Journal of Digital Convergence
Vol. 19. No. 12, pp. 11-20, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.12.011

뉴스 빅데이터를 통해 검토한 대학교육의 토픽 분석

양지연[1], 구정호[2*]
[1]금오공과대학교 응용수학과 부교수, [2]금오공과대학교 경영학과 교수

A Topic Analysis of College Education Using Big Data of News Articles

Ji-Yeon Yang[1], Jeong-Ho Koo[2*]
[1]Associate Professor, Dept. of Applied Mathematics, Kumoh National Institute of Technology
[2]Professor, Dept. of Business Administration, Kumoh National Institute of Technology

요 약 본 연구는 신문기사 빅데이터를 통해 대학교육 관련 보도의 토픽을 추출하고, 토픽별 특징 및 신문사별 보도양상을 분석한다. 2016년-2021년 상반기 주요 중앙지와 지역지의 기사를 빅카인즈를 통해 추출하였고, 잠재디리슐레할당을 이용하여 총 9개의 토픽을 발견하였다. 토픽1과 토픽3은 교육에 대한 대학지원사업에 관련된 것이나 토픽3은 지역대학에 초점이 맞추어져 있다. 토픽2는 코로나19 이후 대학교육, 토픽4는 교수-학습법, 토픽5는 정부정책, 토픽6은 고교교육기여대학 지원사업, 토픽7은 대학교육 비전, 토픽8은 국제화, 토픽9는 입시 등을 논하고 있다. 조선일보, 경향신문, 한겨레는 코로나19 이후 강의, 정부정책 관련, 대학교육에 대한 기사와 논평을 많이 보도한 반면 동아일보, 중앙일보, 한라일보, 부산일보, 대전일보, 경인일보는 대학지원사업, 고교교육기여대학 지원사업 등 광고·홍보성 기사가 상대적으로 많았다. 2016년부터의 관련기사를 신문사별 뿐 아니라, COVID-19 발생 전후로도 분석하여 관련 보도의 토픽 차이를 살펴볼 수 있었다. 사회적으로 주요 관심 사항인 대학교육이 언론에 어떻게 보도되고 있는지 확인함으로써 미래의 대학교육 정책 방향과 미디어의 순기능과 역기능 등 언론의 역할에 대해 고찰할 필요가 있음을 시사한다.

주제어 : 대학교육, 신문기사, 토픽모델링, 잠재디리슐레할당, 텍스트마이닝

Abstract This study extracts topics related to university education through newspaper articles and analyzes the characteristics of each topic and the reporting patterns of each newspaper. The 9 topics were discovered using LDA. Topic 1 and Topic 3 are related to university support projects for education, but Topic 3 is focused on local universities. Topic 2 is about university education after COVID-19, Topic 4 teaching-learning methods, Topic 5 government policies, Topic 6 the high school education contribution university support projects, Topic 7 the university education vision, Topic 8 internationalization, and Topic 9 the entrance exam. The Chosun Ilbo, Kyunghyang, and Hankyoreh reported a lot of articles associated to lectures after COVID-19, government policies, and comments on university education. Relevant articles since 2016 have been analyzed by newspaper type and before/after COVID-19 through which differences in the topics were studied and discussed. These findings would suggest a basic policy guideline for university education and imply that the positive and negative effects of the media need to be considered.

Key Words : University education, newspaper articles, topic modeling, latent Dirichlet allocation, text mining

*This paper was supported by Kumoh National Institute of Technology Research Grant in 2020 (No.20200231001)
*Corresponding Author : Jeong-Ho Koo(jhk2001@kumoh.ac.kr)
Received September 8, 2021 Revised October 31, 2021
Accepted December 20, 2021 Published December 28, 2021

1. 서론

여러 학문 및 산업분야에서 빅데이터 활용이 늘어나면서 정형화되거나 정량화된 자료에 국한하지 않고 언론미디어, 블로그, 게시판, 웹사이트, SNS 등에서 추출한 텍스트를 이용한 비정형화된 자료 분석이 폭 넓게 이루어지고 있다. 그 중 신문기사는 암묵적으로 사회 구성원들의 높은 동의를 나타내는 가치 있는 의제를 보도하고, 사회적으로 논의할 의제 설정 영향력을 갖고 있기 때문에 이를 이용하여 사회적 이슈사항, 트렌드, 화두, 관심사항을 분석한 연구가 지속되고 있다[1-4].

텍스트 마이닝은 텍스트 자료를 정형화하여 특정 주제와 관련하여 의미가 있는 개념이나 특징을 추출하는 도구이며, 텍스트 마이닝에 기반한 토픽모델링은 방대한 양의 텍스트 데이터에 내재되어 있는 토픽들을 추출하고 자동으로 문서들을 분류하는데 유용하다. 이와 같은 이유로 토픽모델링은 다양한 분야에서 생성되는 텍스트 데이터에 적용되어 연구 탐색 및 동향이나 사회적 이슈, 주제 발굴, 트렌드 분석 등에 사용되고 있다[1-11]. 예를 들면, 토픽모델링을 이용하여 뉴스 기사에 인공지능, 태양광 등이 어떻게 보도되고 있는지 토픽을 추출하여 제공하고[1-3], 연구논문 텍스트를 활용하여 국내 연구의 토픽을 밝히고 연구 변화추세를 검토한다[5, 9, 10]. 이뿐만 아니라 코로나19 팬데믹이 선포되어 전 세계적으로 미친 영향이 커지자 SNS에서 수집된 텍스트를 기반으로 코로나 확산 시기에 따른 대중들의 코로나19 관련 관심사 변화 등을 살펴보고 있다[6, 11].

우리나라에서 교육은 사회 구성원들의 가장 주요한 관심사항 중 하나이다. 코로나19로 초, 중, 고, 대학에서 대면수업과 같은 교육방식이 어려워지자, 교육적 측면에서 포스트 코로나를 준비할 수 있도록 코로나19와 연관된 교육부 교육정책에 대한 언론보도 기사를 분석하고, 코로나19 이후 온라인 교육 관련 토픽들의 변화 추이와 언론매체에서 간과하고 있는 교육관련 의제들을 논하고 있다[7-8]. 그러나 코로나19 이전에 언론 미디어에서 대학교육을 어떻게 바라보고 있는지를 검토한 연구는 미흡하다. 언론보도의 가장 큰 특징은 사회구성원들이 갖고 있는 공통의 관심사항과 의견 수렴, 합의가 반영되어 신뢰성이 높다는 것이다. 이와 반대로 광고주와 유착하여 홍보성의 대가를 반영한 기사형 광고가 생산되며, 언론사의 경영상의 문제로 기사형 광고는 증가하는 추세이다. 기사형 광고는 소비자와 구독자가 합리적인 의사결정을 하는데 부정적인 영향을 미쳐 2021년 8월 국회입법조사처는 기사형 광고 규제가 필요함을 제언하였다.

