Journal of Information Technology and Architecture
Vol. 21. No. 2, June 2024, Pages 159-169
DOI http://doi.org/10.22865/jita.2024.21.2.159

Keybert와 Bertopic을 활용한 텍스트마이닝 연구동향 분석

길완제[1] · 장환영[2] · 신인수[3]*

[1]동국대학교 교육대학원 겸임교수
[2]동국대학교 교육학과 교수
[3]동국대학교 교육대학원 AI융합전공 교수

jaygil8755@gmail.com, welcome@dongguk.edu, 9065031@hanmail.net

(2024년 4월 30일 접수; 2024년 6월 24일 수정; 2024년 6월 26일 채택)

요약: 대량의 텍스트 데이터를 분석하고 활용하는 텍스트 마이닝 기법은 공학 분야뿐만 아니라 사회과학과 교육 등 거의 모든 학문 분야에서 널리 사용되고 있다. 특히 최근 대규모 언어 모델의 급속한 발전은 기존 텍스트 마이닝 기법의 한계를 보완하는 혁신적인 방법들을 도입하는 데 기여하고 있다. 본 연구의 목적은 국내 학술 및 학위 논문을 수집하여 최신 텍스트 마이닝 기법을 활용해 분석하는 것이다. 이를 위해 학술연구정보서비스(RISS) 데이터베이스에서 '텍스트 마이닝'을 키워드로 논문을 수집하였고, 수집된 논문들에 대해 키워드 분석과 토픽 모델링을 수행하였다. 키워드 분석에서는 TF-IDF를 활용한 빈도 기반 분석과 BERT 기반의 KeyBERT를 활용한 분석을 비교하였다. 또한, 토픽 모델링 분석에서는 기존 통계 기반의 LDA 기법과 최신 언어 모델인 BERT 기반의 토픽 모델링 기법인 BERTopic을 비교하였다. 그 결과, BERT 기반의 토픽 분석이 응집도(Coherence Score) 점수에서 보다 우수한 성능을 나타냈다. 특히, Bertopic에서 한국어 임베딩 모델과 Keybert 기반의 토픽 추출이 다국어 모델과 문장 기반의 추출보다 더 높은 응집도 점수를 기록하였다. 본 연구는 이러한 결과를 통해 한국어 텍스트 마이닝에서 최신 기법들의 적용과 활용 가능성을 제시하고자 한다.

주제어: 텍스트마이닝, 토픽모델링, 자연어처리, Bertopic, Keybert

An Analysis of Academic Research Trends in Text Mining using Keybert and Bertopic

Wan-Je Gil[1], Hwan-Young Jang[2], and In-Soo Shin[3]*

[1]Dept. of AI Convergence Education, Graduate School of Education, Dongguk University
[2]Dept. of Education, Dongguk University
[3]Dept. of AI Convergence Education, Graduate School of Education, Dongguk University

(Received April 30, 2024; Revised Jun 24, 2024; Accepted Jun 26, 2024)

Abstract: Text mining techniques for analyzing and utilizing large-scale text data are widely used not only in engineering but also in social sciences, education, and almost all academic fields. This study

*Corresponding Author
본 논문은 2016년 교육부의 재원으로 한국연구재단(NRF-2016S1A3A2925401)의 지원을 받아 수행된 연구임

aims to collect and analyze domestic academic and thesis papers on text mining using the latest text mining techniques. For this purpose, papers were collected from the Research Information Sharing Service (RISS) database using the keyword 'text mining', and keyword analysis and topic modeling were conducted on the collected papers. In keyword analysis, frequency-based analysis using TF-IDF and analysis using BERT-based KeyBERT were compared. Additionally, in topic modeling analysis, traditional statistical-based LDA techniques were compared with BERTopic, a topic modeling technique based on the latest BERT language model. The results showed that BERT-based topic analysis demonstrated superior performance in terms of coherence score. Particularly, the topic extraction based on Korean embedding models and Keybert recorded higher coherence scores compared to those based on multilingual models and sentence-based extraction. Through these findings, this study aims to present the applicability and potential of the latest techniques in Korean text mining.

Keywords: Text Mining, Topic Modeling, NLP, Bertopic, Keybert

1. 서 론

대량의 텍스트 데이터에서 의미 있는 정보를 추출하고 패턴을 파악하는 기법인 텍스트마이닝(Text Mining)[1]은 컴퓨터 과학과 자연어처리 분야의 기술 발전에 힘입어 데이터 기반의 의사 결정을 지원하는 중요한 도구이다[2]. 인공지능의 기법과 텍스트마이닝을 연결하는 공학분야의 연구 뿐만 아니라, 텍스트마이닝은 언어 단위의 식별 및 분류를 위한 언어학에서부터 고대 텍스트 재구성을 위한 역사, 감정 분석에 이르기까지 거의 모든 학문 분야에서 활용되고 있다[3].

본 연구의 목적은, 텍스트마이닝를 주제로 한 다양한 분야의 연구결과를 텍스트마이닝의 주요한 기법들을 활용해 분석하는 것이다. 지금까지 텍스트마이닝에서는 TF_IDF(Term Frequencty-Inverse Document Frequency)와 LDA(Latent Diriclet Allocation) 토픽모델링 같이 전통적인 빈도 기반의 분석 방법을 많이 사용하고 있다. 이 방법은 직관적으로 주요 주제나 트렌드를 파악하는데 효과적이지만, 단어의 빈도만을 고려하기 때문에 맥락을 파악 하는데 한계를 가지고 있다.

최근 대규모언어모델(Large Language Model)의 등장으로 텍스트마이닝에 새로운 패러다임이 도입되고 있다. 본 논문에서는 텍스트마이닝 연구동향을 분석하기 위해 논문 수집과 분석을 파이썬 프로그램으로 자동화하는 시스템을 구축하였고, 빈도 기반의 키워드추출과 LDA 토픽모델링을 수행하고 동시에 BERT (Bidirectional Encoder Representations from Trans-formers) 기반의 키워드분석과 토픽모델링도 실행하여 그 결과를 비교하고 분석하였다. 이를 통해 텍스트마이닝 연구동향을 효과적으로 파악하기 위한 최신 기법 소개와 텍스트마이닝 활용에 대한 인사이트를 얻을 수 있는 방법을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 텍스트마이닝의 주요 이론적, 방법론적 배경에 대해서 살펴본다. 제 3장에서는 연구방법과 절차에 대해 제시하고 제 4장에서는 연구결과를 기술한다. 마지막으로 재 5장에서 결론과 시사점을 기술한다.

