JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 12, pp. 83-91, December 2020
https://doi.org/10.9708/jksci.2020.25.12.083

Analysis of Research Topics and Trends on COVID-19 in Korea Using Latent Dirichlet Allocation (LDA)

Seong-Min Heo*, Ji-Yeon Yang*

*Student, Dept. of Applied Mathematics, Kumoh National Institute of Technology, Gumi, Korea
*Associate Professor, Dept. of Applied Mathematics, Kumoh National Institute of Technology, Gumi, Korea

[Abstract]

This study aims to identify research topics and examine the trend of Covid19-related papers on DBpia. Applying latent Dirichlet allocation (LDA), we have extracted seven research topics, each of which concerns "International Dynamics", "Technology & Security", "Psychological Impact", "Biomedical-Related", "Economic Impact", "Online Education", and "Religion-Related". In addition, we used the multinomial logistic model to examine the trend of research topics. We found that the papers mainly cover topics related to "International Dynamics" and "Biomedical-Related" before June 2020, but the topics have become diverse since then. In particular, topics regarding "Economic Impact", "Online Education" and "Psychological Impact" has drawn increased attention of researchers. The findings would provide a guideline for collaboration in Covid19-related research, and could serve as a reference work for active research.

▸ **Key words**: Covid19, text mining, latent Dirichlet allocation, research topic, multinomial logistic model

[요 약]

본 연구에서는 DBpia에 등록된 코로나19 관련 논문을 대상으로 연구 토픽을 밝히고 연구 변화 추세를 검토한다. 잠재 디리슐레 할당(Latent Dirichlet Allocation) 알고리즘을 적용한 결과, 7개의 연구 토픽을 도출하였고, 각 토픽은 "International Dynamics", "Technology & Security", "Psychological Impact", "Biomedical-Related", "Economic Impact", "Online Education", "Religion-Related"에 관한 내용이었다. 또한 다범주 로짓모형을 사용하여 연구 토픽의 추세 변화를 살펴본 결과, 2020년 6월 전에는 국제적 역학관계 및 생물 의학 관련 논문이 주를 이루었다면, 이후에는 다양한 분야로 연구 주제가 확대되었다. 특히 경제적인 영향, 온라인 교육, 심리적인 영향에 관한 연구가 꾸준히 증가함을 확인할 수 있었다. 이러한 결과는 향후 코로나19 관련 공동 연구의 가이드라인을 제시하고, 활발한 연구 활동을 위한 기초자료로 활용될 수 있을 것이다.

▸ **주제어**: 코로나19, 텍스트 마이닝, 잠재 디리슐레 할당, 연구 토픽, 다범주 로짓모형

• First Author: Seong-Min Heo, Corresponding Author: Ji-Yeon Yang
*Seong-Min Heo (cjsm03@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology
*Ji-Yeon Yang (jyang@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology
• Received: 2020. 11. 18, Revised: 2020. 12. 07, Accepted: 2020. 12. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

21세기 들어 동물에서 유래한 바이러스성 호흡기 질환으로 전 세계가 패닉현상을 겪고 있다. 2003년 중국에서 시작하여 20개국 이상에서 퍼져나간 중증급성호흡기증후군(severe acute respiratory syndrome, SARS)의 경우 8,098명의 확진 환자와 774명의 사망자를 발생하였으며, 그 원인이 박쥐에서 나온 바이러스인 것으로 알려졌다[1]. 또 2009년에는 돼지독감 바이러스인 신종플루가 유행하여, 전 세계적으로 수많은 사망자가 발생하였다. [2]에 의하면 신종플루로 인한 사망자수는 최대 57만 5천 400명에 이를 수 있다고 밝혔다[2]. 2012년 사우디아라비아에서 처음 발견된 메르스(middle east respiratory syndrome, MERS)의 경우에는 현재까지 전 세계적으로 총 2,449명의 확진 환자와 858명의 사망자를 발생하였다[3]. 2015년 국내에도 유입되어 186명의 환자와 38명의 사망자를 초래했다[4].

2019년 12월에는 중국 후베이성 우한시에서 원인불명의 폐렴이 발병하였고 원인은 신종 코로나바이러스(severe acute respiratory syndrome coronavirus 2, SARS-CoV-2)로 밝혀지면서, 2020년 2월 세계보건기구(World Health Organization, WHO)는 이를 코로나바이러스 감염증-19(COVID-19, 코로나19)로 명명하였다[5]. 유례없는 감염 확산이 지속되자 WHO는 지난 3월 감염병의 전 세계적 대유행인 팬데믹을 선언한 바 있다[6]. 2020년 11월 8일 현재까지, 전 세계적으로 49,079,663명의 확진자와 1,238,680명의 사망자가 발생하였다[7]. 국내에서도 해외유입과 집단·지역사회의 감염이 지속되고 있으며, 예측하기 어려운 상황이 여전히 진행 중이다. 이러한 코로나19의 장기적인 대유행은 이제 의료, 보건당국 뿐 아니라 정치, 경제, 사회, 교육, 문화 등 다양한 분야에 영향을 미치고 있다.