빅데이터와 관련된 대학교육 연구는 주로 코로나19 이후에 초점을 두고 있으나 코로나19 이전을 포함하여 대학교육이 언론보도에 투영된 사항을 통해 사회적으로 대학교육에 대한 화두 및 이슈, 대학과 연계된 기사형 광고 비중 등을 살펴볼 필요가 있을 것이다. 본 연구에서는 토픽모델링을 활용하여 주요 신문사의 대학교육 관련 텍스트 기사들을 분석함으로써 언론 보도의 주제를 검토하고, 신문사별 특징을 파악하고자 한다. 이를 통해 대학교육의 정책적 방향과 언론이 추구해야 할 사회적 역할에 대한 제언을 할 수 있으리라 기대한다.

본 논문의 구성은 다음과 같다. 제2장에서는 뉴스 빅데이터와 토픽 분석에 관한 선행연구를 검토하고, 제3장에서 텍스트 데이터의 수집과 전처리 방법 및 연구에 사용된 분석 방법에 대해 살펴본다. 제4장에서는 최근 6년간 대학 교육과 관련된 신문기사의 핵심 토픽을 도출하고 토픽별 특징 및 신문사별 보도양상을 검토한다. 마지막으로 제5장에서는 본 연구의 결론과 시사점을 제시하고 있다.

2. 선행연구 검토

2.1 뉴스 빅데이터와 토픽 분석

빅데이터 활용이 범용화 되면서 뉴스 기사에서 추출한 텍스트 자료를 이용하여 사회적 관심사항을 많은 선행연구에서 확인한다. 노설현(2020)[1]은 신문에 보도된 인공지능 기사 텍스트를 대상으로 토픽모델링을 통해 추출된 10개의 토픽들이 인공지능이 기술적인 측면뿐 아니라 사회 전반에 걸친 변화를 반영하고 있음을 보고하였다. 이새미와 홍순구(2020)[12]는 뉴스에 보도된 블록체인과 관련 기사를 이용하여 토픽모델링을 통해 19개의 뉴스 토픽을 도출하였고, 사회적 관심사항이 암호화폐에 치중되어 있음을 나타내었다. 김민정(2020)[13]은 신문기사에서 보도된 웨어러블 관련 텍스트를 대상으로 키워드 추이를 분석한 결과, 최상위 빈도어로 애플, 삼성전자, LG전자가 추출되어 스마트워치, 스마트밴드가 꾸준히 기사로 보도됨을 확인하였다. 또한, IT 전자쇼 등과 같은 차세대 기술 관련 키워드와 융합된 내용이 지속적으로 기사화됨을 나타내었다. 이와 같은 결과는 기사형 광고를 의심할 필요가 있음을 시사한다. 이상숙 등(2020)[3]은 2018년도와 2019년도 2년 동안 11개 신문에 보도된 인공지능(AI) 교육에 대한 대중들의 관심사항을 토픽모델

링을 통해 AI분야의 여성인재 육성, 대학교육과정의 변화, K-12의 소프트웨어 교육 및 교육과정 변화 등 거시적인 정책 지원에 대한 토픽들을 확인하였다.

박일수(2021)[7]는 2020년1월부터 9월까지 코로나19 이후 교육부의 교육정책에 대한 언론보도의 성향을 텍스트 마이닝을 이용하여 분석하였다. 분석 결과, '등교, 개학, 입시, 수업, 보건, 교육, 휴업, 학사, 평가, 보육, 휴교, 학교, 교사' 등이 언론에서 보도한 코로나에 대한 주요 이슈임을 밝히었다. 손민성 등(2021)[14]은 코로나19로 온라인 교육의 중요성이 강조되자 네이버, 다음, 구글에 보도된 뉴스, 블로그, 카페에서 추출한 텍스트 자료를 활용하여 온라인 교육과 관련한 상위 키워드를 분석한 결과, 코로나19 이전에는 학점은행제, 평생교육, 블로그 등에서 코로나19 이후 온라인 개학, 비대면 교육, 실시간, 콘텐츠 제작, 유튜브 등으로 변화함을 발견하였다.

3. 연구방법

3.1 연구자료 수집

본 연구에 사용된 신문기사 텍스트 자료는 한국언론진흥재단에서 제공하는 빅카인즈(BIG KINDS)에서 추출하였다[15]. 2021년 7월 17일 기준으로, 제목에서 "교육", "수업" 또는 "강의"로 검색되는 주요 신문사들의 기사들을 대상으로 삼고 있다. 단, 언론을 통해 대학 교육의 주제를 살펴보고자 하는 본 연구의 목적에 맞게 검색범위를 "대학"이 포함된 기사로 한정하였다. 상세검색 기능에서 검색어 처리는 "형태소 분석"으로 설정하였다. 한편 기간은 2016년 1월 1일 이후로 하였는데, 분석시점 기준으로 2021년 자료를 모두 획득할 수 없는 상황이기 때문에, 매해 동일기간을 설정하여 2016년-2021년의 상반기(1월-6월)로 분석기간을 한정하였다.

신문사의 발행 지역과 성향에 따라 기사의 내용, 주제, 논조가 달라질 수 있는 가능성을 배제할 수 없다. 이에 표본의 편향성(sampling bias)을 사전에 방지하고자 각기 다른 발행지역과 성향을 나타내는 12개 일간지들을 고루 선정하고자 하였다. 보수 중앙지 및 진보 중앙지는 발행부수에 근거하여 대표성을 갖는 조선일보, 중앙일보, 동아일보 및 한겨레와 경향신문을 선정하였다. 지역 일간지로는 권역별로 발행부수, 지명도가 높은 다음의 7개의 일간지 '강원: 강원일보, 광주 및 전라: 광주일보, 대구 및 경북: 매일신문, 대전 및 충청: 대전일보, 부산·울산 및 경남: 부산일보, 인천 및 경기: 경인일보, 제주: 한라일보'를 선택하였다.

3.2 전처리 과정

수집된 기사 텍스트 자료는 키워드를 중심으로 데이터 정제 및 전처리 과정이 수행되었다. 숫자, 기호 등은 사전에 제거되었고 "학년도", "학년", "학기", "서울" 등의 결과에 크게 영향을 미치지 않는 단어들도 삭제되었다. 광고성 단어나 웹페이지 링크, 신문사나 기자 관련 단어들 역시 제거하였고, 띄어쓰기 오류나 명사 추출이 제대로 이루어지지 않은 경우에는 수정, 보완 작업을 수행하였다. 학생들, 교수들 등처럼 복수형은 단수형으로 통일하였다. 제목에 "대학"이 반드시 검색되는 기사로 분석대상을 한정하고 있어서, "대학"은 상대적으로 지나치게 자주 등장함을 확인할 수 있었다. 분석에 크게 영향을 미치지 않는다고 판단되는바 "대학" 단어도 사전에 제거하였다.