2. 이론적 배경

2.1 TF-IDF와 n-gram

TF-IDF는 Term Frequencty-Inverse Document Frequency의 약자로서 전체 문서에 나타난 단어 빈도와 문서 내 단어의 역빈도를 동시에 계산하여 문서 내 용어의 중요성을 계산한다. 즉, 문서에 용어가 나타나는 빈도를 측정 (TF)할 뿐 아니라 여러 문서에서 용어가 얼마나 고유한지를 평가 (IDF)하여 결합함으로써 기존 카운트 기반의 방법을 개선하였다[4]. TF-IDF는 문서의 중요한 단어를 효과적으로 강조하여 텍스트 분류, 추세 예측, 정보 검색, 데이터 클러스터링 등에 독립적인 연구방법으로 활용되고 있고[5], 토픽모델링을 위한 주요 어휘 추출 기법으로 활용되고 있기도 하다[6].

의미 있고 중요한 단어를 추출할 때 하나의 단어 단위로만 수행하는 것이 아니라, 연속된 단어를 하나의 묶음으로 처리하는 방법도 고려해야 한다. n-gram은 통계학 기반의 언어 모델 중 하나로서 다음 단어를 예측할 때 특정 개수(n)의 단어를 하나의 묶음으로 간주한다[7]. n-gram은 n개 항목의 연속된 시퀀스를 의미하며, 이는 단어, 메모 또는 더 큰 시퀀스 내의 다른 요소일 수 있다. n-gram은 인공신경망 기반의 언어모델에 의해 대체되는 듯했지만, 최근 연구에서는 n-gram

기법이 이미지 분류, 기계 번역 등에서 여전히 효과적임이 밝혀지고 있다[8,9].

2.2 LDA Topic Modeling

토픽 모델링은 문서 집합, 즉 말뭉치(Corpus) 안에서 특정 주제를 찾기 위한 비지도학습 기반의 인공지능 기법으로 텍스트마이닝 중에서 가장 많이 활용되고 있다. 토픽 모델링은 확률기반과 비확률 기반으로 구분할 수 있고[10] 확률기반의 토픽모델링 중 가장 대표적인 것이 잠재 디리클레 할당(LDA, Latent Diriclet Allocation)이다[11]. LDA는 토픽의 단어 분포와 문서의 토픽분포를 결합하여 문서 내 단어들이 생성된다고 가정한다. 즉, 단어가 특정 토픽에 속할 확률과 문서에 특정 토픽이 존재할 확률을 결합하여 토픽을 추출한다. LDA는 순서를 고려하지 않는 단어가방(Bag of Words)[12]을 기반으로 하고 있기 때문에 문맥 정보나 문장 구조 등의 순서 정보가 손실되는 단점이 있다.

비확률기반의 토픽 모델링 기법으로는, 선형대수학의 특이값분해를 활용해 토픽을 추출하는 잠재의미분석(LSA)[13]과 주어진 음수를 포함하지 않는 행렬을 두 개의 음수를 포함하지 않는 행렬로 분해하는 알고리즘인 비음수행렬분해(NNMF)[14] 등이 있다.

2.3 Keybert와 Bertopic

Keybert는 BERT 모델을 기반으로 문서 수준에서 키워드를 추출하는 기법으로 세 단계로 진행된다. 먼저 BERT로 문서 임베딩을 수행하고 그 다음으로 n-gram을 사용해 적정한 수의 연속된 단어에 대한 단어 임베딩을 추출한다. 마지막으로 코사인 유사도를 구하여 문서와 가장 유사한 단어를 찾는다. 가장 유사하는 단어는 전체 문서를 가장 잘 설명하는 단어로 식별할 수 있다[15].

Bertopic 또한 BERT 기반 임베딩을 수행하고 클래스 기반의 TF-IDF를 사용한 것이 아이디어의 핵심이다[16]. Bert는 사전 학습된 언어모델로 문장을 이해하고 문맥을 파악하는 능력이 있기 때문에[17] 문맥을 고려한 토픽 추출에 활용될 수 있다.

Bertopic은 세 단계를 거쳐 수행된다. 먼저 문서 임베딩을 수행한다. 그 다음으로 문서 클러스터링을 한다. 문서 임베딩은 고차원이기 때문에 UMAP (Uniform manifold approximation and projection for dimension reduction)[18]을 활용해 차원 축소를 수행한다. 그런 다음 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)[19]과 같은 밀도 기반 클러스터링 기법을 적용하여 토픽 클러스터를 생성하고 이상치를 식별한다. 마지막으로 토픽 표현을 한다. 각 토픽을 나타내기 위해 TF-IDF 점수를 수정하여 클러스터별로 가장 중요한 단어를 추출하는 클래스 기반 TF-IDF로 토픽을 추출한다. <Figure 1>은 Bertopic 알고리즘을 도식화한 것이다.

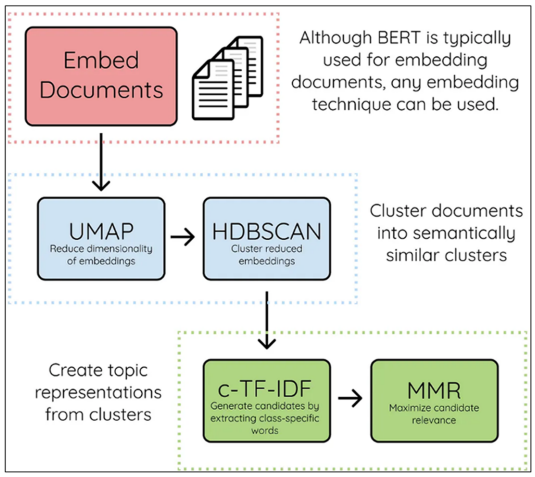

Figure 1. Bertopic Algorithm[20]

Bertopic은 하나의 고정된 알고리즘을 사용하는 것이 아니라 대규모언어모델을 적용해 확장할 수 있다는 점에서 새로운 연구가 활발한 편이다[21].