이에 따라 다양한 분야에서 코로나19 관련 연구가 진행되고 있으며 연구 논문이 출판되고 있다. 임상 역학적 특성과 치료제와 백신 연구([8-10]) 뿐 아니라 사회-경제적(socio-economic), 심리-사회적(psycho-social), 기술적 함의 등을 다루는 연구 역시 활발하다([11-13]). 많은 다양한 연구가 쌓임에 따라 이를 통합하고 체계적으로 분석하는 메타분석 및 체계적인 문헌 고찰이 증가하고 있다. [14]에서는 코로나19 사망관련 인자들에 대한 기존 연구 결과를 계량적으로 병합하여 통합적인 추정치인 교차비(odds ratio)를 계산하는 방법을 제시하고 있다. 반면 [15]는 인공지능(AI)을 활용하는 기존 연구들의 결과를 기반으로, 코로나19 환자들의 X-ray, CT 등 영상 이미지를 생성하고 진단하는 데 필요한 파이프라인을 제공하고 있다. 토픽모델링 중 널리 사용되고 있는 LDA(Latent Dirichlet Allocation)를 이용하여 [16]에서는 펍메드(PubMed)와 아카이브(arXiv)에 올라온 코로나19 관련 논문들을 대상으로 다음과 같은 잠재토픽을 발견하였다. 개인보호장비(personal protective equipment), 종양학(oncology), 분석(analytics), 고위험군(high-risk groups), 공황장애 치료(rehabilitation-panic), 유전체학(genomics), 기관내 삽관-산소투여(Intubation- oxygenation). 반면 [17]은 2000년부터 2020년까지 SARS-CoV, MERS-CoV, COVID-19 등의 코로나 바이러스(coronavirus)와 관련하여 SCIE, SSCI 논문을 대상으로 문헌 조사를 하고 있다. 특히 코로나19 발발 이후의 논문들을 단백질 및 유전자 관련, 발생 관련, 감염 및 증상 관련, 호흡기 바이러스 관련된 논문으로 분류하고 있다. 반면 국내에서도 토픽모델링을 이용하여 코로나19 관련 토픽을 검토한 연구들이 있다 [18]은 LDA를 이용하여 2019년 12월부터 2020년 3월까지 보도된 뉴스데이터를 바탕으로 토픽모델링 분석을 실시하여 총 20개의 토픽을 도출하고 있다. [19]는 바이오아카이브(bioRxiv)와 메드아카이브(medRxiv)에 올라온 논문들을 대상으로 국가별 특징을 밝히고 LDA를 이용하여 역학조사 및 예측모델, 임상연구, 진단·치료제·백신개발연구 등의 3대 분야와 11개의 주요 토픽을 밝히고 있다.

이러한 연구들이 코로나19 관련 문헌 조사 및 토픽 발견에 기여하고 있지만, [16]과 [17]은 2020년 5월, [18]은 2020년 3월, [19]는 2020년 4월까지의 문헌을 분석 대상으로 삼고 있어, 최신 논문을 포함시켜 재분석을 시행할 필요가 있다. 특히 코로나19의 유행이 지속됨에 따라 관련 연구 주제도 확대되었으리라 예상되는 바, 연구 트렌드를 파악하기 위해서는 최근 연구를 포함한 재분석이 반드시 요구된다. 이에 본 연구에서는 지금까지 DBPia에 수록된 코로나19 관련 논문을 전체를 대상으로 국내 연구 트렌드와 토픽들을 분석하고 있다. 대부분의 기존의 연구들이 심사 전 공개 아카이브에 올라온 논문을 대상으로 하는 반면, 본 연구에서는 전문가 심사 후 게재가 이루어진 논문을 대상으로 한다는데 차이가 있다. 또 많은 기존 연구가 생물, 의학 등 특정 분야의 논문(PubMed, bioRxiv, medRxiv 등)들을 대상으로 한다면, 본 연구에서는 분야를 한정시키지 않고 코로나19 관련 논문을 모두 대상으로 한다는 데서 의의가 있다. 코로나19가 여러 분야에 영향을 미치고 있는 상황에서, 이와 같은 다양한 연구 분야로의 확장 분석은 전반적인 연구 트렌드를 발견하고 코로나19를 대응하기 위한 학제 간 공동 연구의 가이드라인을 제시할 수 있을 것이다.

본 연구의 구성은 다음과 같다. 제2장에서는 자료의 수집과 분석 방법, 그리고 연구에 사용된 토픽 모형에 대해 기술하고 있다. 제3장에서는 토픽 모델링의 결과를 제시하고 다범주 로짓모형을 활용하여 연구 토픽이 시간에 따라 어떻게 변화하고 있는지 검토한다. 마지막으로 제4장에서는 본 연구의 결론으로 연구의 의의와 향후 연구방향에 대해 논의한다.

II. Research Methodology

1. Data Collection

본 논문에서는 웹 크롤링을 이용하기 위해 분석 도구 R(ver 3.6.3)을 사용하였다. 특히 동적인 웹 사이트를 원격으로 조작하는 방식으로 데이터를 수집할 때 이용되는 "RSelenium" 패키지와 웹 페이지의 html 구조에 맞춰 다양한 데이터를 수집할 때 이용되는 "rvest" 패키지를 활용하였다. 기타 HTTP 통신과 관련된 패키지인 "httr", html 문서 작성 관련 패키지인 "htmltools"와 "htmlwidgets", JSON 파서 및 생성기 관련 패키지인 "jsonlite", 인코딩 관련 패키지인 "urltools", "readr", "base64enc", 텍스트 전처리 패키지인 "stringr" 등을 웹페이지의 특성에 맞게 적절하게 적용한다면 DBpia 이외 Google 학술 검색, Pubmed 등에서 관련 논문을 검색하고 크롤링하는 것이 크게 어렵지 않으리라 본다.

본 연구 분석에서는 다양한 국내 학술지를 검색할 수 있는 디비피아(DBpia, http://www. dbpia.co.kr/)에서 "코로나19"와 "covid19"를 키워드로 검색하여 수집된 논문을 이용하였다. 전문잡지, 연구보고서 등 여러 자료 유형 중에서 학술 저널을 선택하여 수집하였다. 2020년 10월 10일자에 사회과학 273개, 공학 76개, 의약학 70개, 인문학 63개, 복합학 31개, 예술체육학 16개, 자연과학 13개, 농수해양학 1개 총 543개의 논문이 검색되었고 그 중 초록이 없는 논문, Corona(광환)와 관련된 논문, 한글초록만 있는 논문, 중복되는 논문 등 관련 없는 253개의 논문을 제거하여 290개의 논문에서 제목, 영문초록, 키워드, 학회지, 학회지 종류, 연도 등을 추출하였다.

