1과 같이 토픽의 개수를 4~20개까지 변화시키며 coherence 값을 측정하였다. Coherence 값은 토픽 모델링 결과를 평가하는 방법 중 하나로서 일반적으로 coherence 값이 높을수록 토픽의 일관성이 높다. 그림 1에서 토픽의 개수가 8 일 때, coherence 값이 0.557로 가장 높은 값을 나타내므로, 최적의 토픽의 개수를 8로 결정했다. 그림 2는 토픽 개수가 8일 때, 토픽 모델링의 결과를 pyLDAvis로 시각화 한 것이다. 각각의 버블은 토픽을 나타내며, 버블의 겹치는 부분이 최소이고, 사분면에 골고루 분포할 때 좋은 모델로 본다[9]. 토픽 개수가 8인 토픽 모델은 겹치는 부분이 적고, 사분면에 골고루 분포되어 있으므로 비교적 좋은 모델이라 할 수 있다.

표 1은 LDA 모델링으로 도출한 8개의 토픽들에 대하여, 수식1의 relevance 값을 계산하여 상위 20개 단어를 산출한 것이다. 각 단어의 오른쪽에 있는 값이

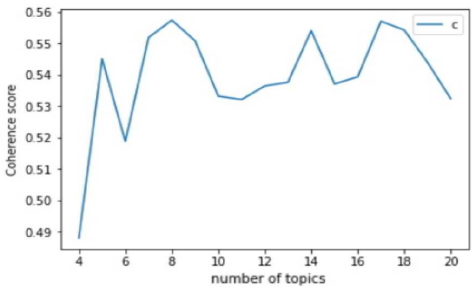

그림 1. 토픽 개수에 따른 coherence 값

Fig. 1. Coherence value by the number of topics

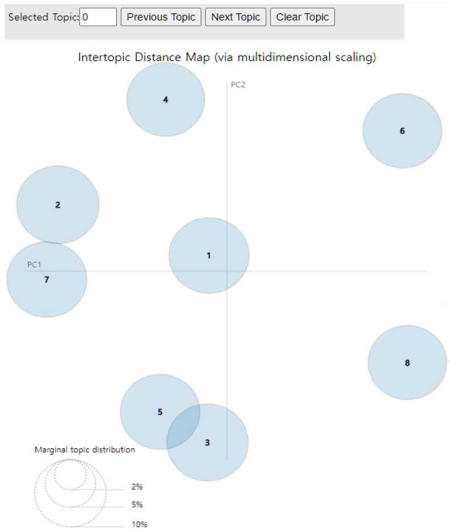

그림 2. LDA 모델의 시각화(토픽개수: 8)

Fig. 2. Visualization of LDA models(number of topics: 8)

relevance 값이고, 값이 작을수록 연관성이 높다. 표 2는 논문의 저자들이 표1의 각 토픽별 상위 20개 단어들을 살펴보고 붙인 토픽의 제목을 보여준다. 토픽 1은 토픽에 기여한 상위 단어들이 industry, company, strategy, market, consumer, business 등이고, 상위 20개 단어들로 볼 때 COVID-19로 인한 산업과 경제의 변화와 충격에 관련한 주제의 논문들로 유추되므로 토픽의 제목을 'Industry & Economy'으로 붙였다. 토픽 2의 상위 단어는 policy, government, system. country, china 등이며, COVID-19와 관련된 정부의 정책관련 주제의 논문들로 유추되므로 토픽의 제목은 'Government Policy'로 정했다. 토픽 3의 상위 단어

표 1. Relevance 값으로 산출한 토픽별 상위 20개 단어
Table 1. Top 20 terms by topic calculated based on relevance value

