



텍스트 마이닝 기반 토픽 모델링을 활용한 AI리터러시 프로젝트 학습자 행위 분석: 대학교육에서 IT비전공자를 대상으로*

Analysis of Learner Behavior in AI Literacy Projects Using Text Mining-Based Topic Modeling: Targeting Non-IT Majors in Higher Education

차현진[†] · 김연성^{††}

Hyunjin Cha[†] · Yeonseong Kim^{††}

요약

본 연구는 IT비전공계열 대학생을 대상으로 개설된 AI리터러시 교과목에서 학습자 행위를 분석하기 위해 텍스트 마이닝 기법을 적용하여 기말 프로젝트의 정성적 결과물을 분석하고자 한다. AI리터러시 기말 프로젝트는 2023-2학기부터 2025-2학기까지 총 7개 분반 246명 학생을 대상으로 수합되었고, 텍스트 마이닝 기반 분석은 키워드 빈도 분석, 토픽 모델링, 네트워크 분석이 시행되었다. 분석한 결과 기말 프로젝트 보고서는 AI를 활용한 문제해결 과정과 머신러닝 기법, 이를 활용한 데이터의 분석과 해석이 이루어진 것을 파악할 수 있었다. 또한 BERTopic 기반 토픽모델링을 통해 학습자들이 관심을 가지고 프로젝트를 수행한 7개의 주제와 관련 머신러닝 기법 및 해석에 관련된 동향을 파악할 수 있었다. 본 결과는 프로젝트 기반 학습에서 학습자가 수행한 정성적 보고서를 기반으로 학습자의 행위를 분석할 수 있으며, 프로젝트 수행 과정에 대한 동향을 파악하고, AI리터러시 교육에 대한 시사점을 도출할 수 있다는 것을 시사한다.

주제어 AI리터러시, 학습자 행위 분석, 토픽 모델링, 텍스트 마이닝, 네트워크 분석, 프로젝트 기반 학습

ABSTRACT

This study aims to analyze learners' behaviors in an AI literacy course offered to non-IT major university students by applying text mining techniques to the qualitative outcomes of their final projects. The final project reports were collected from a total of 246 students in seven course sections offered between the second semester of 2023 and the second semester of 2025. The text mining analysis consisted of keyword frequency analysis, topic modeling, and network analysis. The results showed that the final project reports mainly focused on AI-based problem-solving processes, machine learning techniques, and the analysis and interpretation of data using these techniques. In addition, BERTopic-based topic modeling identified seven major topics that reflected the themes of the projects in which students showed interest, as well as trends related to machine learning methods and their interpretation. These findings suggest that qualitative project reports can be used to analyze learners' behaviors in project-based learning, identify trends in the project implementation process, and provide implications for AI literacy education.

Keywords AI literacy, Student's behavior analysis, Topic Modeling, Text Mining, Network Analysis, Project-based learning

[†]중신회원 순천향대학교 기초공통교양학부 부교수
(교신저자)

^{††}정회원 순천향대학교 컴퓨터공학과 학부생

논문투고 2026년 06월 15일

심사완료 2026년 07월 14일

게재확정 2026년 07월 31일

발행일자 2026년 09월 11일

* 본 논문은 2024년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구지원사업의 지원을 받아 수행된 연구임(NRF-2024S1A5A2A01029079). 또한, 본 연구는 순천향대학교 학술연구비 지원으로 수행하였음

1. 서론

인공지능(AI, Artificial intelligence) 리터러시는 앞으로 AI시대를 살아가는데 모든 사람들에게 요구되는 역량으로 특정 IT전공자 뿐 아니라, 전 인류가 갖추어야 할 기본 역량으로 규명되고 있다[1]. AI 기술은 다양한 분야에 융합하여 활용되고 있으며, 교육 분야에서도 AI를 접목한 변화와 혁신을 요구하고 있다[2]. 특히, AI리터러시와 AI를 융합한 문제해결 등을 통해 미래 인재들이 AI와 협업하여 지속 가능한 사회를 이룰 수 있도록 AX 교육이 필요함을 논의하고 있다[3].

AI리터러시에 대한 개념을 살펴보면 단순히 AI에 대한 이해를 의미하는 차원에서 나아가 AI를 활용하여 윤리적으로 AI시스템을 설계하고 개발하거나, AI를 적용한 문제해결을 수행 할 수 있는 역량을 의미한다[4]. AI리터러시 교육은 학교급과 대상에 따라 초점을 다르게 가져가고 있다[5]. 고등교육에서의 선행 연구에 대한 메타 분석을 기반으로 대학생들을 위한 AI리터러시 교육은 데이터 리터러시, 문제해결능력, 전공 및 경력에 연계한 역량으로 확장되고 있음이 분석되었다[5]. 즉, IT비전공 계열 학생의 경우 자신의 전공과 연계하여 문제를 해결하고 데이터 분석을 통해 부가가치를 창출할 수 있는 교육 방법이 필요하다[2, 6].

문제해결 역량을 함양하는데 프로젝트 기반 학습이 주로 사용되고 있으며, 실제 AI리터러시 교육에서도 프로젝트 기반 학습을 접목하여 다양한 학습의 효과를 논의하는 연구가 늘어나고 있다.

본 연구에서는 대학 교육에서 IT비전공 계열 학생을 대상으로 백워드 설계를 기반으로 전공 및 경력과 연계한 실질적인 문제 해결 활동을 통해 AI리터러시를 함양하는 교과목을 설계·개설하였다[6]. 해당 AI리터러시 교과목은 AI에 대한 이해를 기반으로 AI를 적용해서 문제를 해결할 수 있는 다양한 활동 및 프로젝트, AI 윤리 등으로 구성하였다. AI리터러시 교과목은 2023-2학기부터 2025-1학기까지 총 4개 학기를 운영하면서 총 246개의 기말 프로젝트 보고서를 수합하였다.

기말 프로젝트는 문제 정의, 데이터 수합 및 전처리, 머신러닝 적용 AI모델 생성, AI모델 성능 평가, 데이터 시각화를 통한 해석 및 시사점 도출, 성찰로 구성하였고 6주간 미니 프로젝트를 통해 연습하였다. 기말 프로젝트 수행 과정에 대한 분석을 통해 학습자의 행위 및 수업에 대한 시사점을 도출하기 위해 텍스트 마이닝을 수행하였다. 우선 전체 보고서에 대한 전처리를 수행하고, 키워드 빈도 분석을 통해 학습자들이 프로젝트에서 수행한 키워드를 분석하였다. 또한 BERTopic 기반 토픽 모델링을 수행하여, 수업 설계에 따른 프로젝트에서 학생들이 어떤 주제를 연계하여 문제해결을 수행하였고 프로젝트 과정에서 활용한 머신러닝 기법과 데이터 해석에 대한 동향을 분석하였다. 마지막으로 네트워크 분석을 통해 키워드 연계성을 확인하였다.

본 연구는 다양한 텍스트 마이닝 기법을 적용하여 AI리

터러시 기말 프로젝트의 정성적 결과물을 분석하고자 한다. 텍스트 마이닝 기반 분석은 학습자의 행위와 프로젝트 수행 과정에 대한 동향을 파악하고, AI리터러시 교육에 대한 시사점을 도출하는 것을 목적으로 한다.

2. 선행연구 분석

2.1 AI리터러시 교육과 프로젝트 기반 학습

AI리터러시의 중요성이 확대되면서 학교급에 따라 학습자의 수준과 목표에 따라 AI리터러시 교육 체계를 개발하고 AI리터러시 교육 프로그램을 확대하고 있다[5].

대학 교육 맥락에서 AI 리터러시에 대한 선행연구를 살펴보면, AI리터러시를 다음 Table 1 과 같이 정의하고 있다.

