The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.3.423
32(3), 423–434

Issue analysis of the admission officer system using topic analysis

Younghee Hong[a,1]

[a]Education Policy Institute, Busan Metropolitan City Office of Education

(Received February 11, 2019; Revised March 12, 2019; Accepted April 9, 2019)

Abstract

An important issues in Korea society in 2018 was the revision of the university entrance examination system. Among the discussions, in order to grasp what the issue of admission officer system is, attention was focused on the function of media such as monitoring and criticism as well as the tried topic analysis of related news articles. As a result, the reorganization of the College Scholastic Ability Test (CSAT) was derived and showed the sensitivity of Korean society towards the CSAT. Topics directly related to the admission officer system were the selection factor and fairness of the university entrance examination system in relation to the selection factor.

Keywords: topic analysis, news article, admission officer system

1. 서론

2018년 12월 4일 한국교육과정평가원장은 2019학년도 대학수학능력시험 채점 결과를 브리핑하며 불수능이라는 오명에 대해 고개 숙여 사과하는 모습을 보일 만큼, 한국 사회에서 대입과 관련한 사항은 모든 학부모의 관심사이면서 동시에 민감한 사안 중의 하나이다. 한국은 대입제도가 처음으로 실시된 1945년부터 지금에 이르기까지 무수히 많은 변화 과정을 거쳐 왔다. 대입제도와 관련하여 끊임없이 공통적으로 등장한 화두는 제도의 공정성과 신뢰성을 어떻게 지켜낼 것인가? 고교 교육과정 운영을 어떻게 정상화 시킬 것인가? 그리고 사교육비를 어떻게 하면 줄일 수 있을 것인가? 하는 등의 문제로 귀결된다.

근자에는 2015학년도부터 도입된 학생부종합전형이 고교 교육과정 운영의 정상화와 학생의 다양한 면모를 평가할 수 있다는 측면에서 이를 지지하는 집단과 수능 점수로 줄 세우는 평가방법이 가장 공정하고 신뢰할 수 있다는 점을 강조하는 집단 간의 첨예한 대립이 발생했다. 이는 교육부가 2022학년도 대입제도 개편(안) 마련을 위해 시행한 국민 참여형 공론조사를 통해 여실히 드러났다. 본 연구에서는 이러한 논란의 중심에 서 있는 학생부종합전형에 대한 쟁점이 무엇인가를 파악해 보고자 한다.

학생부종합전형과 관련한 쟁점을 다룬 연구로 Lee (2018)은 학생부종합전형이 대학교육 적격자 선발제도로서 가지는 한계와 비교과 활동을 대비하기 위한 새로운 사교육 수요를 유발할 가능성이 있다고 지적하였다. 가정의 사회경제적 배경 영향에 따른 전형 대비의 유·불리와 이로 인한 교육 불평등의 가능성

[1]Education Policy Institute, Busan Metropolitan City Office of Education, 12 Hwaji-ro Busanjin-gu, Busan 47119, Korea. E-mail: yhhong0120@gmail.com

등에 관한 논란과 문제점을 지적하였다. Joo와 Kim (2017)은 학생부종합전형이 추구하는 가치가 학생의 균등한 교육기회 보장, 대입제도의 다양성, 대학의 자율성 지향 등이라고 할 수 있으나, 실제로 대학의 학생 선발에 대한 자율성이 보장되는가에 대한 문제를 지적하였다.

이러한 연구들은 관련 정책에 대한 문헌과 선행연구 분석을 통해 논리적이고 체계적인 방법으로 접근하였으나, 해당 정책에 대한 여론의 즉시성(immediacy)을 반영하지는 못했다는 점에서 제한점을 가진다. 사회적으로 특정 사안이 발생하거나 정부의 정책이 수립되는 과정에서 개인 또는 단체가 그들의 생각이나 의견, 입장 등을 표출하는 행위들이, 과거에 비해 보다 다양한 방법으로 활발하게 이루어지고 있다. 온라인 매체인 블로그, SNS 등을 통해 손쉽게 자신의 의견을 생산, 공유, 확대하는 것이 보편화 되었고, 언론 또한 이러한 시대적 흐름을 선도하고 있다. 특히 언론 매체의 뉴스 기사는 Yu와 Baek (2016)에서 지적하였듯이 정책의 주요 내용을 국민에게 전달하고 정책에 대한 국민적 여론 형성을 통해 관련 쟁점과 논의들을 심층보도하거나 사설, 칼럼 등을 통해 의견을 개진함으로써 정책 형성과 결정에 영향을 미치는 중요한 역할을 한다.

따라서 본 연구에서는 언론 매체의 감시와 비판이라는 기능에 주목하여 학생부종합전형에 관한 뉴스 기사를 분석함으로써, 여론을 통해 나타난 학생부종합전형과 관련한 주된 쟁점들을 파악해 보고자 한다. 기존의 쟁점 분석이 문헌과 선행연구를 분석하는 것으로 이루어져 왔다면, 본 연구에서는 언론 매체의 뉴스 기사를 분석함으로써 기존 문헌과 선행연구에서 다루지 못한 여론을 반영하였다는 것에 의의를 둘 수 있다. 본 논문의 구성은 2장에서는 텍스트 데이터와 관련한 토픽 분석에 대해 다루고, 3장에서는 실제 온라인 뉴스 기사를 수집하고 전처리하여 분석한 결과를 제시하였다. 4장에서는 주요 결론을 정리하고 한계점을 진단하고자 하였다.

2. 텍스트 데이터의 토픽 분석

비정형데이터의 대표적 형태 중 하나인 텍스트 데이터는 신문 기사, SNS, 공고문, 블로그 등으로부터 얻을 수 있고 기본적으로 자연어처리기술(natural language processing; NLP)을 기반으로 한 텍스트 마이닝을 통해 분석하게 된다.