내용 또는 제목이 동일한 중복기사, 또는 완전히 동일하지는 않지만 유사도가 높으면서 반복적으로 수집된 기사, 인사 및 부고, 동정, 사진 등 의미 있는 결과를 기대하기 어려운 기사는 분석에서 제외되었다[15].

이러한 정제 및 전처리 과정을 거친 후 구조화된 단어-문서 행렬(term-document matrix)을 구축하였으며, 효율적인 데이터 분석을 위해 다른 문서에 있는 단어와 최소 95% 공통적으로 쓰이지 않는 희소 단어들(sparse terms)을 제거하였다.

3.3 토픽모델링 및 토픽개수

토픽모델링의 대표적인 방법 중 하나인 잠재 디리슐레 할당(LDA, Latent Dirichlet Allocation)을 적용하여, 대학 교육 관련 기사에서 나타나는 토픽들을 추출하고자 한다. LDA는 대표적인 확률적 토픽 모델기법 중 하나로서, 토픽내 단어들의 분포 및 문서내 토픽들의 분포의 결합을 가정하여 문서에 내재되어 있는 토픽을 발견하는 과정이라 할 수 있다. 사전확률분포(prior probability distribution)로서 다항분포와 켤레(conjugate) 관계인 디리슐레 분포를 사용하고 있으며, 베이지언 방법을 이용하여 통계적 추론이 이루어진다. LDA와 관련된 더 자세한 내용은 [16, 17]을 참조할 수 있다.

토픽모델링을 적용하기 위해 사전에 토픽의 개수를 결정할 필요가 있는데, 토픽의 개수가 너무 많으면 지나치게 세분화된 토픽들이 등장하는 반면, 너무 적으면 한 토픽에 여러 토픽이 섞이면서 모호한 특성을 나타낼 수 있

다. 이에 토픽의 해석가능성과 그 유용성에 근거하여 결정할 필요가 있다[18]. 토픽의 개수를 사전에 정하기 위한 여러 측도들이 제안되고 있는데, CaoJuan2009 [19]는 토픽들 간의 코사인 유사도가 작아지도록 Arun2010 [20]은 토픽-단어 행렬의 특이값 분포의 대칭적 쿨백-라이블러 발산이 최소화되도록 토픽의 개수를 선택할 것을 제안한다. 반면에 Griffiths2004 [21]과 Deveaud2014 [22]는 우도함수의 조화평균과 토픽 분포간 젠슨-샤논 거리의 평균값이 최대가 되도록 토픽의 개수를 정할 것을 제안하고 있다. 이에 본 연구에서는 이러한 네 가지 측도를 검토하고 해석의 용이성 및 적합성을 감안하여 적절한 토픽의 개수를 정하고자 한다.

3.4 토픽별, 신문사별 특징

토픽 개수를 정하고 난 후, 기사에서 등장하는 상위 1,500개의 단어들을 대상으로 LDA를 적용한다. 추출된 토픽들에서 등장하는 주요 단어들을 파악함으로써 각 토픽의 특징을 파악할 수 있다. 또한 신문사들의 토픽별 기사 편수를 통해 대학 교육 보도기사와 관련한 신문사별 보도 양상을 고찰한다. 또한 광고·홍보성 기사를 분류하고 이러한 기사들에서 등장하는 주요 단어들을 파악하여 어떠한 신문사에서 해당 기사들이 많이 보도되었는지를 살펴본다.

4. 연구결과

4.1 신문사별 보도 건수

2016년 이후 기사들 중, 제목에 "교육" 또는 "수업" 또는 "강의"를 포함하면서 "대학"으로 검색되는 기사 개수는 총 1,208개이었다. 중복되거나 유사한 기사, 대학 교육과 관련이 없는 기사들을 제외하여 1,172개의 기사들을 추출하였으며, 그중 상반기에 보도된 591개의 기사를 최종적인 분석에 사용하였다.

Fig. 1에서 신문사별 상반기 기사 수 및 권역별 대학 개수를 제시하고 있다. 5개의 대표 중앙지와 7개의 지역지의 연도별 보도 건수를 막대그래프로 확인할 수 있다. 지역지의 경우, 해당 지역에 얼마나 많은 대학이 있는지가 보도 건수에 영향을 줄 수 있기 때문에 각 권역별 대학 수를 그림에 추가하였다. 단, 중앙지는 전국을 대표하고 한국 신문시장 전체를 지배하는 특수성을 감안하여 대학 수를 표시하지 않았다. 대학 개수는 통계청 기준(2020년 9월 11일 당시)에 따라 일반대, 전문대, 교육대, 산업대의 개수를 합한 수치이다[23].

그림에서 보듯이 중앙지 중에는 중앙일보와 동아일보가 상대적으로 관련 기사를 많이 보도하였다. 지역지를 대상으로 보도 건수를 반응변수로, 권역별 대학 수를 설명변수로 하는 단순회귀분석을 수행한 결과, 평균적으로 대학 수가 많은 권역일수록 보도 건수가 많았고 10% 유의수준에서 유의하였다. 특히 대학이 적은 강원과 제주 지역에서 발행되는 강원일보와 한라일보의 보도 건수가 적음을 확인할 수 있다.

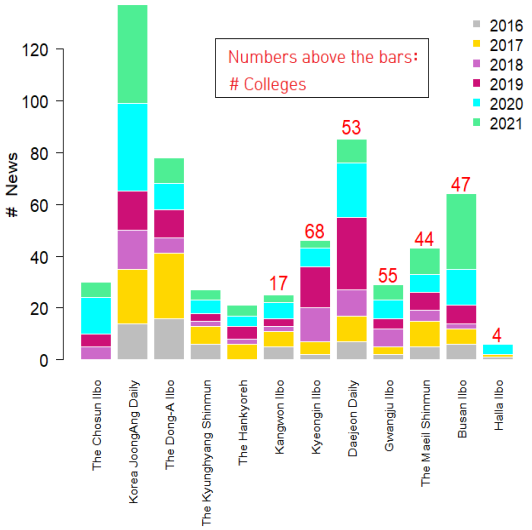

Fig. 1. The number of newspaper articles with the number of colleges

4.2 토픽 개수 선택

토픽 개수를 사전에 결정하기 위해, 앞 장에서 살펴본 네 가지 측도들의 값을 확인하였다. 토픽 개수에 따른 측도들의 값이 Fig. 2에 제시되어 있는데, 위 패널에 있는 Arun2010 [20]과 CaoJuan2009 [19]는 작을수록, 아래의 패널에 있는 Griffiths2004 [21]와 Deveaud2014 [22]는 클수록 최적의 값이다. 네 값의 측도가 모두 최적이 되는 상황은 없으나, CaoJuan2009 [19]의 경우 토픽의 개수가 8에서 9가 될 때 값이 크게 감소하고 9에서 10으로 될 때에는 감소폭이 줄어듦을 확인할 수 있다. 반면 Deveaud2014 [22]는 토픽의 개수가 8에서 9가 될 때 그 값이 크게 증가하고 10 이상부터는 값이 감소함을 볼 수 있다. 이러한 점들과 의미 유용성 및 해석 용이성을 감안하여 토픽의 개수를 9로 선택하였다.