3. 연구 방법

본 논문에서 제안하는 연구방법은 아래 <Figure 2>와 같다.

먼저 학술연구정보서비스(riss.kr)에서 “텍스트 마이

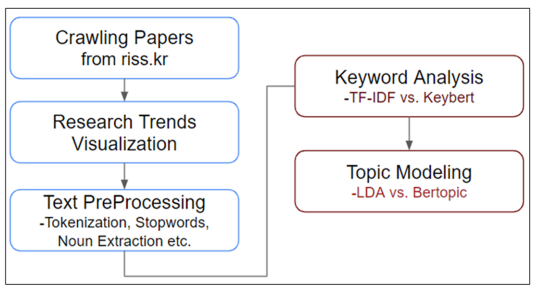

Figure 2. Research Process

닝”이란 키워드로 국내학술논문과 국내학위논문을 수집하였다. 저자, 연도, 기관(대학), 학술지(석박사), 상세링크, 초록 등 7개 항목을 Python 라이브러리인 Requests와 BeautifulSoup으로 자동 수집하였다. 또한 사회과학, 공학, 교육, 인문학, 복합학, 예술체육, 자연과학, 기타, 농수해양, 의약학 등 10개 분야에 대해서도 별도의 데이터로 구분하여 수집하였다. 초록이 없는 논문은 <NoAbs.>로 표시하였고 중복 논문을 제거한 후 csv 파일로 저장하였다. 수집된 논문의 제목에 “텍스트 마이닝”, “텍스트마이닝”, “텍스트 분석”, “텍스트 분석” 중 하나라도 들어있지 않은 논문은 제외하였다.

다음으로 텍스트 마이닝을 위해 텍스트 전처리를 수행하였다. 한글 명사 추출을 위해 Konlpy 라이브러리의 Okt() 클래스를 활용했고 논문 제목과 초록을 대상으로 수행하였다. 보다 의미있는 정보를 추출하기 위해 sklearn의 Countvectorizer와 Tfidfvectorizer를 적용했고 n-gram 범위를 2로 설정하여 연속된 두 개의 단어를 묶어서 적용했다. 또한 한 글자 단어와 불용어를 제거하였다. 불용어는 “텍스트 마이닝”과 같이 너무 자주 등장하거나 “연구”, “중심”, “관한”, “통한”과 같이 의미 없는 단어를 설정했다.

키워드 분석은 TF-IDF로 추출한 상위 60개 명사를 김성근의 2016년 선행 연구[22]를 참고하여 “Mehtod & Technique”, “Data Source”, “Application&Area”로 구분하여 분석하였다. 또한 텍스트마이닝에 관한 논문이 가장 많이 출판된 분야인 공학, 사회과학, 교육 분야의 논문들을 대상으로 TF-IDF 빈도기반 키워드 추출과 BERT기반 Keybert 키워드 추출을 실행하여 주요 키워드들을 워드클라우드로 시각화하여 결과를 비교하였다.

전체 논문을 관통하는 토픽을 알아보기 위해 확률기반의 LDA 토픽모델링을 수행하고 동시에 BERT 기반의 토픽모델링인 Bertopic을 수행하였다. LDA 토픽모델링의 토픽 수 계산을 위해서 응집도(Coherence Score)와 혼란도(Perplexity)를 구하였다. 응집도는 값이 높을수록 의미론적인 일관성이 높다고 평가하고 혼란도는 얼마나 빠르게 수렴하는지를 확인 가능하며 파라미터에 따라 성능 평가를 할 때 주로 사용한다. 일반적으로 혼란도는 낮을수록 성능이 높다고 판단한다[23]. 본 연구에서는 아래 <Figure 3>과 같이 토픽 개수를 6개부터 11개까지 변화하면서 혼란도(위 그림)와 응집도(아래 그림)를 구하여 토픽 수를 정하는데 참고하였다. 응집도는 토픽 수 7개일 때 0.401로 가장 높았고 혼란도는

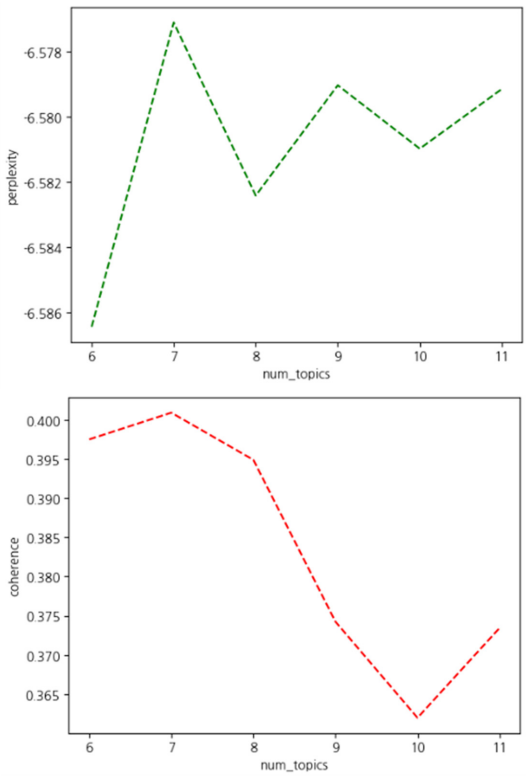

Figure 3. Calculation of Coherence and Perplexity for the Number of Topics

토픽 수 6일 때 -6.586으로 가장 낮았다. 토픽 수를 산정하고 모델 성능을 평가할 때 혼란도 보다 응집도가 보다 일반적이므로 응집도를 기준으로 토픽 수를 7개로 정하였다.