Ranking	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8
1	industry -1.458	policy -1.25	education -0.889	health -1.108	service -1.107	patient -1.03	life -1.278	effect -0.79
2	company -1.516	government -1.361	class -1.003	disease -1.453	information -1.238	vaccine -1.708	world -1.52	relationship -1.249
3	strategy -1.619	system -1.396	student -1.02	work -1.604	technology -1.452	treatment -1.762	society -1.53	difference -1.556
4	market -1.623	country -1.45	face -1.45	worker -1.604	type -1.453	coronavirus_disease_2019 -1.802	community -1.686	behavior -1.604
5	consumer -1.655	china -1.756	school -1.524	child -1.625	development -1.492	group -1.818	people -1.724	intention -1.651
6	business -1.69	issue -1.854	university -1.55	pandemic -1.689	data -1.613	risk -1.87	era -1.782	stress -1.675
7	product -1.751	state -1.882	teacher -1.566	safety -1.723	medium -1.624	infection -1.899	church -1.797	level -1.725
8	impact -1.826	korea -1.91	learning -1.601	response -1.847	tourism -1.657	vaccination -1.972	concept -1.943	data -1.856
9	change -1.832	disaster -1.944	online -1.634	care -1.865	platform -1.708	day -1.976	theory -2.02	influence -1.87
10	management -1.843	law -1.948	experience -1.665	family -1.875	characteristic -1.761	rate -2	meaning -2.059	anxiety -1.91
11	model -1.874	security -1.976	program -1.822	facility -1.894	design -1.825	symptom -2.006	perspective -2.084	variable -1.91
12	demand -1.915	crisis -2.023	learner -1.853	home -1.899	network -1.853	hospital -2.026	situation -2.093	questionnaire -1.926
13	growth -1.96	measure -2.053	difficulty -1.887	area -1.946	user -1.869	year -2.033	culture -2.124	attitude -1.935
14	customer -1.993	cooperation -2.113	interaction -1.908	center -1.958	content -1.881	results -2.058	pandemic -2.172	group -1.94
15	increase -2.04	organization -2.151	environment -1.957	distancing -1.963	performance -1.965	sars_cov -2.1	change -2.192	depression -1.968
16	food -2.047	act -2.169	satisfaction -1.964	city -1.974	field -2.066	test -2.15	crisis -2.249	job -1.976
17	consumption -2.108	order -2.19	practice -1.976	prevention -1.976	topic -2.066	outcome -2.156	nature -2.281	total -1.979
18	sale -2.166	protection -2.307	activity -2.011	outbreak -1.98	image -2.082	age -2.156	context -2.292	correlation -2.018
19	period -2.212	regulation -2.314	lecture -2.09	contact -1.992	space -2.097	review -2.232	place -2.323	college_student -2.023
20	distribution -2.224	damage -2.347	training -2.21	mask -2.031	stage -2.158	case -2.256	phenomenon -2.345	sport -2.035

는 education, class, student, face, school 등이며, COVID-19로 인한 온라인 교육관련 주제의 논문들로 유추되므로 제목은 'Education'으로 붙였다. 토픽 4의 상위 단어는 health, disease, work, worker, child, pandemic 등으로 COVID-19로 인한 보건관련 이슈들이 주를 이루므로, 제목은 'Health'로 붙였다. 토픽 5의 상위 단어는 service, information, technology, type, development 등으로 COVID-19로 인한 정보서비스 기술과 관련된 단어들이므로 제목은 'Information Service Technology로 붙였다. 토픽 6의 상위 단어는 patient, vaccine, treatment, coronavirus_disease_2019 등으로 COVID-19로 인한 의료 관련 단어들이므로 제목은 'Medical Treatment'로 했다. 토픽 7의 상위 단어는 life, world, society, community, people 등과 같이 COVID-19와 관련된 생활과 사회적 이슈와 관련된 단어들이므로 제목은 'Life & Society'라고 붙였다. 마지막으로 토픽 8의 상위 단어들은 effect, relationship, difference, intention, stress 등의 COVID-19로 인한 영향을 분석하는 이슈들과 연관된
단어들이므로 토픽의 제목은 'Impact Analysis'로 정했다.

표 2는 각 토픽으로 분류된 논문들의 수와 그 비율을 보여준다. 8개 토픽 중에서 가장 많은 비중을 차지하는 토픽은 '토픽 3. Education' 으로, 1842개의 논문이 이에 속하며, 전체 논문들 중 17.38%가 교육 토픽에 속한다. 그 뒤로 '토픽 8. Impact Analysis', '토픽 6. Medical Treatment'가 많은 비중을 차지하며 각각 14.63%와 13.49%의 논문이 해당 토픽으로 분류되었다.

