아시아교육연구 17권 3호
Asian Journal of Education
2016, Vol. 17, No. 3, pp. 409-436.

텍스트 빅데이터 분석 기법을 활용한 대학구조개혁 평가의 쟁점 분석*

김지은(金智恩)**
백순근(白淳根)***

논문 요약

이 연구는 텍스트 빅데이터 분석 기법(text big data analytics)을 활용하여 교육부 보도 자료와 신문 기사에 나타난 대학구조개혁 평가와 관련된 주요 쟁점들을 분석한 것이다. 2013년 1월 1일부터 2016년 4월 30일 사이에 공개된 대학구조개혁 평가 관련 교육부의 보도 자료 25개와 국내 10대 종합일간지의 기사 625개를 수집하여 토픽 모델링(topic modeling) 기법 중 잠재 디리쉴레 할당(latent Dirichlet allocation) 알고리즘을 활용해 토픽을 추출하고 쟁점을 분석하였다.

분석 결과, 첫째, 교육부 문서에서는 3개, 신문 기사에서는 7개의 잠재된 토픽이 나타났고, 대학구조개혁 평가가 진행되는 과정에 따라 주로 다루는 토픽이 변화하고 있음을 확인하였다. 둘째, 교육부 문서와 신문 기사에서 유사한 세 가지 토픽(추진 배경, 운영 방안, 결과 활용)이 나타났지만 중점을 두는 내용에 차이가 발견되었고, 신문 기사에서는 이 외에 4개의 토픽(평가 방식, 고등교육의 질, 대학의 책무성, 대학의 홍보)이 추가로 나타나 대학구조개혁 평가와 관련한 다양한 관심과 쟁점을 확인할 수 있었다.

이 연구는 축적되어 있는 많은 양의 텍스트 자료를 빅데이터 분석 기법을 활용하여 객관적, 종합적으로 대학구조개혁 평가에 대한 주요 쟁점들을 구체화하였다는데 의의가 있으며, 정책 이해 주체 간 유사점과 차이점을 살펴봄으로써 향후 대학구조개혁 평가 정책의 발전 방향 등에 대한 시사점을 제공하고 있다.

주요어 : 대학구조개혁 평가, 텍스트 빅데이터 분석, 토픽 모델링, 잠재 디리쉴레 할당, 교육부 보도 자료, 신문 기사

* 이 논문은 2016년도 한국교육학회 연차학술대회(2016. 6. 25)에서 발표된 원고를 수정·보완한 것임.
** 서울대학교 교육학과 박사수료(교신저자)
*** 서울대학교 교육학과 교수

Ⅰ. 서론

2015년 교육부는 전국의 298개 대학들(일반대, 산업대, 전문대)을 대상으로 대학구조개혁 평가를 시행하였다. 대학구조개혁 평가는 학령인구의 급감에 대비하고 대학 간 균형발전과 고등교육의 경쟁력 확보를 위하여 대학 규모를 줄이면서 대학교육의 질을 높이기 위한 대학 구조개혁 추진계획(교육부, 2014a)에 따라 도입되었고, 대학의 특성을 고려하고 평가부담을 완화할 수 있도록 고등교육기관의 운영 및 교육과정의 필수 요소에 대해 정량평가와 정성평가를 모두 활용하는 형태로 시행되었다(김정민, 2014; 교육부, 한국교육개발원, 2015).

대학구조개혁 평가는 모든 대학을 평가 결과에 따라 5등급으로 분류하여 등급별로 대학 입학정원 감축 및 재정 지원 정도에 차이를 두는 구조개혁 조치에 활용할 것으로 명시(교육부, 2014a)되어 대학 현장에 직접적인 영향을 주기 때문에 시행 과정과 평가 결과 활용 등에 대해 이해당사자들의 다양한 이견 제시와 찬반 논란(한겨레, 2015.08.31.; 중앙일보. 2015.09.02.)이 발생하는 등 많은 쟁점들이 여전히 남아 있다(남궁근, 2014; 이영, 2014; 이기종, 2015; 이영학, 2016).

다수의 선행연구들(남궁근, 2014; 이영, 2014; 이원근, 2014; 강창동, 2015; 반상진, 2015; 박순진, 2016; 신정철, 2016; 이영학, 2016)은 정부의 대학구조개혁 정책이나 대학구조개혁법에 관심을 갖고 논의를 진행하면서 대학구조개혁 평가에 대한 내용은 부분적으로만 포함시키고 있다. 또한 대학구조개혁 평가가 대학구조개혁의 핵심적인 내용의 하나임에도 불구하고 이에 초점을 맞춘 연구들은 부족한 실정이다. 대학구조개혁 평가 방안을 수립하기 위하여 한국교육개발원에서 수행된 김미란 외(2014)의 연구를 제외하고, 평가지표 특성만을 다룬 연구들(김성열, 오범호, 2014; 강성환, 한대희, 2015)과 한국교육평가학회 세미나의 토론 자료들(김신영, 2014; 서민원, 2014; 이기종, 2014; 이원석, 2014), 그리고 이기종(2015)의 연구 정도가 대학구조개혁 평가 자체에 대한 쟁점과 시사점 등을 제시하고 있다. 그러나 이 연구들도 대학구조개혁 평가의 일부 지표에만 초점을 두거나 한정된 문헌에 기초한 내용 분석 및 방안 제시 등으로 제한된 범위의 분석 및 쟁점 도출이 주를 이루었다. 따라서 대학구조개혁 평가에 초점을 맞추어 보다 충분한 정보를 이용하여 객관적이고 종합적인 분석을 수행할 필요가 있다.

신문기사와 같은 언론 자료는 교육에 대한 담론 형성에 중요한 역할을 하는 것으로 알려져 있다(Blackmore & Thorpe, 2003; 하연섭, 2010; 황하성, 손승혜, 장윤재, 2012). 국가인적자원개발에 대한 언론의 관점을 분석한 임경수, 이희수(2009)와 교육정책과 언론의 관계를 다룬 김병주, 김태완, 김은아(2006), 하연섭(2010), 박성태(2011) 등의 연구들은 언론 자료를 이용하는 것이 실제 교육현상에 대한 사회적 관심을 반영하여 보다 폭넓은 시사점을 제시할 수 있음을 보인 바 있다. 아울러 대학구조개혁 평가에 대한 연구에서도 신문기사를 이용한 연구가 수행되었는

데, 이기종(2015)은 구조개혁 또는 구조조정이라는 키워드와 연결된 신문기사 30개를 이용해 의미 연결망 지도를 제시하여 대학구조개혁 평가를 위한 제언을 도출하였다. 다만 이 연구는 실제 대학구조개혁 평가를 다룬 신문기사 중 극히 일부만을 이용하였기 때문에 전반적인 사회적 관심을 반영하였다고 보기에는 한계가 있다. 따라서 보다 많은 언론 자료를 이용한 분석을 수행하여 어떤 관심과 논의가 이루어지고 있는지 확인할 필요가 있다.

한편, 최근에 컴퓨터, 데이터베이스 및 관련 소프트웨어의 기능이 급격하게 발달하면서 숫자나 범주형 자료와 같은 정형 자료(structured data)뿐만 아니라 책, 보고서, 신문 기사, 법원 판례 및 판결문, 웹 페이지, 이메일 등 수많은 다양한 문서자료, 비디오, 이미지 등의 비정형 자료(unstructred data)의 생성, 저장이 폭발적으로 늘어나고 있다(Gan et al., 2014; Daniel, 2015), 이러한 자료를 빅데이터(big data)라는 용어로 표현하는데, 이는 기존 일반적인 데이터베이스 소프트웨어가 수행할 수 있는 데이터 수집 방식이나 저장, 관리 및 분석의 정도를 뛰어넘는 대량의 데이터를 의미(Manyika et al., 2011)한다. 최근에는 자료의 크기로만 빅데이터를 정의하는 것이 아니라 복잡하고 다면적이며 덜 구조화된 특징을 의미하는 다양성(variety)과 어떠한 정보를 발견하는데 사용될 수 있는가라는 가치(value)에도 주목하고 있다(Einav & Levin, 2014; Daniel, 2015). 빅데이터는 기존의 데이터, 이론, 방법과 해석을 발전시키고 더욱 견고하게 만드는 잠재력을 갖고 있다고 인식되고 있어(White & Breckenridge, 2014), 이를 분석하기 위한 다양한 분석 기법들이 발전하고 있다(Gan et al, 2014; George et al., 2014; Daniel, 2015; HanChen, MaoShan & Peng, 2016). 이러한 빅데이터 분석 기법(analytics)은 실시간 자료이며 이전에 많이 다루어지지 않은 새로운 형태의 관찰 자료인 빅데이터로부터 기존의 분석 방식으로 접근할 수 없었던 유용한 정보와 의미를 빠른 속도로 찾아낼 수 있다는 장점이 부각되고 있다(이재성, 홍성찬, 2014; Daniel, 2015; Zakir, Seymour & Berg, 2015).

특히 비정형 자료의 대표적 형태인 텍스트(text) 자료를 이용하는 텍스트 빅데이터 분석 기법은 텍스트 형태의 많은 양의 정보들을 제한된 시간 내에 효과적으로 분석하여, 기존의 방법론이 가지는 연구자의 주관성 개입 문제를 최소하면서도 의미 있는 지식을 찾아낼 수 있다는 장점이 강조되면서 그 중요성과 활용 사례가 점차 증가하고 있다(정다미 외, 2013; Hannigan, 2015; Matthies & Corners, 2015; Moreno & Redondo, 2015; Guo et al., 2016). 텍스트 분석 기법은 일반적으로 텍스트 마이닝(text mining), 텍스트의 지식 발견(knowledge-discovery in text, KDT)이라는 용어와 혼용되고 있으며 비정형 텍스트로부터 중요한 정보나 지식을 추출해 내는 과정을 의미한다(Moreno & Redondo, 2015). 텍스트 빅데이터 분석 기법 중 잠재적 정보(latent intelligence)를 추출해 내는 무감독 학습(혹은 자율학습, unsupervised learning methods) 방법인 토픽 모델링(topic modeling)은 대용량의 텍스트에 내재된 주제나 토픽을 발견할 수 있는 자동화된 텍스트 분석 기법(Steyvers & Griffiths, 2004, 2007; Blei, 2012; HanChen, MaoShan &

Peng, 2016)으로, 토픽 모델링을 이용한 많은 연구들은 신문기사를 분석에 이용하여 전체 내용으로부터 핵심적이고 의미 있는 토픽들을 추출하고 토픽의 변화 경향이나 작성 주체에 따른 차이를 확인하였다(e.g., Newman et al., 2006; Yang, Torget & Mihalcea, 2011; DiMaggio, Nag & Blei, 2013; 강범일 외, 2013; 정다미 외, 2013; 이호엽 외, 2014).