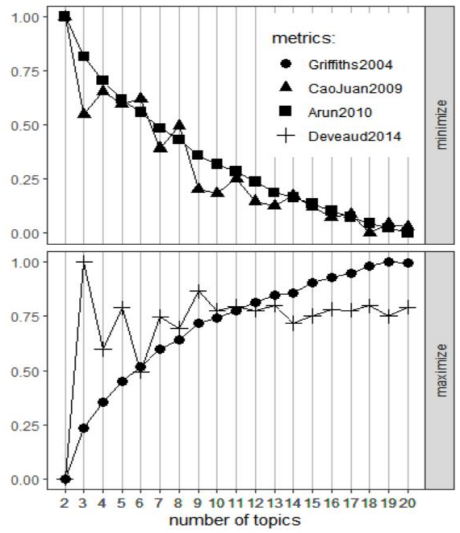

Fig. 2. Selection of the number of topics

4.3 LDA를 통한 토픽 추출 및 토픽의 특징

분석 기사 텍스트 자료에 LDA를 적용하여 9개의 토픽을 추출하였다. 먼저 각 토픽별 상위단어를 살펴봄으로써 토픽에 담긴 주제와 내용을 파악하고자 한다. Fig. 3에서는 각 토픽별 주요 단어를 막대그래프로 보여주고 있는데, 이때 토픽의 넘버링은 해당 토픽의 크기에 따라 결정되었다. 즉 토픽1은 전체 기사들 중에서 차지하는 비중이 가장 큰 토픽이며, 토픽9는 가장 비중이 작은 토픽이다.

토픽1의 주요 단어는 인재, 지원, 프로그램, 혁신, 취업, 산업, 창업 등임을 확인할 수 있다. 해당 토픽에 속하는 기사들은 주로 대학에서 시행되는 사업 및 프로그램에 관한 내용임을 알 수 있다. 반면 토픽2는 온라인, 대면, 코로나19, 개강 등의 단어들이 주로 나타나, 코로나19 이후의 대학 강의에 대한 논의가 주를 이루었다.

토픽3은 사업, 지원, 선정 등 토픽1과 일부 내용이 겹치지만, 지역, 공동, 협력 등의 단어들이 등장하고 토픽1보다는 좀 더 지역대학 관련기사가 많음을 확인하였다.

토픽4에는 강의, 온라인, 사이버 등의 단어가 두드러지게 나타났다. 토픽2 역시 온라인 강의방법에 대한 논의가 이루어지나 이는 코로나19에 대한 대응책의 일환으로 다루어진 반면, 토픽4는 주로 새로운 교육 패러다임에 기반한 교수법에 관한 내용이라는 데 차이가 있다.

토픽5에는 교육부, 등록금, 강사, 강사법, 재정, 정책, 정원 등이 주로 나타났으며, 주로 대학등록금정책, 강사

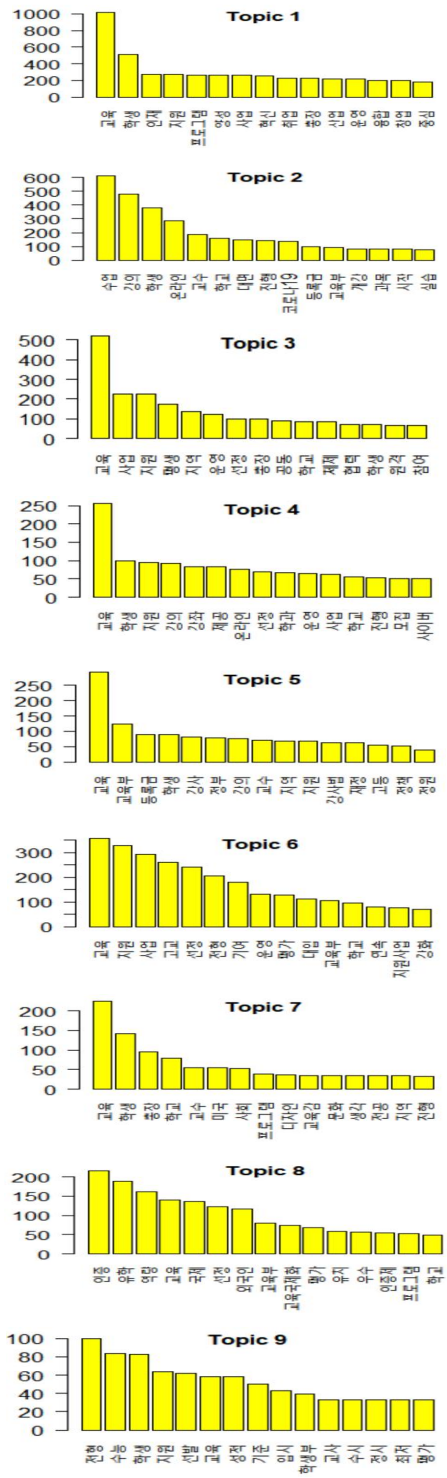

Fig. 3. Key words in each topic

법, 대학정원정책 등 정부의 대학 관련 정책과 지침에 관련한 기사들이 이 토픽에 속하였다. 반면 토픽6은 고교교육 기여대학 지원사업에 관한 기사보도가 주를 이루었다.

토픽7, 토픽8, 토픽9는 상대적으로 비중이 작았는데, 10% 미만의 기사들이 각각 이 토픽들에 포함되었다. 토픽7은 대학 교육 비전, 선거공약에 관한 내용으로 나타났으며, 토픽8은 교육국제화, 외국인 유학생에 관한 내용이었음을 확인하였다. 마지막으로 토픽9는 오직 20개의 기사만이 여기에 속하였는데, 대학 입시 전형과 관련한 내용으로 나타났다.