토픽모델링의 결과를 객관적으로 비교해보기 위해 Bertopic도 7개로 주제를 구분하였다. Bertopic을 수행하기 위해 연구자가 직접 정해야 하는 하이퍼파라미터들이 꽤 많은데 특히 영어가 아닌 한국어를 대상으로 수행하기 위해서는, 한국어에 적합한 토큰화 방법과 임베딩 모델선택이 중요하다. 본 연구에서는 Konlpy의 Okt() 클래스를 sklearn의 CountVectorizer의 tokenizer로 설정하였다. 제목과 초록에 있는 단어들 중 두 글자 이상의 명사를 추출하였고 보다 의미있는 단어를 선정하기 위해 n_gram의 범위를 1에서 2까지로 설정하였다. 또한 불용어(Stopwords)를 설정하여 의미 없는 단어들을 배제하고 너무 자주 등장하거나 희소하게 나타나는 단어들을 이상치로 처리하여 제외하였다. 한국어 임베딩을 위해 허깅페이스의 Sentence Transfer Model 중에서 ‘jhgan/ko-sroberta-multitask’을 활용했고

Table 1. Experimental Set Up for Topic Modeling

Algorithm	Experimental Set Up
LDA	number of Topic = 7 passes = 15 tokenizer = Okt() coherence='c_v'
Bertopic	number of Topic = 7 vectorizer_model = CountVectorizer(tokenizer=Okt(), n_gram=(1,2)) coherence='c_v' embedding_model = "jhgan/ko-sroberta-multitask" or "paraphrase-multilingual-MiniM-L12-v2" representation_model = KeyBERTInspired() or None

[24] 다국어 임베딩 모델인 'paraphrase-multilingual-MiniM-L12-v2'도 적용하여 비교하였다. 또한 Keybert의 하이퍼파라이터 중에서 representation_model에 Keybert 아이디어를 바탕으로 확장된 개념의 키워드 추출방식인 KeyBERTinspired를 적용했을 때와 안했을 때를 구분하여 실험하였다. Keybert는 각 토픽의 중심을 효과적으로 나타낼 수 있는 키워드를 제공해줄 수 있기 때문에 토픽모델링의 품질을 높이고 토픽의 해석 가능성과 더 나은 토픽요약을 가능하게 한다.

응집도를 구하기 위한 방법으로 LDA와 Bertopic c_v와 umass가 있는데 상호 객관적인 비교를 위해 모두 c_v를 사용하여 토픽의 일관성을 측정하였다.

LDA와 Bertopic을 수행할 때 사용한 주요 파라미터와 설정은 <Table 1>과 같다.

4. 연구 결과

4.1 연구동향 분석

2000년부터 2023년까지 수집된 논문은 학술 3,927편, 학위 1,791편 등 전체 5,339편이었다. 학술 논문의 제목에 "텍스트 마이닝", "텍스트마이닝", "텍스트 분석", "텍스트분석", "데이터 마이닝", "데이터마이닝" 중 하나 이상이 포함된 논문들만 선별한 결과 1,727편이었고, 중복된 논문들까지 제거한 최종 분석대상 학술 논문은 1,677편이었다. 마찬가지 방법으로 학위논문도 전처리 과정을 거쳐 선정된 논문은 500편이었다. 따라서 학술 논문과 학위 논문 전체 5,339편 중에서 2,177편의 논문을 최종 분석대상으로 선정하였다.

텍스트마이닝에 관한 논문 편수를 연도별로 확인해보면, <Figure 4>와 같이 2016년 100편으로 연간 100편을 넘어섰고 그 후로 꾸준히 매년 증가하여 2023년에는 456편으로 많은 연구자들이 선호하는 핫한 토픽인 것을 알 수 있다. 학위 논문에 비해 학술 논문의 수가 2배 이상 많은 것으로 나타났다. 텍스트 마이닝에 대한 논문 수가 꾸준히 증가하고 있는 이유로는, 학술 논문을 비롯해 소셜 미디어, 뉴스, 상품평 등 디지털 텍스트의 양이 비약적으로 증가하였고 키워드 분석, 토픽 모델링, 네트워크분석 등 자연어처리 관련 기술과 기법이 다양하게 발전하고 오픈 소스와 온라인 서비스 형태로 제공되고 있어 비전공자도 비교적 쉽게 사용할 수 있게 되었기 때문인 것으로 분석된다.

<Figure 5>에서 학문 분야에 따른 논문 수를 살펴보

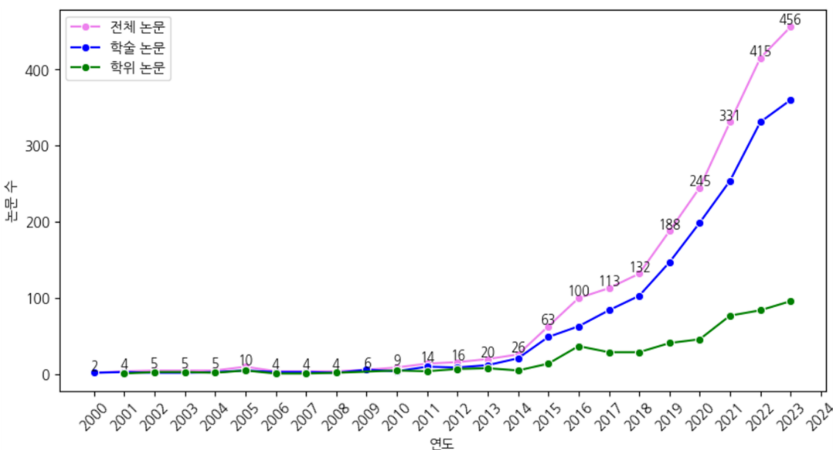

Figure 4. Number of papers by year

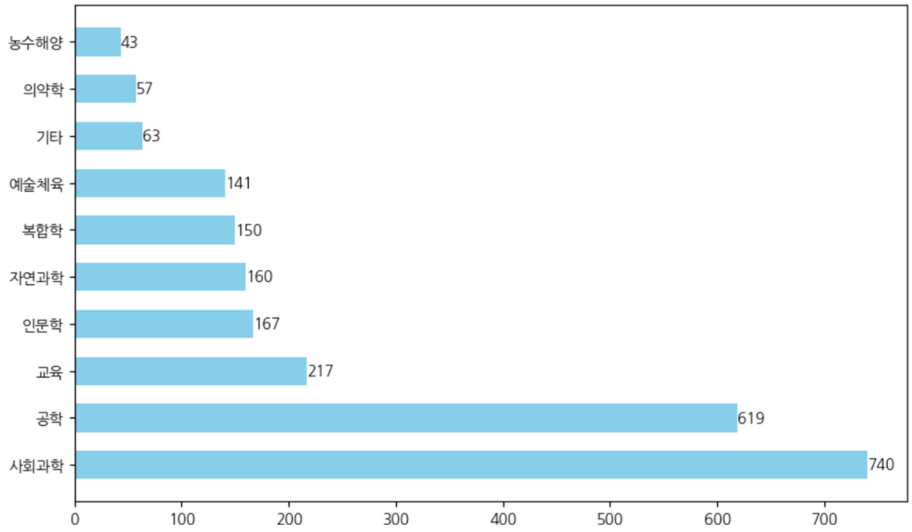

Figure 5. Number of Papers by Academic Fields

면, 사회과학 740편, 공학 619편, 교육 217편, 인문학 167편, 자연과학 160편, 복합학 150편 순이었으며 그 외 예술체육, 농수해양, 의약학 등 거의 모든 학문 분야에서 텍스트마이닝이 다뤄지고 있음을 확인할 수 있다.