이 연구는 텍스트 빅데이터 분석 기법을 활용해 대학구조개혁 평가에 대한 교육부의 보도 자료들과 주요 신문들의 기사들을 분석하여 대학구조개혁 평가에 대한 쟁점들을 정리한 것이다. 이 연구에서는 기존 연구들이 일부 문헌이나 제한된 연구 대상으로부터 얻은 정보들에만 기초한 것과는 달리, 그동안 교육부와 주요 신문들이 제공했던 많은 양의 광범위한 관련 자료들을 토픽 모델링이라는 대표적인 텍스트 빅데이터 분석 기법을 활용하여 분석함으로써 대학구조개혁 평가에 대한 쟁점들을 체계적이고 객관적으로 분석하고자 하였다. 이 연구의 주요 연구 문제는 다음과 같다. 첫째, 대학구조개혁 평가와 관련하여 교육부의 보도 자료들과 주요 신문들의 기사들에서 주로 어떤 토픽들이 나타나며, 또 시간의 흐름에 따라 어떻게 변화하는가? 둘째, 대학구조개혁 평가와 관련하여 교육부의 보도 자료들과 주요 신문들의 기사들에서 추출된 토픽 간 유사점과 차이점을 비교하였을 때 어떠한 쟁점들이 나타나는가?

Ⅱ. 이론적 배경

1. 대학구조개혁 평가의 주요 내용 및 진행 과정

교육부는 대학교육의 질을 제고하고 학령인구 감소에 따른 대학 입학자원의 급격한 감소에 대비하기 위하여 '2015년 대학 구조개혁 평가 기본계획' 발표한 후 지속적으로 대학구조개혁을 추진하고 있다. 대학구조개혁의 핵심은 대학의 특성화 및 대학 교육의 질을 제고하기 위한 새로운 평가체제 도입(교육부, 2014a)으로, 2015년 시행된 대학구조개혁 평가는 고등교육기관으로서 갖추어야할 필수 요소들을 중심으로 정량지표와 정성지표를 함께 이용한 절대평가 방식의 평가이며, 최근 3년간(2012년~2014년) 대학의 지속적인 노력을 반영하고, 일반대와 전문대를 분리하여 평가하고, 대학의 특성이나 여건을 고려할 수 있는 지표를 사용해 평가의 공정성과 합리성을 확보하기 위하여 노력하였다는 특징이 있다(교육부, 2014b; 교육부, 한국교육개발원, 2015).

일반대, 산업대, 전문대 전체를 평가 대상으로 하지만, 특정 종파 종교지도자 양성대학과 같이 특수성이 확인되는 일부 대학의 경우 평가 대상에서 제외하였고, 일반대학의 경우 2단계 평가, 전문대학의 경우 단일평가를 시행해 5개 등급(A, B, C, D, E등급)으로 구분하였다. '2015년 대학구조개혁 평가 추진 계획'(교육부, 2014a)에 따르면 강제적인 정원 조정이 예고되었으나, 관련

법률안이 아직까지 국회에서 통과되지 않은 관계로 강제적인 정원 감축은 이루어지지 않고 있다. 현재 상위 세 개 등급(A, B, C등급)에 해당하는 대학의 경우 재정지원 사업 평가 등과 연계하여 자율적인 구조개혁을 추진하도록 하는 반면, 하위 등급(D, E등급)의 경우 재정지원 제한 조치와 대학별 맞춤형 컨설팅을 통해 구조개혁을 추진하도록 진행하고 있다(교육부, 2015a).

대학별 맞춤형 컨설팅은 2015년 11월부터 4년제 대학 32개 대학과 전문대 34개 대학을 대상으로 시행되었다(교육부, 2015b; 한국대학신문, 2015.11.09). 한국교육개발원 대학평가본부 주관으로 전문가 5명 내외로 구성된 컨설팅 위원단을 파견해 대학관계자와 컨설팅을 실시하여 대학의 학사 구조, 특성화, 학생지원 등 운영 전반에 대한 개선 등이 이루어질 수 있도록 유도하였다. 컨설팅은 1차와 2차로 나뉘어 진행되었는데, 그 기간 동안 해당 대학은 학내 구성원 등 의견 수렴을 통해 이행과제 의견서를 제출하고 컨설팅 위원단과 면접을 통한 이행 과제 합의를 거쳐 최종 구조개혁 이행 계획서를 수립, 제출하도록 하였다. 이후 과제를 이행하는지에 대한 점검을 통해 교육부와 대학구조개혁위원회에서 단계적으로 2017년 정부 재정지원제한에서 제외할 것인지 여부를 검토하도록 되어 있다.

대학구조개혁 평가와 관련된 구체적 업무 진행 과정의 일시와 내용을 요약하여 제시하면 <표 1>과 같다.

<표 1> 대학구조개혁 평가 진행 과정

일 시	내 용
2013년 8월 1일	대학구조개혁위원회 발족 및 회의 개최
2014년 1월 29일	「대학 구조개혁 추진계획」 발표
2014년 4월 7일	「대학평가 및 구조개혁에 관한 법률안」(김희정 의원 대표 발의) 공청회 개최
2014년 5월~ 9월	대학구조개혁평가 방안 정책연구
2014년 9월 30일	대학구조개혁평가 1차 공청회
2014년 11월 11일	대학구조개혁평가 2차 공청회
2014년 12월 1일	한국교육개발원 내 대학평가본부 신설
2014년 12월 24일	「2015년 대학구조개혁 평가 기본계획」 확정 발표
2014년 12월 26일	2015년 대학구조개혁 평가 기본계획 설명회 개최
2015년 4월 3일	1단계 자체평가보고서(정성평가용) 제출
2015년 4월 28~30일	1단계 일반대 면접 평가 시행
2015년 5월	1단계 정량평가용 지표별 평가자료 및 증빙자료 제출 완료 (일반대 4일, 전문대 11일)
2015년 6월	전문대 면접 평가 시행, 2단계 자체평가 및 자료 제출
2015년 7월	2단계 서면 및 현장 평가 실시
2015년 8월 31일	대학구조개혁 평가 결과 발표
2015년 10월 27일	대학별 맞춤형 컨설팅 설명회 개최
2015년 11월	대학별 맞춤형 컨설팅 실시 시작
2016년 2월	컨설팅 대상 학교의 구조개혁 이행 계획서 제출
2016년 3월	구조개혁 이행 계획서 실행 시작

2. 대학구조개혁 평가에 대한 선행연구 분석

대학구조개혁 평가는 현재 고등교육에서 중요한 이슈이며 논란의 대상임에도 불구하고 이에 대한 연구는 매우 제한적으로 이루어져 왔으며, 선행연구들에서는 세 가지 연구 방법들을 주로 활용하였고, 몇 가지 특징적인 쟁점들이 있는 것으로 보고되었다.

1) 선행연구에서 활용된 주요 연구 방법들

대학구조개혁 평가와 관련된 선행연구들에서는 다음의 세 가지 연구 방법들을 주로 활용하였다.

첫째, 의견 조사 방법을 활용한 정책 제언 도출에 초점을 맞춘 연구가 이루어졌다. 김미란 외(2014)의 연구는 대학구조개혁 평가를 시행하기 전에 관련 정책의 방향을 제시하기 위해 고등교육에 대한 기존 구조개혁 및 평가 정책의 한계와 문제점을 분석하고 면담, 설문조사를 통해 전문가 및 관련 이해 관계자의 의견을 수렴하여 향후 대학구조개혁 평가의 지향점에 대한 정책 제언을 제시하였다. 이태희, 김종인(2015)은 인적자원개발 관점에서 대학구조개혁 평가에 대한 메타평가 준거를 개발하였는데, 이를 위해 문헌 연구, 전문가 좌담회, 조사연구를 수행하였다.

둘째, 모의평가 방법을 활용한 평가지표 개발에 대한 연구가 이루어졌다. 김성열, 오범호(2014)는 대학구조개혁 평가 지표에 대한 실증 연구를 수행하여, 66개 대학의 정량지표 자료를 활용한 모의평가를 실시하고 그 결과에 근거해 대학별 정원감축 규모를 추정한 후 이러한 분석 결과를 바탕으로 대학 구조개혁 평가 모형을 어떻게 설계하느냐에 따라 지역별·설립별 감축규모가 달라짐을 확인하였다. 강성환, 한대희(2015)는 전문대학 주요 평가지표에 대한 지역별 차이와 2011년도부터의 변화 추이를 분석하여 전문대학의 정량지표 값의 수준이 상승하고 지역 간 격차가 감소하는 평준화 현상이 나타났음을 보고하였다.

셋째, 문헌 분석이나 의미연결망지도 분석을 활용한 쟁점 분석 연구가 이루어졌다. 강창동(2015)은 문헌연구를 통해 대학구조개혁 평가 목적의 모순성, 내용의 오류성, 결과의 정치성에 대한 쟁점을 부각한 바 있다. 한편, 이기종(2015)은 대학구조개혁 평가 자체에 초점을 맞추어 구체적 배경과 평가의 개요를 소개하고 관련된 쟁점과 문제점 등을 고찰했는데, 특히 구조개혁 또는 구조조정이라는 표제어와 연결된 30개 신문기사에 대한 의미 연결망 지도 분석을 수행하여 도출된 표제어간 연결 중심성 정보를 기초로 대학구조개혁 평가 시행 방안에 대한 제언을 제시하였다.

이러한 선행연구들은 대학구조개혁 평가의 개선과 발전 방향을 도출하기 위하여 각각 서로 다른 연구 방법을 활용하여 의미 있는 시사점을 제시하였지만, 실증자료 분석 연구의 경우는 대학구조개혁 평가의 일부 평가지표에만 초점을 맞추어 논의하였기 때문에 대학구조개혁 평가

에 대한 부분적인 논의에 그쳤다는 한계가 있다. 또한 의견 조사와 문헌 분석 연구들의 경우는 연구자들의 상황에 따라 임의적으로 수집된 매우 제한적인 자료나 정보만을 이용하여 대학구조개혁 평가에 대해 연구하였다는 한계가 있다.

2) 선행연구에서 분석된 주요 쟁점들

대학구조개혁 정책에서의 평가와 대학구조개혁 평가에 대한 선행 연구들은 주로 다음의 세 가지 쟁점들을 중심으로 논의와 제언을 진행하였다.