표 2. 토픽의 제목(label), 논문 편수와 비율
Table 2. Topic labels, number and ration of papers

Topic Number	Label	Number of Papers	Ratio
1	Industry & Economy	1277	12.05%
2	Government Policy	1304	12.30%
3	Education	1842	17.38%
4	Health	911	8.60%
5	Information Service Technology	1197	11.29%
6	Medical Treatment	1430	13.49%
7	Life & Society	1087	10.26%
8	Impact Analysis	1551	14.63%

4.2 연구 분야별 토픽 분석

논문 정보를 수집한 KCI는 표 3과 같이 주요 카테고리 8개(공학, 농수해양학, 복합학, 사회과학, 예술체육학, 의약학, 인문학, 자연과학)을 가지고 있으며, 주요 카테고리별로 세부 카테고리가 총 135개이다.

본 논문에서는 먼저 그림 3과 같이 주요 카테고리별로 토픽들의 분포를 산출했다. 사회과학 분야에서 전체 논문의 43.7%에 해당하는 총 4,637개의 COVID-19 논문을 발표하였고, 연구된 토픽은 Government Policy, Education, Industry & Economy, Impact Analysis, Information Service technology, Health, Life & Community, Medical Treatment

표 3. 주요 카테고리별 논문 개수와 비율

Table 3. Number and proportion of papers by main category

Major Category	Number of Detailed Categories	Number of Papers	ratio
공학	20	721	6.8%
농수해양학	7	86	0.8%
복합학	8	1287	12.1%
사회과학	22	4637	43.7%
예술체육학	12	833	7.9%
의약학	35	1750	16.5%
인문학	21	998	9.4%
자연과학	10	287	2.7%
Total	135	10599	100.0%