Khan과 Kanth (2016)에 의하면 텍스트 마이닝은 일차적으로 텍스트 요약, 문서 검색과 같은 정보 추출을 수행할 수 있고 텍스트 분류, 문서 군집화, 언어 인식, 핵심문구 식별 등 문서의 유사성을 측정하는 방식으로 확대되기도 한다. 또, Jeong 등 (2013)에서 보듯이 정확하게 파악하기 어려운 사회 현상 및 이슈와 관련된 다양한 논의를 파악하는데 유용하게 사용된다.

텍스트 마이닝을 위한 구체적인 분석 방법으로는 주제어 분석, 단어 연관성 분석, 단어 군집화 분석, 텍스트 네트워크 분석, 토픽 분석 등 다양한 방법이 활용되고 있다.

2.1. 뉴스 기사를 이용한 분석

스마트 기기의 대중화 시대와 함께 데이터 발생량은 폭증하고 있다. 대중은 과거 뉴스를 소비하는 것이 주된 역할이었다면, 지금은 소비와 함께 재생산을 통해 데이터의 발생 속도를 가속화 시키고 있다. 이러한 시대적 상황 속에서 뉴스 기사도 정보 전달이라는 1차적 기능을 넘어서 대중과의 소통을 통한 여론 형성은 물론이고, 특정 사안에 대한 집단적 혹은 개인적 태도나 행동을 취하도록 하는 역할까지도 하고 있다. SNS를 통한 뉴스 기사의 작성과 공유가 과거에 비해 대중화 되고 확산됨에 따라 일명 가짜 뉴스의 양산이라는 폐해는 지양해야 할 바가 분명하나, 감시와 비판이라는 언론의 순기능은 분명 존재한다.

이러한 언론의 순기능에 주목하여 정책을 수립하고 집행하는 과정에서는 여론을 파악하여 이를 반영하

는 과정이 반드시 필요하고, 실제로 이러한 연구들은 활발하게 진행되고 있다. Park (2017)은 누리과정과 관련한 온라인 뉴스를 분석한 결과 유아교육과 보육 서비스의 전반적인 질 관리와 공공성 강화, 누리과정 예산의 안정성 등이 국민의 요구임을 파악하였다. Jang과 Park (2017)은 언론보도 분석을 통해 해양수산 정책 분야의 사회이슈를 식별하고 향후 비중 있게 다뤄져야 할 정책 이슈를 도출하였다. Kim과 Baek (2016)은 교육부 보도 자료와 신문 기사 분석을 통해 대학구조개혁 평가와 관련된 주요 쟁점들을 구체화하고, 정책 이해 주체 간의 유사점과 차이점을 분석하였다. 새로운 연구방법은 많은 시간이 소요되는 전통적 조사방법이 가지는 한계인 여론을 즉시에 반영할 수 없다는 단점을 보완할 수 있다는 점에서 주목할 여지가 있다.

2.2. 토픽 분석

토픽 분석은 문서에 숨겨져 있는 주제들을 찾아내기 위해 문서에 출현하는 단어의 빈도를 기반으로 문서를 분류하기 위해 개발된 분석방법이다. 하나의 문서에는 여러 개의 주제가 함께 등장할 수 있다는 것을 가정하되, 개별 문서에서 비중이 가장 큰 주제를 대표 주제로 할당하여 문서들을 분류한다. 이후 비슷한 주제로 분류된 문서의 공통 논점을 도출하고 이를 통해 특정 주제와 관련된 이슈를 파악할 수 있게 된다.

토픽 분석 모형의 시초는 Deerwester 등 (1990)에서 제안한 잠재 의미 분석(latent semantic analysis; LSA) 모형으로 대량의 텍스트 문서에서 발생하는 단어들 간의 연관관계를 분석하여 잠재적인 의미 구조를 도출하고자 하는 시도이다. 단어-문서 행렬(term-document matrix; TDM)을 행렬의 특이값 분해(singular value decomposition; SVD)를 활용하여 3개의 독립적인 행렬로 분해하게 되고, 분해된 행렬의 계수에 따라 단어-문서 행렬의 차원(dimension)을 축소하게 된다. 다시 말해 텍스트 문서 집합을 내용의 유사도에 따라 여러 개의 소집단으로 분할하게 되고, 문서 집합의 연관성은 동시 출현 빈도가 높은 단어들을 기준으로 유사한 문서를 추출하게 된다.

Hofmann (1999)에 의해 LSA 모형에 확률 개념을 도입한 probabilistic LSA (PLSA)로 발전하였다. 이후 확률 모형에 대한 모수 추정의 베이지안적 방법을 적용한 연구가 Blei 등 (2003)에 의해 개발되었는데, 이들이 제안한 잠재 디리클레 할당(latent Dirichlet allocation; LDA) 모형이 현재 토픽 분석의 대표적 모형으로 활용되고 있다.

LDA 모형을 설명하기 위해 문서의 집합은 n개의 개별 문서로 구성되어 있고 j번째 문서에는 n_j개의 단어가 포함되어 있다고 가정하자 (단, $j = 1, \ldots, n$). 문서를 표현할 수 있는 전체 단어들의 집합인 단어사전은 $\{1, \ldots, W\}$로 표시하고 j번째 문서에서 i번째로 등장하는 단어는 $x_{ji} \in \{1, \ldots, W\}$로 나타낸다 (단, $i = 1, \ldots, n_j$).

j번째 문서는 해당 문서에 등장하는 단어들의 집합인 $\boldsymbol{x}_j = (x_{ji})_{i=1}^{n_j}$로 표시되고, 분석 대상인 문서의 집합은 $\boldsymbol{x} = (x_j)_{j=1}^{n}$로 나타낼 수 있다. 이때 문서에 포함된 단어의 순서는 모형에 영향을 미치지 않고, 출현한 단어의 빈도만이 모형에 영향을 끼친다.

k번째 토픽($k = 1, \ldots, K$)에 대한 단어들의 확률분포는 $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kW})$로 나타내고, 토픽들의 집합을 $\{1, \ldots, K\}$로 표시하면 이에 대한 확률분포 $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK})$는 j번째 문서가 각 토픽을 어떤 확률 값으로 갖는지를 나타낸다. 확률적 토픽모형에서 관측된 자료의 우도 함수는 다음과 같다.