첫째, 평가와 평가지표의 객관성, 공정성 확보가 필요함을 주장하였다. 이원근(2014)과 반상진(2015), 강창동(2015)은 대학구조개혁 평가와 평가지표의 공정성과 객관성, 형평성, 신뢰성 확보 및 단기간의 일괄적인 평가로 대학 소재 지역 등 특성 차이를 어떻게 반영할 수 있을 것인가 등에 대해 의문을 제기하였다. 이기종(2015)의 경우 현행 평가 준거가 이론적 근거를 갖추지 않아 타당성이 부족함을 지적하면서 대학구조개혁 평가에 대한 객관성과 공정성이 담보되어야 함을 주장하였다. 평가지표 측면에서 김신영(2014)과 이영(2014)은 정성평가를 활용하는 평가지표가 상당한 비중을 차지하므로 객관성이나 신뢰성을 어떻게 확보할 것인가에 대해 주의할 것을 지적한 반면, 이영학(2016)은 정량지표만으로 대학교육의 질을 확보할 수 없음을 지적하면서 정성지표를 적극적으로 반영할 것을 주장하였다.

둘째, 누가 평가의 주체가 되어 어떻게 운영해야 하는가에 대한 문제를 제기하였다. 남궁근(2014)은 교육부가 정부주도의 대학평가체제와 평가기구 설립이 필요하다고 주장하는 반면, 대학은 획일적이고 강압적인 평가에 반발하여 최소한의 범위에서 한시적으로 이루어져야 함을 주장하고 있다고 지적하였다. 이기종(2015) 또한 시장 주도의 구조개혁은 대학의 경쟁력 차이에 기초하여 이루어지므로 대학의 질적 제고를 자연스럽게 유도할 수 있는 반면 지방대학의 도태를 야기할 수 있어, 정부 주도의 대학구조개혁이 이루어져야 한다는 주장과 서로 대립하고 있음을 지적하였다. 반면 반상진(2015)은 대학구성원이 중심이 되는 전문가 집단이 자율적으로 개혁할 수 있도록 하는 방안이 필요하다고 주장하였고, 신정철(2016)도 대학 상황에 대한 지식과 대학 평가에 축적된 경험이 있는 기관이 평가를 주관하도록 할 것을 주장하였다.

셋째, 평가의 목적과 결과 활용에 대한 의문을 제기하였다. 김춘란(2014)은 대학구조개혁 평가가 대학특성화와 교육의 질을 향상시키는 목적을 갖고 결과를 활용하고자 하여 기존의 대학 평가와 차이가 있음을 주장한 반면, 윤지관(2014), 서민원(2014), 강창동(2015) 등은 정부가 일방적인 대학구조개혁 평가를 통해 정치적 의도를 갖고 평가 결과를 대학에 대한 관리·감독기능을 강화하여 통제하려는 목적으로 사용하고자 한다고 지적하였다.

선행연구들에서 거론된 대학구조개혁 평가에 대한 주요 쟁점들은 지속적으로 논의될 필요가

있으며, 이러한 쟁점들이 교육부의 보도 자료들과 주요 신문들의 기사에서도 나타났는지, 나타났다면 언제 어떻게 나타났는지 혹은 이 외의 또 다른 쟁점들이 존재했는지 등을 확인할 수 있다면 대학구조개혁 평가에 대한 여러 이해당사자들의 다양한 쟁점들을 종합적으로 고려하여 더 나은 논의를 위한 기초를 제공할 수 있을 것이다.

3. 텍스트 빅데이터 분석 기법

빅데이터는 거대한 데이터의 집합을 의미하며, 규모, 다양성, 복잡성, 속도 증가 등의 특성을 갖고 있다(정지선, 2011). 빅데이터 분석에서는 적은 표본으로 큰 모집단을 예측하거나 추론하는 전형적인 통계학을 따르지 않고 확보된 빅데이터로부터 의미를 찾는 것에 초점을 맞추는데, 이는 빅데이터 속의 패턴 혹은 모형을 찾아내기 위한 데이터 분석과 알고리즘 개발을 포함하는 데이터베이스 속의 지식 발견(knowledge-discovery in databases)이라 할 수 있다(Daniel, 2015). 일반적으로 다른 분야에서는 자동적으로 해석 가능한 패턴을 찾아내거나 예측하는 형태로 지식 발견이 주로 시행되는 반면, 사회과학 분야에서는 데이터에서 모형을 발견하거나 더 크고 다양한 형태의 자료를 분석하기 위한 분석 기법을 적용하는 방식이 주로 시행되고 있다(Romero & Ventura, 2010).

텍스트 빅데이터 분석 기법은 비정형 혹은 무정형의 특징을 지닌 자연어 텍스트를 특정한 목적에 유용한 정보 추출을 위해 분석하는 과정으로(Witten, 2005), 기존 정보 이상의 새로운 시각과 예측을 위한 패턴 혹은 전환점을 보여줄 수 있다는 장점이 있다(Reardon, 2014). 일반적으로 책이나 글, 웹페이지 등에 나타나는 문서나 소셜 네트워크 서비스, 블로그의 게재글, 댓글 등에 나타나는 텍스트 등을 대상으로 분석하는데, 기본적으로 이러한 텍스트에 나타나는 단어를 이용하여 분석한다. 하지만 텍스트에 나타나는 단어만을 이용하면 여러 문서에 존재하는 복잡성을 효율적으로 고려할 수 없으므로 문서와 단어만을 이용하여 모형을 만드는 것이 아니라, 문서와 문서에 나타나는 단어 사이에 토픽(topic) 계층이 잠재적으로 존재한다고 가정하고, 이 잠재정보를 추출하여 모형을 만드는 토픽 모델링을 주로 이용한다. 토픽 모델링은 텍스트 빅데이터 분석 기법 중 최근에 정립, 발전하고 있는 가장 주목받는 방법으로(Jockers, 2014), 정책 및 의사결정 연구에서 다양하게 사용되고 있다(HanChen, MaoShan & Peng, 2016).

토픽 모델링에 대해 부연 설명하자면, 텍스트는 위계적 구조를 가지고 있는데, 여러 개의 문서의 종합을 의미하는 코퍼스(corpus)를 가장 상위 수준이라고 하면, 바로 아래 하위 수준은 문서가 되고, 문서 안에 포함된 단어들은 최하위 수준이 된다. 모수를 추정하는 통계적 과정은 가정되는 확률 분포가 다를 뿐 개념적으로는 위계적 선형 모형과 다르지 않다(Blei, Ng & Jordan, 2003). 이 때 문서와 그 안의 단어 사이에 잠재적인 토픽을 가정하여, 문서는 여러 가지 토픽이

존재하는 혼합체이고, 문서 안에 포함되어 있는 토픽은 확률 분포를 따라 생성된다는 아이디어에 기초하는 것이 토픽 모델링이다(Blei, 2012; Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2004, 2007). 즉, 어떠한 문서를 단어자루(a bag of words)로 간주하고 그 안에 다수의 토픽이 존재하는데, 잠재된 각 토픽은 일정한 단어들이 그 순서나 배열에 상관없이 뭉쳐서 나타나므로 이를 파악할 수 있다고 가정한다. 따라서 어떠한 토픽 안에서 함께 등장한 횟수가 많은 단어들은 그 특정 토픽에서 가중치가 높게 나타나고 이러한 관찰 정보를 근거로 잠재된 토픽을 추론한다. 토픽 모델링은 토픽의 수를 정하면 토픽에 포함된 단어들과 각 단어들이 그 토픽에 속할 확률 및 각 문서가 어떤 토픽으로 구성되는지를 자동적으로 산출해 주므로 연구자의 사전 지식이나 분류 작업(labeling) 없이 대량의 문서를 빠르고 명료하게 처리할 수 있는 특징이 있다(남춘호, 2016).

이러한 토픽 모델링을 활용한 연구들이 국내외에서 최근 다양하게 나타나고 있다. 예컨대, 신문자료를 이용하여 토픽 모델링을 활용한 연구로는 언론매체의 정파성을 분석한 강범일 외(2013)의 연구와 법률 관련 기사와 법 개정 간의 관계를 분석한 이호엽 외(2014)의 연구 등이 있다. Grimmer(2010)는 미국 상원들의 보도 자료를 수집하여 의원들이 강조하는 의제가 무엇이며 어떤 방식으로 업무를 홍보하는지 분석하였고, 많은 양의 일기자료에서 의미 있는 토픽을 추출하기 위해 활용하거나(남춘호, 2016), 공공기관 문서, 신년사, 외교문서를 사용한 경우(백영민 외, 2014; Shirota, Hashimoto, & Sakura, 2014; 박종희 외, 2015; Das, Sun, & Dutta, 2016), 학문 분야의 논문을 모두 수집하여 특징적인 주제와 연구 경향을 살펴본 연구(박자현, 송민, 2013; 송혜지 외, 2013; Wang et al., 2016), 강의 계획서의 내용을 자료로 이용한 연구(Apaza et al., 2014) 등 다양한 텍스트 데이터에 대한 활용연구가 이루어졌다. 이러한 연구들은 LDA 알고리즘 기반의 토픽 모델링을 활용하여 텍스트 빅데이터를 이용해 중요 토픽을 추출하고 그 실제적 의미와 변화를 확인하는 것이 효과적임을 보고하고 있다.

Ⅲ. 연구 방법

1. 연구 자료

이 연구는 2013년 1월 1일부터 2016년 4월 30일 사이에 공개된 교육부의 보도 자료들과 주요 신문들의 기사들을 분석 자료로 사용하였다[1]. 교육부 문서는 교육부 홈페이지 (http://www.moe.go.kr)

1) 교육부는 2012년도에 21개교의 경영부실대학을 지정하는 등 평가를 통해 대학구조개혁을 추진해왔으며 특히 2013년 업무계획(교육부, 2013)을 통해 대학에 대한 상시적 구조개혁을 위해 새로운 평가 체제를 도

의 통합검색을 이용해 검색되는 보도 자료들 중 “대학구조개혁 평가”용어를 정확히 포함하는 25개 문서를 직접 다운받아 텍스트 자료로 전환하여 연구에 이용하였다. 25개 중 교육부 업무나 정책 등에 대한 보도 자료는 13개, 신문 기사에 대한 정정 및 해명 자료는 12개였다. 신문 기사는 국내 10대 종합일간지(경향신문, 국민일보, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보)가 제공하는 기사 전문으로, 역시 동일한 용어를 포함하는 총 625개 기사들을 수집하여 데이터로 이용하였다. 각 일간지 홈페이지에서 기사 전문을 수집하였는데, 홈페이지에 접근이 제한적이거나 연구가 설정한 기간 안의 기사를 충분히 제공하지 않는 동아일보, 문화일보, 국민일보, 한국일보의 경우는 검색포탈 “네이버 뉴스” (http://news.naver.com)와 한국언론진흥재단이 제공하는 “빅카인즈” (http://www.kinds.or.kr) 사이트에서 수집하였다. 신문기사 전문은 컴퓨터 프로그래밍 언어인 파이썬(python)의 뷰티풀수프(beautifulSoup) 모듈을 이용하는 웹크롤링(web-crawling) 기법을 적용하여 출처, 제목, 게시일, 인터넷 연결정보인 링크(link), 본문을 자동으로 수집하였고, 교육부 문서와 마찬가지로 “대학구조개혁 평가” 용어를 포함하는 기사만을 추출하였다. 총 글자 수가 500자 미만인 기사는 속보, 인사, 포토 및 영상 관련 뉴스로 파악이 되어 분석 자료에 포함시키지 않았다. <표 2>는 기사 수집 결과를 신문사별로 분류하여 제시한 것이다.