Table 1. Previous studies about AI literacy definition'

Reference	Definition
Ng et al. (2021)[7]	the ability to understand and utilize AI in various contexts, and to monitor and reflectively think about AI application tools, as well as to apply and evaluate them.
Southworth et al.(2023) [8]	the ability to understand, utilize, evaluate, and ethically harmonize AI.
Yuan, Tasi, & Chen (2024)[9]	the ability to effectively and ethically utilize products incorporating AI technology, encompassing K-12 and higher education, as well as teacher and general education.
Chee et al. (2024)[5]	a holistic competency to utilize AI across all aspects of life, critically evaluate AI technology, and interact effectively with AI

특히, AI리터러시 교육이 강조되면서 AI리터러시에 대한 개념을 기반으로 학교급별 갖추어야 할 역량 교육 틀을 개발해왔다. Chee et al.(2024)[5]는 AI리터러시 역량 요소 도출에 대한 선행연구를 메타 분석함으로써 대학 교육에서 강조하고 있는 필수적인 역량 요소를 도출하였다. 고등교육에서는 데이터와 알고리즘 리터러시, 문제해결 능력, 직업 연계 능력, AI 윤리 등이 강조되는 것으로 분석하였다.

최근 많은 대학에서는 IT비전공 계열 학생들을 대상으로 AI리터러시 교육을 강화하고 있다[10-12]. 앞서 고등교육 기관에서 강조하는 기본 AI리터러시 교육은 과거 SW 코딩 교육에서 나아가 AI와 함께 살아가는 시대에 예측하기 어려운 윤리적 이슈와 함께 자신의 전공 분야 및 직업과 연계하여 AI를 활용한 문제해결이 강조되고 있다[3]. 본 연구에서는 Chee et al.(2024)[5]이 제시한 고등교육 차원에서의 AI리터러시 역량 프레임워크를 기반으로 데이터 리터러시를 기반으로 직업 연계 문제해결을 수행할 수 있는 프로젝트 기반 학습(Project-based learning)을 IT 비전공 계열 대학생들을 대상으로 설계하였다[6].

Li et al. (2024)[13]는 AI리터러시 교육에서 PBL은 매우 활동적이며, 학생 중심의 교수 방법으로 학생들에게 자율성과 구성주의적 결과를 도출할 수 있는 기회를 제공한다고 강

조한다. 선행연구[7, 14]에서도 PBL의 과정에서 학생들은 실질적인 현실세계 문제를 기반으로 융합적 정보를 활용하여 문제해결을 수행하는 과정으로 인지적, 메타인지적(전략적), 정서적 효과를 가질 수 있음을 보여주고 있다. 해외에서도 대학 교육에서 단순히 기술적 측면을 배우는 수준에서 나아가 지속 가능한 미래를 위해 AI활용 프로젝트 기반 문제해결 방법을 제안한다[15, 16]

특히, 대학 교양에서는 다양한 전공의 학생들이 다양한 목적과 진로를 가지고 AI를 활용한 문제해결을 수행하는 만큼 동일한 목적과 데이터를 활용하는 것도 자신이 관심을 가지고 있는 목적과 데이터를 활용해 문제해결을 수행하는 PBL의 과정은 직업 세계에서의 AI리터러시를 강조한 Chee et al.(2024)[5]의 관점에서 매우 효과적인 접근이 될 수 있다.

하지만, PBL이 가지는 인지적 및 정서적 이점에도 불구하고 학생들이 수행한 프로젝트 결과물을 통해 AI리터러시 역량 함양의 효과성을 분석하는 것은 쉽지 않다. 본 연구에서는 이러한 관점에서 학생들이 수행한 프로젝트 결과물을 기반으로 어떠한 주제와 AI기법을 활용하여 어떤 결과를 도출하였는지 과정을 통해 효과성을 평가하는 방법으로 텍스트 마이닝 기법을 활용하고자 한다.

2.2 텍스트 마이닝 및 토픽 모델링

PBL의 수행 결과는 PBL을 통해 함양하고자 하는 역량과 창의적 문제해결 역량, 협업 역량을 별도의 양적 평가 도구를 활용하여 평가한다[17]. 텍스트 마이닝 기법은 이러한 비정형 텍스트 데이터로 구성된 결과물을 기반으로 의미 있는 주제와 구조를 도출하는 방법으로 최근 다양한 연구 분야에서 동향을 파악하거나 담론으로 구성된 내용을 중심으로 유의미한 통찰을 얻어내는데 활용되고 있다.

본 연구에서도 AI리터러시 교육에서 수행된 프로젝트 보고서를 통해 학생들이 수업 설계에서 고려한 수업 목표, 내용, 성취의 과정을 통해 AI리터러시를 함양했는지를 분석하기 위해 텍스트 마이닝 기반 토픽 모델링을 수행하고자 한다.

토픽 모델링은 문서 집합으로부터 유의미한 통찰을 도출하고 잠재적 패턴과 구조를 추출하는 텍스트마이닝 방법이다. 토픽 모델링을 수행하는데 가장 많이 활용하는 방법으로 오픈소스 기반 정량적 평가법을 적용할 수 있는 잠재적 디리클레 할당(LDA, Latent Dirichlet Allocation) 방법이 대표적이다[18]. LDA는 문서 집합에서 단어를 통해 어떤 잠재적인 토픽들이 생성되는지를 생성적 확률 모델로 도출하는 방법으로 간결하면서 핵심적인 주제 단위를 도출하고 대규모 텍스트 문서 집합 처리에 적합한 것으로 논의되고 있다. 하지만, LDA는 단어의 순서 또는 의미적 관계를 포착하는데 한계가 논의되었고(길완재 외, 2024), 최근 BERTopic 기반의 토픽 모델링 방식이 비정형 텍스트 데이터의 해석을 높여주고 의미적인 구조를 탐구할 때 유의미한 방법으로 논의되고 있다[19].

학생들이 수행한 프로젝트의 비구조화된 특성과 구성주의적 과정 평가와 함께 AI리터러시의 복합적 개념을 이해하기 위해 학생들이 수행한 보고서를 기반으로 군집화하는데 텍스트 마이닝 기반 토픽 모델링은 유의미한 통찰을 제공할 수 있다[20].

3. 연구 방법

3.1 AI리터러시 교육 프로그램 및 PBL 적용

본 연구에서는 백워드 모형을 기반으로 설계된 AI리터러시 교육 프로그램을 2023-2학기부터 2025-1학기까지 총 4개 학기 7개 분반의 학생을 대상으로 적용하였다(저자, 2025). 교육 프로그램은 S대학교 중핵 교양 교과목으로 운영되어 2, 3, 4학년 학생들 중 인문·사회·자연 계열 학생들을 대상으로 AI리터러시 수업으로 진행되었다.

Table 2. Problem-solving activities with AI in the AI literacy class

W	Area	Data	Machine Learning
5	Education	Titanic Survival	Supervised learning (Classification)
		Academic Warning	
		Math Grade	Hierarchical Clustering
6	Culture	Steve Jobs Speech	Text mining- Word clouds, sentiment analysis, statistics
		pop song lyrics	
		Guardians articles	
7	Economy	Customer churning	Supervised learning (Classification)
		Kangnam House prices	Supervised learning (linear regression)
		Iowa House prices	
9	Environment	Garbage	- Supervised learning (Classification)
		Birds	- Hierarchical Clustering
10	Health	Stroke	- Supervised learning (Classification) PCA
		BMI	- Supervised learning (linear regression)PCA
11	Business	Mall customer	K-means clustering
		Car renting service	

AI리터러시 교육 프로그램은 Chee et al.(2024)이 제시한 개념을 기반으로 AI에 대한 이해, AI도구를 활용한 다양한 활동, 노코드 툴(Orange3)를 활용한 데이터 기반 AI 적용(교육, 경제, 비즈니스, 문화, 환경, 건강 영역의 데이터를 기반으로 머신러닝을 적용한 문제해결), AI 윤리를 배우고 마지막으로 AI문제해결 프로젝트를 수행하는 15주차 수업으로 구성되었다(저자, 2025). 특히 5주차부터는 학생들이 기말 프로젝트를 수행하기 위해 필요한 각 영역별 데이터를

활용해서 지도학습에서 분류 및 회귀에 해당하는 문제해결, 텍스트 마이닝과 이미지 분석을 활용한 지도학습과 거리 기반 비지도학습(계층적 그룹화), 건강에서는 지도학습의 분류와 회귀 수행 후 성능을 높이기 위한 주성분 분석을 통한 차원 축소(PCA), 비지도학습을 통해 고객 그룹화(K-means Clustering) 등을 수행하여 반복적으로 AI 활용 문제해결 연습을 수행하였다.