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{j=1}^{n} \prod_{i=1}^{n_j} p(x_{ji}|\boldsymbol{\theta}_j, \boldsymbol{\phi}) = \prod_{j=1}^{n} \prod_{i=1}^{n_j} \left(\sum_{k=1}^{K} \theta_{jk} \phi_{k x_{ji}} \right)$$

이 때, $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)_{j=1}^{n}$, $\boldsymbol{\phi} = (\phi_k)_{k=1}^{K}$이다.

LDA 모형에서 $\boldsymbol{\theta}_j$와 $\boldsymbol{\phi}_k$에 대한 사전분포는 각각 $D(\alpha, \ldots, \alpha)$와 $D(\beta, \ldots, \beta)$인 k차원의 디리클레 분포

를 따른다. LDA 모형을 계층적으로 나타내기 위해 x_{ji}가 어떤 토픽에서 추출되었는지를 나타내는 잠재변수를 z_{ji}로 가정하면, 다음과 같이 표현할 수 있다.

$$p\left(z_{ji}=k|\boldsymbol{\theta}_j\right)=\theta_j k,$$
$$p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\phi})=\prod_{j=1}^{n}\prod_{i=1}^{n_j}p\left(x_{ji}|z_{ji},\boldsymbol{\phi}_{z_{ji}}\right)$$

이 때, $p(x_{ji}=\omega|z_{ji}=k,\phi_k)=\phi_{k\omega}$, $\boldsymbol{z}=(\boldsymbol{z}_j)_{j=1}^{n}$, $\boldsymbol{z}_j=(z_{ji})_{i=1}^{n_j}$이다.
따라서, $(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{z})$의 사후분포는 다음과 같이 구할 수 있다.

$$\begin{aligned}
&P(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{x})\\
&\propto P(\boldsymbol{x}|\boldsymbol{\phi},\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{\theta})P(\boldsymbol{\phi})P(\boldsymbol{\theta})\\
&\propto\left[\prod_{j=1}^{n}\prod_{i=1}^{n_j}\prod_{\omega=1}^{W}\phi_{k\omega}^{I(x_{ji}=\omega)}\right]\left[\prod_{j=1}^{n}\prod_{i=1}^{n_j}\prod_{k=1}^{K}\theta_{jk}^{I(z_{ji}=k)}\right]\left[\prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{\omega=1}^{W}\phi_{k\omega}^{\beta-1}\right]\left[\prod_{j=1}^{n}\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{jk}^{\alpha-1}\right],
\end{aligned}$$

여기서 $I(\cdot)$는 지시함수이며, 사후분포는 계산의 복잡성을 해결하기 위해 붕괴된 깁스 샘플러(collapsed Gibbs sampler)를 통해 근사시키게 된다.

이제 수집된 자료에 대하여 토픽 분석을 적용하기 위해서는 연구자가 사전에 토픽의 개수를 사전에 결정해 주어야 하는 문제가 남아 있다. 토픽의 수는 분석 결과의 해석가능성에 영향을 주기 때문에 수집된 자료와 해당 분야에 대한 이해가 부족한 상태에서는 결정이 어려울 수 있다. 토픽의 개수를 결정하는 방법에 대한 선행연구들을 살펴보면 다음과 같이 정리할 수 있다.

첫 번째 방법은 연구자가 해석가능성과 해당 분야에의 적합성 등을 고려하여 적당한 수를 지정하는 방법이다 (Kang 등, 2013; Bae 등, 2013; Grant 등, 2013). 두 번째 방법으로는 Blei 등 (2003)이 제안한 복잡도(perplexity) 지수를 활용하는 방법이 있는데, 이는 단어의 우도함수에 대한 기하평균으로 계산이 가능하다. 마지막으로 통계프로그램 R의 'ldatuning' 라이브러리에서 제공하는 3가지 기준 값을 활용하는 것으로, Griffiths와 Steyvers (2004), Cao 등 (2009), Arun 등 (2010)이 제안한 통계량을 토대로 결정 할 수 있다. Griffiths와 Steyvers (2004)는 붕괴된 깁스 샘플링 알고리즘으로 사후분포를 추정하여 로그-우도함수의 조하평균이 최댓값을 가지는 경우가 토픽의 수로 결정되고, Cao 등 (2009)은 토픽 분포들 간의 cosine similarity의 평균값을 추정하여 그것이 최소가 되는 경우에 토픽의 수로 결정할 수 있다고 하였다. Arun 등 (2010)은 각 토픽에 대하여 단어가 나타날 확률을 나타낸 행렬의 singular values와 코퍼스(corpus) 내에서 각 토픽 분포 사이의 symmetric Kullback-Liebler divergence가 최소화 될 때 토픽의 수로 결정할 수 있다고 하였다. 세 번째 방법은 'ldatuning' 라이브러리를 통해 한번에 얻을 수 있기는 하지만 최댓값 혹은 최솟값을 판단하기에 그 기준이 다소 모호한 면이 있다.

따라서 본 연구에서는 보다 엄격하게 토픽의 수를 결정하기 위해 일차적으로 복잡도 지수를 고려하였다. 복잡도 지수는 분석 대상 자료를 임의로 훈련용 자료(training data)와 테스트 자료(test data)로 나눈 후, 훈련용 자료로 만든 토픽 모델을 테스트 자료에 적용하여 구할 수 있으며 다음과 같이 정의된다.

$$\text{Perplexity}(D_{\text{test}})=\exp\left\{-\frac{\boldsymbol{\Sigma}_{d=1}^{M}\log p(w_d)}{\boldsymbol{\Sigma}_{d=1}^{M}n_d}\right\}.$$

단, D_{test}는 테스트 자료에 포함된 M개의 개별 문서의 집합이며, w_d는 테스트 자료에 포함된 문서에 등장하는 단어를 의미한다.