Fig. 4. Intertopic distance map(IDM)

Fig. 4는 9개 토픽들을 2개의 차원(PC1, PC2)으로 축약하여 보여주는 그림으로 이를 통해 토픽들의 크기 및 토픽들 간의 거리를 알 수 있다[24]. 그림에서 보듯이, 토픽1은 가장 큰 토픽으로 토픽3과 겹쳐지는 내용이 있음을 알 수 있다. 두 토픽에 해당하는 기사들의 제목을 확인한 결과, 두 토픽 모두 대학 지원 사업에 관련한 내용이지만 토픽3은 지역사회 공동대응, 지역대학의 평생교육 등 좀 더 지역대학의 지원사업에 중점이 맞추어져 있었다. 그 다음으로 큰 토픽2는 앞에서 언급했듯이 코로나19 이후 대학 강의에 관한 내용으로 다른 토픽들과 상당히 떨어져 있어 다른 토픽들과의 유사도가 낮음을 알 수 있다. 토픽5와 토픽9가 대학 입시 정원과 관련하여 다소 중복되는 내용이 나타나기도 하지만, 전반적으로 9개의 토픽에 해당하는 원들이 고르게 배치되어 있어 사전에 정한 9개 토픽 개수가 적절하였다고 판단된다.

이 9개 토픽의 코로나19 전후 특징을 검토하기 위해 코로나19 전후로 데이터를 나누어 토픽모델링을 적용하였다. 코로나19 이전은 2016년-2019년 상반기 336개 기사들을 대상으로 하며, 코로나19 이후는 2020년-2021년 상반기 255개 기사들을 대상으로 한다. 비교의 용이성을 위해 모든 데이터를 사용했을 때처럼 토픽의 개수를 k=9로 선택하였다. 코로나19 전 또는 후에 나타난 토픽들과 이미 밝혀진 9개 토픽들 간 매칭은 등장하는 주요 단어 및 해당 기사들의 멤버쉽(membership)을 비교해서 확인하였다. 각 기사별로 토픽의 확률분포를 구한 후, 가장 큰 확률을 가지는 토픽을 해당 기사의 대표 토픽으로 선택함으로써 각 기사의 멤버쉽을 부여하였다.

코로나19 이전(-2019년)의 336개 기사만을 대상으로 토픽모델링을 적용했을 때 주된 토픽은 대학지원사업 관련 내용으로, 모든 자료를 이용했을 때 추출된 토픽1과 거의 동일한 단어들로 구성되어 있었다. 또한 336개 기사만을 대상으로 한 분석에서 대학지원사업 토픽으로 분류된 기사들의 87% 이상이 모든 자료를 이용했을 때의 토픽1에 속함을 확인하였다. 그 다음으로 토픽4(교수-학습법), 토픽5(정부정책), 토픽6(고교교육기여대학 지원사업), 토픽9(입시), 토픽3(지역대학 지원사업), 토픽7(대학교육 비전)에 해당하는 내용 순으로 비중이 컸다. 반면에 토픽2(코로나19 이후 대학교육), 토픽8(국제화)은 코로나19 전에 명확하게 나타나지 않았다. 이 두 개의 토픽의 경우, 여러 토픽의 내용이 혼재되어 있었는데 특히 토픽4(교수-학습법), 토픽7(대학교육 비전) 내용이 많았다.

한편 코로나19 이후(2020년-)의 255개의 기사만을 대상으로 하였을 때, 가장 주된 토픽은 코로나19 이후 대학교육 관련된 토픽2이었다. 다음으로 토픽1(대학지원사업), 토픽3(지역대학 지원사업), 토픽6(고교교육기여대학 지원사업), 토픽8(국제화), 토픽5(정부정책), 토픽4(교수-학습법) 순으로 비중이 큰 것으로 나타났다. 반면에 토픽7(대학교육 비전), 토픽9(입시)는 코로나19 이후에 뚜렷하게 나타나지 않았다.

이를 통해 코로나19 전후로 등장하는 토픽의 차이를 살펴보면, 코로나19 이전에는 대학지원사업이 가장 비중이 큰 토픽이었다면, 이후에는 코로나19로 인한 대학교육이 가장 큰 비중을 차지하였다. 또 코로나19 이전에는 명시적으로 나타나지 않았던 토픽2(코로나19 이후 대학교육)와 토픽8(국제화)이 코로나19 이후에는 주요 토픽으로 등장하였다. 코로나19 이후에는 이전과 비교하여 상대적으로 대학교육 비전, 교수-학습법의 비중이 낮고, 지역대학 지원사업, 국제화의 비중이 높은 것으로 나타났다.

다음은 광고·홍보성 기사의 특징을 살펴보고, 이러한

기사들을 제외했을 때 나타나는 토픽들의 특징을 검토하였다. 분석 대상 기사들의 제목과 내용을 살펴보면, 기사의 제목에 특정 대학의 이름이 나타날 경우 해당 학교에 대한 광고 또는 홍보성 기사가 많음을 확인하였다. 이에 따라 제목에 대학명이 나오면 광고·홍보성 기사로 분류하였다. 대학을 소개하고 홍보하는 기사 시리즈와 제목에 모집, MOU, 출범, 협약, 체결, 선정, 표창, 추진단, 사업단 등의 단어가 나타날 경우도 광고·홍보성 기사로 분류하였다. 총 269개의 광고·홍보성 기사(45.5%)와 322개의 비광고·비홍보성 기사(54.5%)를 분류할 수 있었다.

광고·홍보성기사에서 나오는 주된 단어는 교육, 지원, 사업, 운영, 학교, 역량, 프로그램, 평가, 고교, 인증, 기여, 유학, 양성, 평생이었으며, 대학지원 사업 관련 토픽1과 토픽3, 고교교육 기여대학 지원사업 관련 토픽6, 국제화 관련 토픽8에 이러한 기사들이 많았다. 반면 코로나19 이후 교육 관련 내용의 토픽2와 대학 관련 정부 정책의 토픽5, 입시 관련 내용의 토픽9는 광고·홍보성 기사가 적은 것으로 나타났다. 중앙지로는 동아일보, 중앙일보가 분석 기사의 40% 이상, 지역지로는 한라일보, 광주일보, 대전일보, 부산일보, 매일신문이 50% 이상, 경인일보가 40% 이상 이러한 기사들로 구성되어 있음을 확인할 수 있었다.

광고·홍보성 기사를 제외한 322개의 기사만을 대상으로 토픽모델링을 적용하였는데, 비교의 용이성을 위해 이번에도 k=9로 선택하였다. 322개의 기사를 대상으로 발견된 토픽과 이미 밝혀진 9개 토픽들 간 매칭은 주요 키워드와 멤버쉽 비율을 확인해서 이루어졌다. 비광고·비홍보성 기사만을 대상으로 했을 때 주된 토픽은 대학지원사업 관련 내용으로, 모든 기사를 이용했을 때 추출된 토픽1과 거의 유사한 단어들로 구성되어 있었다. 또한 기사의 약 80%가 기존의 토픽1에 속함을 확인하였다. 하지만 전체 토픽들 중에서 이 토픽이 차지하는 비중은 기존 24%에서 18%로 감소하였다.