<표 2> 신문사별 대학구조개혁 평가 기사 수

경향 신문	국민 일보	한국 일보	문화 일보	서울 신문	세계 일보	한겨레	조선 일보	중앙 일보	동아 일보	계
47	73	66	24	63	96	55	78	45	78	625

2. 자료에 대한 사전 처리 및 분석

교육부 문서는 반복적으로 나타나는 표 형태의 기본 정보, 즉 작성자, 작성처 등과 그림을 제거하고, 그 이외의 표들은 표 프레임 없이 내용을 확보하여 모든 정보를 텍스트 파일로 만들었다. 신문 기사는 인터넷 사이트를 통해 수집되었기 때문에 신문사별 고유 문구나 연결되는 기사 제목 등이 포함되는 경우가 있어 파이썬 프로그램을 이용해 제거하고 기사 본문만 텍스트 파일로 만들었다.

수집한 모든 텍스트 파일을 R 프로그램을 이용하여 기호, 숫자, 영문을 제거하고 일부 혼용되는 단어[2]를 일치시킨 후 한글자연어처리 KoNLP 패키지를 이용해 명사만 추출[3]하였다. 추출된

입하기로 명시하였다. 따라서 이 연구는 현행 대학구조개혁 평가에 대한 도입과 논의가 본격적으로 전개된 2013년 1월 1일을 자료 수집의 시점으로 설정하였다.

2) “학생들”을 “학생”, “구조조정 평가”를 “구조개혁 평가”, “사립대학”을 “사립대” 등과 같이 동일한 의미이나 다른 표현이 되어 있는 단어

단어 중 일반명사, 고유명사가 아닌데도 오분류 되어 나타난 단어[4]나, 단위, 순서 등을 나타내는 단어[5] 등을 불용어(stopwords)로 처리하고, 띄어쓰기나 맞춤법이 맞지 않아 동일한 단어임에도 다른 단어로 인식하는 경우를 하나로 일치시켰다[6]. 더불어 거의 대부분의 문서에 나타나는 "대학", "평가", "구조개혁", "교육"의 네 단어는 그 단어의 등장 자체로 토픽을 도출하는데 도움이 될 만한 차별적 정보를 갖고 있지 않고, 모든 문서에서 단 한번만 등장하는 단어 또한 큰 도움을 주지 못하기 때문에 분석을 위한 자료에서 제외하였다[7].

사전 처리를 수행한 자료에 대해 R 프로그램의 tm, stringr, topicmodels, lda패키지를 이용하여 단어-문서 행렬(term-document matrix)과 문서-단어 행렬(document-term matrix)을 만들고 토픽의 수를 결정하기 위하여 Grün & Hornik(2011)이 제안한 복잡도(perplexity)를 구하였다. 복잡도는 임의로 연구 자료를 훈련 자료(training data)와 시험 자료(test data)로 나누어 훈련 자료로 얻은 모델을 시험 자료에 적용하였을 때의 질을 수치로 나타낸 것으로 복잡도가 낮을수록 더 좋은 모델임을 의미한다(Battisti, Ferrara & Salini, 2015). 토픽 모델링에서 문서는 많은 수의 토픽을 동시에 포함할 수 있고, 각 문서마다 포함된 토픽이 서로 다를 수 있다는 것을 전제한다. 따라서 전체 문서를 모았을 때 실제로 수많은 토픽이 존재할 수 있기 때문에 모든 내용을 포함하는 많은 수의 토픽이 필요한 것인지, 아니면 관련된 전문가들에 의해 쉽게 이해되고 타당화될 수 있는 제한된 수의 토픽으로 요약할 필요가 있는 것인지에 대한 결정이 이루어져야 한다(Battisti, Ferrara & Salini, 2015). 이 연구는 대학구조개혁 평가에 대해서 중요하면서도 분명하게 나타난 주요 토픽을 추출하고자 하므로 제한된 수의 토픽이 필요하고, 이러한 토픽의 수 결정을 위해서 복잡도를 구하여 이 값을 최소화하는 토픽의 수를 결정하였다.

이후 결정된 토픽의 수에 따라 잠재된 토픽을 추출하도록 잠재 디리쉴레 할당(latent Dirichlet allocation: LDA) 알고리즘 기반 토픽 모델링 분석을 수행하였다. LDA 알고리즘은 베이지안 통계학에 기초하여 토픽의 확률이 디리쉴레 분포(Dirichlet distribution) 형태의 사전 분포를 따른다는 것을 가정하고, 개별 문서는 토픽의 확률적 분포로 표현되고 각 토픽은 단어의 확률적 분포로 표현되는 구조를 이용해 문서 안의 단어들이 어떤 특정 토픽에 포함될 확률을 계산하는 생성 확률 모델(generative probabilistic model)이다. LDA는 통계적 추론의 문제를 단순하게 만들어

3) "요즘 대학들이 학과 통폐합 등 강도 높은 구조조정을 진행 중이다."와 같은 문장은 한글자연어처리 패키지를 이용하면 "요즘 대학 학과 통폐합 강도 구조조정 진행"으로 명사만 추출된다.

4) "마침내", "이제" 등과 같은 부사

5) "첫째", "둘째" 등의 순서를 나타내는 단어

6) "지원 체제"는 "지원체제"로, "서울특별시", "서울시", "서울", "서 울시"를 "서울"로 일치시킴

7) 대부분의 문서에서 나타나는 네 단어와 한 번만 나타나는 단어들을 모두 포함하여 분석을 시도하여도 토픽의 개수에는 변화가 없었으나, 네 단어가 토픽을 설명하는 단어 목록의 상위에 위치하고 두 개 이상의 토픽의 상위에 나타난 경우도 있어 토픽의 특성을 분명하게 파악하는데 도움이 되지 않는다고 판단하였다. 강범일 외(2013), 박종희 외(2015), 남춘호(2016) 등도 동일한 이유로 빈번하게 나타난 단어들을 제거하여 분석하고 있다.

간결하게 데이터의 차원을 축소하는데 유용하며, 해석가능하고 의미적으로 일관성 있는 토픽을 만들어낸다는 장점을 가지는 텍스트 분석 방법으로 알려져 있다(Steyvers & Griffiths, 2007; Mimno & McCallum, 2008). 특히 LDA는 여러 개의 토픽이 내재되어 있는 많은 문서들을 분석에 사용할 수 있고, 같은 의미를 지닌 서로 다른 단어나 문맥에 따라 다른 의미를 가지는 단어들을 분리 혹은 통합하여 효과적으로 다룰 수 있다는 특징이 있으며(Born, Scheihing, Guerra, & Cárcamo, 2014), 디리쉴레 분포를 이용하기 때문에 추출된 토픽 간에 독립성이 두드러지므로 추출된 토픽에 여러 자료에서 공동으로 나타나는 단어보다 특징적인 단어들이 나타나게 된다(김규하, 박철용, 2015).

LDA는 베이지안 모형이므로 사후 확률의 근사치를 주어진 자료로부터 반복적으로 추론하여 토픽을 추출하는데, R 프로그램에서는 variational expection-maximization(VEM) 추정법과 markov chain monte carlo (MCMC) 기법들 중 하나인 붕괴된 기브스 표집(collapsed Gibbs sampling) 을 추론 방식으로 제공한다. 이 연구에서는 Grün & Hornik(2011)의 연구에서 보도자료를 이용한 예와 동일하게 반복회수를 1000으로 설정한 붕괴된 기브스 표집을 이용하여 토픽모델링을 수행하였다.[8)]

Ⅳ. 연구 결과

1. 교육부 문서와 신문 기사의 주요 사용 단어

총 25개의 교육부 문서에 나타난 명사는 총 1,823개였고, 수집된 625개 주요 신문들의 기사에는 총 10,711개 명사가 사용된 것으로 나타났다. 다음의 <표 3>은 교육부 문서와 신문 기사에서 사용빈도가 높은 30개 단어들을 제시한 것이다. “발표”, “결과”, “정부”, “정원” 등 17개 단어가 교육부 문서와 신문기사에서 모두 빈번하게 사용된 것으로 나타났다.

8) Grün & Hornik(2011)의 연구에서 토픽의 수를 결정하기 위해 제공하는 복잡도(perplexity)를 고려하여 동일한 조건에서 VEM과 붕괴된 기브스 표집을 이용한 토픽 추출 결과를 비교하였는데, 붕괴된 기브스 표집을 적용할 때 더 적은 수의 토픽이 추출되었다. 따라서 토픽의 해석 가능성을 고려하여 Grün & Hornik(2011)의 연구 결과에서와 같이 더 적은 수의 토픽을 추출할 것으로 기대되는 붕괴된 기브스 표집 추론 방식을 연구에 적용하였다.

<표 3> 교육부 문서와 신문 기사에 나타난 빈도가 높은 단어

교육부 문서		신문 기사	
문서 수	단어	문서 수	단어
20	계획, **내용**, 발표	300개 이상	정부, 정원, 결과, 지원, 발표
19	결과	281~299개	감축, **학교**
16	전문, 지방, 지표	261~280개	사업
15	방안, **예정**, 정부, 정원, 추진	241~260개	**입학, 교수**, 추진, 계획
14	대상, 사업, **연계**, 재정지원, 지원	221~240개	**총장**, 재정지원, **제한, 사회**, 전문, **경쟁**, 운영
13	감축, **노력, 도입**, 운영, 특성화, **현장**	201~220개	지표, 특성화, **국가**, 대상, 방안
12	**교육과정, 반영, 실시, 여건, 질, 개최, 검토**	181~200개	**학과, 지역, 필요**, 지방, **장학금, 문제, 취업**

* 굵은 글씨는 빈번하게 사용된 단어 30개 중 각 문서에서만 나타난 것들을 표시한 것임.

교육부 문서에서는 “내용”, “예정”, “연계”, “노력”, “도입”, “반영”, “개최”, “검토” 등 정부의 대학구조개혁 평가 관련 홍보 및 내용 전달을 위한 단어가 빈번하게 사용된 것으로 나타났고, 신문 기사에서는 “학교”, “입학”, “교수”, “총장”, “경쟁”, “학과”, “장학금”, “취업” 등 대학 현장과 관련이 높은 단어가 빈번하게 사용된 것으로 나타났다.