Blei 등 (2003)에 의하면 복잡도가 낮은 모델이 분석 대상 문서를 보다 잘 분류한 것으로 판단할 수 있다. 그러나 Zhao 등 (2015)에서 지적하였듯이 동일한 분석 자료에 대해서도 초기 값을 어떻게 주느냐에 따라 복잡도 지수의 값이 안정적으로 나타나지 않는 문제가 발생한다. 따라서 Zhao 등 (2015)에서 제안한 복잡도 변화율(rate of perplexity change; RPC)을 활용하여 자료의 특성과 초기 값에 의한 민감함을 보완하여 토픽의 개수를 결정하고자 한다.

복잡도를 계산하는 방법으로 k-fold cross validation을 적용하는데, 분석 대상 자료를 임의로 k개의 부분 데이터 셋, $S_1, \ldots, S_k$로 나누고 토픽의 개수에 대한 후보군을 $t_1 \leq \cdots \leq t_r$로 표시하자. 이 때 통상적으로 고려되는 k 값은 경험적으로 5 또는 10이며, 이를 결정하기 위한 정형화된 방법은 없어 (Kuhn과 Johnson, 2013; James 등, 2013) 본 연구에서는 k 값을 10으로 하여 적용하였다. 복잡도를 계산하기 위해 $(k-1)$개의 훈련용 자료로 LDA 모형을 적합 시켜 시험용 자료에 적용한다. 이러한 과정은 k번 반복하게 되는데, 각각의 부분 데이터 셋 S_i는 $(k-1)$번은 훈련용 자료로, 1번은 시험용 자료로 활용된다. 임의의 토픽 수 t_i에 대해 k개의 복잡도 평균값을 최종적으로 사용하게 되고, 토픽의 개수를 변화시켜 가며 복잡도 값의 변화를 관측하게 된다. r개의 토픽 개수 후보군에 대한 복잡도의 평균을 $P_1, \ldots, P_r$로 표시하면, 임의의 토픽 개수 t_i $(1 \leq i \leq r)$에 대한 RPC은 다음과 같이 정의된다.

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

토픽의 수를 점차적으로 증가시키며 RPC를 계산하게 되는데, RPC가 감소하다가 처음으로 증가하는 지점을 기준으로 토픽의 수를 결정할 수 있다 (Zhao 등, 2015).

토픽 분석을 활용한 국내 연구는 2012년 이후부터 SNS 데이터, 특히 트위터 자료에 대한 분석이 교육, 사회, 정치 등 다양한 분야에서 활발하게 진행되고 있다. Kim과 Yoon (2016)은 토픽 모델 기반의 SNS 이슈 분석 기술에 세계 지도 시각화를 결합하여 각 국가별 특정 주제와 관련한 관심 이슈와 그 분포의 변화 추이를 분석하였다. Choi와 An (2015)은 온라인 토론장에 게시된 글을 수집하여 토픽 모델링을 적용함으로써 원자력 발전에 대한 한국인의 인식을 조사하였다. 또, Cho 등 (2015)은 교통카드 데이터에 대하여 LDA 기법을 적용하여 청주시 버스 승객들의 이동패턴을 분석하여 보였다.

3. 데이터 수집 및 분석

3.1. 데이터 수집과 전처리 과정

학생부종합전형에 대한 언론보도 자료의 데이터 수집은 한국언론진흥재단에서 운영하는 빅카인즈를 활용하였고, 수집 기간은 2014년 1월 1일부터 2018년 8월 31일로 설정하였다. 데이터 수집 기간에 대한 근거는 학생부종합전형이 시행된 시기가 2015학년도 대입전형부터로 2014학년도 수시모집 전형기간이 끝난 2014년 1월 1일을 시작기준일로 삼았다. 데이터 수집을 위한 검색어는 대학입학전형이며 빅카인즈를 통해 검색된 기사의 세부적인 내용 검토를 통해 학생부종합전형에 관한 신문기사만을 분석대상으로 하였다. 제외된 기사의 내용은 주로 특정 대학의 입학전형을 소개하거나 사교육 기관의 홍보내용 등이며, 동일한 내용의 중복된 기사도 1개만 남기고 나머지는 삭제하였다.

기사 수집 시 선택한 언론사는 한국ABC협회의 2017년 일간신문 163개사에 대한 발행부수와 유료부수 인증결과 상위 20위권에 속하는 신문사 중에서 빅카인즈를 통해 수집이 가능한 10개의 채널(매일경제, 한국경제, 한겨레, 경향신문, 문화일보, 한국일보, 국민일보, 서울신문, 부산일보, 국제신문 등)을 대상으로 하였다. 최초 수집된 기사는 총 2,108건이며 이 중 연구 주제와 직접적인 관련이 없는 기사들, 예를 들면 각 대학별 대입전형 모집요강, 사교육 기관의 대입 관련 광고, 단순히 특정 행사를 홍보하기 위

Table 3.1. Number of final collected news articles

채널	검색 건수	채널	검색 건수
경향신문	107	한국일보	122
국민일보	98	매일경제	136
문화일보	48	한국경제	224
서울신문	169	국제신문	85
한겨레	137	부산일보	114
최종 분석 대상			1,240

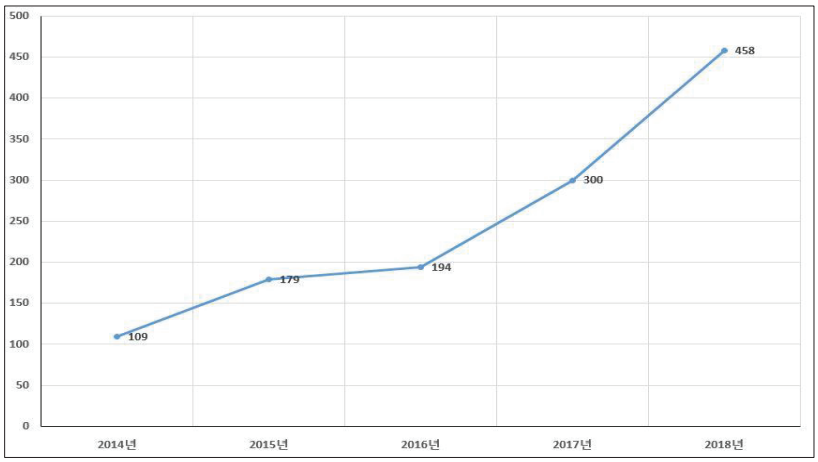

Figure 3.1. change in the number of news articles.