그 다음으로 토픽2(코로나19 이후 대학교육), 토픽5(정부정책), 토픽7(대학교육 비전), 토픽3(지역대학 지원사업), 토픽9(입시), 토픽4(교수-학습법)에 해당하는 내용 순으로 비중이 컸다. 하지만 토픽6(고교교육기여대학 지원사업), 토픽8(국제화) 관련 토픽은 뚜렷이 나타나지 않았다.

광고·홍보성 기사를 제외했을 때, 전반적으로 토픽1(대학지원사업), 토픽3(지역대학 지원사업), 토픽4(교수-학습법)의 비중이 작아지고 토픽6(고교교육기여대학 지원사업), 토픽8(국제화)은 명확하게 등장하지 않았다. 즉 이러한 토픽들이 상대적으로 광고·홍보성 기사들을 많이 포함하고 있음을 시사한다. 반면 토픽2(코로나19로 인한 대학교육), 토픽5(정부정책), 토픽9(입시)는 비광고·비홍보성 기사만을 대상으로 했을 때 비중이 커짐을 확인하였다.

4.4 신문사별 토픽 비중

각 신문사의 토픽별 보도 건수 비율을 살펴봄으로써 대학 교육과 관련하여 신문사별 보도 양상을 파악하고자 한다. Table 1은 신문사별로 9개 토픽에 속하는 기사 수의 비율(%)을 보여준다.

많은 신문사에서 대학 지원 사업 내용이 주인 토픽1 또는 토픽3에 포함되는 기사를 제일 많이 보도한 반면, 3개의 신문사, 조선일보, 경향신문, 한겨레는 코로나19에 따른 대학 교육 변화와 관련된 토픽2와 정부의 대학 관련 정책 내용의 토픽5에 관한 내용이 주를 이루었다. 반면에 동아일보, 중앙일보를 비롯한 중앙지는 토픽1에, 한라일보, 부산일보, 대전일보, 경인일보는 토픽3과 토픽6에 해당하는 기사가 많았다.

기타 토픽4의 비중은 동아일보, 중앙일보, 경인일보 순으로 나타났는데, 특정 사이버대학에 대한 소개기사를 비롯하여 K-무크, KOCW, 플립드러닝 등 대학의 교수-학습방식을 논의한 기사들이 여기에 속하였다.

토픽7의 경우 전체 토픽에서 차지하는 비중이 크지는 않다. 매일신문, 조선일보, 경향신문 순으로 비중이 높았으며, 논평 및 기고 형태의 기사가 많이 포함되었다. 특히 대학교육의 비전, 기초교육, 창의성교육, 교양교육 등에 대한 의견 제시 및 논의가 이루어지고 있다.

한편 교육국제화 및 외국인 유학생에 관한 토픽8의 경우, 지방지를 대상으로 보면 대학 수 대비 외국인 학생 수 비율이 높은 권역의 대전일보, 경인일보에서 그 비중이 높았다. 한국교육개발원 교육통계서비스(2020년 4월 1일 기준, [25])에서 제공하는 외국인 학생 수와 통계청에서 제공하는 대학 수(2020년 9월 11일 기준, [23])를 이용하여 계산한 대학 수 대비 외국인 학생 수는 제주 552.75명, 대전 및 충청 373.26명, 인천 및 경기 316.78명 순이다. 토픽8에 속하는 보도 건수를 반응변수로 하고 대학 수 대비 외국인 학생 수를 설명변수로 하는 단순회귀분석을 수행한 결과, 추정치가 양수이지만 통계적으로 유의하지는 않았다. 토픽9는 비중이 아주 작으며(총 591개 기사 중 20개의 기사), 대학 입시 전형 관련 기사들 중 제목에 대학, 교육(또는 강의 또는 수업) 단어가 들어가 있는 일부 기사가 추출된 것으로 확인된다.

Table 1. Proportions of newspaper articles by newspaper

(%)	T1	T2	T3	T4	T5	T6	T7	T8	T9
The Chosun Ilbo	10.0	43.3	6.7	0.0	20.0	0.0	20.0	0.0	0.0
Korea JoongAng Daily	33.6	12.4	11.0	11.7	4.4	13.1	3.7	8.0	2.2
The Dong-A Ilbo	44.9	14.1	5.1	18.0	6.4	2.6	6.4	1.3	1.3
The Kyunghyang Shinmun	7.41	22.2	14.8	7.4	25.9	3.7	14.8	3.7	0.0
The Hankyoreh	4.8	23.8	14.3	4.8	38.1	9.5	0.0	0.0	4.8
Kangwon Ilbo	16.0	20.0	16.0	8.0	8.0	4.0	4.0	0.0	24.0
Kyeongin Ilbo	17.4	6.5	21.7	10.9	2.2	15.2	8.7	10.9	6.5
Daejeon Daily	18.8	8.2	22.4	9.4	1.2	9.4	10.6	17.7	2.4
Gwangju Ilbo	20.7	6.9	17.2	3.5	17.2	10.3	10.3	6.9	6.9
The Maeil Shinmun	7.0	20.9	14.0	7.0	4.7	16.3	23.3	4.7	2.3
Busan Ilbo	17.2	15.6	23.4	7.8	4.7	14.1	6.3	9.4	1.6
Halla Ilbo	16.7	16.7	33.3	0.0	0.0	33.3	0.0	0.0	0.0

5. 결론 및 논의

본 연구에서는 신문기사 빅데이터를 통해 대학교육 관련 보도의 토픽을 추출하고, 토픽별 특징 및 신문사별 보도양상을 검토하고 있다. 빅카인즈의 텍스트 자료를 이용하여 9개의 토픽을 추출하였다. 이러한 토픽들은 토픽간 거리 지도(Intertopic distance map)에서 첫 번째, 두 번째 주성분 공간(PC1 vs PC2)에 고르게 배치되어 있었고 토픽의 개수가 적절히 선택되어졌음을 확인할 수 있었다. 추출된 토픽1과 토픽3은 모두 대학교육과 관련하여 교육부에서 주도하는 대학지원사업에 연관된 내용으로 중복되는 부분이 있으나 토픽3은 지역대학에 초점이 맞추어져 있다. 이는 정부 주도하의 각종 사업이 대학재정에 큰 영향을 미치기 때문에 가장 비중 높게 보도됨을 의미한다. 토픽2는 코로나19 이후 대학 교육에 관련한 내용이며, 토픽4는 사이버 대학, 온라인 강의 등 대학의 교수-학습방식을 소개, 논의하고 있다. 토픽5는 정부의 대학 관련 정책에 대한 주제이며, 토픽6은 고교교육 기여대학 지원사업에 관련한 내용이다. 토픽7은 대학교육의 비전, 기초교육, 창의성교육, 교양교육 등의 내용이 주를 이루며 토픽8은 대학교육 국제화와 관련된 내용이다. 반면에 토픽9는 비중이 작지만 주로 대학 입시 전형과 관련한 내용이었다. 토픽7, 8, 9의 비중이 상대적으로 낮다는 것은 대학의 질을 결정하는 교육내용을 다루는 부분이 작음을 시사하고, 대학입시가 민감성이 높은 이슈이나 대학교육과 관련해서는 연관성이 약한 영역이며 코로나19가 대학 교육에 미치는 영향이 매우 큼을 의미한다.