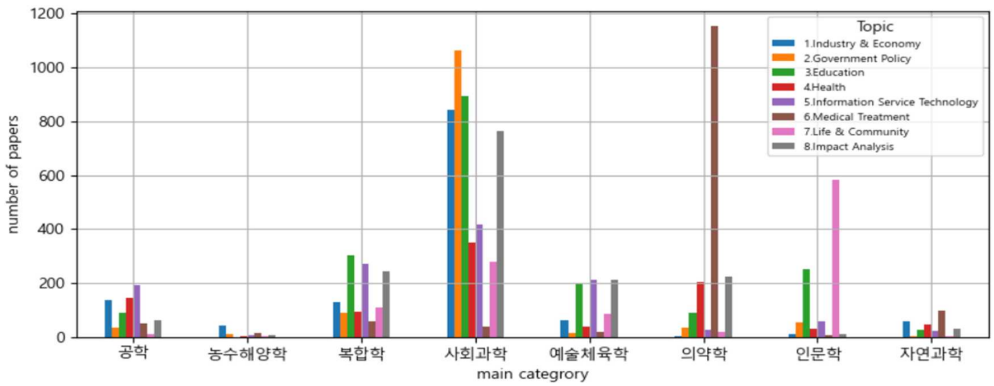

그림 3. 주요 연구 카테고리별 토픽의 분포

Fig. 3. Distribution of topics by major research categories

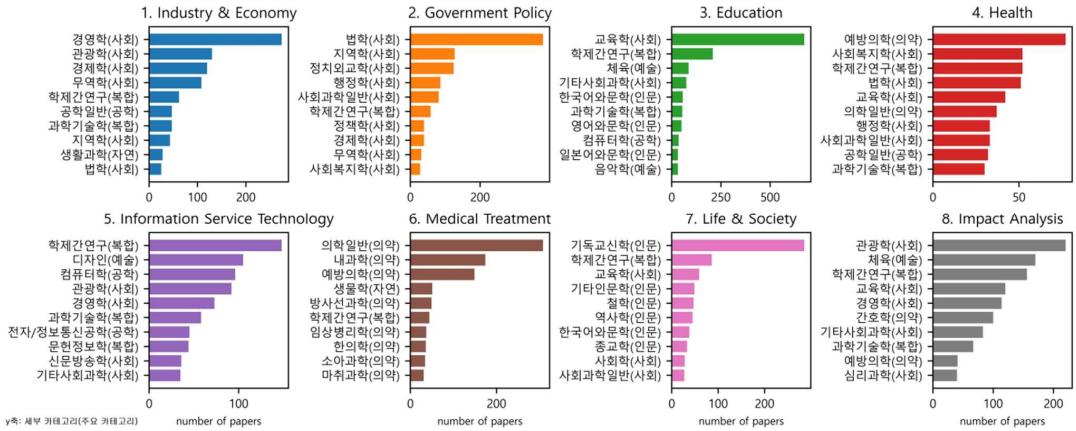

그림 4. 토픽별로 연구가 많이 이루어진 10개의 세부 카테고리

Fig. 4. 10 Subcategories that have been heavily researched by topic

순으로 연구가 이루어졌다. 의약학 분야에서는 Medical Treatment 토픽이 가장 많이 연구되었고, 그 뒤로 Impact Analysis와 Heath 토픽의 연구가 많이 이루어졌다. 인문학에서는 Life & Community와 Education, 복합학에서는 Education, Information Service Technology, Impact Analysis 토픽이 많이 연구되었다. 공학 분야에서는 Information Service Technology, Heath, Industry & Economy 토픽이 많이 연구되었고, 예술체육학 분야에서는 Information Service Technology, Impact Analysis, Education 토픽 연구가 많이 이루어졌다. 자연과학과 농수해양학 분야는 COVID-19 분야의 연구가 다른 분야에 비해 미비하였다. 이로 볼 때, 자연과학과 농수해양학이 다른 분야에 비해 COVID-19와 관련된 연구 이슈들이 상대적으로 적었다는 것을 알 수 있다. 한편, 사회과학 분야는 의약학 분야에서 발표한 논문의 약 2.6배, 전체 논문의 43.7%에 달하는 많은 논문이 발표되었다. 사회과학의 세부 연구분야가 경영, 경제, 관광, 교육, 국제/지역개발, 무역, 법학, 사회복지, 신문방송, 심리과학, 정책, 지역, 행정 등인 것으로 미루어 볼 때, 유례없던 COVID-19 팬데믹으로 인한 사회적인 영향 또한 전례가 없었기 때문에 이에 따른 사회적인 현상들에 대한 연구가 급증한 것으로 유추할 수 있다.

세부 카테고리는 총 135개에 달하므로, 세부 카테고리별로 토픽 모델링 결과를 분석하는 것은 의미가 없다. 따라서 본 논문에서는 그림 4와 같이 각 토픽별로 연구가 많이 이루어진 연구 세부 카테고리와 세부 카테고리가 소속한 주요 카테고리 정보를 분석하여 제시한다. Y축 레이블의 괄호안의 정보는 세부 카테고리가 속한 주요 카테고리 정보를 나타낸다. 첫 번째로 Industry & Economy 토픽을 많이 연구한 세부 분야는 사회과학의 경영학, 관광학, 경제학, 무역학에서 많이 연구되었다. 그 뒤로 복합학의 학제간연구, 공학의 공학일반, 복합학의 과학기술학, 사회과학이 지역학, 자연과학의 생활과학, 사회과학의 법학 분야에서 많이 연구되었다. Government Policy 토픽은 사회과학의 법학에서 주로 연구가 되었으며, 그 뒤로 지역학, 정치외교학, 행정학, 사회과학일반에서 많이 연구되었다. Education 토픽은 사회과학의 교육학 분야에서 주로 연구가 이루어졌고, 그 뒤로 복합학의 학제간연구, 예술체육학의 체육 분야에서 많이 연구되었다. Health 토픽은 의약학의 예방의학, 사회과학의 사회복지학, 복합학의 학제간연구에서 많이 연구되었으며, 다양한 분야에서 골고루 연구가 이루어진 것을 알 수 있다. Information Service Technology는 복합학의 학제간연구에서 가장 많이 이루어졌고, 예술체육학의 디자인 분야, 공학의 컴퓨터학, 사회과학의 관광학에서 많이 이루어졌다. Medical Treatment 토픽은 의약학의 의학일반과 내과학, 예방의학에서 주로 이루어졌으며, 그 뒤로 자연과학의 생물학 분야가 뒤를 이은다. Life & Society 분야는 인문학의 기독교신학 분야에서 연구가 주로 이루어졌으며, Impact Analysis는 사회과학의 관광학, 예술체육학의 체육, 복합학의 학제간연구 등 다양한 분야에서 연구가 이루어졌다.