특히, 기말 프로젝트를 수행하기 전에 기말 프로젝트에서 수행해야 하는 각 단계에 대한 연습을 교육, 경제, 비즈니스, 문화, 환경, 건강의 영역의 데이터를 기반으로 미니 프로젝트를 시행하였다. 미니 프로젝트에서는 교수자가 선정한 주제를 기반으로 문제의 필요성 및 목적, 데이터수합 및 전처리 과정을 설명하고, 수업 시간에 Orange3를 활용해 머신러닝을 적용한 AI모델 개발, 다양한 머신러닝 기법에 따른 성능 분석, 최적의 모델 선택, 데이터 시각화 위젯을 활용한 해석 및 시사점 도출 등의 과정을 반복적으로 시행하였다.



Figure 1. Steps in the Problem-solving Project with AI(Developed by MiriCavas AI)

학생들은 6주간에 걸쳐 수행된 미니 프로젝트에서의 다양한 지도학습 및 비지도학습 방법 중 하나의 방법을 기반으로 유사하게 기말 프로젝트를 수행할 수 있도록 가이드하였다.

기말 프로젝트는 한글 또는 워드 형식으로 수합되었고, 한 학기 동안 배운 내용을 종합하여 자신만의 문제를 정의하고, 관련 데이터를 수합 및 전처리를 수행, 문제해결을 위한 지도학습 또는 비지도학습을 통한 AI모델 개발, AI모델의 성능 분석, 관련 데이터를 시각화함으로써 시사점 도출 및 해석을 수행하도록 하였다.

3.2 텍스트마이닝 분석 방법

수합된 기말 프로젝트 보고서는 텍스트 마이닝 분석을 위

해 전처리 과정에서 불용어, 표현 통일(주요 키워드에서 한글 또는 영어 단어를 통일, 단수 또는 복수 통일 등), 머신러닝 기법 등 하나의 통일된 용어로 처리(RandomForest, MachineLearning, LogisticRegression, ConfusionMatrix, WordCloud 등)하는 과정을 거쳐 문서별 키워드 빈도 분석과 토픽 모델링을 수행하였다.

토픽 모델링은 보고서의 특성상 데이터 수집, 전처리, 머신러닝 모델 적용, 예측 결과 해석, 시각화, 시사점 서술 등 다양한 내용을 포함하고 있기 때문에 단순 빈도 분석만으로는 전체적인 구조를 파악하기 어렵다. 이러한 관점에서 비정형 텍스트 데이터의 해석을 높여주고 의미적인 구조를 탐구할 때 유의미한 방법으로 논의되고 있는 BERTopic 기법을 활용하였다. BERTopic 기반 토픽 모델링을 수행하기 위해 임베딩용 전처리를 수행하였다. 임베딩용 전처리는 문장 구조를 최대한 보존하면서 이상 기호와 반복 문장부호를 정리하는 방식으로 수행되었다.

BERTopic 분석은 파이썬을 활용해서 수행되었고, 토픽 표현용 추가 전처리를 통해 불용어와 숫자-only 토큰을 제거하였다. 문서 임베딩에는 네 가지 SentenceTransformer 모델을 사용하였으며, UMAP을 이용해 차원을 축소하고, HDBSCAN을 통해 문서를 군집화하였다. 또한 Default, KeyBERT, MMR 표현 방식을 함께 적용하여 토픽 키워드의 대표성과 다양성을 비교하였다. 임베딩 모델 비교에서는 paraphrase-multilingual-MiniLM-L12-v2, distiluse-base-multilingual-cased-v2, xlm-roberta-base-nli-stsb-mean-tokens, snunlp/KR-SBERT-V40K-klueNLI-augSTS 등 네가지 임베딩 모델을 비교하였다. 이는 학생 보고서의 특성상 데이터 분석 도구, 시사점 도출, 평가지표 및 머신러닝 용어 활용, 시각화 기법 등 전문 용어와 보고서용 용어, 영문과 한국어가 혼용되어 있는 점을 고려하여 네가지 모델의 결과를 비교하기 위함이다. 본 연구에서는 네 가지 임베딩 모델의 결과를 비교하면서 반복적으로 나타나는 주제 구조와 수업 설계에 따른 토픽별 해석이 가능한 snunlp/KR-SBERT-V40K-klueNLI-augSTS 임베딩 모델을 연구자간 논의를 통해 확정하였다.

또한, UMAP 차원 축소 후 HDBSCAN 클러스터링 과정에서 min_cluster_size를 10으로 설정하여 최소 10개 이상의 학생 보고서 문서가 모여야 하나의 토픽 군집으로 인정할 수 있도록 설정하였다. BERTopic에서는 같은 군집에 속한 문서들을 하나의 큰 문서처럼 결합한 뒤, 해당 토픽에서 특징적으로 많이 나타나는 단어를 계산한다. 이때 사용되는 핵심 방식이 c-TF-IDF이다.

Table 3. A procedure of the BERTopic-based topic modeling¹

Step	Phase / Task	Technical Description
1	Data Collection	Acquisition of raw project report text data.
2	Text Cleaning	Removal of report templates, headers, footers, and redundant symbols.

Step	Phase / Task	Technical Description
3	Terminology Standardization	Normalization and unification of machine learning and data analytics terms.
4	Data Persistence (Preprocessing)	Saving the preprocessed documents optimized for embedding text generation.
5	Data Loading	Importing the preprocessed document dataset into the BERTopic pipeline.
6	Topic Representation Preprocessing	Additional text refinement tailored for final topic representation extraction.
7	Vector Embedding Generation	Generating document vector embeddings across different embedding models.
8	Dimensionality Reduction	Applying UMAP (Uniform Manifold Approximation and Projection) for feature reduction.
9	Density-Based Clustering	Conducting document clustering utilizing HDBSCAN.
10	Topic Keyword Extraction	Extracting distinctive topic keywords using the BERTopic framework.
11	Representation Comparison	Comparative analysis of topic representations using Default, KeyBERT, and MMR methods.
12	Result Export & Network Mapping	Exporting document-topic assignments, topic metadata, and co-occurrence network data.

본 연구에서는 각 토픽을 하나의 키워드 목록으로만 해석하지 않고, 세 가지 표현 모델(Default Representation, KeyBERTInspired, MMR)을 비교하였다. 세 가지 표현모델에서 Default Representation은 기본 c-TF-IDF의 통계적 대표성, KeyBERTInspired는 의미적 적합성, MMR은 키워드 다양성을 중심으로 제시한다. 본 연구는 세 가지의 표현 모델을 비교하여 키워드 간 상대적 순위와 해석 가능성을 중심으로 토픽모델링을 도출하였다.

또한, 토픽별 도출된 키워드 간 관계를 분석하여 학생들의 프로젝트 결과물에 대한 의미있는 해석을 이끌어내기 위해 네트워크 분석을 실시하였다. 네트워크 분석에서는 토픽별 키워드 간 관계를 시각적으로 확인하기 위해 공출현 네트워크도 생성하였다. 각 토픽에 대해 상위 키워드 50개를 추출하고, 해당 토픽에 속한 문서들 안에서 키워드들이 함께 등장하는 빈도를 계산하였다.