코로나19 전후로 토픽 변화가 있는지를 살펴본 결과, 코로나19 이전에는 토픽1 대학지원사업이 주를 이루어 기사의 87% 이상을 차지하였고, 토픽 비중 순서가 토픽4, 토픽5, 토픽6, 토픽9, 토픽3, 토픽7 순으로 변화하였다. 그러나 토픽2와 토픽8은 명확하게 나타나지는 않았다. 지역대학 중심의 사업 비중이 코로나 이전에는 작음을 알 수 있고, 교육과 관련된 교수법 등 패러다임이 중요하게 다루어지고 있다. 코로나19 이후에는 토픽2가 가장 크게 나타났고 토픽1, 3, 6, 8, 5, 4 순으로 바뀌었으며 토픽7, 9는 뚜렷하지 않은 것으로 나타났다. 이는 코로나19를 대학 교육이 어떻게 대응하는지 사회적 요구에 맞게 가장 중요하게 다루나 토픽5 교육부 정책이나 토픽4 교수법의 비중이 상대적으로 약화됨을 알 수 있다. 즉, 코로나19 이전에는 대학지원사업이 가장 비중이 큰 토픽이었다면, 이후에는 코로나19로 인한 대학교육이 가장 큰 비중을 차지하며 코로나19 이전에는 명시적으로 나타나지 않았던 토픽2(코로나19 이후 대학교육)와 토픽8(국제화)이 코로나19 이후에는 주요 토픽으로 등장하였다.

광고·홍보성 기사와 비광고·비홍보성 기사가 주로 다루는 토픽 영역을 살펴본 결과, 광고·홍보성 기사는 대학지원 사업 관련 토픽1과 토픽3, 고교교육 기여대학 지원사업 관련 토픽6, 국제화 관련 토픽8에 이러한 기사들이 많았다. 반면 코로나19 이후 교육 관련 내용의 토픽2와 대학 관련 정부 정책의 토픽5, 입시 관련 내용의 토픽9는 광고·홍보성 기사가 적은 것으로 나타났다. 즉, 광고·홍보성 기사는 대부분 대학재정에 긍정적으로 영향을 미치는 내용으로 대학은 정부 지원에 크게 의존할 뿐 아니라 대학 입학률, 이미지 형성 등에 이러한 사업들이 작용함을 시사한다. 반면 광고홍보성 기사를 제외한 경우 토픽1, 2, 5, 7, 3, 9, 4 순으로 변경되었고, 토픽6, 8은 명확하지 않은 것으로 나타났다. 즉, 토픽2(코로나19로 인한 대학교육), 토픽5(정부정책), 토픽9(입시)는 비광고·비홍보성 기사만을 대상으로 했을 때 비중이 커져 코로나19가 대학교육에 미치는 영향은 근본적으로 다루어야 하는 사회적 화두임을 함축하고 있다.

신문사별로 살펴보았을 때에는 조선일보, 경향신문, 한겨레가 상대적으로 광고·홍보성 기사가 적었으며, 이들 신문사는 코로나19이후 강의 관련 기사, 정부 정책 관련 기사, 대학교육에 대한 논평을 많이 보도하였다. 반면에 동아일보, 중앙일보, 한라일보, 부산일보, 대전일보, 경인일보는 대학 지원 사업, 고교교육 기여대학 지원사업 등

상대적으로 광고·홍보성 기사가 많았다. 전체적으로 봤을 때, 기사의 45% 가량이 광고·홍보성이었다. 여론 형성이라는 언론의 역할과 책임을 감안할 때, 단순한 사실을 보도하는 기사나 광고·홍보성 기사보다는 대학 교육과 관련하여 활발히 문제 제기를 하고 논평하고 해결책을 제시할 수 있는 언론의 모습이 보다 필요함을 의미한다.

본 연구는 첫째, 2016년부터 2021년까지 신문기사에 보도된 대학교육과 관련하여 토픽모델링을 이용하여 토픽들을 추출하여 사회적으로 대학교육과 관련된 공통의 주요 관심사항과 이슈사항이 무엇이었는지 살펴보았다는 점에서 의미가 있을 것이다. 둘째, 코로나19 이전 대학교육과 관련된 화두와 코로나19 이후 화두가 변화한 것을 통해 미래에 나아가야할 대학교육 정책 방향이나 대응전략 등을 추론하고 파악하는데 유용할 것이다. 셋째, 광고·홍보성 기사 대부분이 정부 주도하에 실시되는 각종 사업들로 대학재정과 관련된 점을 미루어 볼 때 대학교육과 관련한 실질적인 대학교육 콘텐츠, 대학교육 미래 비전과 방향성, 입학자원 감소에 따른 대학교육의 변화, 사회가 요구하는 대학교육 화두 등을 미디어에서 진정성 있게 논하여 광고성 기사 확대를 제어할 필요가 있음을 제언한다는 점에서 의미가 있을 것이다.

향후 더 폭넓은 분석을 위해서, 기사의 제목이 아니라 기사의 내용을 중심으로 분석 대상을 확장하고 적합한 분석대상 기사를 선정할 수 있을 것이다. 또한 본 연구에서는 상반기(1월-6월)의 기사만을 대상으로 하지만, 향후 연구에서 하반기(7월-12월)의 기사들을 포함할 때에도 잠재토픽의 특징 및 양상이 유사한지 검토할 수 있을 것이다. 하반기에는 대학입시 관련 기사, 교육부 관련 정책 관련 토픽 비중이 커지지 않을까 예상해본다. 또한 다른 토픽모델링 알고리즘을 적용·비교해봄으로써, LDA 알고리즘에 의해 이루어진 토픽별 기사 분류가 얼마나 안정적인지(robust)한지 검증할 수 있을 것이다. 향후 더 많은 일간지를 선정하여 분석대상에 포함시킴으로써 표본의 대표성을 확대할 수 있을 것이다. 특히 신문사의 발행지역별, 성향별로 나누어 분석하여 그 차이를 검토한다면 흥미로운 연구가 될 것으로 기대한다.