데이터 전처리 과정으로 구두점, 공백, 기호, 불용어를 제거하였고, 추가적으로 소문자 변환, 유사어, 파생어를 같은 품사로 변환하였다. 반면 출현 빈도가 매우 높지만 분석 결과에는 영향을 미치지 않는 단어들(coronavirus, disease, korea 등)은 사전에 제거하였고 너무 적게 출현해 의미가 없는 빈도수 10 이하의 단어들도 제거하였다. 그 결과, 초록에서 566개의 단어를 얻었다.

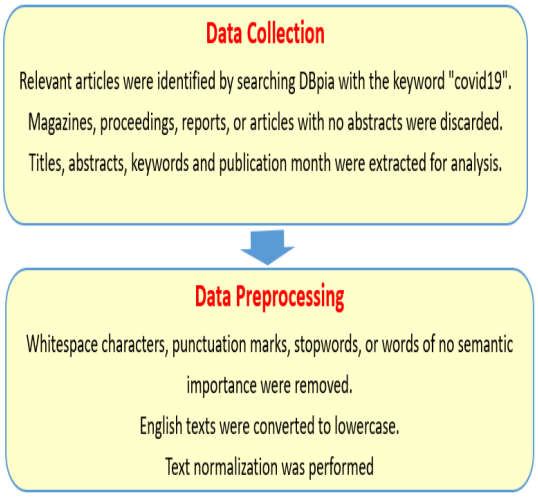

Fig. 1. Data Collection and Data Preprocessing

2. Topic Model

본 연구에서 연구 토픽을 분석하기 위해 사용된 토픽 모델링은, 텍스트 데이터의 숨겨진 토픽을 자동으로 추출하기 위한 방법으로 텍스트에서 사용된 단어들의 등장 패턴을 바탕으로 문서 내에 어떤 토픽이 포함되어 있고 토픽 간 비중이 어떤지를 모형화하여 살펴본다[20]. 최근 대용량의 비정형화된 텍스트 데이터가 증가하면서, 문서의 잠재적인 의미 구조를 발견하고 문서 집합을 자동 분류하는 데 유용하게 쓰이고 있다.

이 토픽모델링을 실제로 실행하기 위해 본 연구에서는 LDA를 활용하고 있다. LDA는 토픽 내 단어 분포와 문서 내 토픽 분포의 결합으로 특정 문서 내 단어들이 생성된다고 가정한다[21-22]. 토픽을 선택하는 과정 및 토픽 내 단어를 선택하는 과정은 다항(multinomial) 분포를 따른다고 가정하고, 켤레(conjugate) 관계인 디리슐레(Dirichlet)를 각 문서에 포함된 주제의 사전분포로 사용한다. 추론의 대상인 베이지언 사후분포는 보통 해석적으로 해를 찾을 수가 없기 때문에, 사후분포에서 추정하고자 하는 모수들을 조건부 확률로부터 순차적으로 추출하는 깁스 샘플링(Gibbs sampling)이 사용되고 있다[23].

토픽 모델링은 텍스트 마이닝에서 연구에서 많이 사용하는 방법으로, 대표적인 방법에는 LDA, sLDA, HDP 토픽모델, PAM, STM 등이 있다. 본 연구에서는 LDA 알고리즘을 적용하여 적당한 주제를 찾고자 한다. LDA(Latent Dirichlet Allocation, 잠재 디리슐레 할당)는 다양한 문서에서 각 문서에 대한 잠재적인 주제를 찾아내기 위한 절차적 확률 분포 모델이다[20]. 토픽의 단어 비중과 문서의 토픽 비중이라는 두 가지 변수의 결합 확률분포에 따라 문서의 토픽을 찾는 과정이며, 0과 1사이의 값을 가지는 연속

다변수 확률분포인 디리슐레(Dirichlet) 분포를 따른다[24]. LDA 방법에서 문서를 생성하는 방식을 다음과 같이 가정하고 있다. 우선 K개의 토픽을 선택한다. 그 다음, 어떤 분포를 따르는 특정한 토픽이 무작위로 선택되며, 각각의 토픽들은 다양한 단어가 출현할 확률을 갖고 있다. 이때 문서는 확률 분포에 따라 특정한 단어들이 선택되어 생성되는 방식으로 구성된다[25]. 이 내용을 수식으로 표현하면 Fig. 2와 같다. $p(t|d)$는 문서 d의 토픽 t에 대한 비중을 뜻하며 ($\sum t\, p(t|d) = 1$), $p(w|t)$는 토픽 t에 등장하는 단어 w의 비중을 뜻한다($\sum w\, p(w|t) = 1$). 두 확률을 결합한 $p(w|d)$는 문서 d에서 어떤 단어 w가 등장할 것인지에 대한 비중을 뜻하며, 여기에서 높은 비중을 가지는 단어들의 토픽이 해당문서의 토픽으로 할당된다[26].

$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d)$$

Fig. 2. Latent Dirichlet Allocation Model

3. Topic Number Selection

토픽 모형을 적용하기 위해서 사전에 토픽의 개수를 결정할 필요가 있다. 토픽의 개수가 너무 많으면 지나치게 세분화된 주제들이 나타날 가능성이 높은 반면, 토픽의 개수가 너무 적으면 여러 주제가 섞여 특성이 모호해질 수 있다. 이에 도출된 토픽의 유용성과 해석가능성에 따라 토픽의 개수를 결정할 필요가 있다[21-22]. 토픽 수의 결정을 위한 여러 연구들 중, [27]과 [28]은 각각 로그 우도의 조화 평균 값과 토픽 분포간의 젠슨-샤논 거리가 최대화되도록, [29]와 [30]은 각각 토픽 분포 간의 코사인 유사도와 토픽-단어 행렬로부터 구한 특이값 분포의 대칭적 쿨백-라이블러 거리가 최소화되도록 토픽 수를 선택한다.