Table 3은 본 보고서에서 BERTopic 기반 토픽모델링 및 토픽별 네트워크 분석을 수행한 절차이다.

3.3 연구 참여자

본 연구에 참여한 연구 참여자는 2023-2학기 1개 분반, 2024-1학기부터 2025-1학기 2개 분반씩 개설하여 운영된 AI리터러시 교양 교과목 총 7개 분반으로 학생들은 자발적으로 본 교과를 선택하였다. 마지막 기말 프로젝트는 본 수업의 개선을 위한 연구로 활용될 수 있음에 동의를 받았고, 동의한 학생들의 데이터를 중심으로 분석하였다. 본 연구에 참여한 학생은 총 246명의 기말 프로젝트 보고서를 분석하였다. 학기별 참여자는 다음 Table 2와 같다.

Table 4. Number of Participants in each ter

Term	Num of S	Num of Class
2023-2	35	1
2024-1	72	2
2024-2	64	2
2025-1	75	2

총 7개 학기에 참여한 학생들을 단과대별로 분석한 결과, 인문사회 계열(32.5%)로 가장 많은 비율을 차지하였고, 다음으로 글로벌경영대학(19.5%), 자연과학대학(17.5%), 비IT계열 공학(16.7%), 보건계열(13.8%)로 구성된 것으로 분석되었다.

Table 5. Num of students by college

College	Num	%
Humanities & Social Sciences	80	32.5
Business	48	19.5
Natural Sciences	43	17.5
Engineering	41	16.7
Medical	34	13.8
Total	246	100

4. 연구 결과

4.1 기말 프로젝트 보고서 키워드 빈도 분석

본 연구에서는 학생들이 수행한 기말 프로젝트의 주요 키워드를 분석하기 위해 키워드 빈도 분석을 우선 시행하였다. 키워드 빈도 분석은 상위 105개를 중심으로 도출하였다. <Table 6>는 그 중 상위 60개 키워드의 빈도수를 보여주고, Figure 2.는 상위 100개 키워드를 중심으로 도출한 워드 클라우드를 보여준다.

키워드 분석 결과는 학생들은 기말 프로젝트 보고서에서 데이터(1위, 1397)를 분석(4위, 669)해서 가장 높은(3위, 674) 성능(27위, 219)을 가진 AI(30위, 210) 예측(2위, 1116) 모델(5위, 604)을 확인(6위, 524) 것을 목적으로 한 것에 초점을 맞추고 있음을 보여준다. 학생들은 지도학습(12위, 302)에서 분류(20위, 249) 모델을 가장 많이 적용했으며 AUC(104위, 85) 값과 정확도(57위, 137), 혼동행렬(94위, 97) 비교를 주로 수행한 것으로 드러났다. 또한, 지도학습 방법으로는 naive byes(14위, 291), random forest(15위, 285), neural network(24위, 230), Adaboost(31위, 207), logistic regression(48위, 149) 순으로 가장 많이 적용한 것으로 나타났다. 비지도학습(42위, 157)에서는 클러스터링(87위, 101)과 PCA(82위, 108)에 대한 키워드 빈도수가 높은 것으로 나타났다.

Table 6. Keyword frequency in students' reports

	word	frequency		word	frequency
1	data	1397	31	AdaBoost	207
2	prediction	1116	32	diverse	206
3	high	674	33	age	200
4	analysis	669	34	basis	200
5	model	604	35	stress	195
6	confirmation	524	36	ratio	193
7	sleep	475	37	scatter plot	193
8	value	408	38	case	189
9	use	406	39	health	189
10	group	327	40	variable	182
11	image	304	41	diabetes	176
12	supervised learning	302	42	unsupervised learning	157
13	effect	298	43	exercise	156
14	naive bayes	291	44	time	155
15	random forest	285	45	file	154
16	result	277	46	low	152
17	kaggle	275	47	distribution	149
18	person	271	48	Logistic Regression	149
19	pre processing	262	49	photo	149
20	classification	249	50	level	147
21	customer	244	51	test	146
22	many	233	52	type	145
23	related	233	53	problem	144
24	neural network	230	54	patient	143
25	people	227	55	information	141
26	factor	222	56	relationship	138
27	performance	219	57	Accuracy	137
28	topic	219	58	obesity	134
29	feature	212	59	possibility	134
30	AI	210	60	disease	131

주제 단어는 건강 쪽 단어가 가장 많이 등장하는 것을 볼 수 있다. sleep(7위), stress(35위), health(39위), exercise(43위), diabetes(175위), BMI(71위), heart disease(84위), depression (85위) 등이 가장 눈에 띄었다.

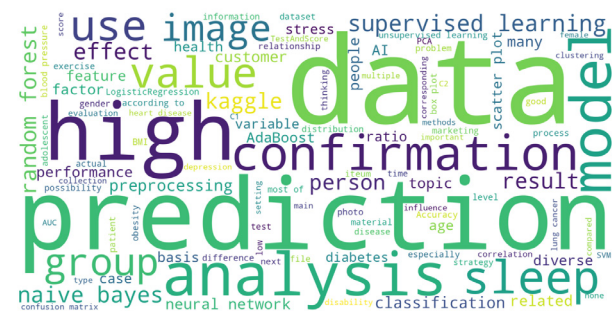


Figure 2. Wordclouds in project reports

4.2 BERTopic 기반 토픽 모델링 분석 결과

BERTopic 기반 토픽 모델링을 수행한 결과 7개의 토픽으로 분류되었다.

Table 7. Topics modeling results based BERTopic

Topic	Count	%	Main keywords top 7
0	59	24.0	data, model, neural network, prediction, Naive Bayes, diabetes, obesity
1	33	13.4	Sleep disorder, adolescence, data, model, impact, mental health, academic performance, depression
2	28	11.4	Customer, Data, consumption, marketing, product, strategy
3	24	9.80	data, similar, Random forest, Idol, Prediction, Topic, Word Clouds
4	22	8.90	data, type, Regional analysis, model, Clustering, Learning, Random Forest
5	15	6.10	Frequently appearing, data, new article, preprocess, prediction, score, favorite
6	11	4.5	play, data, snow, positive, contents, popular

학생들이 마지막 기말 프로젝트에 제출한 성찰 보고서를 분석한 결과는 토픽은 총 7개로 나뉘어 지는 것으로 분석되었다.

4.2.1 토픽 1: 질병 예측 및 건강 위험 요인 분석

BERTopic 기반 토픽 모델링에서 토픽 1은 AI 활용 문제 해결 프로젝트에서 질병예측 및 건강 위험 요인 관련 분석에 관한 토픽들로 구성된 것을 확인할 수 있었다. 학생들은 자연과학(17.5%)과 보건 계열(13.8%) 학생들이 전체 약 31%를 차지하고 있지만, 인문사회(32.5%) 및 글로벌 경영대학(19.5%) 학생들이 더 많은 분포를 차지하는 것에 비해 전반적으로 질병 및 건강에 관심을 가지고 문제해결을 수행한 학생들이 많다는 것을 알 수 있었다.

질병 예측 및 건강 위험 요인에 대한 분석은 주로 지도학습을 통해 수행되었으며 Neural Network, Naive Bayes 방법을 제일 많이 언급했는데, 이는 앞서 전체 키워드 분석에서 다른 지도학습 방법도 많이 언급된 것에 비해 질병 예측 및 건강을 주제로 가중치가 높다는 것은 성능에서 가장 높은 모델로 추측할 수 있다.

AI리터러시 건강 주간에서는 지도학습을 중심으로 분류와 회귀 분석을 수행하고 건강 데이터의 예측률을 높이는 것이 중요하기 때문에 주성분 분석을 통한 차원 축소 방법인 PCA를 활용한 성능을 높이는 연습을 수행하였는데 실제 PCA의 가중치가 27위로 나오면서 많은 학생들이 PCA를 함께 수행한 것으로 분석되었다.