4.2 키워드 분석

논문 제목과 초록의 텍스트를 대상으로 키워드를 추출하였다. sklearn의 Tfidfvectorizer 클래스를 사용하였고 n-gram_range(2,2)로 설정하여 연속된 2개의 명사를 추출하고 90% 이상의 문서에 모두 등장하는 단어는 제외하였다. 이러한 방식으로 추출된 상위 60개 단어를 아래 <Figure 6>과 같이 3가지 주제로 분류하여 분석하였다.

텍스트마이닝의 방법과 기법(Method&Technology)에 대한 키워드로는, 토픽모델링이 월등히 많았고, 데이터 수집, 의미 연결, 주요 키워드, 단어 빈도, 뉴럴 네트워크, 연결 중심성, 워드클라우드, 동시 출현 등을 추출하였다. 텍스트마이닝의 연구대상(Data Source)에 대한 키워드로는, 뉴스기사, 소셜미디어, 메타 버스, 학술 논문, 외부 데이터, 온라인 리뷰, 소비자 인식, 온라인 뉴스 등이 선정되었다. 텍스트마이닝의 적용분야(Application&Area)에 대한 키워드로는, 문제해결, 영향요인, 동향 파악, 정보 제공, 정보 추출, 기술 발전, 전략 수립, 인식 변화, 마케팅 전략 등을 뽑을 수 있었다.

학문 분야별 키워드 분석은 TF-IDF 기법과 Keybert 기법을 함께 수행하여 서로 비교하였고 한 눈에 결과를 파악해보기 위해 워드클라우드로 시각화하였다. 각 분

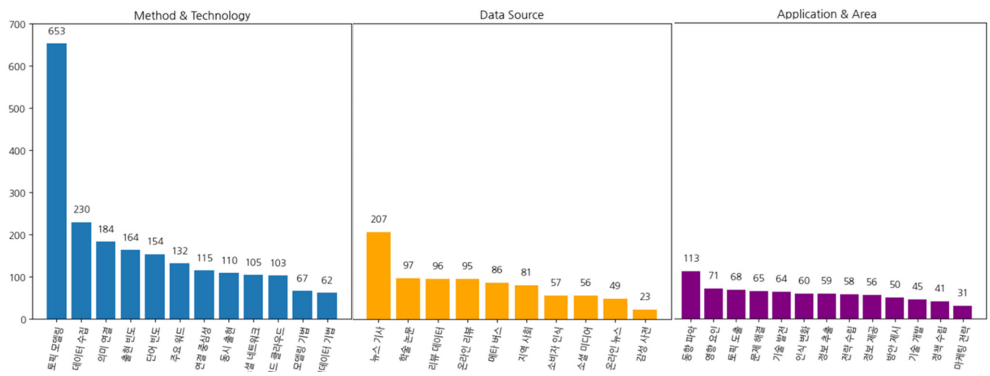

Figure 6. Number of Keywords by Subject

Figure 7. Keywords from Engineering Field

Figure 8. Keywords from Social Science Field

야의 논문 제목과 초록에 나타난 단어들을 대상으로 분석하였다.

빈도 기반의 키워드 추출에서는 학문 분야별로 키워드의 차이가 많이 드러나지 않은 반면, Keybert에서는 확연히 구분되는 키워드들이 추출되었다.

<Figure 7>은 공학분야의 키워드 분석 결과이다. TF-IDF 분석에서는 데이터, 기술, 이용, 기법, 분류, 워드, 방법, 동향, 예측, 추출과 같은 단어들이 보이는데 다른 학문 분야에서 나타나는 키워드와 크게 차이가 없었다.

반면, Keybert 기법을 통한 분석에서는 행정안전부, 안전모, 안정성, 공공안전, 자연재해, 위험관리, 자연재해, 정보과학, 위험부담과 같이 확연히 구별되는 키워드들이 추출되었다. 이는 문서 임베딩과 단어 임베딩을 통해 문서와 가장 유사한 단어를 키워드로 선정하는 Keybert 알고리즘이 잘 적용된 결과로 해석할 수 있다.

사회과학 분야와 교육 분야의 논문들에서 비슷한 양상의 결과가 나타났다.

<Figure 8>은 사회과학 분야의 키워드 분석 결과이다. 사회과학 분야의 키워드 분석에서는 TF-IDF 결과 토픽, 데이터, 워드, 이용, 빅데이터, 리뷰, 사회, 정책, 소비자방안, 기업과 같이 비교적 이 분야의 특징적인 단어들이 추출되었고 Keybert에서는 데이터마이닝, 컴퓨터과학, 정보검색, 정보산업, 프로토타입, 정보처리와 같이 다른 분야에서는 나타나지 않는 단어들 위주로 선정된 것을 확인하였다.

교육 분야의 키워드 분석에서는 <Figure 9>와 같다. TF-IDF 결과 토픽, 사회, 인식, 동향, 모델링, 단어, 주제와 같이 일반적인 키워드와 함께 교육과정, 교육, 학습, 교사 수업, 학생 등 교육 분야에 특화된 단어들도 주요 키워드로 선정되었다. Keybert에서는 수업, 교과목, 학습분석, 단원학습, 교육실습, 교육과정평가원, 교육공학, 교육사상, 이러닝과 같이 다른 분야에서는 나타나지 않는 독특한 키워드들 위주로 선정된 것을 확인할 수 있다.