본 연구에서도 이 네 가지의 측도를 활용하여 타당하면서도 해석이 용이하도록 적절한 토픽의 개수를 정하는 것을 목표로 한다. 토픽의 개수를 선택한 후 영문 초록에 LDA를 적용하여 토픽을 도출하고, 출현 단어를 검토하여 잠재 토픽의 특성을 반영할 수 있도록 토픽의 이름을 정하고 있다.

4. Research Trend Analysis

지난 2월부터 현재까지 기간은 비교적 짧지만, 코로나19의 대유행은 여러 분야의 연구자들의 위기의식을 고취하고 연구 참여를 유도하였으리라 예상된다.

이에 관련 연구의 토픽의 변화가 어떻게 이루어지고 있는지 다범주 로짓모형을 이용하여 살펴보고자 한다. 6월에 게재된 논문의 수가 급증한 점을 감안하여 6월을 기준으로 전, 후로 나눈 후, 각 토픽이 등장하는 패턴의 변화를 검토한다.

III. Results

1. Number of Publications

본 연구의 분석대상은 2020년 10월 10일 기준으로 DBpia에서 검색되는 코로나19 관련 총 543편의 논문 중 앞 장에서 기술한 기준을 적용하여 추출한 290편의 논문이다. Fig. 3에서 보듯이, 상당수의 논문이 2020년 6월에 게재되었는데, 108편(37.2%)의 논문이 이 때 게재되었다. 2020년 1월부터 5월까지 게재된 논문은 79편(27.2%), 7월 이후 게재된 논문은 103편(35.5%)이다.

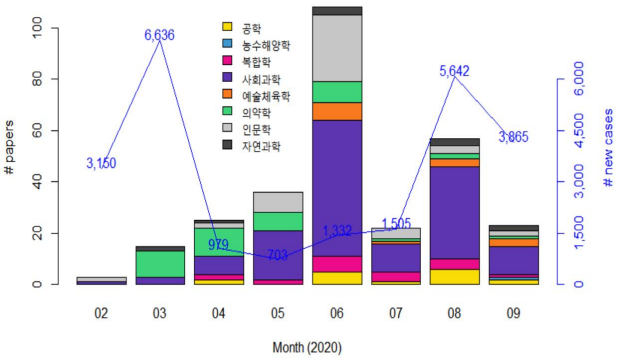

Fig. 3. The number of publications and the number of COVID-19 cases over time in Korea

특이한 점은 의약학으로 분류되는 학회지에 게재된 논문의 개수가 3월 이후 증가하다가 7월 이후 크게 감소한다는 점이다. 물론 진단, 백신 및 치료제 개발, 유전자 관련 연구는 의약학 뿐 아니라 생물학, 생물공학 등의 분야에서도 많이 이루어지고 있으리라 예상되어, 학회지 분류가 정확한 기준이 되기 어렵다.

반면 사회과학으로 분류되는 학회지에 게재된 논문 수는 꾸준히 증가하여 6월 이후에는 50% 이상을 차지한다. 하지만 사회과학에는 다양한 세부 분야가 포함되어 있어서 구체적으로 어떠한 주제들이 논의되고 있는지 학회지 분류만으로 파악하기는 무리가 있다.

대상이 되는 총 186개의 학회지 중에서 주요 학회지를 기간별로 살펴보면 다음과 같다. 5월 이전에는 30% 가량의 논문이 "예방의학회지", "보건교육건강증진학회지", "Pediatric Infection and Vaccine", "Infection and Chemotherapy", "아시아연구" 에 게재되어, 이 기간에는 주로 의학 또는 아시아 지역학 관련 연구가 이루어졌음을

알 수 있다. 6월에는 비교적 다양한 분야의 학회지가 등장하여, 해당 기간의 약 25%의 논문이 "JOURNAL OF BACTERIOLOGY AND VIROLOGY", "중국지식네트워크", "언어와정보", "한국스포츠학회", "신학과세계", "법학연구", "The Journal of Asian Finance, Economics and Business"에 게재되었다. 코로나19 관련 연구가 의학, 지역학, 스포츠학, 종교, 법학, 경제학 등으로 확대되었음을 확인할 수 있다. 이러한 추세는 7월 이후에도 계속 이어져, "통상법률", "한국엔터테인먼트산업학회논문지", "한국컴퓨터정보학회논문지", "교양교육연구", "인적자원개발연구", "한국산학기술학회논문지", "한국예술연구", "한국콘텐츠학회논문지" 등의 학회지에 25%의 논문이 게재되었다.

한편 질병관리본부에서 발표하고 있는 국내 월별 코로나19의 확진자 수를 검토하였다. 올 3월에 6,000명 이상의 확진자가 발생하였고, 이후 주춤하다가 8월에 다시 증가하였다. 논문을 작성하고 심사, 게재하기까지의 시간을 고려하면, 3월의 1차 유행은 6월에 크게 증가한 논문 수를 어느 정도 설명한다. 물론 저자마다 집필 기간이 다르고 학회지마다 심사 기간이 다르기 때문에 확진자 수와 게재 논문 수와의 상관관계를 정확히 파악하기는 어렵지만, 평균적으로 2~3개월의 지연시간(lag time)을 가정한다면 위기 상황이 발생한 시점에 다양한 분야의 연구자들이 현 위기를 인식하고 적극적으로 대응방안을 모색하고 있는 것으로 해석할 수 있을 것이다.

2. Number of Topics

연구 토픽의 개수를 선택하기 위해 앞 장에서 살펴본 네 가지 측도를 검토하였다. 영문 제목 또는 영문 키워드에는 포함된 단어가 그렇게 많지 않아 충분한 정보를 활용하기 위해 영문 초록을 이용하였다. Fig. 4에서는 토픽의 개수에 따른 네 가지 측도의 값을 보여주고 있다. 네 가지 측도 모두 0과 1사이의 값을 갖도록 표준화시킨 값이다.