Table 8. Keywords and weight in Topic 1

	topic	Weight		topic	Weight
1	Data	0.0368	16	Drug	0.0103
2	Model	0.0308	17	Lung Cancer	0.0101

	topic	Weight		topic	Weight
3	Neural Network	0.0261	18	Scatter Plot	0.0101
4	Prediction	0.0222	19	Heart Disease	0.0099
5	Naive Bayes	0.0172	20	Accuracy	0.0094
6	Diabetes	0.0171	21	Exists	0.0092
7	Obesity	0.0158	22	Clustering	0.009
8	Impact	0.0157	23	Predicting	0.0088
9	Random Forest	0.0153	24	Patient	0.0088
10	Value	0.0133	25	AUC Value	0.0086
11	BMI	0.013	26	Blood Pressure	0.0084
12	AdaBoost	0.0127	27	PCA	0.0083
13	Health	0.0125	28	Level	0.0082
14	Because	0.0114	29	Risk of Developing	0.0082
15	Cholesterol	0.0104	30	Index	0.0081

성능 분석에서 AUC 값과 Accuracy 비교를 수행하였고 데이터 시각화에서는 Scatter plot을 가장 많이 활용한 것으로 해석된다.

4.2.2 토픽2: 건강, 심리 생활 습관 분석

학생들이 수행한 기말 프로젝트의 2번째 토픽은 주로 건강, 심리, 생활 습관과 관련된 주제로 분석이 수행된 것을 확인할 수 있었다.

두 번째 토픽을 살펴보면, 주요 건강 심리 대상이 청소년들이라는 것을 알 수 있다. 토픽 2에서는 수면 장애에 대한 분석이 제일 많이 이루어졌으며, 성적, 우울증, 스트레스, 디지털 기술과 같은 우리가 청소년의 건강 심리와 관련된 단어들이 상관(correlation)과 함께 등장하는 것으로 분석되었다.

이 주제에서도 naive bayes 머신러닝 기법이 가장 많이 언급이 되었고 scatter plot을 시각화한 해석이 이루어진 것을 알 수 있다. 특히 예방이나 제안 등을 통해 분석 결과를 바탕으로 다양한 예방 전략이 제안이 이루어졌음을 알 수 있다.

Table 9. Keywords and weight in Topic 2

	topic	Weight		topic	Weight
1	Sleep Disorder	0.0465	16	Stress	0.0094
2	Adolescents	0.0430	17	Random Forest	0.0091
3	Data	0.0339	18	Digital Technology	0.0091
4	Model	0.0242	19	Depressive Disorder	0.0091
5	Impact	0.0187	20	Affecting	0.009
6	Mental Health	0.0163	21	Naive Bayes	0.0089
7	Academic Performance	0.0147	22	Supervised Learning	0.0087
8	Substance Use	0.0106	23	Correlation	0.0087
9	Depression	0.0104	24	Obstructive Sleep Apnea	0.0086
10	Topic	0.0101	25	Index	0.0081
11	Alcohol	0.01	26	Scatter Plot	0.0083

	topic	Weight		topic	Weight
12	Insomnia	0.0098	27	Prediction Model	0.0082
13	Sleep Duration	0.0097	28	Suggests	0.008
14	Crime	0.0096	29	Blood Pressure	0.0078
15	Sleep Apnea	0.0094	30	Prevention	0.0077

4.2.3 토픽3: 고객, 마케팅, 소비자 행위 분석

학생들이 수행한 기말 프로젝트의 3번째 토픽은 고객 데이터를 중심으로 마케팅 전략이나 소비자 행동 분석이 주를 이루고 있음을 확인할 수 있었다.

Table 10. Keywords and weight in Topic 3

	topic	Weight		topic	Weight
1	Customer	0.0583	16	Marketing Strategy	0.0107
2	Data	0.0339	17	Model	0.0106
3	Purchase	0.0244	18	Advertising	0.0104
4	Consumption	0.0223	19	Strategy	0.0102
5	Marketing	0.0202	20	Consumer	0.0101
6	Product	0.0201	21	Romance	0.0097
7	Strategy	0.0198	22	C3 Cluster	0.0096
8	Kaggle	0.0182	23	Drama	0.0095
9	Online	0.0148	24	Which	0.0094
10	Shopping	0.0137	25	Preference	0.0091
11	Product Item	0.0136	26	Influencer	0.0091
12	Topic	0.0117	27	Preference	0.0083
13	Promotion	0.0114	28	Group	0.0082
14	Box Plot	0.0112	29	Genre	0.0081
15	Pattern	0.0108	30	Loan	0.0081

토픽 3은 비지도학습을 중심으로 고객, 소비자, 온라인 쇼핑 데이터 및 소비자 행동 패턴(pattern)을 분석하고, 그에 적절한 마케팅 전략과 프로모션 제품 등을 제안하는 주제로 구성되었다는 것을 알 수 있다.

고객 데이터 분석은 11주차 비즈니스 영역에서 k-means 클러스터링을 중심으로 수행하였고, 데이터 시각화를 통해 고객의 그룹을 이해하고 적절한 전략을 도출하는 미니 프로젝트를 수행한 절차를 학생들이 잘 따르고 있음을 보여준다. 우선 가장 많이 사용한 데이터 시각화 방법은 Box plot으로 나타났다.

비지도 학습에서는 특별히 머신러닝 방법론을 언급하지는 않았지만 Cluster라는 용어를 사용한 것으로 비지도 학습임을 알 수 있었고, 캐글이 많이 언급된 것으로 데이터를 수집한 장소가 캐글이 주된 소스임을 알 수 있다.

4.2.4 토픽 4: 문화 콘텐츠, 장르, 감성 분석

토픽 4에서는 학생들이 텍스트 분석에서 아이돌 그룹의 노래 가사를 중심으로 주제가 구성되어 있음을 알 수 있다.

수업 시간에 텍스트 분석은 국내가요 50곡의 가사를 분석한 적이 있는데 학생들은 유사하게 아이돌 그룹의 가사를 분석하고 워드 클라우드를 통해 경향성을 살펴보고, 장르의 등급(Rating) 및 평점(average)과 사랑(love)이라는 특정 단어 통계, 순위 분석 등에 대한 논의가 이루어진 것으로 볼 수 있다.

Table 11. Keywords and weight in Topic 4

	topic	Weight		topic	Weight
1	Data	0.0404	16	Lyrics	0.0106
2	Similar	0.0191	17	Group / Cluster	0.0193
3	Random Forest	0.0179	18	Product	0.0095
4	Idol	0.0157	19	Marketing	0.0091
5	Customer	0.0157	20	AdaBoost	0.0091
6	Prediction	0.014	21	Excel	0.0088
7	Topic	0.0136	22	Rating	0.0163
8	Word Cloud	0.0134	23	Scatter Plot	0.0086
9	Test	0.0218	24	Mean / Average	0.0162
10	Wine	0.0127	25	Number	0.0082
11	Fruit	0.0125	26	Naive Bayes	0.0082
12	SVM	0.0124	27	Supervised Learning	0.008
13	AUC Score	0.0119	28	Understanding	0.0077
14	Neural Network	0.0207	29	Image	0.015
15	Orange	0.0116	30	love	0.0075

4.2.5 토픽 5: 회귀분석 수행

토픽 5에서 등장하는 주제는 주로 지도학습의 회귀 분석에 해당하는 용어들이 많이 나타났다. 우선 회귀 분석을 위해 미니 프로젝트에서는 강남 집값과 아이오와 집값을 다루었는데 일부 학생들은 역시나 집값(house price)이나 주식(stock price)을 주제로 다루고 있었다. 데이터에 대한 분석은 지역별 특성이나 인기도, 투자 등에 대한 언급이 있었다. 주로 머신러닝 기법으로는 logistic regression과 random forest가 언급되었다.