한 기사 등은 분석 대상에서 제외하여 총 1,240건의 기사를 최종 분석대상으로 하였고 각 채널별로 수집된 기사의 수는 Table 3.1과 같다.

분석 대상으로 삼은 기사를 통해 학생부종합전형에 대한 언론의 관심도를 알아볼 수 있다. 2014년도에는 월별 10건 내외의 기사가 보도 되었으나, 2015년도와 2016년도에는 월별 20건 내외의 기사가 보도되었다. 언론의 관심이 폭발적으로 증가하는 시기는 2017년 8월경으로 2021학년도 수능 체제 개편이 유예됨으로 인해 주목을 받았고, 2018년 4월에도 국가교육회의를 통한 2022학년도 대입 제도 개편에 관한 논의가 시작되면서 관심이 급증하였다.

수집된 텍스트 데이터는 자연어처리과정을 거쳐 분석이 가능한 형태로 변환하는 작업이 필요한데 전처리 과정(preprocessing)과 형태소 분석이라는 두 단계가 필요하다. 이를 위해서는 먼저 수집된 문서를 일정한 전자형식의 구조를 가지는 텍스트 집합인 코퍼스로 변환하여 전처리 과정을 거치게 된다.

전처리 과정은 먼저 분석에 불필요한 단어나 어구를 삭제하는 작업으로 시작되는데, 예를 들면 신문기사와 관련 없는 광고성 문구, 특정 홈페이지나 SNS 주소, 관련 뉴스 기사 링크, 이메일 주소, 문장부호, 각종 특수기호, 불필요한 숫자 등을 제거한다.

또, 분석 시 의미를 부여할 필요가 없거나 분석 목적과 관계없는 단어인 불용어(stopword)를 제거하였다. 예를 들면 검색어로 활용된 학생부종합전형, 대입이라는 단어의 경우 특별히 차별적 의미를 제공하지 못하기 때문에 삭제하였고, 특정 대학교의 명칭을 나타내는 단어(예: 서울대 등), 시점을 나타내는 단어(예: 지난해, 올해 등), 특정 인물의 이름(예: 문재인, 김상곤 등), 직위명(예: 대통령, 장관 등), 사

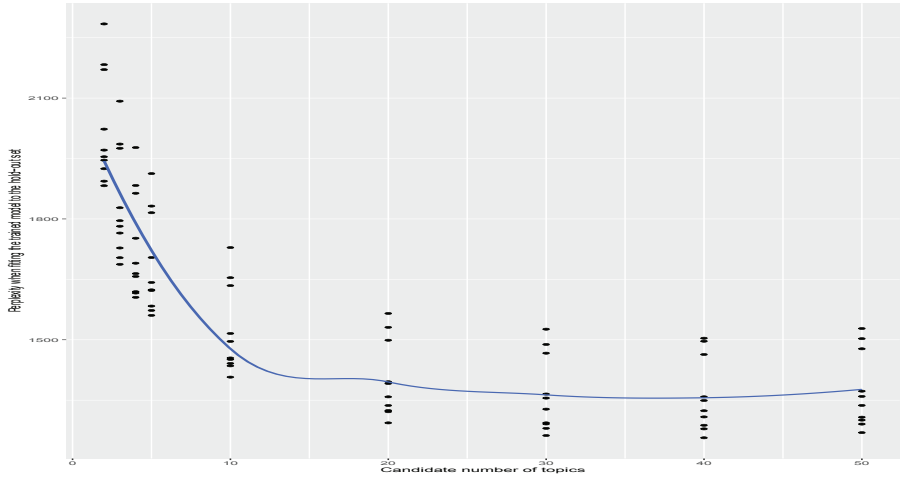

Figure 3.2. 10-fold cross-validation of topic modelling.

교육업체명(예: 진학사 등), 단위나 순서를 나타내는 단어(예: 명, 만 등) 등을 삭제하였다.

전처리 과정의 두 번째 단계로는 동일하거나 유사한 의미를 가진 단어이지만 다르게 표현된 단어들을 통일하는 정규화 작업이 필요한데, 예를 들면 학교생활기록부, 생기부 등은 '학생부'로, 대학수학능력시험, 대수능 등은 '수능'으로 정규화 하였다. 이러한 전처리 작업을 위해 통계프로그램 R의 'tm' 라이브러리를 활용하였다.

이상의 불용어 제거와 단어 정규화 작업은 한 번의 실행으로 모든 것이 정리되는 것이 아니라 계속적인 반복 작업을 통해 새로이 등장하는 불용어를 제거하고 정규화 과정을 거치게 된다.

다음 단계로 한글 형태소 분석을 통해 토픽 분석에 필요한 명사 추출(extract noun) 작업이 일반적으로 필요한데, 통계프로그램 R의 'KoNLP' 라이브러리의 extractNoun() 함수를 활용하면 짧은 시간 안에 처리가 가능하다. 빅카인즈를 통하면 명사만을 추출할 수 있으므로 본 연구에서 형태소 분석은 생략하였다.

최종적으로 길이가 2 이상인 단어들만 이용하여 22,363 × 1,240의 단어-문서행렬을 구성하여 분석에 사용하였다.

3.2. 토픽 분석 결과

토픽 분석을 위해 통계프로그램 R에서 'topicmodels' 라이브러리의 LDA 함수를 활용하였고, RPC을 이용하여 적합한 토픽의 수를 탐색하였다. 먼저 10-fold cross validation을 적용하여 토픽의 개수를 $2, 3, 4, 5, 10, 20, 30, 40, 50$으로 변화시켜 가며 복잡도를 계산한 결과는 Figure 3.2와 같고, 대략적인 토픽의 수는 20개 이내로 결정될 것이 예상된다. 정확한 토픽의 수를 결정하기 위하여 토픽의 수를 2부터 30까지 1씩 증가시키며 RPC을 계산한 결과는 Figure 3.3과 같다. 토픽의 수를 증가시키면 RPC의 값은 전반적으로 작아지는데, RPC의 값이 처음으로 반등하게 되는 경우는 토픽의 수를 5로 하였을 때로 확인된다.