4.3 시계열 분석

COVID-19 관련 연구 토픽들의 시계열 분석을 위해 그림 5와 같이 먼저 토픽별로 논문의 발행 월에 따른 논문의 개수의 변화를 그래프로 표현했다. 그림 5에서 월별 발행된 논문의 개수 그래프가 들쑥날쑥한 이유는 논문을 발표하는 각각의 논문지가 일정 시간적인 간격을 두고 발행되는데 보통 월간, 월격간, 계간, 연간 등의 간격으로 발행되기 때문이다. 논문지마다 발행 간격이 다양하기 때문에 월별로 발행되는 논문의 개수의 격차가 발생되며 특히, 6월과 12월에 더 많은 논문이 발행된다. 따라서 월별 발행된 논문의 개수만으로는 토픽의 추세를 명확히 파악하는 것이 어렵다. 따라서 본 논문에서는 시계열 분해를 사용하여 토픽의 추세를 분석한다.

그림 6은 Python 패키지 statsmodels의 tsa에서 제공하는 시계열 분해를 사용하여 추세 패턴을 추출한 것이다. Government Policy와 Life & Society 토픽은 2020년까지 증가 추세였으나 2021년도에는 정체성을 보였고, 다른 6개의 토픽은 2021년까지 상승세를 보였다. 특히 가장 상승세가 두드러진 것은 Impact Analysis 토픽이고, Industry & Economy, Education, Information Service Technology 토픽