Cao et al.[29]와 Arun et al.[30]에서는 측도의 값이 작을수록, Griffiths et al.[27]와 Deveaud et al.[28]는 측도의 값이 클수록 최적이지만, 이 모든 기준을 만족하는 토픽 수는 없는 것으로 보인다. 단 Cao의 측도의 경우, 개수가 7일 때 크게 감소하다가 8이 되면 다시 증가하여 극소적인 최솟값을 보인다. Deveaud의 측도 역시 개수가 7일 때 증가하다가 8이 되면 감소하는 것을 확인할 수 있다. 이러한 결과와 기존 코로나19 관련 연구의 결과를 감안하여 토픽의 개수 7을 선택하였다.

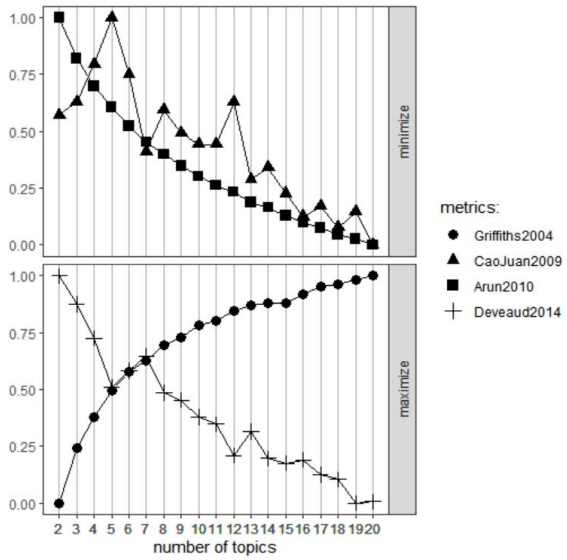

Fig. 4. The number of topics indicated by four metrics

3. Topic Identification

토픽의 개수를 7개로 선택한 후, 영문 초록에 LDA 알고리즘을 적용하였다. Fig. 5는 각 토픽별 상위 30개의 단어에 대한 워드 클라우드이다. 첫 번째 토픽에 등장하는 주요 단어는 china, global, government, policy, country, international, nation, strategy, world, economy, crisis, disaster, change 등으로 나타나, 세계적인 위기와 변화에 따른 국제적 역학관계가 주된 내용으로 판단된다. 이에 "International Dynamics"라고 이름 붙였다.

두 번째 토픽에서는 medical, information, datum, work, protect, personal, technology, smart, law, regulation, prevent, security, legal 등의 단어가 많이 등장하고 있다. 코로나19 발생 이후에 의료, 직장 등에서의 정보통신기술 변화와 그에 따르는 개인정보 보안 및 법적 보호에 관한 내용으로 보인다. 그래서 두 번째 토픽을 "Technology & Security"라 지정하였다.

반면 세 번째 토픽에서는 effect, datum, behavior, stress, psychological, program, level, conduct, affect, anxiety, experience, identity, coping 등의 단어로 유추해볼 때 이 토픽에 속하는 논문들은 주로 코로나19로 인한 심리, 감정적인 변화를 연구하고 있다. 이에 "Psychological Impact"라고 지정하였다.

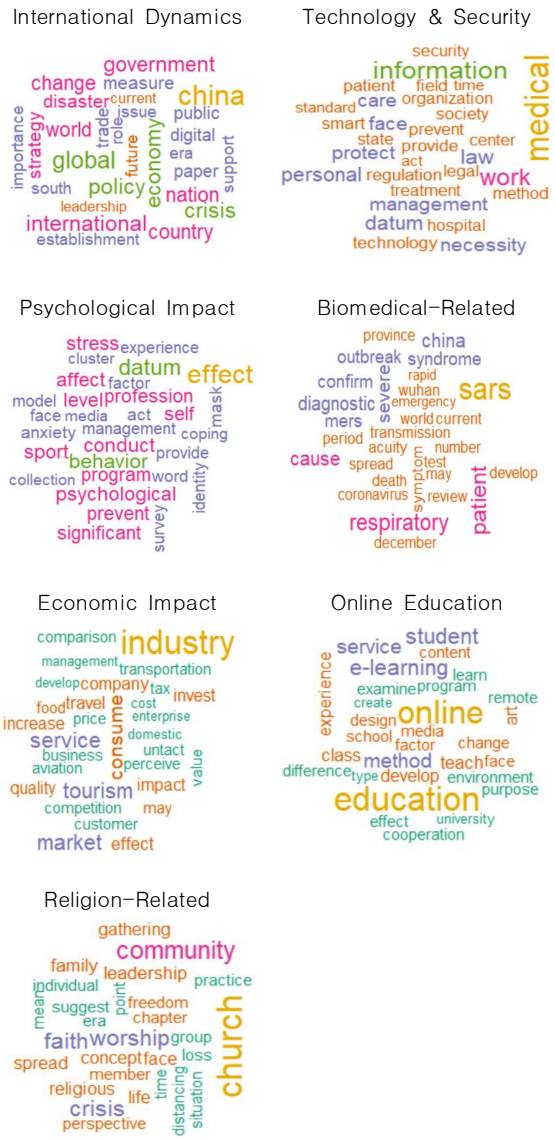

Fig. 5. Word clouds of keywords by topic using LDA

네 번째 토픽에서 많이 등장하는 단어는 sars, patient, respiratory, cause, diagnostic, severe, outbreak, confirm, symptom, syndrome, mers, transmission, spread, death, test, coronavirus로 코로나 바이러스의 특징, 전파, 역학, 임상 양상 등의 내용으로 추측된다. 따라서 이 토픽을 "Biomedical-Related"라 이름 붙였다.