토픽의 수를 5개로 하여 분석 대상 문서를 분류해 보면, 각 토픽별로 나타날 확률이 높은 단어와 해당 토픽이 대표적으로 나타난 기사 내용을 바탕으로 Table 3.2와 같이 토픽명을 부여할 수 있다. LDA 모형

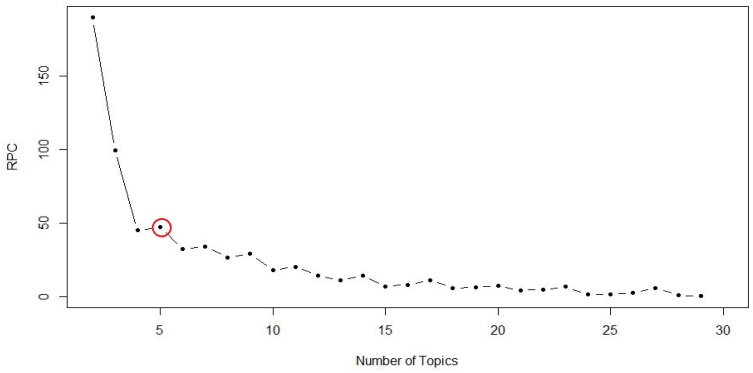

Figure 3.3. The Rate of perplexity change.

Table 3.2. Some keyword of 5 topics

	토픽명	주요 키워드	관련 기사 수(%)
1	학생부종합전형 선발요소	학생, 활동, 학생부, 자기소개서, 평가, 면접, 동아리	237(19.1)
2	수능체제 개편 논의	수능, 교육부, 절대평가, 확대, 정시모집, 폐지, 개편, 정책	305(24.6)
3	대입전형 지원 전략	수시모집, 논술전형, 정시모집, 수험생, 반영, 비중, 인원	285(23.0)
4	대입전형의 공정성	학생, 교사, 학부모, 일반고, 자율형사립고, 사교육, 학원	217(17.5)
5	대입전형 개편 논의	정시모집, 교육부, 국가교육회의, 공론화, 비율, 확대, 의견	196(15.8)

에서 각 문서는 여러 개의 토픽을 포함할 수도 있다는 것을 가정하므로 개별 문서마다 각 토픽이 나타날 확률(문서-토픽 확률)을 계산할 수 있고, 개별 문서에 나타난 각 토픽의 출현 확률을 모두 합하면 1이 된다. 따라서 개별 문서에 대한 토픽의 확률이 가장 큰 것이 해당 문서에 나타날 가능성이 가장 높은 것이므로 이를 해당 문서의 대표 토픽으로 판단할 수 있다. 다시 말해 각 토픽별 주요 키워드와 개별 문서의 대표 토픽을 토대로 토픽명을 결정하게 된다.

각 토픽별 주요 키워드를 워드 클라우드로 표현하면 Figure 3.4와 같다. 토픽분석 결과 가장 큰 비중을 차지하는 토픽1의 경우, 주요 키워드 중 학생부, 자기소개서, 면접 등은 대부분의 대학에서 학생부종합전형을 통해 학생을 선발하고자 할 때의 방법인 학교생활기록부와 자기소개서를 통한 서류평가와 면접에 대한 논의들을 다루고 있음을 알 수 있다. 학교생활기록부에 나타난 학생의 활동 중에서도 특히 동아리 활동이 중요하게 평가되고 있어 이에 대한 자세한 기록이 중요하고, 자기소개서를 통해 학생의 역량이 드러날 수 있도록 하는 방법 등에 대한 논의들이 주요 기사에서 등장함을 확인할 수 있다. 따라서 토픽 1을 '학생부종합전형의 선발요소'라고 명명하였다.

토픽2는 2015 개정 교육과정이 학교 현장에 도입됨으로 인해 수능 체제에 대한 개편의 필요성을 제기하는 기사들이 도출되었다. 2021학년도 수능 체제 개편에 관한 논의가 진행되던 2017년, 교육부는 수능 과목 조정을 포함하여 '일부 과목 절대평가(안)'과 '전 과목 절대 평가(안)' 중 택일 할 예정이었으나 이에 대한 선택을 1년 유예하면서 학생, 학부모에게 큰 혼란을 안기며 언론의 질타를 받았다. 또, 2018학년도부터 적용된 수능 영어영역 절대평가 확대에 대한 논의가 가세하여 주요 토픽으로 등장하게 되었다. 따라서 이러한 논의들을 포괄하여 토픽2를 '수능체제 개편 논의'로 명명하였다.

토픽3에서는 수시모집과 정시모집 관련한 전형에 대한 구체적 언급들이 다루어지고 있다. 수험생이 수시모집과 정시모집을 지원함에 있어 논술전형, 학생부종합전형 등 각 전형별로 요구하는 세부적인 사항, 예를 들면 세부 전형별 반영 요소와 선발인원, 내신 성적 반영 비중 등에 대해 자세히 다루고 있다. 수험