2. 교육부 문서와 신문 기사의 토픽 분석

LDA 토픽 모델링을 수행하기 위해 복잡도를 고려해 교육부 문서들의 적절한 토픽 개수를 결정한 결과 3개가 가장 적절한 것으로 나타났다.[9] 다음의 <표 4>는 교육부 문서들에 대해 토픽 모델링 분석을 수행한 결과, 추출된 토픽에 포함된 15개 단어를 제시한 결과이다. 각 토픽에 나타날 확률이 높은 단어순서대로 제시하였고 이와 함께 해당 토픽이 가장 대표적으로 나타난 문서의 수를 제시하였다.

<표 4> 교육부 문서에 대한 토픽과 포함된 단어 및 해당 문서 수

토픽명		포함 단어	문서수
토픽1	추진 배경	방안, 노력, 실시, 참여, 국가, 중심, 개선, 부담, 비율, 사회, 인구, 필요, 학과, 질, 대학교육	6
토픽2	운영 방안	지방, 계획, 운영, 현장, 교육과정, 반영, 개최, 의견, 협의, 고려, 연구, 위원회, 적극, 적용, 제한	5
토픽3	결과 활용	내용, 발표, 관련, 결과, 지표, 예정, 추진, 대상, 사업, 재정지원, 정부, 정원, 주요, 지원, 감축	14

9) Grün & Hornik(2011)의 연구에서와 같이 10등분 교차검증(10-fold cross validation) 방식으로 토픽의 수를 2개부터 50개까지 설정하여 복잡도를 구한 후, 값이 감소하는 폭이 완만해지는 지점의 토픽의 수를 선택하였다.

토픽 모델링을 적용한 결과는 단어들의 집합만을 제공하므로(강범일 외, 2013; Matties & Corners, 2015) 토픽명은 단어들이 갖는 의미적인 연계성을 고려하여 연구자들이 부여한 것이다. 교육부 문서들의 첫 번째 토픽에는 “방안”, “노력”, “실시”, “개선”, “인구”, “필요”, “질”, “대학교육” 등의 단어가 포함된 것으로 나타났는데, 이는 대학구조개혁 평가의 추진 배경 및 필요성을 나타내는 맥락에 등장하는 단어들로 판단되어 해당 토픽명을 ‘추진 배경’으로 명명하였다. ‘추진 배경’ 토픽을 나타내는 문서의 개수는 6개였다. 두 번째 토픽은 “지방”, “계획”, “현장”, “교육과정”, “반영”, “협의” 등의 단어가 포함되어 있는데 의미와 맥락을 고려할 때 대학구조개혁 평가 시행 혹은 운영과 관련한 내용을 나타내어 “운영 방안”으로 토픽명을 정하였고, 이에 해당하는 문서의 개수는 5개로 나타났다. 가장 많은 문서에서 나타난 토픽은 세 번째 ‘결과 활용’ 토픽으로 모두 14개 문서가 이에 해당되었다. ‘결과 활용’에는 “발표”, “관련”, “예정”, “지원”, “사업”, “재정지원”, “감축” 등의 단어가 포함되어 있다.

주요 신문들의 기사에 대한 토픽 모델링 분석에서 복잡도를 고려하였을 때 적절한 토픽의 수는 7개인 것으로 나타났고, 추출된 7개 토픽에 나타날 확률이 높은 단어 15개를 순서대로 제시한 결과는 다음의 <표 5>에 제시하였다.

<표 5> 신문 기사에 대한 토픽과 포함된 단어 및 해당 문서 수

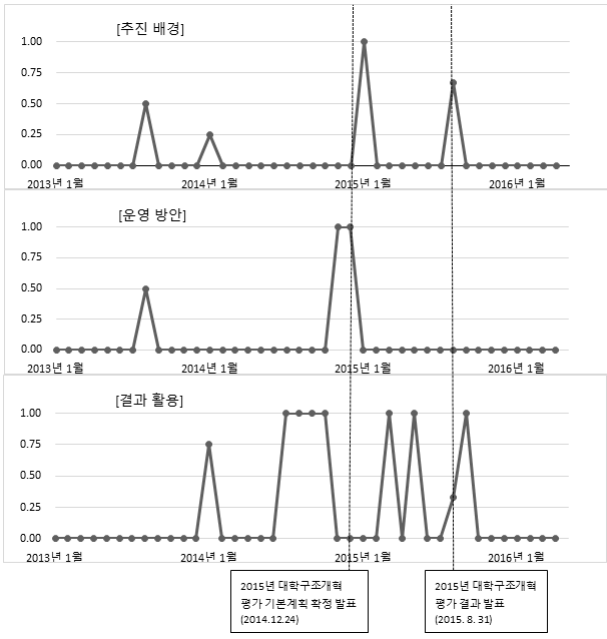
토픽명		포함 단어	문서 수
토픽1	추진 배경	필요, 경쟁, 사회, 감소, 고교, 문제, 인구, 졸업생, 고등교육, 정책, 상황, 강조, 질, 방향, 현실	88
토픽2	평가 방식	지표, 감축, 정원, 지방, 발표, 방식, 정량, 수도권, 지적, 단계, 사립대, 확보, 계획, 취업률, 지역	110
토픽3	운영 방안	지원, 추진, 과정, 계획, 시행, 제도, 등록금, 예정, 개선, 학기, 학과, 장관, 대상, 운영, 확대	65
토픽4	결과 활용	정부, 제한, 재정지원, 결과, 발표, 전문, 장학금, 국가, 재정, 내년, 퇴출, 감축, 학자금, 정원, 국회	90
토픽5	고등교육의 질	특성화, 운영, 인재, 우수, 역량, 취업, 양성, 사업, 협력, 산학, 프로그램, 분야, 세계, 선정, 강화	111
토픽6	대학의 책무성	총장, 학교, 교수, 구성, 요구, 촉구, 관계자, 국립대, 책임, 주장, 사퇴, 정상, 반발, 학내, 반대	109
토픽7	대학의 홍보	선정, 신입생, 학과, 재학생, 학부, 설명, 모집, 장학금, 취업, 성적, 선발, 졸업, 전공, 수준, 포함	52

첫 번째 토픽은 교육부 문서에서와 같이 ‘추진 배경’으로 토픽명을 정하였는데, 포함된 단어를 살펴보면 “필요”, “감소”, “문제”, “인구”, “고등교육”, “질” 등 대학구조개혁 평가의 핵심 배경인 학령인구 감소와 고등교육의 질 제고와 관련된 단어들로 구성되어 있다. 두 번째 토픽은 “지표”, “지방”, “정량”, “수도권”, “지적”, “단계”, “취업률” 등이 2단계 평가를 실시하며 지방대학

과 수도권 대학 등 대학 소재지 특성을 반영하는 대학구조개혁평가 방식을 나타내고 있어 '평가 방식'으로 토픽명을 정하였다. 같은 방식으로 세 번째 토픽부터 순서대로 '운영 방안', '결과 활용', '고등교육의 질', '대학의 책무성', '대학의 홍보'로 토픽명을 정하였다. '고등교육의 질' 토픽은 대학의 특성화와 우수 인재 양성 등 대학교육의 질 제고와 맥락을 같이하는 단어로 구성되어 있으며, '대학의 책무성' 토픽은 대학구조개혁 평가 결과 하위 등급(D, E 등급)으로 판정되어 이에 대한 총장 및 대학 본부, 관계자의 사퇴 등 일련의 과정을 겪은 대학들과 연관된 단어들로 구성되어 있는 것으로 나타났다. 반면에 '대학의 홍보' 토픽은 대학구조개혁 평가 결과 A 혹은 B등급을 받은 지방 대학의 경우 대학 홍보를 위해 대학구조개혁 평가 결과를 기사 내용에 포함시키는 경우가 있었는데 이 때 나타나는 단어들로 주로 구성되어 있었다. '평가 방식', '고등교육의 질', '대학의 책무성' 토픽이 나타난 신문기사의 수가 100개 이상으로 나타났고, '대학의 홍보'와 '운영 방안' 토픽을 주로 다룬 신문기사의 수가 상대적으로 다소 적었다.

3. 교육부 문서와 신문 기사의 토픽 변화

다음의 [그림 1]과 [그림 2]는 교육부 문서에서 추출된 3가지 토픽과 신문 기사에서 추출된 7가지 토픽별로 각 토픽이 가장 분명하게 나타난 문서가 해당 시기 전체 문서 중에서 차지하는 비율이 시간에 따라 어떻게 변화하는지를 나타낸 그래프이다.



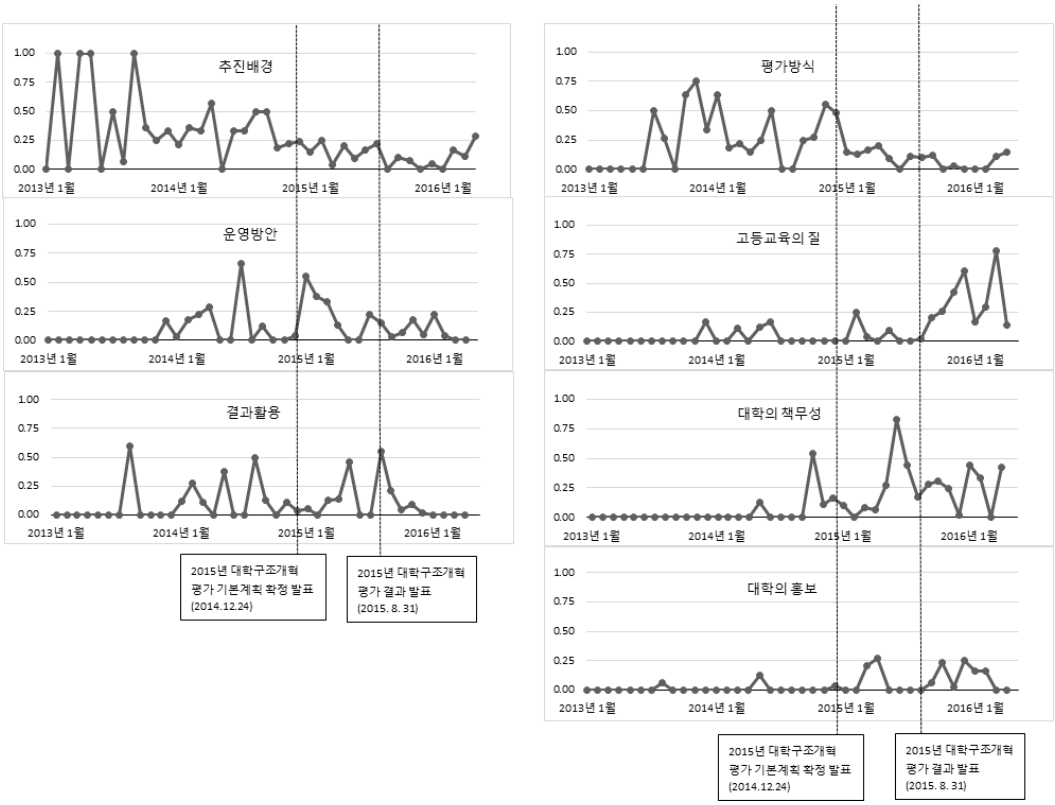
[그림 1] 교육부 문서 토픽별 토픽 비율의 변화

[그림 1]에 나타난 바와 같이 ‘추진 배경’ 토픽의 경우 대학구조개혁위원회가 발족된 2013년 8월과 대학구조개혁 추진계획이 발표된 2014년 1월에 교육부 문서에 등장하였다가 2014년 12월 기본 계획의 확정 발표와 2015년 8월 대학구조개혁 평가 결과의 발표 시기에 상대적으로 높은 비율로 나타나고 있다. ‘운영 방안’은 대학구조개혁 평가 기본계획이 확정되어 구체적인 평가 방식과 지표가 발표된 시기에 교육부 문서에서 가장 두드러지게 나타나는 토픽이었다.