REFERENCES

[1] S. Noh. (2021). A Analysis of Issues Related to Artificial Intelligence Based on Topic Modeling. *Journal of Digital Convergence, 18(5),* 75-87.
DOI : doi.org/10.14400/JDC.2020.18.5.075

[2] J. Ki & S. Ahn. (2020) Application of Sentiment Analysis and Topic Modeling on Rural Solar PV Issues : Comparison of News Articles and Blog Posts. *Journal of Digital Convergence, 18(9),* 17-27.
DOI : doi.org/10.14400/JDC.2020.18.9.017

[3] S. S. Lee, I. Yoo & J. Kim (2020). An analysis of public perception on Artificial Intelligence(AI) education using Big Data: Based on News articles and Twitter. *Journal of Digital Convergence, 18(6),* 9-16.
DOI : doi.org/10.14400/JDC.2020.8.6.009

[4] S. M. Kim. (2020). Analysis of Press Articles in Korean Media on Online Education related to COVID-19. *Journal of Digital Contents Society, 21(6),* 1091-1100.
DOI: https://doi.org/10.9728/dcs.2020.21.6.1091

[5] S. M. Heo & J. Y. Yang. (2021). A Convergence Study on the Topic and Sentiment of COVID19 Research in Korea Using Text Analysis. *Journal of the Korea Convergence Society, 12(4),* 31-42.
DOI : dx.doi.org/10.15207/JKCS.2021.12.4.031

[6] S. Yoon, S. Jung & Y. A. Kim. (2021). Trend Analysis of Corona Virus(COVID-19) based on Social Media, *Journal of Korea Academia- Industrial cooperation Society, 22(5),* 317-324.
DOI : 10.5762/KAIS.2021.22.5.317

[7] I. S. Park. (2021). Analysis of press articles in Korean media on education policy of the Ministry of Education related to COVID-19. *Teaching Practicum Research, 3(1),* 10-21.
http://www.riss.kr/link?id=A107781888

[8] S. M. Kim. (2020). Analysis of Press Articles in Korean Media on Online Education related to COVID-19. *Journal of Digital Contents Society, 21(6),* 1091-1100.
DOI: https://doi.org/10.9728/dcs.2020.21.6.1091

[9] J. Kim, H. S. Na & K. H. Park. (2021). Topic Modeling of Profit Adjustment Research Trend in Korean Accounting. *Journal of Digital Convergence, 19(1),* 125-139.
DOI : doi.org/10.14400/JDC.2021.19.1.125

[10] S. M. Kim & Y. J. Kim. (2020). Research Trend Analysis on Living Lab Using Text Mining. *Journal of Digital Convergence, 18(8),* 37-48.
DOI : doi.org/10.14400/JDC.2020.18.8.037

[11] S. K. Park, H. J. Lee & B. G. Lee (2021) Exploring Social Issues of On-demand Delivery Platform Participants. *Journal of Digital Convergence, 19(7),* 79-85.
DOI : doi.org/10.14400/JDC.2021.19.7.079

[12] S. M. Lee & S. G. Hong. (2020). Policy agenda proposals from text mining analysis of patents and news articles. *Journal of Digital Convergence, 18(3),* 1-12.
DOI : doi.org/10.14400/JDC.2020.18.3.001

[13] M. J. Kim (2020). Analyzing the Trend of Wearable Keywordsusing Text-mining Methodology. *Journal of Digital Convergence , 18(9),* 181-190.

DOI : doi.org/10.14400/JDC.2020.18.9.190

[14] M. S. Shon, M. J. Im & K. H. Park (2021). A Study on Consumer perception changes of online education before and after COVID-19 using text mining. *Journal of Digital Convergence, 19(1),* 29-43.
DOI : doi.org/10.14400/JDC.2021.19.1.029

[15] BIG KINDS, News Bigdata & Analysis. Korea Press Foundation.
https://www. bigkinds.or.kr

[16] D. M. Blei, A. Y. Ng & M. I. Jordan. (2003). Latent dirichlet allocation, *The Journal of Machine Learning Research, 3,* 993-1022.
https://dl.acm.org/doi/10.5555/944919.944937

[17] S. M. Heo & J. Y. Yang. (2020). Analysis of Research Topics and Trends on COVID-19 in Korea Using Latent Dirichlet Allocation. *Journal of The Korea Society of Computer and Information, 25(12),* 83-91.
DOI : 10.9708/jksci.2020.25.12.083

[18] M. L. Jockers & R. Thalken. (2014). Text analysis with R for students of literature,. New York: Springer.
DOI : 10.1007/978-3-319-03164-4

[19] J. Cao, T. Xia, J. Li, Y. Zhang, & S. Tang. (2009). A density-based method for adaptive lda model selection, *Neurocomputing, 72(7),* 1775-1781.
DOI: 10.1016/j.neucom.2008.06.011

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

[21] T. L. Griffiths & M. Steyvers. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences. 101, suppl 1,* 5228-5235.
DOI: 10.1073/pnas.0307752101

[22] R. Deveaud, E. SanJuan, & P. Bellot. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique. 17(1),* 61-84.
DOI: 10.3166/DN.17.1.61-84

[23] *KOSIS KOrean Statistical Information Service Statistics Korea,* https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1YL21181

[24] C. Sievert & K. Shirley. (2014). LDAvis: A method for visualizing and interpreting topics. Conference: Workshop on Interactive Language Learning, Visualization, and Interfaces at the Association for Computational Linguistics. 63-70.
DOI:10.13140/2.1.1394.3043

[25] KESS, Korean Educational Statistics Service,
https://kess.kedi.re.kr

양 지 연(Ji-Yeon Yang) [정회원]

· 2010년 7월 : Univ. of Illinois Urbana-Champaign 통계학 박사
· 2010년 7월 ~ 2011년 6월 : Claremont McKenna College 방문조교수
· 2011년 7월 ~ 2014년 2월 : MD Anderson Cancer Center 연구원
· 2014년 3월 ~ 현재 : 금오공과대학교 응용수학과 부교수
· 관심분야 : Bayesian analysis, big data analytics and computational statistics.
· E-Mail : jyang@kumoh.ac.kr

구 정 호(Jeong-Ho Koo) [정회원]

· 2009년 8월 : 성균관대학교 일반대학원 회계학과 (경영학 박사)
· 1998년 2월 : 성균관대학교 일반대학원 회계학과 (회계학 석사)
· 2012년 2월 ~ 현재 : 금오공과대학교 경영학과 교수
· 관심분야 : 원가행태, 이익조정, 회계, 교육, Big data analytics
· E-Mail : jhk2001@kumoh.ac.kr