수업 시간에는 linear regression과 오차에 대한 설명이 시행되었었는데 회귀 분석을 학생들이 가장 이해하기 어려워했다. 본 주제에서도 R2이나 오차값에 대한 언급이 없고 오히려 지도학습의 분류 모델이나 클러스터링을 추가적으로 시도해 본 것을 알 수 있었다.

Table 12. Keywords and weight in Topic 5

	topic	Weight		topic	Weight
1	Data	0.0515	16	Stock Price	0.0106
2	Type	0.0219	17	Test	0.0104
3	Regional Analysis	0.0182	18	Average	0.01
4	Model	0.0163	19	Classification	0.0098
5	Clustering	0.0161	20	Box Plot	0.0096

	topic	Weight		topic	Weight
6	Group / Cluster	0.0153	21	House Price	0.0094
7	Learning	0.0139	22	Chungnam	0.0093
8	Real Estate	0.0136	23	Orange	0.0089
9	Random Forest	0.0129	24	Scatter Plot	0.0087
10	Search	0.0123	25	Price	0.0086
11	Logistic Regression	0.0115	26	Image	0.0086
12	Popularity	0.0115	27	Child	0.0078
13	Investment	0.011	28	Accuracy	0.0076
14	Service	0.0109	29	AI	0.0076
15	Strategy	0.0106	30	AdaBoost	0.0075

4.2.6 토픽 6: 예술 및 체육 뉴스 기사 분석

토픽 6에서 등장하는 주제는 뉴스 기사 중 학생들이 관심을 가지는 야구, 영화, 축구, 스포츠 등 뉴스 기사 분석이 주제를 알 수 있다. 6주차에 수행된 문화 관련 텍스트 분석에서는 노코드 머신러닝 툴인 오렌지 3에서 무료로 검색이 가능한 가디언스에 접속해서 특정 기간에 검색어를 중심으로 뉴스 기사를 검색해서 데이터셋 구성이 가능함을 알려주었다.

실제로 학생들에게 가디언스가 영국의 저널지로 다양한 국제 기사를 가져올 수 있고 검색하는 방법에 대한 설명과 함께 전처리 방법 등을 설명하였다. 오렌지 3가 전처리 과정에서 불필요한(unnecessary) 단어를 메모장(notepad)에 넣어 제거하도록 되어 있는데 이 과정도 수행된 것으로 나타났다.

실제 학생들은 관심 있는 국제 뉴스를 검색해서 전처리를 수행하고, 가장 자주 등장하는 단어 등을 살펴보고, 감정 분석(sentiment analysis)을 수행한 것을 알 수 있다. 특히 해외 및 최신이라는 단어도 뉴스 기사에서 주로 활용하는 단어로 나타났다.

Table 13. Keywords and weight in Topic 6

	topic	Weight		topic	Weight
1	Frequently Appearing	0.0723	13	Screen	0.0106
2	Data	0.0289	14	Performed	0.0102
3	News Article	0.0192	15	Player	0.0101
4	Preprocessing	0.0189	16	Sentiment Analysis	0.0096
5	Prediction	0.0187	17	Work	0.0096
6	Score	0.0157	18	Sport	0.0096
7	Favorite	0.0151	19	Football	0.0094
8	Baseball	0.0131	20	Kaggle	0.0093
9	Clustered	0.0131	21	Overseas	0.0092
10	Curious	0.0125	22	Recent	0.0091
11	Called	0.0122	23	Notepad	0.0091
12	Viewing	0.0112	24	Unnecessary	0.009

4.2.7 토픽 7: 영화, 드라마, 게임 등 분석

마지막 토픽 7에서 등장하는 주제도 학생들이 관심있는 분야의 콘텐츠를 텍스트 분석을 주제로 이루어져 있다는 것을 알 수 있다.

Table 14. Keywords and weight in Topic 7

	topic	Weight		topic	Weight
1	Genre	0.0532	12	Movie	0.0137
2	Play	0.0375	13	Warning	0.0137
3	Snow	0.0319	14	Text	0.0117
4	Data	0.0275	15	Review	0.0115
5	Positive	0.0248	16	Emotion	0.0103
6	Content	0.024	17	Drama	0.0188
7	Box Office Success	0.0216	18	Prediction	0.0102
8	Positive	0.0194	19	Game	0.0098
9	Negative	0.0177	20	Damage	0.0098
10	User	0.0148	21	Sentiment	0.0097
11	Evaluation	0.0145	22	Word Cloud	0.009

토픽 7의 공출현율과 토픽 중 가중치가 낮은 것을 고려했을 때, 일부 학생들에게 관심 있는 주제였음을 알 수 있다. 하지만, 별도의 토픽으로 분류된 것을 보면 많은 학생들이 영화, 드라마, 게임 등에 관심을 가지고 관련 장르 분석, 감성분석으로 긍정성과 부정성 분석, 사용자 리뷰, 워드 클라우드로 경향 분석 등을 수행한 것을 알 수 있다.

4.3 기말 프로젝트 키워드 네트워크 분석

기말 프로젝트를 전체에 대해 단어간 영향력과 연관성을 분석하기 위해 네트워크 분석을 실시하였다.

네트워크 분석 결과(Figure 4)를 살펴보면, 매개중심성(Degree centrality), 근접중심성(Closeness centrality), 연결중심성(Between centrality), 가중중심성(Weighted centrality)으로 구성된다. Degree Centrality 분석 결과 Data(0.8649)가 가장 높은 중심성을 나타냈으며, Prediction(0.5946), Model(0.4865), Learning(0.2703), Supervised Learning(0.2297) 순으로 나타났다. 이는 학생들의 프로젝트가 데이터를 기반으로 머신러닝 모델을 구축하고 예측을 수행하는 공통적인 문제해결 과정을 중심으로 구성되었음을 의미한다.

Betweenness Centrality에서는 Data(0.7268)가 가장 높은 값을 보여 다양한 키워드를 연결하는 핵심 매개 역할을 수행하는 것으로 나타났다. Prediction(0.2218), Model(0.2299), Classification(0.0104) 등이 상대적으로 높은 값을 나타내 머신러닝 모델 구축과 결과 해석 과정에서 여러 주제를 연결하는 중심 개념으로 기능하였다. Closeness Centrality 역시 Data(33.3991), Prediction(31.7069), Model(30.6629), Learning(28.0053) 순으로 높게 나타나 전체 네트워크에서 다른 키워드와 가장 가까운 위치에 존재하는 핵심 개념임을 확인하였다.