‘결과 활용’은 교육부 문서에서 가장 많이 나타난 토픽으로, ‘추진 배경’이나 ‘운영 방안’ 토픽과 같이 특정한 교육부 발표 시기의 문서에 주로 등장하는 것이 아니라 2014년 1월 대학구조개혁 추진계획이 발표된 이후 지속적으로 등장하는 것으로 나타났다.

[그림 2]에 나타난 신문 기사 토픽의 변화 양상을 살펴보면 ‘추진 배경’ 토픽은 대학구조개혁 추진 계획이 발표되고 공청회 등을 통해 대학구조개혁 평가 시행을 준비하던 2013년과 2014년 신문 기사에 나타난 비율이 두드러지게 높았다가, 대학구조개혁 평가 기본계획이 확정 발표된 이후에는 관심 토픽으로 등장한 비율이 상대적으로 현저하게 낮아졌다. ‘운영 방안’ 토픽의 경우는 대학구조개혁 평가 공청회와 기본계획 확정 발표 시기에 나타난 신문 기사에 나타난 비율이 가장 높았고, ‘결과 활용’의 경우 2013년 8월 대학구조개혁위원회가 발족한 시기 이후 꾸준히 나타나다가 2015년 8월 대학구조개혁 평가 결과 발표 이후에 신문 기사에 나타난 비율이 급격하게 감소한 것으로 나타났다.

‘결과 활용’ 토픽과 유사하게 ‘평가 방식’ 토픽도 2014년 12월 대학구조개혁 평가 기본계획이 확정 발표되기 이전에는 꾸준하게 등장하는 토픽이었지만 발표 이후 신문 기사에서 다뤄진 비율이 상당히 낮아진 것으로 나타났다. ‘고등교육의 질’의 경우 2015년 8월 대학구조개혁 평가 결과 발표 이후에 두드러지게 나타나는 토픽이었고, ‘대학의 책무성’ 토픽의 경우 2015년 5월 대학구조개혁 평가 1단계 평가가 시행된 이후 신문 기사에서 다루어진 비율이 급격히 높아졌고 2015년 8월 평가 결과 발표 이후에는 신문 기사에 지속적으로 주요 토픽으로 나타나는 것으로 나타났다. ‘대학의 홍보’ 토픽은 대부분 2014년 12월 대학구조개혁 평가 기본계획 확정 발표 이후에 주로 등장하였는데 2015년 8월 평가 결과 발표 이후 차지하는 비율은 높지 않지만 신문 기사에서 꾸준히 나타나고 있다.



[그림 2] 신문 기사 토픽별 토픽 비율의 변화

4. 교육부 문서와 신문기사 비교

토픽 모델링은 문서에는 직접적으로 나타나지 않지만 작성자가 특정한 토픽 하에서 단어를 선택하고 글을 작성하는 것을 가정하기 때문에, 분석을 통해 얻은 통계적으로 나타날 확률이 높은 단어들의 집합은 잠재된 토픽의 내용과 의미를 추론할 수 있는 정보가 된다.

교육부 문서와 신문 기사를 이용하여 대학구조개혁 평가에 대해 이들 두 주체가 어떠한 인식을 갖고 있는지 살펴본 결과, ‘추진 배경’, ‘운영 방안’, ‘결과 활용’ 토픽이 모두 나타나 문서에 잠재된 관심의 유사성을 발견하였다. 하지만 ‘추진 배경’과 ‘운영 방안’ 토픽을 구성하는 단어들에 다소 차이가 있는데, 교육부의 ‘추진 배경’ 토픽에는 “노력”, “실시”, “참여”, “국가”, “필요” 등의 단어들이 포함되어 대학구조개혁 평가의 관련 주체들의 노력과 필요성 인식을 표현하고 있는 것으로 나타난 반면, 신문기사는 “경쟁”, “감소”, “고교”, “문제”, “방향”, “현실” 등 현재 직면하고 있는 고등교육의 문제 환경을 주로 다루고 있는 것으로 나타났다.

‘운영 방안’ 토픽에서도 교육부는 “현장”, “반영”, “개최”, “협의”, “적극” 등 소통과 공유를

의도하는 운영방안을 주로 문서에서 언급하는 반면, 신문 기사에서는 “지원”, “추진”, “과정”, “시행”, “개선” 등 대학구조개혁 평가를 진행하는 과정 자체를 표현하고 있는 것으로 나타났다. ‘결과 활용’ 토픽의 경우는 “발표”, “결과”, “재정지원”, “정원”, “감축” 등 다수의 동일한 단어가 공통적으로 사용되는 확률이 높아 교육부 문서와 신문 기사에 나타난 대학구조개혁 평가의 결과에 대한 인식 및 표현은 유사한 것으로 나타났다.

신문 기사에서는 이 세 가지 토픽 이외에도 ‘평가 방식’, ‘고등교육의 질’, ‘대학의 책무성’, ‘대학의 홍보’라는 네 가지 토픽이 추출되어 대학구조개혁 평가에 대한 관심이 보다 다양하게 나타나고 있음을 확인할 수 있었다.

V. 요약 및 논의

이 연구는 대학구조개혁 평가와 관련된 주요 쟁점들을 분석하기 위하여, 관련 교육부 보도자료와 주요 신문들의 기사들을 자료로 하여 텍스트 빅데이터 분석 기법을 활용해 문서 안에 내재된 토픽을 추출하고 변화 양상과 내용 등의 특징을 살펴보았다. 교육부 문서에서는 ‘추진 배경’, ‘운영 방안’, ‘결과 활용’의 3가지 토픽이 추출되었고 이 중 ‘결과 활용’ 토픽을 다룬 문서의 수가 가장 많았다. 신문 기사에서는 이에 더하여 ‘평가 방식’, ‘고등교육의 질’, ‘대학의 책무성’, ‘대학의 홍보’의 총 7가지 토픽이 추출되었고, ‘고등교육의 질’, ‘평가 방식’, ‘대학의 책무성’ 토픽을 포함한 기사가 상대적으로 많이 나타났다. 시간의 흐름에 따라 문서에서 우세하게 나타나는 토픽이 달라짐을 확인할 수 있었는데, 교육부 문서의 경우 ‘결과 활용’ 토픽이 가장 많이 지속적으로 나타났다. 반면에 신문 기사의 경우에는 대학구조개혁 평가 시행 이전 초기 단계에서는 ‘추진 배경’ 토픽이 주로 나타나 학령인구의 감소와 대학교육의 경쟁력 확보라는 대학구조개혁의 필요성에 대한 논의가 주로 이루어진 것으로 나타났다. 그리고 대학구조개혁 평가의 시행과 관련한 공청회 및 시행 계획 발표 시기 이전에는 ‘평가 방식’에 대한 토픽이 주로 나타났다. 대학구조개혁 평가가 실제로 실시되어 결과가 발표되기 이전까지의 시기에는 ‘운영 방안’, ‘결과 활용’ 토픽에 초점을 둔 신문 기사들이 상대적으로 많았고, 특히 대학구조개혁 평가 결과 발표를 앞두고 ‘대학의 책무성’ 토픽을 다룬 기사들이 많아졌는데 대학구조개혁 평가의 2단계 서면·현장 평가 실시 시기와 맞물려 대학의 반발과 책임에 대한 논의가 일어났음을 시사하였다. 대학구조개혁 평가 결과가 발표된 이후에는 주로 ‘고등교육의 질’, ‘대학의 책무성’, ‘대학의 홍보’ 토픽을 다룬 신문 기사가 많은 것으로 나타났다. 이는 대학구조개혁 평가가 대학에 미친 영향과 향후 발전 방향에 대한 논의가 이 시기 신문들의 주요 관심이었음을 나타낸다.

교육부 문서와 신문 기사 간 관심을 갖는 토픽 간 유사점과 차이점을 살펴본 결과, ‘추진 배

경', '운영 방안', '결과 활용'에 관련한 토픽은 교육부 문서와 주요 신문들의 기사에서 모두 추출되어 공통의 관심사인 것으로 나타났지만, '추진 배경'과 '운영 방안'에 대한 토픽들에서 의미를 표현하기 위해 사용하는 단어의 구성이 다소 차이가 있었다. 또한 교육부 문서의 '추진 배경' 토픽에서 "국가", "중심"이 주요 사용 단어로 나타난 반면, 신문 기사에서 추출된 토픽에서는 대학구조개혁 평가의 주체로서 논의되었던 대학, 교수, 전문가 등의 단어가 분명하게 나타나지 않았다. 이는 남궁근(2014)과 이기종(2015)이 언급한 바와 같이 교육부와 대학 중 누가 대학구조개혁 평가를 주도하는 것이 바람직한가에 대한 이견이 쟁점으로 존재함을 확인한 것으로, 다만 교육부 문서에 비해 신문 기사는 이에 대해 명확하게 표현하고 있지 않은 것으로 보인다. 교육부 문서와 신문 기사에서 '결과 활용' 토픽에 다수의 동일한 단어들이 나타났지만, 윤지관(2014), 서민원(2014), 강창동(2015) 등의 주장처럼 대학에 대한 통제와 관리, 감독과 연관될 수 있는 "제한", "퇴출", "학자금" 등의 단어들은 신문 기사에서 주로 나타나 교육부 문서보다 대학에 대한 제재 측면을 강조하였음을 시사하였다.