Figure 9. Keywords from Education Field

4.3 토픽 모델링

LDA 토픽모델링 수행 결과는 아래 <Table 2>와 같다. 각 토픽에 속한 주요 단어들을 기반으로 토픽 이름

Table 2. Results of LDA Topic Modeling

	Topic 이름	주요 단어
Topic1	대상(기업)	기술정보, 기업데이터, 사용자리뷰, 고객, 감성
Topic2	분야(언론)	사회뉴스, 정부정책, 언론뉴스, 이슈, 기사, 코로나
Topic3	활용 유형	문서분류, 정보추출, 단어정보, 모델적용
Topic4	분야(교육)	교육, 학습, 인공지능 교육과정, 학생, 학습자, 수학
Topic5	기법 유형	토픽모델링, 워드클라우드, 빈도분석, 네트워크분석
Topic6	분야(마케팅)	소비자, 디자인, 공간, 문화,온라인, 감성, 소비
Topic7	분야(안전)	관광, 안전, 재난, 사고, 지역
Coherence sore		0.538

을 정하였다. 텍스트 마이닝의 분야와 활용 유형, 기법, 그리고 분야과 대상 등이 7개 토픽의 주요 내용이다.

Topic1의 주요 단어로는 기술정보, 기업데이터, 사용자리뷰, 고객, 감성 등 텍스트 마이닝의 대상이 되는 데이터 종류에 대한 단어들이 선정되었다. Topic2에서는 사회뉴스, 정부정책, 언론뉴스, 이슈, 기사, 코로나와 같이 언론 분야 주제가 한 데 묶였다. Topic3에서는 문서분류, 정보추출, 단어정보, 모델적용과 같이 텍스트 마이닝의 활용 유형에 해당하는 단어들이 선택되었다. Topic4에서는 교육, 학습, 인공지능 교육과정, 학습자, 수학 등 교육 분야에 해당하는 키워드들이 특징적으로 나타났다. Topic5에서는 토픽모델링, 워드클라우드, 빈도분석, 네트워크분석과 같이 텍스트 마이닝의 기법 유형을 하나의 토픽으로 그룹화하였다. Topic6은 소비자, 디자인, 공간, 문화, 온라인, 감성, 소비와 같이 마케팅에 관한 토픽이고, Topic7은 관광, 안전, 재난, 사고, 지역 등 안전 분야의 토픽이다. LDA 모델의 응집도 점수는 토픽 수가 7일 때, 0.538로 나왔다.

Bertopic을 활용한 토픽모델링 결과는 <Table 3>과 같다.

Table 3. Results of Bertopic

Model Name	Embedding Model	Representation Model	Coherence Score
Ko-Keybert	jhgan/ko-sroberta-multitask	KeyBERTInspired( )	0.67149
Ko-Base	jhgan/ko-sroberta-multitask	None	0.56619
Multi-Keybert	paraphrase-multilingual-MiniM-L12-v2	KeyBERTInspired( )	0.54826
Multi-Base	paraphrase-multilingual-MiniM-L12-v2	None	0.49433

Table 4. Major Keyword of Bertopic Results

	Ko-Keybert	Ko-Base	Multi-Keybert	Multi-Base
Topic1	'빅데이터', '리뷰', '댓글', '데이터', '평가', '토픽 모델링', '기사', '토픽', '논문', '동향 이용'	'이용", '국내', '데이터', '빅데이터', '한국', '인식', '비교', '네트워크', '뉴스', '논문	'온라인', '글쓰기', '소셜', '동향 이용', '콘텐츠', '토픽 모델링', '언론 기사', '이러닝', '리뷰', '디지털'	'동향', '온라인', '비교', '워드', '네트워크', '인식', '토픽', '기사', '분야', '빅데이터'
Topic2	'데이터', '리뷰', '정보', '토픽 모델링', '방법론', '동향 기법', '도출', '평가', '분류'	'인식', '온라인', '리뷰', '코로나', '비교', '기사', '변화', '빅데이터', '탐색'	'동향 이용', '정보', '인공 지능', '기법 이용', '교육 동향', '온라인', '댓글', '리뷰', '사용자', '기술'	'인식', '국내', '한국', '교육', '빅데이터', '변화', '뉴스', '탐색', '기사'
Topic3	'정책 인민일보', '전략 북한', '중국학 지적', '중국 전략', '북한 경제정책', '인민일보 기사'	'특허', '기술', '특허 정보', '정보', '이용 특허', '방법', '이용', '방법론', '모델', '접근'	'관광 정보', '비교 관광', '온라인 관광', '이용 관광지', '관광지 선택', '관광지 만족도', '지속 관광', '관광지', '호텔 고객'	'비교', '리뷰', '내용', '기법 이용', '조선', '기사', '부동산', '동향', '주택', '감사'
Topic4	'빅데이터 태권도', '태권도원 빅데이터', '태권도 기사', '태권도 동향', '태권도 변화'	'북한', '중국', '김정은', '시기', '김정은 시기', '경제', '경제정책', '발전', '시진핑', '시기 중국'	'특허 정보검색', '특허 정보', '특허 기술', '특허 데이터', '특허 분류', '로드맵 특허', '방법 특허'	'관광', '호텔', '항공', '교통', '관광지', '온라인', '선택', '영향', '이미지', '양상'
Topic5	'예측 중국영화', '영화 마케팅', '흥행 예측', '영화 리뷰', '영화 온라인', '영화 스크립트'	'영화', '흥행', '관객', '리뷰', '흥콩', '흥행 예측', '영화 흥행', '영화 리뷰', '예측'	'데이터마이닝 이용', '데이터마이닝 기법', '방법론 데이터마이닝', '데이터마이닝 기술'	'특허', '기술', '특허 정보', '정보', '이용 특허', '방법', '방법론', '접근', '모델 특허', '특허 기술'
Topic6	'기법 데이터마이닝', '데이터마이닝 기법', '방법론 데이터마이닝', '데이터마이닝 이용'	'데이터', '데이터마이닝', '데이터 기법', '데이터마이닝 이용', '데이터마이닝 기법'	'이용 코로나바이러스', '코로나 백신', '코로나바이러스 감염증', '코로나바이러스', '접종 동향'	'데이터마이닝', '데이터마이닝 기법', '시설 지표', '인문', '인문 전산학', '폐교 시설'
Topic7	'빅데이터 노인학', '빅데이터 노인장', '기법 빅데이터', '댓글 빅데이터', '빅데이터', '동향 노인학'	'일본어', '일본어 교육', '조선 교육', '기법 일본어', '교육 요람', '요람', '조선', '영어'	'영화 마케팅', '영화 수용', '영화 리뷰', '예측 영화', '영화 온라인', '데이터마이닝 영화'	'영화', '흥행', '흥콩', '영화 리뷰', '영화 흥행', '흥행 예측', '관객', '리뷰'