다섯 번째 토픽의 경우, industry, market, tourism, service, company, invest, consume, food, travel, impact 등의 단어가 주로 나타나, 코로나19가 경제 또는 산업에 미치는 영향에 관한 내용임을 유추할 수 있다. 이에 "Economic Impact"라고 지정하였다.

여섯 번째 토픽은 education, online, student, e-learning, service, method, teach, class, change, content, school 등의 단어를 포함하여 코로나19에 따른 온라인 수업(Online Education)으로 명명하였다.

마지막으로 일곱 번째 토픽에서 등장하는 주요 단어는 church, community, worship, faith, religious, chapter, spread, gathering 등으로 종교 모임과 신앙생활과 관련된 내용으로 추측된다. 이에 "Religion-Related"로 이름 붙였다.

4. Research Trend by Topic

코로나19 관련 논문의 게재가 급증한 2020년 6월을 전후로 해서 연구 토픽의 추세 변화가 있는지를 살펴보았다. Fig. 6은 6월 전, 6월, 6월 후 각 시점에 대해서 7개의 연구 토픽이 나타날 확률을 보여준다. 반면 Table 1에서는 설명변수를 시간 더미로, 반응변수를 연구토픽으로 둔 다범주 로짓모형의 결과이다. 시간 더미에서 6월이 비교 시점(reference time period)이다.

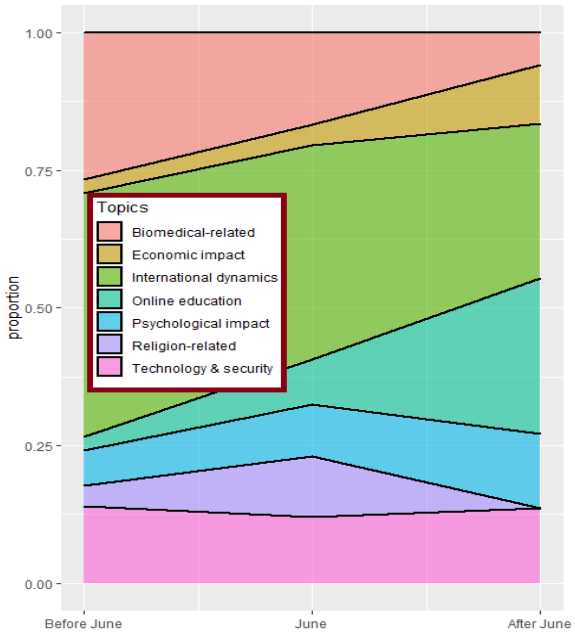

Fig. 6. Topic probability at each time period

Fig. 6과 Table 1을 보면, 생물 의학 관련 연구는 계속 감소하는 추세이다. 반면 경제적인 영향과 관련된 연구는 계속 증가하는데, 6월 후의 증가는 통계적으로도 유의하였다(p값=0.029).

국제적 역학관계와 관련된 연구는 계속해서 감소하는 추세이기는 하지만, 각 시점에서 연구 주제의 상당 부분을 차지하고 있다. 6월 전에는 44%, 6월에는 38.9%로, 6월 후에는 28.2%의 논문이 국제적 역학 관련 연구로, 온라인 교육과 더불어 여전히 가장 많은 연구진들의 관심을 끌고 있다. 온라인 교육 관련 연구는 6월 이후 크게 증가하여(p

값=0.001) 새로운 교육 패러다임에 대한 연구가 활발히 이루어지고 있음을 확인할 수 있다.

심리적인 영향과 관련된 연구도 지난 2월 이후 꾸준히 증가하고 있지만, 그 증가 추세가 통계적으로 유의하지는 않다. 종교 관련 연구는 6월에 크게 증가하였다가(p값=0.079), 그 이후 다시 감소하고 있다. 아마 초기 지역사회의 감염과 관련하여, 종교의 역할 및 방향에 관한 논의가 이루어지지 않았나 추측된다. 반면 정보통신기술과 보안 관련 연구는 큰 변화 없이 12%~14%대의 연구를 차지하고 있다 (6월 전 13.9%, 6월 12.0%, 6월 후 13.6%).

Table 1. Results of multinomial logistic regression. The reference topic is "nter national Dynamics".

	Estimates	P-values
Technology & Security		
Intercept	-1.173	<0.001
Before June	0.015	0.974
After June	0.444	0.328
Psychological Impact		
Intercept	-1.435	<0.001
Before June	-0.511	0.389
After June	0.707	0.140
Biomedical-Related		
Intercept	-0.847	0.003
Before June	0.336	0.394
After June	-0.728	0.169
Economic Impact		
Intercept	-2.351	<0.001
Before June	-0.511	0.568
After June	1.382	0.029
Online Education		
Intercept	-1.540	<0.001
Before June	-1.322	0.105
After June	1.540	0.001
Religion-Related		
Intercept	-1.253	<0.001
Before June	-1.204	0.079
After June	-14.736	0.979

IV. Conclusions

코로나19는 우리 삶의 많은 부분을 바꾸어 놓고 있다. 전 세계인의 건강을 위협할 뿐 아니라 정치 및 경제의 위기, 심리적 불안, 교육 공백 등 여러 부면에서 영향을 미치고 있다. 이에 다양한 분야의 연구진들이 현 위기 상황을 분석하고 대응 전략을 마련하는 노력을 기울이고 있다. 특히 올 6월에는 코로나19 관련 논문이 무려 108편이나 DBpia에 등록되어 연구진들의 적극적인 위기 대응과 전략 마련을 엿볼 수 있다.