더불어 주요 신문들의 기사에서는, 교육부의 문서 분석에서 드러나지 않았던 토픽들이 추가로 나타났는데, 대학구조개혁 평가의 '평가 방식' 자체와 대학교육의 질 및 나아가야 할 방향에 대한 논의가 담겨져 있는 '고등교육의 질', 평가결과 발생한 대학의 반발과 진통, 수습 과정에 대한 논의가 담겨있는 '대학의 책무성' 토픽에 주로 관심을 갖고 있는 것으로 나타났다. 신문 기사 분석에서 나타난 '평가 방식' 토픽은 선행연구들에서 나타나는 대학구조개혁 평가에 대한 쟁점들 중 하나와 유사한 내용을 포함하는 것으로 보이지만, 선행연구에서 주장하는 공정성, 객관성 등 평가 방식의 방향성에 대한 단어들의 사용은 뚜렷하게 나타나지 않고 평가지표, 정원감축, 정량 평가, 평가 방식, 2단계 평가 등의 표현들이 나타나 어떻게 대학구조개혁 평가가 시행되는지에 대한 정보 전달 수준의 토픽인 것으로 나타났다.

한편, 선행연구에서의 쟁점과 비교하여 교육부와 주요 신문들이 갖는 대학구조개혁 평가에 대한 쟁점들을 정리하면 다음과 같다. 첫째, 대학구조개혁 평가에 대한 교육부와 주요 신문들의 토픽 분석에서 유사한 부분도 있었지만 주로 차이가 나타났고, 일부 쟁점에 주요한 관심을 갖고 있는 것으로 나타났다. 즉 각각의 토픽에 사용된 단어들을 통해 분석해 보았을 때 교육부 문서는 평가의 주체에 대해 상대적으로 더 강조하고 있음에 반해, 주요 신문들의 기사는 대학에 대한 통제, 대학이 직면한 문제 상황 등을 상대적으로 더 강조함으로써 선행연구들에서 나타난 대학구조개혁 평가의 목적과 결과 활용 측면의 쟁점을 부각시키고 있었다. 이는 대학구조개혁 평가와 관련된 신문기사의 표제어 간 의미 연결망을 분석하여 퇴출, 재정지원제한, 부실 대학 등의 연결 정도가 강하게 나타났다고 보고한 이기종(2015)의 연구 결과와 유사하다. 이외의 쟁점들, 즉 평가 지표에 대한 논의, 평가 운영 방식에 대한 의견 등은 교육부나 신문 기사의 토픽으로 분명하게 나타나지 않았는데, 이들 쟁점에 대해서는 정보 전달 수준에 그치고 있었음을 시사하

였다. 박세훈(2015), 신정철(2015), 이기종(2015) 등도 지적한 바와 같이 대학구조개혁 정책은 지속적으로 추진되어야하기 때문에 대학구조개혁 평가에 대한 적극적이고 다각적인 논의가 필요하다. 따라서 대학구조개혁 평가 정책의 개선·발전을 위해 교육부 보도 자료와 신문 기사들의 관심과 쟁점이 좀 더 다양하게 제기되고 종합적으로 논의될 필요가 있다.

둘째, 대학구조개혁 평가에 대한 쟁점은 정책 진행 단계 및 시기에 따라 변화하는 것으로 나타났다. 대학구조개혁 평가와 관련하여 중요하고 필요한 쟁점이라면 지속적으로 논의될 필요가 있는데, 교육부와 신문 기사가 꾸준히 관심을 갖는 쟁점은 분명하게 나타나지 않았고 시기에 따라 관심을 갖는 토픽들이 변화하여 수시로 쟁점이 달라지는 것을 확인할 수 있었다. 예컨대, 대학구조개혁 평가 기본 계획이 확정되기 전에는 '추진 배경'이나 '평가 방식'에 높은 관심을 보이다가, 그 이후에는 '대학의 책무성'이나 '고등교육의 질'에 높은 관심을 보였다. 이러한 결과는 주요 신문들이 특정 쟁점을 지속적으로 논의하기보다 새로운 정보의 전달이나 일반 대중의 관심에 초점을 맞추어 수시로 쟁점을 바꾸고 있음을 시사하였다. 강창동(2015), 박세훈(2015), 박순진(2015), 신정철(2015), 이기종(2015) 등이 제기한 바와 같이 대학구조개혁 평가의 내용이나 운영에 대한 꾸준한 관심과 논의들이 필요함에도 불구하고 교육부 보도 자료나 신문 기사에서는 뚜렷하게 나타나지 않으므로, 향후 대학구조개혁 평가 정책의 개선·발전을 위해 이러한 쟁점들에 관심을 유지하고 지속적으로 논의할 필요가 있다.

이 연구는 기존의 선행연구가 다루지 못했던 많은 양의 문서 자료들을 분석하여 대학구조개혁 평가에 대한 관심을 객관적이고 종합적으로 분석하였다는 점에 의의가 있다. 분석에 활용한 토픽 모델링은 통계적으로 나타날 확률이 높은 단어들의 집합으로 문서에 잠재된 토픽의 내용을 추론할 수 있는 정보를 제공하는데, 이를 이용하여 대학구조개혁 평가에 대한 주요 관심과 쟁점이 무엇인지를 구체적으로 확인할 수 있었다. 또한 시간에 따라 관심과 쟁점이 어떻게 변화하는지, 주체에 따라 유사점과 차이점이 있는지 등을 비교하여 교육부와 주요 신문들 간에 부분적인 인식의 차이가 있음도 확인할 수 있었다. 주요 신문들이 대학구조개혁 평가에 대해 더 다양한 관심을 갖고 있었으며, 특히 대학구조개혁 평가 결과 발표 이후 대학이 지향해야 하는 교육의 질이나 책무성에 대한 논의가 증가하고 있다는 것은 사회적 관심과 변화의 방향을 확인한 것으로, 향후 이에 부응하는 방향으로 대학구조개혁 평가가 지속적으로 개선·발전해 나가야 할 것이다.

끝으로, 이 연구에서는 교육부 문서의 수가 상대적으로 부족하여 교육부와 주요 신문들 간 토픽의 특징이나 변화를 보다 세밀하게 비교하기 어려웠다. 아울러 신문 기사의 경우 10대 종합일간지만을 분석 대상으로 삼았기 때문에 TV, 그 외의 신문이나 잡지, 인터넷 뉴스 등의 자료를 분석 대상에 포함시키지 않았다. 향후 연구에서는 더 많은 관련 문서와 언론 매체의 자료들을 대상으로 자료의 종류나 작성자의 특성에 따라 대학구조개혁 평가에 대한 쟁점을 세밀하게 비교, 분석할 필요가 있다. 그리고 이 연구에서는 교육 정책 연구를 위하여 텍스트 빅데이터 분석

기법의 활용 가능성을 탐색하기 위한 시도로서 토픽 모델링의 가장 일반적 형태인 LDA 알고리즘을 적용하였지만, 향후 연구에서는 토픽 모델링 알고리즘의 사전(prior) 및 사후(posterior) 분포의 종류와 특성을 고려하고, 토픽 간 상관을 가정하는 상관 토픽 모델(correlated topics model), 작성 시간이나 작성자 등 문서의 특성을 모형에 포함시켜 분석하는 디리쉴레 다항 회귀 토픽 모형(Dirichlet-multinomial regression topic models) 등 다양한 형태의 토픽 모델링 모형들을 적용하여 분석할 필요가 있다.

참고문헌

강범일, 송민, 조화순(2013). 토픽 모델링을 이용한 신문자료의 오피니언 마이닝에 대한 연구. **한국문헌정보학회지, 47**(4), 315-334.

강성환, 한대희(2015). 전문대학 평가지표 개선을 위한 주요 정량지표 변화분석 연구. **직업교육연구, 34**(4), 151-168.

강창동(2015). 정부의 「대학구조개혁」정책에 관한 비판적 연구. **한국교육학연구, 21**(4), 275-306.

교육부(2013). 행복교육, 창의인재 양성 - 2013년 국정과제 실천계획. (2013.3.28).

교육부(2014a). 대학 교육의 질 제고 및 학령인구 급감 대비를 위한 대학 구조개혁 추진계획. (2014.1.28.).

교육부(2014b). 2015년 대학구조개혁평가 기본 계획(안). (2014.12).

교육부(2015a). [설명자료] 대학평가서 D·E 등급 나와도 정원 강제로 못 줄인다. (2015.5).

교육부(2015b). 대학구조개혁 평가결과 발표. (2015.8.31.).

교육부, 한국교육개발원(2015). **2015년 대학 구조개혁 평가 대학 담당자 설명회 자료집**. 연구자료 CRM-48. 서울: 한국교육개발원.

김규하, 박철용(2015). 토픽 모형 및 사회연결망 분석을 이용한 한국데이터정보과학회지 영문초록 분석. **한국데이터정보과학회지, 26**(1), 151-159.

김병주, 김태완, 김은아(2006). 교원평가제에 대한 신문의 보도태도 분석. **한국교원교육연구, 23**(1), 349-371.

김미란, 이정미, 김정민, 서영인, 심우정(2014). **대학 구조개혁 평가 방향 정립을 위한 대학평가 운영 실태 분석**. 한국교육개발원 현안보고 OR 2014-07.

김성열, 오범호(2014). 합리적 대학 구조개혁 평가모형 설계를 위한 제안: 시뮬레이션 결과를 중심으로. **교육정치학연구, 21**(4), 49-68.

김신영(2014). 대학구조개혁 평가지표 및 평가절차에 대한 토론. **2014년 한국교육평가학회 세미나 자료집: 대학구조개혁 평가방안의 타당성**, 31-35.

김정민(2014). 대학구조개혁 평가 방안. **2014년 한국교육평가학회 세미나 자료집: 대학구조개혁 평가방안의 타당성**, 3-19.

김춘란(2014). 고등교육의 경쟁력 강화와 대학구조개혁 추진. **The HRD Review, 2014년 7월**, 74-83.

남궁근(2014). 대학 구조개혁정책의 쟁점과 과제. 2014 세계행정학술회의 교육부 특별세션 발제문.

남춘호(2016). 일기자료 연구에서 토픽모델링 기법의 활용가능성 1검토. **비교문화연구, 22**(1), 89-135.

박성태(2011). 사회 갈등적 공공이슈에 대한 언론의 보도태도연구: 정권교체기 보수와 진보언론의 교육정책 관련 보도태도 분석. **한국공공관리학보, 25**(3), 97-118.

박세훈(2015). "대학구조개혁법, 어떻게 할 것인가?"에 대한 토론. **제3회 대학구조개혁법 토론회 자료집: 대학구조개혁법, 어떻게 할 것인가?** 37-42.

박순진(2016). "대학구조개혁법, 어떻게 할 것인가?"에 대한 토론. **제3회 대학구조개혁법 토론회 자료집: 대학구조개혁법, 어떻게 할 것인가?** 49-56.

박자현, 송민(2013). 토픽 모델링을 활용한 국내 문헌정보학 연구동향 분석. **정보관리학회지, 30**(1), 7-32.