결국, 중심성 분석 결과를 종합하면 학생들의 프로젝트

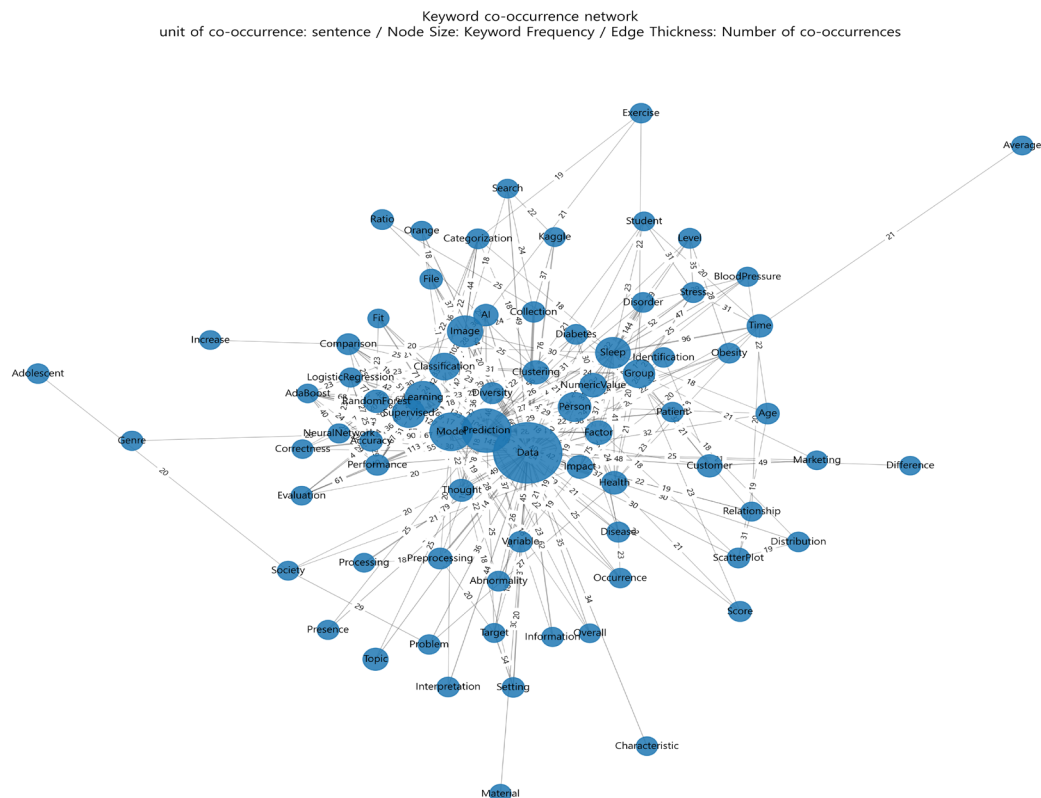


Figure 3. Network Analysis in students' project reports

Table 15. Keywords frequency in reflection reports

	word	f		word	f
1	class	1099	31	department	198
2	data	964	32	result	194
3	thought	846	33	progress	193
4	utilization	733	34	prediction	188
5	project	680	35	model	187
6	word	627	36	future	181
7	AI	610	37	first	175
8	analysis	579	38	attempt	175
9	technology	436	39	last	172
10	process	415	40	plan	170
11	help	388	41	mini	169
12	problem	385	42	program	163
13	literacy	742	43	daily life	159
14	interest	434	44	idea	158
15	performance	356	45	job	157
16	understanding	328	46	interpretation	152
17	related	325	47	directly	146
18	solution	297	48	problem solving	144
19	goal	293	49	career	142
20	major	284	50	material	139
21	field	267	51	housing	136
22	lecture	261	52	name	133
23	society	260	53	experience	128
24	learning	253	54	most	124
25	activity	246	55	application	123
26	item	245	56	person	122
27	Orange	242	57	processing	120
28	meaning	234	58	reason	119
29	ethics	231	59	actual	106
30	topic	211	60	difficulty	99

5. 논의 및 시사점

본 연구에서 비전공 IT계열의 대학생을 대상으로 AI기초 역량(AI이해, AI활용, AI윤리)에 나아가 실질적인 문제를 해결하는데 자신의 전공과 연계하여 AI 머신러닝 적용 및 데이터 해석을 통해 문제해결을 수행하는 프로젝트를 설계하였다. 이는 고등교육에서 AI리터러시 교육의 방향에 대한 메타 분석을 수행한 결과에서 데이터 리터러시, 문제해결, 경력 관련 역량을 강조한 것을 기반으로 설계한 것이다[5].

본 연구에서 모든 학생은 동일한 프로젝트 절차를 기반으로 과제를 수행하였다. 따라서 문제 정의, 데이터 전처리, 머신러닝 적용, 성능평가 등의 요소가 보고서에 공통적으로 나타나는 것은 과제 설계의 영향을 반영한다. 그럼에도 불구하고 BERTopic 기반 토픽 모델링의 결과는 수업 설계에 따른 프로젝트에서 학생들이 어떤 주제를 연계하여 문제해결을 수행하였고, 프로젝트 과정에서 활용한 머신러

닝 기법과 데이터 해석에 대한 동향을 분석하는데 중요한 방법이 될 수 있음을 보여주었다. 특히, BERTopic 분석결과 수업 설계 방향이 동일했음에도 불구하고, 동일한 프로젝트 구조 안에서 학생들이 어떠한 실제 문제를 선택하고 어떤 데이터와 머신러닝 기법을 활용하여 문제를 해결했는지에 대한 학습자의 관심과 적용 양상을 탐색하는데 시사점을 제시하였다.

다양한 토픽은 학생들의 전공과 경력에 연계된 것이면서 머신러닝 기법과 데이터 시각화 툴을 통한 데이터의 이해, 고차원적인 융합을 통한 문제해결 과정이 각 토픽별 키워드에 포함된 것으로 분석할 수 있었다. 다섯 번째 토픽의 주제인 회귀 분석에서는 수업에서 의도한 방향으로 토픽 키워드가 드러나지 않는 것을 통해 회귀 분석에 대한 이해가 좀 더 필요하고, 학생들이 회귀 분석에 대한 문제해결 연습이 더 필요함을 보여준다. 수업의 개선 방향도 토픽 모델링을 통해 분석될 수 있음을 알 수 있다. 학생들이 관심을 가지는 텍스트 분석을 통해 오렌지 3의 툴이 가지는 한계를 최소화하고 좀 더 다각적으로 적용될 수 있는 사례와 미니 프로젝트가 설계될 필요가 있음을 보여준다.

본 연구를 통해 도출된 학생들의 관심 주제와 토픽은 향후 AI리터러시 수업을 연계한 각 전공별 AX 융합 수업 등에서 데이터를 활용한 문제해결에 시사점을 줄 수 있다. 최근 대학교육에서 AI리터러시가 강조되고 AX교육을 통해 각 분야에 전문성을 가지고 AI를 적용한 혁신과 부가가치를 창출할 수 있는 인재에 대한 필요성이 요구되고 있다는 점에서 본 연구에서 제시한 문제해결 프로젝트의 토픽과 방법론은 해당 데이터를 다루는 분야에 시사점을 줄 수 있다.

토픽모델링은 이러한 경향성이나 전반적인 활동에 대한 이해를 돕는 반면, 개별적인 학생에 대한 이해를 돕는 것에는 한계가 있다. 즉, 성찰 보고서의 키워드 분석에서도 나타났듯이 학생들은 다양한 경험을 통해 이 수업에 문제해결 과정에 참여했고, 장점과 한계점을 느꼈으며 어려움을 토로하였다. 하지만 개별 학생별 어떠한 어려움이 있고, 프로젝트 과정에서 학생들의 사고를 촉진하기 위해 어떤 스캐폴딩 전략이 필요한지에 대한 구체화된 개별화 전략을 도출하기는 어렵다.

그럼에도 불구하고 텍스트 마이닝 방법을 활용한 기말 프로젝트의 토픽 및 키워드 분석은 향후 학생들의 정성적 결과물을 평가하고 이를 기반으로 수업을 설계하고 개선 방향을 도출하는데 도움을 줄 수 있음을 본 연구 결과를 통해 알 수 있다. 또한, 실질적인 문제해결 과정을 문제 정의부터 데이터 수집, 전처리 및 머신러닝 적용, 데이터 시각화를 통한 해석 및 결과 도출의 과정을 텍스트 마이닝 한 이번 연구 결과는 향후 AI리터러시 수업을 타 전공과 융합하거나 IT 비전공 계열 학생들에게 적용하기 위한 시사점을 제시했다는 점에서 의의가 있다.

6. 결론 및 제언

6.1 결론

본 연구에서는 비IT 전공 계열 대학생의 AI리터러시를 함양하기 위한 목적으로 개설된 수업에서 AI활용 문제해결 프로젝트 보고서를 기반으로 학습자의 행위를 분석하고자 하였다.

총 4개 학기, 7개 분반 수업에서 246명의 대학생이 제출한 AI활용 문제해결 프로젝트 보고서를 분석하기 위해 키워드 빈도 분석을 시행하였다. 또한, 학생들의 보고서의 토픽을 분류하여 어떻게 토픽별 과제가 수행되었는지를 분석하기 위해 토픽 모델링을 수행하였다. 우선, 보고서의 키워드 빈도 분석을 통해 학생들이 프로젝트에서 중심으로 다루어야 할 데이터를 기반으로 지도학습 또는 비지도학습을 적용하여 AI모델을 개발하고, 성능 분석을 통해 최적의 모델을 찾고자 노력했다는 것을 확인할 수 있었다. 또한, 학생들이 주로 활용한 머신러닝 기법을 확인하였고 주제 단어들을 확인할 수 있었다.