본 연구에서는 이러한 연구진들의 노력의 산물인 연구 논문을 이용하여, 코로나19 관련 연구의 토픽들을 밝히고 각 토픽들이 어떠한 추세인지를 보이는지를 검토하였다. 토픽 모델링 기법 중 LDA 알고리즘을 DBpia에 등록된 국내 논문 290편에 적용한 결과 총 7개의 연구토픽을 발견하였고, 토픽 내 주요 단어들을 검토함으로써 각 토픽은 "International Dynamics", "Technology & Security", "Psychological Impact", "Biomedical-Related", "Economic Impact", "Online Education", "Religion-Related"에 관한 내용임을 밝힐 수 있었다. 또한 다범주 로짓모형을 사용하여 연구 토픽의 추세 변화를 살펴보았다. 주목할 만 한 점은 6월 전에는 국제적 역학관계 및 생물 의학 관련 논문이 주를 이루었다면, 이후에는 다양한 분야로 연구 주제가 확대되었다. 특히 경제적인 영향, 온라인 교육, 심리적인 영향에 관한 논문이 꾸준히 증가함을 확인할 수 있었다.

토픽모델링 특히 LDA를 사용하여 코로나19 관련 연구 토픽 및 동향을 탐색하고 있는 선행연구들이 문헌 조사 및 토픽 발견에 기여하고 있지만, 이들은 2020년 상반기까지의 문헌을 대상으로 하기 때문에 최근 문헌을 포함시켜 재분석을 시행할 필요가 있다. 본 연구에서는 비교적 최근까지의 문헌을 분석대상으로 하여 그 결과를 제시하고 있다는 점에서 독창성이 있다 하겠다. 또한 대부분의 기존 연구들이 심사 전 아카이브에 올라온 논문을 대상으로 하지만, 본 연구에서는 전문가 심사 후 게재가 이루어진 논문을 대상으로 한다는데 차이가 있다. 그리고 기존 논문과 달리 생물, 의학 등 특정 분야의 문헌에 한정시키지 않고 본 연구에서는 코로나19 관련 국내 논문을 모두 대상으로 한다는 데 의의가 있다. 특히 코로나19의 유행이 장기화됨에 따라 여러 부면에서 관련 연구들이 이루어지고 있으리라 예상되기 때문에, 본 연구에서처럼 분석 문헌의 확대는 반드시 필요하다고 판단된다.

본 연구에서는 코로나19 관련 연구 토픽을 발견하고 시기별 트렌드를 분석하고 있다. 이는 앞으로 코로나19 관련 연구 전략을 수립하는 데 필요한 기초자료로 활용되고, 관련 정부 정책의 방향성을 제시하는 데 기여할 수 있을 것이다. 특히 본 연구 결과는 최근에 상대적으로 주춤한 생물 의학 관련 국내 연구를 활성화시킬 수 있는 방안이 모색되어야 함을 시사한다. 치료제나 백신개발 관련 연구 뿐 아니라, 의료 현장에서의 역학과 임상 연구 등이 지속적으

로 쌓일 때 신뢰성 있는 결과를 확보할 수 있으며 효과적인 의료지침을 개발할 수 있다. 이를 위해 향후 연구지원 및 인력확보 정책이 필요할 것으로 생각된다. 또한 국가 위기 상황에서 이를 극복하기 위한 다양한 연구의 활성화가 절실히 요구되는 지금, 본 연구 결과는 전반적인 연구 트렌드를 확인하고 학제 간 공동 연구의 가이드라인을 제시할 수 있을 것이다. 특히 후속 연구 주제를 결정하는 데 기초자료로 쓰일 것으로 기대된다.

보다 정확한 연구 동향을 파악하기 위해 향후 더 많이 축적된 관련 문헌을 바탕으로 시계열 분석이 가능하겠다. 만일 국외 논문까지 포함시킬 수 있다면 방대한 양이겠지만, 코로나19 관련 연구의 세계적인 추세를 살펴보고, 국가별 특징을 살펴볼 수 있어 흥미로운 연구 과제가 될 것이다. 또한 분석 대상을 논문에 한정시키지 않고 SNS, 언론 기사 등으로 확대하여 분석한다면 사회적으로 형성되어 있는 관련 의제를 파악하는 데 도움이 될 것이다. 추가적으로 본 논문에서 LDA를 이용하여 발견한 토픽을 다른 토픽모델링으로 검증하는 작업과 토픽 그룹별로 감성분석을 시행함으로써 분야별 연구진들의 감성 강도 비율을 파악하는 작업을 수행할 수 있을 것이다.

마지막으로 이 코로나19 위기를 효과적이고 지혜롭게 극복하기 위해 다각도 연구 결과의 융합 및 연구진들의 협력이 절실히 필요함을 강조하고 싶다.

ACKNOWLEDGEMENT

This research was supported by Kumoh National Institute of Technology (202001950001).

REFERENCES

[1] Centers for Disease Control and Prevention(CDC), https://www.cdc.gov/sars/

[2] F. S. Dawood, A. D. Iuliano, C. Reed, et al., "Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study," The Lancet infectious diseases, Vol. 12, No. 9, pp. 687-695, Sep. 2012. DOI: 10.1016/S1473-3099(12)70121-4

[3] World Health Organization, "Middle East respiratory syndrome coronavirus (MERS-CoV) - The Kingdom of Saudi Arabia," Jul. 2019. https://www.who.int/csr/don/ 24-july-2019-mers-saudi-arabia/en/

[4] Korea Disease Control and Prevention Agency, http://www.cdc.go.kr/contents.es?mid=a20301020709

[5] World Health Organization, "WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020," Feb. 2020. https://www.who.int/dg/speeches/detail/who- director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020

[6] World Health Organization, "WHO Director-General's Opening Remarks at the Media Briefing on Covid-19 - 11 March 2020," Mar. 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[7] K. F. Yuen, X. Wang, F. Ma, and K. X. Li, "The Psychological Causes of Panic Buying Following a Health Crisis," International Journal of Environmental Research and Public Health, Vol. 17, No. 10, pp. 3513-3536, May 2020. DOI: 10.3390/ijerph17103513

[8] N. Zhu, D. Zhang, W. Wang, et al., "A novel coronavirus from patients with pneumonia in China, 2019," New England Journal of Medicine, Vol. 382, No. 8, pp. 727-733, Feb. 2020. DOI: 10.1056/NEJMoa2001017