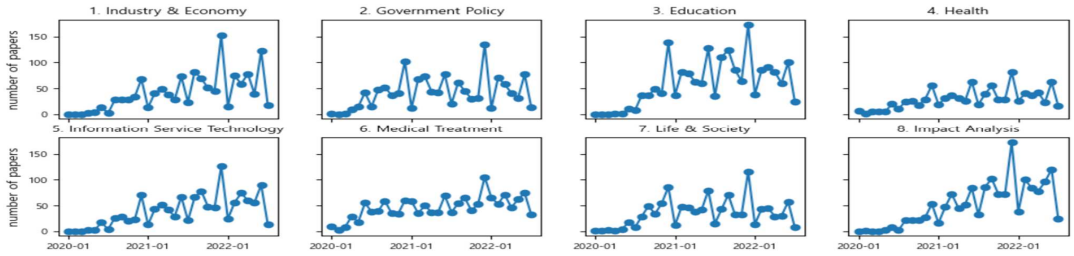

그림 5. 토픽별 발행 월에 따른 논문 개수

Fig. 5. Number of papers by publication month by topic

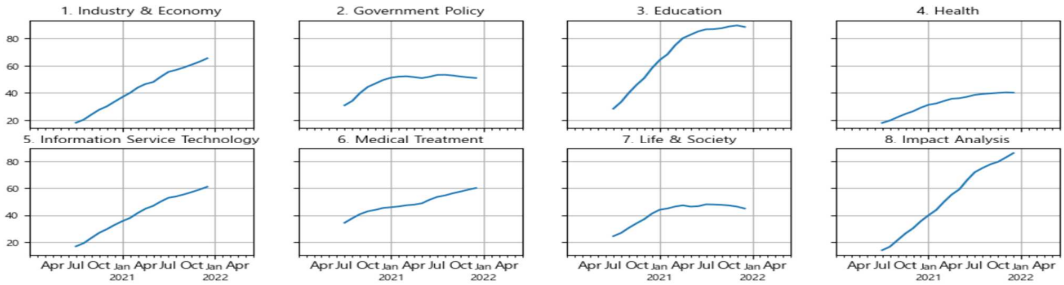

그림 6. 토픽별 시계열 분해: 계절성 그래프

Fig. 6. Time series decomposition by topics: seasonal graph

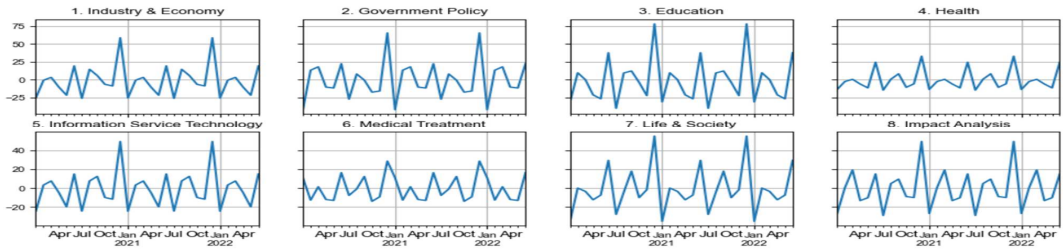

그림 7. 토픽별 시계열 분해: 계절성 그래프

Fig. 7. Time series decomposition by topics: seasonality graph

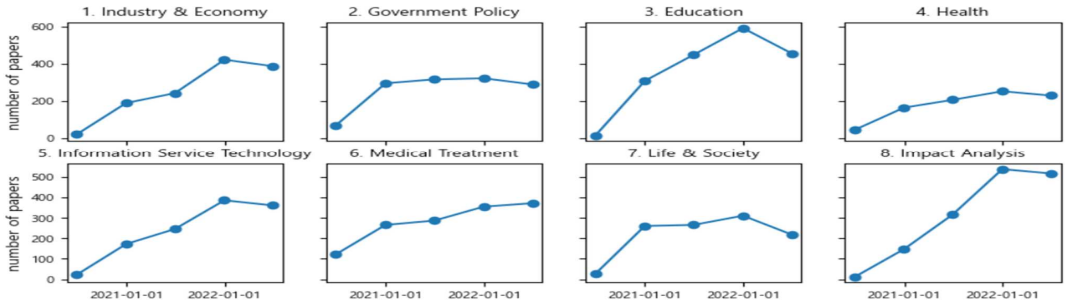

그림 8. 토픽당 반기 논문 개수

Fig. 8. Number of semi-annual papers by topics

분야가 상승 폭이 큰 것으로 보아 많은 연구가 급속도로 이루어진 것으로 보인다. Education 토픽은 2021년 하반기부터 상승폭이 둔화되었다. Health, Medical Treatment 토픽 또한 꾸준히 상승 패턴을 보였으나, Heath 토픽은 2021년 하반기부터 상승세가 둔화되었다.

그림 7은 시계열 분해를 사용하여 추세를 제거한 후 계절성을 추출한 것이다. 모든 토픽이 1년 주기로 일정한 빈도로 반복되는 패턴을 보인다. 국내에 코로나 연구 논문이 발표된 이후 2년 6개월간의 COVID-19 관련 연구 토픽을 분석한 것이므로 년 단위로 계절성을 파악하기에는 기간이 너무 짧다. 시계열 분해로 추세를 분석했을 때는 2021년 12월 하반기까지의 추세만이 추출되었다. 추가적으로 2022년 6월 상반기까지의 추세를 좀 더 파악하기 위해 6개월 단위로 각 토픽별로 논문 발행 개수를 합산하여 그림 8로 나타내었다. 그림 8를 살펴보면, 2022년 상반기에는 Medical Treatment 토픽을 제외하고 모두 하강 추세를 보였다. 가장 급격한 하강 추세 보인 것은 Education 분야이다. 2022년 상반기부터는 온라인 수업을 많이 하던 대학까지 대면 수업이 이루어지면서 COVID-19 관련 교육 토픽의 연구가 많이 적어진 것을 알 수 있다. 또한 life & Society 토픽 또한 하강 패턴이 두드러지는데 이는 2022년도 상반기에 거리두기 완화가 이루어지면서 사회 활동과 일상 생활이 점점 정상화 된 것과 관련이 있다고 할 수 있겠다.

5. 결론 및 향후 연구

본 논문은 학술연구자들이 COVID-19 관련 논문의 전체적인 연구 동향을 파악할 수 있도록 KCI 사이트에서 수집한 2020년 1월부터 2022년 7월까지의 총 10,599편의 COVID-19 관련 논문 정보를 LDA 토픽 모델링으로 분석한 결과를 제시하였다. 총 8개의 토픽, Industray & Economy, Government Policy, Education, Heath, Information Service Technology, Medical Treatment, Life & Society, Impact Analysis이 도출되었다. 학술연구자들이 자신의 관심 연구분야의 토픽을 쉽게 파악할 수 있도록 토픽 모델링의 결과를 주요 연구 카테고리별로 분석하였고, 토픽별로 연구가 많이 이루어지는 세부 연구 카테고리 정보를 분석하여 제시했다. 또한 시계열 분해를 사용하여 토픽들의 추세(trend)를 분석하였다. 분석 결과 Government Policy와 Life & Society 토픽은 2020년까지 증가 추세였으나 2021년도에는 정체성을 보였고, 다른 6개의 토픽은 2021년까지 상승세를 보였다. 2022년 상반기부터는 Medical Treatment 토픽을 제외한 모든 토픽에서 감소 추세를 보였다. 특히 Education과 Life & Society 토픽의 감소 추세가 컸다.