Bertopic은 주요 하이퍼파라미터들을 변경해가면서 그 결과를 비교해보았다. 먼저 한국어 임베딩 모델을 사용했을 때와 다국어 임베딩 모델을 사용했을 때이다. 한국어 임베딩 모델을 사용했을 때 응집도 점수가 0.566이었고 다국어 임베딩 모델을 사용했을 때 0.494였다. 이러한 결과는 LDA 0.538과 비교해봤을 때 한국어 모델은 LDA보다 더 우수했고 다국어 모델은 LDA 결과가 더 좋았다. Bertopic의 표현모델 하이퍼파라미터를 KeyBERTInspired()를 사용한 경우에는, 모두 LDA보다 점수가 높았고 같은 임베딩 모델에서도 사용하지 않을 때보다도 높은 응집도 점수를 보였는데, 특히 한국어 모델에 KeyBERTInspired()를 같이 사용하면 0.6714로 비교적 다른 모델들과 큰 차이를 보이며 가장 우수한 성능을 나타냈다.

<Table 4>는 각 Bertopic 수행 결과 추출된 키워드들이다. 실험 결과를 가공하지 않은 상태에서 각 토픽별 키워드들을 표시한 것이다. 이는 사후 작업을 통해서 토픽들을 병합하거나 토픽 수를 변경하여 더 의미있는 결과를 도출할 수도 있다는 의미인 것이다. 본 연구에서는 LDA와 Keybert를 비교해보고 특히, 한국어 텍스트에서 어떤 설정이 더 우수한 지 알아보는 것이 우선 과제이기 때문에 사후 작업은 생략하였다.

5. 결론 및 시사점

본 연구에서는 텍스트마이닝에 관한 연구동향 분석, 키워드 분석, 토픽 모델링을 수행하였다. 2000년부터 2023년까지 5,339편을 파이썬 라이브러리를 활용하여 자동 수집하였고 전처리 과정을 거친 학술 논문은 1,677편과 학위 논문 500편이었다. 2016년 선행연구와 비교해 볼때, 2016년 이후 연구기관이나 학문분야가 공학 중심에서 경영, 교육 등 인문사회과학분야로 확대되는 경향을 보였다.

키워드 분석을 통해서 연구방법 및 기술, 데이터 소스, 적용영역을 확인하였다. 텍스트 마이닝 연구방법에서도 2016년 연구에서는 오피니언, 온톨로지, 클러스팅이 상위 빈도였다면, 이번 연구에서는 토픽모델링, 의미연결 등으로 변화하였다. 특히 토픽모델링을 활용한 연구가 압도적으로 많았다. 데이터 소스도 트위터, 블로그, 소셜미디어 중심에서 뉴스기사, 메타버스, 학술논문 등으로 다양화되었다. 적용영역도 단백질, 트렌드, 마케팅에서 문제해결, 영향요인, 동향파악 등으로 변화되었다. 학문 분야별 키워드 분석은 TF-IDF기법과 Keybert 기법을 함께 수행하여 서로 비교한 결과, TF-IDF와 n-gram을 활용한 키워드 추출은 여전히 직관적이고 의미 있는 결과를 보여주지만, Keybert 또한 분야별 키워드 추출에서 특징을 보다 분명하게 드러내는 성과를 나타냈다.

마지막으로 LDA와 Bertopic을 활용한 토픽모델링을 수행하였다. LDA는 보다 전통적이고 단순하며 계산적으로 가벼운 방법을 제공하고, Bert는 더 깊은 언어 패턴을 파악하여 조금 더 맥락적인 결과를 제시할 수 있지만 GPU를 사용해야 하고 연산 시간이 더 많이 소요되는 단점이 있다. Bert 기반의 토픽모델링에 관한 연구는 아직 많지 않고 객관적으로 성능을 평가할 수 있는 방법론이 확실하지 않지만[25], 통계 기반의 LDA 토픽모델링과 Bertopic을 상호 비교하여 사용함으로써 상호 보완할 수 있는 결과를 제시할 수 있다.

특히, Bertopic을 한국어 텍스트에 적용할 시, 한국어 임베딩 모델과 다국어 임베딩 모델을 비교하여 성능을 분석할 필요가 있고, Keybert에 기반한 토픽 추출의 가능성에 대해서도 탐색해볼 필요가 있다. 본 연구 결과에서는, 한국어 임베딩 모델을 사용했을 때 응집도 점수(0.566)가 다국어 임베딩 모델을 사용했을 때 응집도 점수(0.494)보다 높았고, KeyBERTInspired()를 사용한 경우에는, 모두 LDA(0.538)보다 점수가 높았고 같은 임베딩 모델에서도 사용하지 않을 때보다도 높은 응집도 점수를 보였는데, 특히 한국어 모델에 KeyBERTInspired()를 같이 사용하면 0.6714로 비교적 다른 모델들과 큰 차이를 보이며 가장 우수한 성능을 나타냈다.

이러한 분석을 통해 살펴본 이 연구의 기여는 다음과 같다.

첫째, 텍스트마이닝에 대한 국내 연구동향을 광범위하게 분석하여 기존 선행연구에 비해 최근에 나타나고 있는 다양한 연구주제와 연구대상 등을 파악하였다. 둘째, 학술연구에서 활용되는 주요 텍스트 마이닝 기법을 소개하고, Keybert 및 Bertopic 등 최근 연구되고 발전하고 있는 텍스트 마이닝 기법을 소개하고 적용하였다. 셋째, 논문 수집부터 토픽모델링 분석까지 Python 코드로 자동화하여 비전공자도 연구동향과 텍스트 마이닝 분석에 적용할 수 있음을 확인하였다.