박종희, 박은정, 조동준(2015). 북한 신년사 텍스트 분석, 1946-2015. **한국정치학회보, 49**(2), 27-61.

반상진(2015). 대학구조개혁정책의 쟁점과 대응 과제에 관한 연구: 학령인구 감소에 대한 새로운 대학구조개혁 패러다임 탐색. **공학교육연구, 18**(2), 14-26.

백영민, 최문호, 장지연(2014). 한미 정권교체에 따른 주한 미대사관 외교문서의 주체와 감정표현 변화. **언론정보연구, 51**(1), 133-179.

서민원(2014). 대학구조개혁평가 방안의 타당성에 대한 토론. **2014년 한국교육평가학회 세미나 자료집: 대학구조개혁 평가방안의 타당성,** 23-30.

송혜지, 박경수, 정혜은, 송민(2013). 텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석. **2013 한국정보관리학회 학술대회 논문집, 20**, 47-50.

신정철(2016). 대학구조개혁과 대학평가. **제3회 대학구조개혁법 토론회 자료집: 대학구조개혁법, 어떻게 할 것인가?** 3-20.

윤지관(2014). [시평] 대학구조조정과 평가: 대학의 폐허화, 이대로 방치할 것인가, 대학 구조조정의 정치학. **안과밖, 36**(1), 143-162.

이기종(2014). 대학구조개혁평가에서의 이슈. **2014년 한국교육평가학회 세미나 자료집: 대학구조개혁 평가방안의 타당성,** 36-39.

이기종(2015). 대학 구조 개혁 평가의 배경, 쟁점 및 대안. **교육평가연구, 28**(3), 933-954.

이영(2014). 대학 구조개혁 정책의 쟁점과 과제. **The HRD Review, 2014년 11월**, 22-32.

이영학(2016). "대학구조개혁과 평가"에 대한 토론. **제3회 대학구조개혁법 토론회 자료집: 대학구조개혁법, 어떻게 할 것인가?** 43-47.

이원근(2014). 대학 구조개혁의 쟁점과 과제. **The HRD Review, 2014년 11월**, 2-5.

이원석(2014). "대학 구조개혁 평가"에 대한 프로그램 평가 이론적 검토. **2014년 한국교육평가학회 세미나 자료집: 대학구조개혁 평가방안의 타당성**, 40-43.

이재성, 홍성찬(2014). 기업의 빅데이터 적용방안 연구-A사, Y사 빅데이터 시스템 적용 사례. **인터넷정보학회논문지, 15**(1), 103-112.

이태희, 김종인(2015). 대학 구조개혁평가에 대한 메타평가 준거 개발 연구: 인적자원개발 관점의 적용. **2015 한국인사관리학회 및 한국인사조직학회 추계학술대회 발표논문집**, 33-60.

이호엽, 곽주은, 최두원, 김창욱(2014). 토픽모델링과 인과관계 분석을 활용한 법률 관련 기사와 법 개정 간의 관계 분석. **2014 대한산업공학회/한국경영과학회 춘계공동학술대회 논문집**, 1605-1620.

임경수, 이희수(2009). 국가인적자원개발에 관한 언론의 관점(2000년~2007년). **인력개발연구, 11**(2), 1-25.

정다미, 김재석, 김기만, 허종욱, 온병원, 강미정(2013). 사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안. **지능정보연구, 19**(3), 1-20.

정지선(2011). **신가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략. 새로운 미래를 여는 빅데이터 시대**. 한국정보화진흥원, 빅데이터 전략연구센터.

하연섭(2010). 정책아이디어, 틀 짓기, 사회적 담론 형성의 관계에 관한 연구: 교육정책을 중심으로. **행정논총, 48**(2), 189-215.

황하성, 손승혜, 장윤재(2012). 교육 보도에 있어서 정보원, 뉴스 선정, 취재 관행에 관한 연구. **사회과학연구, 19**(1), 247-277.

중앙일보(2015.09.02.). 교육부 "대학평가 지방大 반발? 소모적 논쟁".

한겨레(2015.08.31). [사설] 지방대 차별 논란만 부추긴 '대학 구조개혁 평가'.

한국대학신문(2015.11.09.). 대학들"정원 감축에만 초점 맞춘 것 아니냐"불만.

Apaza, R. G., Cervantes, E. V., Quispe, L. C., & Luna, J. O. (2014). Online courses recommendation based on LDA. In *1st Symposium on Information Management and Big Data* (pp. 42-48).

Battisti, F. D., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: a topic model approach. *Scientometrics, 103*, 413-433. DOI 10.1007/s11192-015-1554-1.

Blackmore, J., & Thorpe, S. (2003). Media/ting change: the print media's role in mediating education policy in a period of radical reform in victoria, Australia. *Journal of Education Policy, 18*(6), 577-595.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning and Research, 3*, 993-1022.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Born, J., Scheihing, E., Guerra, J., & Cárcamo, L. (2014). *Analysing Microblogs of Middle and*

High School Students Participating in Kelluwen. EC-TEL 2014, Graz, Austria. DOI:10.1007/978-3-319-11200-8_2. Retrieved from http://www.inf.uach.cl /wp-content/uploads/2014/10/bitacora_paper.pdf.

Daniel, B. (2015). Big data and analytics in higher education: opportunities and challenges. *British Journal of Educational Technology, 46*(5), 904-920.

Das, S., Sun, X., & Dutta, A. (2016). Text mining and topic modeling on compendium papers form transportation research board annual meetings. In *Transportation Research Board 95th Annual Meeting* (No. 16-3009).

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41*(6), 570-606.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(1), 1-24.

Gan, Q., Zhu, M., Li, M., Liang, T., Cao, Y., & Zhou, B. (2014). Document visualization: an overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*(1), 19-36.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*(2), 321-326.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. *Political Analysis, 18*(1), 1-35.

Grün, B., & Hornik, K. (2011). Topicmodels: an R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-13.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332-359.

HanChen, J., MaoShan, Q., & Peng, L. (2016). Finding academic concerns of the Three Gorges Project based on a topic modeling approach. *Ecological Indicators, 60,* 693-701.

Hannigan, T. (2015). Close encounters of the conceptual kind: disambiguating social structure from text. *Big Data & Society, July-December,* 1-6. DOI: 10.1177/2053951715608655.

Jockers, M. L. (2014). *Text analysis with R for students of literature*. Switzerland: Springer International Publishing.

Matthies, B., & Corners, A. (2015). Computer-aided text analysis of corporate

disclosures-demonstration and evaluation of two approaches. *The International Journal of Digital Accounting Research, 15*, 69-98.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburghh, C., & Byers, A. H. (2011). Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute Report, May*. Retrieved from http://www.mckinsey.com/insights /business_technology/big_data_the_next_frontier_for_innovation.

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th annual conference on uncertainty in artificial intelligence*, Helsinki, Finland.

Moreno, A., & Redondo, T. (2015). Text analytics: the convergence of big data and artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence, 3*(6), 57-64.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In *International conference on intelligence and security informatics* (pp. 93-104). Springer Berlin Heidelberg.

Reardon, S. (2014). Text-mining offers clues to success. *Nature, 509*, 410.

Romero, C. R., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics, Part C(Application and Reviews), 40*(6), 601-618.

Shirota, Y., Hashimoto, T., & Sakura, T. (2014). Extraction of the financial policy topics by latent dirichlet allocation. In *TENCON 2014-2014 IEEE Region 10 Conference* (pp. 1-5). IEEE.

Steyvers, M., & Griffiths, T. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228-5235.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis: A road to meaning* (pp. 427-448). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, S., Ding, Y., Zhao, W., Huang, Y., Perkins, R., Zou, W., & Chen, J. J. (2016). Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC Public Health, 16,* 279. DOI 10.1186/s12889-016-2932-1.

White, P., & Breckenridge, R. S. (2014). Trade-offs, limitations, and promises of big data in social science research. ***Re**view of Policy Research, 31*(4), 331-338.

Witten I. A. (2005). Text mining. In M. P. Singh (ed.), *The practical handbook of internet*

computing (pp.14-1-14-23). Boca Raton, FL: Chapman & Hall/CRC Press.

Yang, T. I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96-104). Association for Computational Linguistics.

Zakir, J., Seymour, T., & Berg K. (2015). Big data analytics. *Issues in Information Systems, 16(*2), 81-90.

* 논문접수 2016년 8월 2일 / 1차 심사 2016년 9월 9일 / 게재승인 2016년 9월 21일

* 김지은: 연세대학교 교육학과를 졸업하고 동대학원 교육학과에서 석사학위를 취득하였으며, 미국 Texas A&M 대학교에서 '교육 연구, 측정 및 통계'를 전공으로 박사과정을 수료하였다. 현재 서울대학교 사범대학 교육학과에서 '교육측정 및 평가' 전공으로 박사과정을 수료하였다.
* E-mail: ookjk9312@snu.ac.kr

* 백순근: 서울대학교 사범대학 교육학과를 졸업하고 동대학원 교육학과에서 석사학위를 취득하였으며 미국 버클리대학교(UC Berkeley에서 '교육측정 및 평가' 분야 박사(Ph.D)를 취득하였다. 현재 서울대학교 사범대학 교육학과 교수로 재직 중이며, 주요 저서로는 '수행평가의 원리', '학위논문 작성을 위한 교육연구 및 통계분석', '白교수의 百가지 교육이야기' 등이 있다.
* E-mail: drl00@snu.ac.kr

Abstract

Analysis of Issues on the College and University Structural Reform Evaluation Using Text Big Data Analytics

Kim, Ji-Eun*
Baek, Sun-Geun**

The purpose of this study is to analyze the issues on the college and university structural reform evaluation revealed by the Government press releases and newspaper articles, using text big data analytics. The topic modeling, especially latent Dirichlet allocation algorithm, was applied to extract and analyze the issues amongst 25 press releases of the Ministry of Education(MOE) and 625 articles in 10 major daily newspapers from January 1st 2013 to April 30th 2016.

According to the analysis result, three issues were found from the documents of MOE, and seven issues from newspaper articles. In addition to this, the MOE press releases and newspaper articles represented 3 similar issues, i.e. 'background', 'management plan', and 'application of result', even though their main focuses were found to be different. There were also 4 additional issues, i.e. 'evaluation methods', 'the quality of higher education', 'the accountability of university', and 'public relations of university', only from newspaper articles showing the varying interests.

This study discloses that analyzing the text data using text big data analytics could be an effective method to find out the social issues in relation to the educational policy, such as the college and university structural reform evaluation, and to guide the advanced direction towards the future policy.

Key words: College and University Structural Reform Evaluation, Text Big Data Analytics, Topic Modeling, Latent Dirichlet Allocation(LDA), Government Press Release, Newspaper Articles

* Corresponding author, Ph. D Candidate, Seoul National University
** Professor, Seoul National University