향후 COVID-19 관련 국내 논문뿐만 아니라, 국외 논문, 언론 뉴스 기사 등을 추가로 분석하여 COVID-19 관련 연구 토픽에 영향을 미치는 요인과 상관관계 등을 분석하는 연구가 필요하다. 이러한 향후 연구는 연구자들이 COVID-19 관련 연구 토픽의 동향을 다각도로 이해하고, 파악하여 연구 주제를 찾는데 더 도움이 될 것으로 기대한다.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[2] Sang-Mi Kim, “Analysis of Press Articles in Korean Media on Online Education related to COVID-19", Journal of Digital Contents Society, 21(6), pp.1091-1100, 2020.

[3] Yoon Young Mi, Kim Seong Kwang, Kim Hye Kyeong, Kim Eun Joo, Jeong Yuneui, “Comparison of Topics Related to Nurse on the Internet Portals and Social Media Before and During the COVID-19 era Using Topic Modeling", Journal of Muscle and Joint Health, 27(3), pp.255-267, 2020.

[4] Jinsol Kim, Donghoon Shin, Hee-Woong KIM, “Analysis of Major COVID-19 Issues Using Unstructured Big Data”, Knowledge Management Research, 22(2), pp.145-165, 2021.

[5] Seong-Min Heo, Ji-Yeon Yang, “Analysis of Research Topics and Trends on COVID-19 in Korea Using Latent Dirichlet Allocation (LDA)”,

Journal of The Korea Society of Computer and Information, 25(12), pp.83-91, 2020.
[6] Seong-Min Heo, Ji-Yeon Yang, "A Convergence Study on the Topic and Sentiment of COVID19 Research in Korea Using Text Analysis", Journal of the Korea Convergence Society,12(4), pp.31-42, 2021.
[7] Kci, https://www.kci.go.kr/
[8] Eunhoe Kim, Yu Hwa Suh, "A Method of Calculating Topic Keywords for Topic Labeling", Journal of The Korea Society of Digital Industry and Information Management, 16(3), pp. 25-36, 2020.
[9] Carson Sievert and Kenneth E. Shirley, "LDAvis: A method for visualizing and interpreting topics", Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA, pp. 63-70, June 27, 2014.
[10] Time series decompose, https://otexts.com/fppkr/decomposition.html
[11] Time Series analysis tsa, https://www.statsmodels.org/stable/tsa.html
[12] Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I. J, "STL: A seasonal-trend decomposition procedure based on loess", Journal of Official Statistics, 6(1), pp. 3-33. 1990.

저자약력

김 은 희 (Eun-Hoe Kim) [정회원]

- 2013년 3월 ~ 현재 : 서일대학교 소프트웨어공학과 조교수
- 2007년 8월 ~ 2012년 2월 : 숭실대학교 정보미디어기술연구소, 지능형로봇연구소 전임연구원
- 2006년 8월 : 숭실대학교 컴퓨터학과 (공학박사)
- 1998년 8월 : 숭실대학교 컴퓨터학과 (공학석사)
- 1993년 2월 : 숭실대학교 전자계산학과(공학사)

〈관심분야〉 분산처리, IoT, 빅데이터, 그린네트워킹

서 유 화 (Yu-Hwa Suh) [정회원]

- 2019년 3월~현재 : 숭실대학교 베어드교양대학 조교수
- 2016년 3월~2019년 2월 : 서일대학교 정보통신공학과 조교수
- 2016년 2월 : 숭실대학교 컴퓨터학과(공학박사)
- 2007년 11월~2009년 10월 :정보통신산업진흥원 연구원
- 2005년 8월 : 숭실대학교 컴퓨터학과(공학석사)
- 2003년 2월 : 숭실대학교 컴퓨터학부(공학사)

〈관심분야〉 그린네트워킹, 유무선네트워크, 인공지능