이 연구의 한계점으로는 학술연구정보서비스 하나의 데이터 베이스에 텍스트 마이닝이란 하나의 키워드만 사용하였다는 점이다. 텍스트 마이닝을 포괄하는 키워드를 활용하고 디비피아, 국회도서관 등 다른 데이터베

이스를 통해 교차 검증을 통해 데이터를 수집할 필요가 있다. 향후 연구에서는 공학, 사회과학, 인문학 학문 분야별로 시간의 흐름에 따라 텍스트 마이닝 적용 기법상의 변화를 조사해 볼 필요도 있다. 동일한 학문 분야별로 국내와 국외의 텍스트 마이닝 기법 및 활용 분야 등에 대한 비교연구도 수행된다면 국내 텍스트 마이닝 연구의 발전에 도움이 될 것이라 생각된다. 또한 마지막으로 토픽 모델링을 평가할 때 최신 LLM을 활용한 평가 방법을 도입하여 보다 객관적이고 합리적인 평가 지표를 제시할 필요가 있다. ChatGPT 등 최신의 대규모 언어모델을 활용하여 프롬프트 엔지니어링[26,27]으로 시기별, 분야별 연구 주제와 연구 방법, 연구목적을 요약하고 분석하는 연구동향 자동화 서비스 시스템을 후속 연구 과제로 선정하였다.

REFERENCES

[1] A. Jadhav, P. Jagtap, S. Gurav, S. Jadhav, N. Jadhav, and A. Akkalkot, A survey on text mining - Techniques, application, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 9, No. 3, pp. 338-343, 2023.

[2] S. Gnanavel, V. Mani, M. Sreekrishna, R.S. Amshavalli, Y. Reta, G. N. Duraimurugan, and S. Rao, Rapid text retrieval and analysis supporting latent dirichlet allocation based on probabilistic models, Mobile Information Systems, 2022(1), 6028739, 2022.

[3] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities, R&D Management, Vol. 50, No. 3, pp. 329-351, 2020.

[4] A. Aizawa, An information-theoretic perspective of tf–idf measures, Information Processing & Management, Vol. 39, No. 1, pp. 45-65, 2003.

[5] S. Qaiser, and R. Ali, Text mining: Use of TF-IDF to examine the relevance of words to documents, International Journal of Computer Applications, Vol. 181, No. 1, pp. 25-29, 2018.

[6] Z. Zhuo, Q. Jiaohua, X. Xuyu, T. Yun, L. Qiang, N. Neal, and Xiong, News text topic clustering optimized method based on TF-IDF Algorithm on Spark, Computers Materials & Continua, Vol. 62, No. 1, 2020.

[7] W. B. Cavnar, and J. M. Trenkle, N-gram-based text categorization, InProceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, Vol. 161175, pp. 14, 1994.

[8] H. Li, D. Cai, J. Xu, and T. Watanabe, N-gram is back: Residual learning of neural text generation with n-gram language model, arXiv preprint arXiv:2210.14431, 2022.

[9] R. Lv, J. Guo, W. Rui, X. Tan, Q. Liu, and T. Qin, N-Gram nearest neighbor machine translation, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[10] K. Pooja, and P. Bansal., Topic modeling: A comprehensive review, EAI Endorsed transactions on scalable information systems, Vol. 7, No. 24, 2019.

[11] D. M. Blei, Probabilistic topic models, Communications of the ACM, Vol. 55, No. 4, pp. 77-84, 2012.

[12] Y. Zhang, R. Jin, and Z. H. Zhou, Understanding bag-of-words model: A statistical framework, International Journal of Machine Learning and Cybernetics, Vol. 1, pp. 43-52, 2010.

[13] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, Effective comparison of LDA with LSA for topic modelling, In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1245-1250, 2020.

[14] D. Lee, and H. S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, 13, 2020.

[15] B. Issa, M. B. Jasser, H. N. Chua, and M. Hamzah, A comparative study on embedding models for keyword extraction using KeyBERT method, In 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET), pp. 40-45, 2023.

[16] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794, 2022.

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2018.

[18] L. McInnes, J. Healy, and J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802. 03426, 2018.

[19] D. Deng, DBSCAN clustering algorithm based on density, In 2020 7th international forum on electrical engineering and automation (IFEEA), pp. 949-953, 2020.

[20] Bertopic Algorithm Hompage image, https://maartengr.github.io/BERTopic/img/algorithm.png, 2024.
[21] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, Using BERTopic model for abstracts classification, Jurnal Media Informatika Budidarma, Vol. 7, No. 3, 2023.
[22] S. G. Kim, H. J. Cho, and J. Y. Kang, The status of using text mining in academic research and analysis methods, Korea Institute of Enterprise Architecture, Vol. 13, No. 2, pp. 317-329, 2016.
[23] L. P. Gurdiel, J. M. Mediano, and J. A. Quintero, A comparison study between coherence and perplexity for determining the number of topics in practitioners interviews analysis, IV Iberoamerican Conference of Young Researchers in Economy and Management, 2021.
[24] Sentence Transformer Model, HuggingFace Homepage, https://huggingface.co/jhgan/ko-sroberta-multitask, 2024
[25] Z. Wang, J. Chen, and H. Chen, Identifying interdisciplinary topics and their evolution based on BERTopic, Scientometrics, pp. 1-26, 2023.
[26] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, and D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382. 2023.

길완제(Wan-Je Gil)

2023년 국립공주대학교 컴퓨터공학과에서 공학박사를 취득하였다. 2020년부터 동국대학교 교육대학원 인공지능 융합전공 겸임교수로 근무하고 있다. 현재 (주)엠알티인터내셔널 대표이사로 재직 중이고 관심 분야는 인공지능 교육, 텍스트 분석, 지능형 로봇, 자연어 처리 등이다.

장환영(Hwan-Young Jang)

2008년 인디아나 대학교 블루밍턴에서 교육학 박사를 취득하였다. 2010년부터 동국대학교 교육학과에 교수로 재직하고 있다. 관심분야는 가치중심 HRD, 성과향상, 교육서비스 과학 등이다.

신인수(In-Soo Shin)

2009년 플로리다 주립대학교에서 교육통계 박사를 취득하였다. 2019년부터 동국대학교 교육대학원 인공지능 융합 전공 주임교수로 재직하고 있다. 관심분야는 메타분석, 교육통계, 인공지능 교육, 텍스트 마이닝 등이다.