[9] S. P. Kaur and V. Gupta, "COVID-19 Vaccine: A comprehensive status report," Virus Research, Vol. 288, Oct. 2020. DOI: doi: 10.1016/j.virusres.2020.198114

[10] Y. Jeong, "2019 Novel Coronavirus Disease Outbreak and Molecular Genetic Characteristics of Severe Acute Respiratory Syndrome-Coronavirus-2," Journal of Bacteri- ology and Virology, Vol. 50, No. 1, pp. 1-8, Jan. 2020. DOI: 10.4167/jbv.2020.50.1.001

[11] M. Nicola, Z. Alsafi, C. Sohrabi, et al., "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," International Journal of Surgery, Vol. 78, pp. 185-193, Jun. 2020. DOI: 10.1016/j.ijsu.2020. 04.018

[12] D. Banerjee and M. Rai, "Social isolation in Covid-19: The impact of loneliness," International Journal of Social Psychiatry, Vol. 66, No. 6, pp. 525-527, Sep. 2020. DOI: 10.1177/0020764020922269

[13] S. Jun and J. Kim, "Theoretical Background and Prospects for the Untact Industry," Journal of New Industry and Business, Vol. 38, No. 1, pp. 96-116, Jun 2020. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09405985

[14] W. Kim, J. Han, and K. E. Lee, "Predictors of Mortality in Patients with COVID-19: A Systematic Review and Meta-analysis," Korean Journal of Clinical Pharmacy, Vol. 30, No. 3, pp. 169-176, Sep. 2020. DOI: 10.24304/kjcp. 2020.30.3.169

[15] F. Shi, J. Wang, J. Shi, et al., "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19", IEEE Reviews in Biomedical Engineering, Apr. 2020. DOI: 10.1109/RBME. 2020.2987975

[16] A. Ebadi, P. Xi, S. Tremblay, et al., "Understanding the temporal evolution of COVID-19 research through machine learning and

natural language processing," arXiv preprint arXiv:2007.11604, 2020.

[17] H. Zhang and R. Shaw, "Identifying Research Trends and Gaps in the Context of COVID-19," International Journal of Environmental Research and Public Health, Vol. 17, No. 10, pp. 3370-3386, May 2020. DOI: 10.3390/ijerph171033 70

[18] T. Kim, "COVID-19 News Analysis Using News Big Data : Focusing on Topic Modeling Analysis," The Journal of the Korea Contents Association, Vol. 20, No. 5, pp. 457-466, May 2020. DOI: 10.5392/JKCA.2020.20.05.457

[19] Korea Institute of Science and Technology, "KISTI DATA INSIGHT," Vol 12, Apr. 2020. http://mirian.kisti.re.kr/ insight /insight.jsp

[20] D. M. Blei, "Probabilistic topic models," Communications of the ACM, Vol. 55, No. 4, pp. 77-84, Apr. 2012. DOI: 10.1145/2133 806.2133826

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, Vol. 3, pp. 993-1022, Jan. 2003. DOI: 10.1162/jmlr.2003.3.4-5.993

[22] J. Y. Yang, "Convergence Study on Research Topics for Thyroid Cancer in Korea," Journal of the Korea Convergence Society, Vol. 10, No. 2, pp. 75-81, Feb. 2019. DOI: 10.15207/JKCS.201 9.10.2.075

[23] G. Casella and E. I. George, "Explaining the Gibbs sampler," The American Statistician, Vol. 46, No. 3, pp. 167-174, Feb. 2012. DOI: 10.1080/00031305.1992.10475878

[24] S. Yoon and M. Kim, "Topic Modeling on Fine Dust Issues Using LDA Analysis," Journal of Energy Engineering, Vol. 29, No. 2, pp. 23-29, May. 2020. DOI: 10.5855/ENERGY. 2020.29.2.023

[25] J. Sim and H. Kim, "A Searching Method for Legal Case Using LDA Topic Modeling," Journal of the Institute of Electronics and Information Engineers, Vol. 54, No. 9, pp. 67-75, Sep. 2017. DOI: 10.5573/ieie.2017.54.9.67

[26] S. Moon, S. Chung, and S. Chi, "Topic modeling of news article about international construction market using latent Dirichlet allocation," Journal of the Korean Society of Civil Engineers, Vol. 38, No. 4, pp. 595-599, Aug. 2018. DOI: 10.12652/Ksce.2018. 38.4.0595

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics. Proceedings of the National academy of Sciences," Vol. 101, No. suppl 1, pp. 5228-5235, Apr. 2004. DOI: 10.1073/pnas.0 307752101

[28] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," Document numérique, Vol. 17, No. 1, pp. 61-84, Jun. 2014. DOI: 10.3166/DN.17.1.61-84

[29] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," Neurocomputing, Vol. 72, No. 7, pp. 1775-1781, Mar. 2009. DOI: 10.1016/j.neucom .2008.06.011

[30] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," Pacific-Asia Conference on Knowledge Discovery and Data Mining, Vol. Part I, pp. 391-402, Jun. 2010. DOI : 10.1007/978-3-642-13657-3_43

Authors

Seong-Min Heo is an undergraduate student of Applied Mathematics at Kumoh National Institute of Technology, Korea. He has been working as an undergraduate researcher in Applied Statistics Laboratory, Kumoh National Institute of Technology since 2017. His research interest is the application of statistical methods, such as text mining, data mining and machine learning, in big data analytics.

Ji-Yeon Yang received MS and PhD degrees in Statistics from University of Illinois at Urbana-Champaign, in 2006 and 2010, respectively. She joined the faculty in 2014 and is currently an Associate Professor of the Department of Applied Mathematics at Kumoh National Institute of Technology, Gumi, Korea. She is interested in big data analytics, Bayesian analysis and computational statistics.