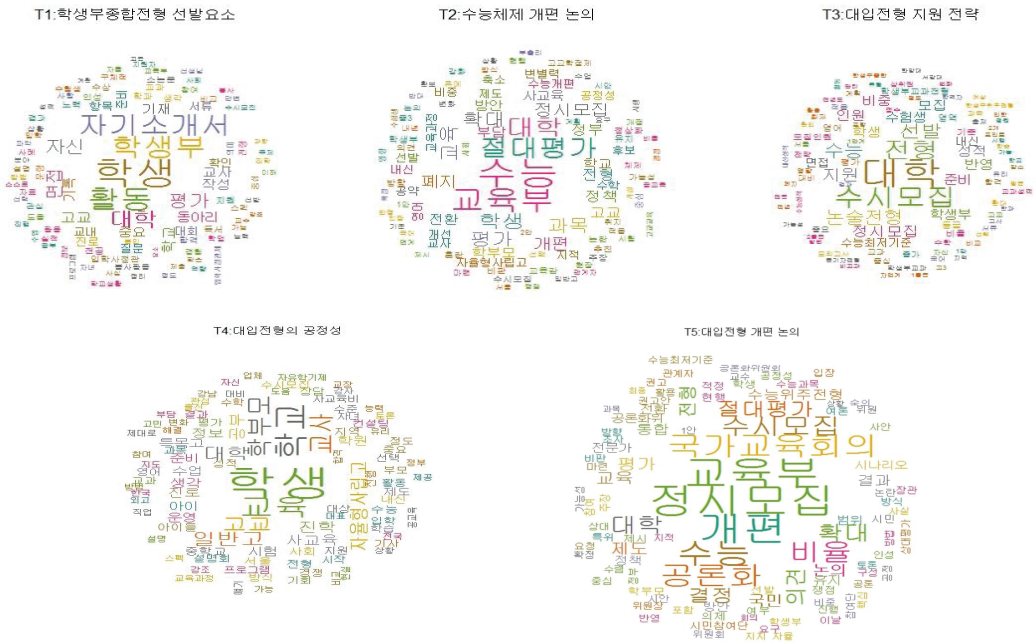

Figure 3.4. 5 topics obtained by latent Dirichlet allocation modeling.

생은 이러한 사항들을 잘 파악하고 본인의 강점과 약점을 분석하여 지원할 것을 조언하는 내용들이 다루어지고 있다. 따라서 토픽 3은 '대입전형 지원 전략'으로 명명 가능하다.

토픽4에서는 교육의 주체라고 할 수 있는 학생, 교사, 학부모가 주요 키워드로 전면에 등장하고, 고교 유형에 해당하는 일반고, 자율형사립고가 연이어 등장하면서 고교 체제에 따른 대입전형의 유불리 논란과 함께 특목고 폐지에 대한 찬반여론 또한 팽팽하게 가시화되었음이 확인된다. 특히, 학생부종합전형에서 중요한 학교생활기록부와 자기소개서 작성을 위한 학원의 컨설팅, 면접 대비를 위한 모의면접 수강 등과 관련한 사교육의 폐해들이 드러나면서 대입전형의 공정성이 도마에 오르게 되었다. 이에 토픽4를 '대입전형의 공정성'으로 명명하였다.

토픽5는 2018년 한 해 동안 우리 사회를 가장 뜨겁게 달군 이슈 중 하나인 2022학년도 대입제도 개편(안)에 관한 논의를 다루고 있다. 2018년 기준 중학교 3학년 학생이 치르게 될 대입제도의 개편과 관련하여, 교육부에서는 그간 유지해 온 수시모집 확대 기조와 함께 등장한 학생부종합전형 선발 비율의 증가에 대한 반대급부로 수능위주전형의 선발 비율 확대 요구가 거세게 일었다. 특히, 교육부는 대입제도 개편 작업을 국가교육회의에 위임하고 이는 다시 대입제도개편특별위원회에 위임되면서 대입개편 논의 위한 국민참여형 공론화위원회까지 출범시키며 뜨거운 논의가 이어졌다. 따라서 토픽5를 '대입전형 개편 논의'로 명명하였다.

통계프로그램 R의 'LDAvis' 라이브러리를 활용하여 Table 3.2에서 도출된 토픽들의 관계를 시각화해 보면 Figure 3.5와 같다. 이는 Sievert와 Shirley (2014)에 의해 제안되었으며, 개별 토픽을 나타내는 원의 중심은 MDS 알고리즘에 따라 2차원으로 배치되고 Jensen-Shannon divergence를 사용하여 토픽 간 거리를 계산한다. 각 토픽이 차지하는 원의 면적은 분석 데이터 내에서 해당 토픽의 상대적 중요도(relative prevalence)를 의미한다. Figure 3.5에서 같은 사분면에 속한 토픽들은 의미적으로 연관성

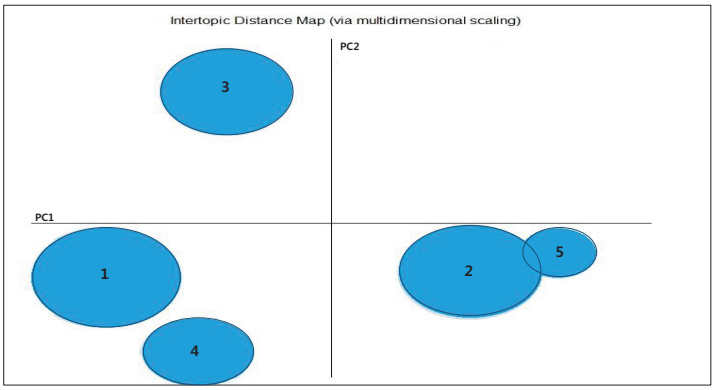

Figure 3.5. Visualization of 5 topics.

을 갖고 있음을 알 수 있는데, 토픽1과 토픽4가 같은 제3사분면에 속하면서 근접하게 위치하고 있어 학생부종합전형과 관련한 공정성 논의가 연관되어 있음을 확인 할 수 있다. 또, 토픽2와 토픽5가 제4사분면에 동시에 속하면서 매우 밀접한 관련이 있음이 확인되는데, 이는 대입전형 개편 논의에서 가장 큰 비중을 차지하는 것은 수능체제 개편 논의임을 재차 확인 할 수 있다.

4. 결론

학생부종합전형에 대한 토픽분석을 시도하고자 경향신문, 부산일보 등 10개의 채널에 실린 4년 8개월간의 기사를 수집하여 총 1,240건을 분석한 결과, 최종적으로 5개의 주제가 도출되었다. 토픽의 수를 결정하기 위해 복잡도 지수를 활용하되, 계산 값의 불안정성을 보완하기 위해 Zhao 등 (2015)이 제안한 RPC을 기준으로 하였다. 토픽의 수가 결정되면 해당 토픽별로 주요 키워드와 관련 기사를 검토하여 토픽명을 결정할 수 있는데, 해당 분야에 대한 전문 지식이 없는 경우에는 토픽명 부여에 다소 어려움을 겪을 수도 있다.

토픽 분석 결과로 가장 큰 비중을 차지하는 것은 학생부종합전형의 선발요소와 관련된 논의였으며, 다음으로 수능체제 개편에 대한 논의가 대입전형 개편 논의보다 더 주요 토픽으로 등장하여 우리 사회에서 수능 시험이 얼마나 민감하고 중요한 사안으로 다루어지고 있는지를 보여 주었다. 또 토픽들의 관계를 시각화하여 표현해 본 결과 학생부종합전형과 대입전형의 공정성이 연관되어 있으며, 대입전형 개편 논의에서 큰 비중을 차지하는 것은 수능체제 개편 논의임이 확인되었다.

다시 말해 학생부종합전형에 대한 쟁점에서도 수능체제 개편이 중요한 논의의 대상임을 확인하였으며, 대입전형의 공정성에 대한 기대가 크다는 것이 확인되었다.

그러나 온라인 뉴스 기사의 수집 및 분석이 가지는 한계점으로는 뉴스 기사에 대한 저작권법 강화에 따라 기사 원문에 대한 수집이 원천적으로 불가능해져서 상용화 서비스를 이용하지 않는 한, 기사의 키워드 수집이 가능한 채널이 제한적이기 때문에 보다 폭넓은 여론 반영에는 다소 한계가 있다.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., and Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: some observations. In Zaki, M.J., Yu, J. X., Ravindran, B., Pudi, V. (Eds), *Advances in Knowledge Discovery and Data Mining. PAKDD 2010.*

Lecture Notes in Computer Science, (Vol. 6118, pp. 391–402), Springer, Berlin, Heidelberg.

Bae, J. W., Son, J. E., and Song, M. (2013). Analysis of twitter for 2012 South Korea presidential election by text mining techniques, *Journal of Intelligence and Information Systems*, **19**, 141–156.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive LDA model selection, *Neurocomputing*, **72**, 1775–1781.

Cho, A., Lee, K. H., and Cho, W. S. (2015). Latent mobility pattern analysis of bus passengers with LDA, *Journal of the Korean Data and Information Science Society*, **26**, 1061–1069.

Choi, H. D. and An, J. W. (2015). How does the general public understand science and technology issues?: a case on the nuclear power issue using topic modeling approach, *Journal of Technology Innovation*, **23**, 152–175.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41**, 391–407.

Grant, S., Cordy, J. R., and Skillicorn, D. B. (2013). Topicsmodels: an R package for fitting topic models. *Journal of Statistical Software*, **40**, 1–30.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences 101*, **supply 1**, 5228–5235.

Grun, B. and Hornik, K. (2011). Topicsmodels: an R package for fitting topic models. *Journal of Statistical Software*, **40**, 1–30.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Jang, D. H. and Park, K. M. (2017). Policy issue discernment of the ocean and fishery field by the press: focus on the big data analysis on journal articles of central daily news papers, *The Journal of Maritime Business*, **37**, 195–203.

Jeong, D. M., Kim, J, S., Kim, G. N,, Heo, J. U., and On, B. W. (2013). A proposal of a keyword extraction system for detecting social issues, *Journal of Intelligence and Information Systems*, **19**, 1–23.

Joo, Y. H. and Kim, S. C. (2017). A study on the comprehensive school report policy for the university admission in Korea, *The Journal of Educational Administration*, **35**, 141–168.

Kang, B. I., Song, M., and Jho, W. S. (2013). A study on opinion mining of newspaper texts based on topic modeling, *Journal of the Korean Library and Information Science Society*, **47**, 315–334.

Khan, R. A. and Kanth, S. (2016). Text mining: knowledge discovery from unstructured data, *Artificial Intelligent Systems and Machine Learning*, **8**, 71–77.

Kim, J. E. and Baek, S. G. (2016). Analysis of issues on the college and university structural reform evaluation using text big data analytics, *Asian Journal of Evaluation*, **17**, 409–436.

Kim, S. H. and Yoon, J. W. (2016). Analysis system for SNS issues per country based on topic model, *Journal of KIISE*, **43**, 1201–1209.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, Springer, New York.

Lee, S. J. (2018). An analysis of conflicting issues on the 'School Performance Records'-centered reform of college admissions system in South Korea, *Journal of Learner-Centered Curriculum and Instruction*, **18**, 923–944.

Park, C. H. (2017). Big data analysis on the demand for the Nuri Curriculum policies based on word clouds and social network analysis, *Journal of Early Childhood Education*, **37**, 79–91.

Sievert, C. and Shirley, K. (2014). LDAvis:a method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

Yu, Y. L. and Baek, S. G. (2016). Issue analysis of the related mass media's news articles on the 2015 revised national curriculum using automated text analysis, *The Journal of Curriculum and Evaluation*, **19**, 127–156.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling, *BMC Bioinformatics*, **16**, S8.

토픽 분석을 이용한 학생부종합전형의 쟁점 분석

홍영희[a,1]

[a]부산광역시교육청 교육정책연구소

(2019년 2월 11일 접수, 2019년 3월 12일 수정, 2019년 4월 9일 채택)

요 약

지난 2018년, 우리사회를 뜨겁게 달구었던 이슈 중 하나로 대입제도 개편에 관한 논쟁을 꼽을 수 있겠다. 그 중에서도 학생부종합전형에 대한 쟁점이 무엇인가를 파악하기 위해 감시와 비판이라는 언론의 기능에 주목하여 관련 뉴스 기사에 대한 토픽 분석을 시도해 보았다. 그 결과 수능체제 개편 논의가 비중있는 주제로 등장하여 수능시험에 대한 한국 사회의 민감성을 보여 주었다. 학생부종합전형과 직접적 관련이 있는 주제로는 학생부종합전형의 세부적인 선발 요소에 대한 논의가 등장하였고, 대입전형의 공정성에 관한 논의와 밀접한 관계를 보였다.

주요용어: 토픽 분석, 뉴스 기사, 학생부종합전형

[1](47119) 부산광역시 부산진구 화지로 12, 부산시교육청 교육정책연구소. E-mail: yhhong0120@gmail.com