단, 키워드 분석의 특성상 단어로 구성되어 있기 때문에 키워드 빈도 분석과 워드 클라우드를 분석하는데 교수자의 경험과 프로젝트 보고서에 대한 이해가 필요함을 알 수 있었다. 교수자의 주관적 해석을 최소화하기 위해, 토픽모델링 과정에서 BERTopic 방법을 사용함으로써 임베딩 모델간 차이를 비교하고, 차원 축소와 클러스터링 과정을 통해 학생들의 다양한 보고서의 주제와 내용 중 유사 토픽을 추출하였다. 마지막으로 세가지 표현 모델을 통해 토픽의 통계적 적합성, 의미적 적합성, 키워드 다양성 등의 지표를 활용하여 토픽을 추출하였다. 이러한 과정에서 수업을 운영한 교수자와 머신러닝 전문가의 비교 및 반복적인 논의 과정을 통해 최종 토픽을 도출하였다.

BERTopic 기반 토픽 모델링을 통해 총 7개의 토픽이 도출되었다. 첫 번째 토픽은 질병 예측 및 건강 위험 요인 분석을 주제로 수행한 것을 확인할 수 있었다. 가장 높은 공출현율과 가중치를 가지고 있는 토픽으로 학생들이 가장 데이터를 구하기 쉽고 관심을 가지고 해석하기에 질병과 건강 데이터의 접근성이 높음을 알 수 있다. 또한, 질병과 건강의 문제를 해결하는데 가장 많이 활용된 방법이 지도학습이면서 neural network와 naive bayes라는 것을 확인할 수 있었고, scatter plot이 데이터 해석에 많이 활용되었음을 알 수 있다. 성능 분석이 AUC, accuracy, Confusion matrix를 중심으로 수행되었으며, 성능을 높이기 위해 PCA를 시도한 학생들도 많다는 것은 본 수업에서 AI를 활용한 문제해결과정이 점차적으로 수행되었음을 알 수 있다.

두 번째 토픽은 청소년 건강 및 심리, 생활 습관을 분석한 주제가 많다는 것을 알 수 있다. 이는 직접적으로 대학생의 생활과 관련 있는 주제로 우리나라 대학생들이 청소년 시기에 가졌던 수면 문제, 성적, 우울증, 스트레스, 디지털 기술의 연계성 등 관심을 가지고 원인 분석을 수행하여 상관관계를 이해하고 예측을 수행하는 문제해결이 많았다는 것을 알

수 있다. 이 문제에서는 Naive Bayes 기법이 가장 많이 언급되었고 예방이나 제언 등을 해석의 관점에서 작성한 것으로 분석되었다.

세 번째 토픽은 소비자 및 고객을 분석하여 마케팅과 전략을 제시하기 위한 주제로 구성되었음을 알 수 있다. 최근 국내에서도 전자상거래에서의 구매 패턴을 분석하고 고객을 그룹화하여 고객에게 적절한 마케팅 전략과 프로모션 제품을 제안하는 일반적인 AI 문제해결 주제로 학생들이 이 부분에서도 관심을 가지고 수행했음을 알 수 있다.

수업에서도 비즈니스 분야의 데이터 분석을 비지도학습으로 수행했기 때문에, 학생들도 기말에서 비지도학습을 수행하여 클러스터링을 주로 수행한 것으로 분석되었다. 또한, Box Plot를 통해 클러스터간 차이를 분석하여 고객 전략화를 도출한 것을 알 수 있다. 본 수업에 참여한 학생 중 글로벌 경영대학 학생들 및 비즈니스에서 AI를 활용한 고객 전략화에 대한 커리어 연계가 수행될 수 있고, AI리터러시 수업 목적에 대한 성취가 이루어졌음을 알 수 있다.

네 번째 토픽은 문화 콘텐츠, 장르 및 감성 분석에 대한 주제였고, 장르에 대한 분석, 등급 및 평점 등을 예측하거나 사랑이라는 텍스트 분석 및 이미지 분석 등 다각적으로 문화 콘텐츠 분석이 수행된 것을 알 수 있었다. 특히, 다양한 지도 학습 방법이 언급됨으로써 일반적인 분류 모델을 찾기가 쉽지 않았음을 알 수 있다. 본 수업을 수강하는 학생 중 디지털 애니메이션, 한국문화콘텐츠학과, 중국어학과, 영미학과 등 다양한 문화 콘텐츠에 관심을 가질만한 학생들이 참여했는데, 여러 가지 시도가 있었음을 알 수가 있다.

다섯 번째는 회귀 분석이다. 학생들이 수업 당시 수행한 카훗형의 퀴즈에서 가장 많이 틀린 부분이 회귀분석이었다. 또한, 회귀분석을 기반으로 실습할 때, 지도학습의 AUC 또는 accuracy 등과 비교하여 성능 평가 방법의 차이를 이해하지 못해 질문이 가장 많았다. 가장 어려워했던 머신러닝 기법이 회귀 분석으로 지도학습의 분류에 대한 이해는 머신러닝포키즈나 티처블 머신으로 이해력을 높였다. 하지만, 회귀 분석은 의외로 오렌지3를 통해 성능 분석과 오차율을 최소화하는 개념에 대한 이해가 어려운 것으로 논의되었다. 실제로 학생들은 회귀 분석에 해당하는 키워드가 오차분석이나 linear regression 등보다는 분류나 클러스터링에 연계된 방식으로 수행된 것을 확인할 수 있었다. 이는 여전히 회귀 분석에 대한 이해가 부족하고 기말 프로젝트로 적절하게 연계되어 문제해결 과정이 수행되지 않았음을 시사한다.

여섯 번째 토픽은 예술 및 체육 뉴스 기사 분석이고 마지막 토픽은 영화, 드라마, 게임 등에 대한 텍스트 분석으로 이루어져 있음을 알 수 있다. 이 2개의 토픽 모두 토픽 중 가중치가 높지 않았지만 학생들이 텍스트 분석에 관심을 가지고 있고 저널 기사 등을 통해 자신이 관심을 가지는 예술 및 스포츠, 엔터테인먼트와 연계한 분석에 흥미를 가지고 수행했음을 알 수 있다.

본 수업에서는 오렌지3의 특성상 텍스트의 전처리가 쉽지 않고 한국어 분석이 쉽지 않았지만 학생들의 관심은 예술

및 스포츠, 엔터테인먼트 주제를 가지고 인간이 이해할 수 있는 텍스트를 통해 의미있는 해석을 시도한 것으로 해석할 수 있다.

학생들이 수행한 기말 프로젝트에 대한 토픽 모델링은 기존 프로젝트 보고서가 참여한 학생의 경험을 평가하는 방식에서 나아가 전반적인 기말 프로젝트 교육 방법과 활동이 AI리터러시에 어떻게 기여할 수 있음을 전반적으로 평가할 수 있는 도구가 될 수 있음을 시사한다.

6.2 제언

본 연구에서는 텍스트 마이닝 기반 토픽 모델링을 통해 학습자 행위를 분석하였다. 대학 교육에서 정량적 평가보다는 학습자의 경험을 바탕으로 프로젝트 기반 학습 및 실질적인 문제해결 활동을 정성적으로 분석함으로써 학습자의 행위를 분석하는 것은 매우 중요한 교육 분야에서의 AI 적용이다.

AI시대를 맞이하여 미래 사회에서는 융합적 사고를 바탕으로 AI와 협업하여 지속가능한 사회를 이끌 수 있는 인재를 요구하고 있다. 대학 교육에서 프로젝트 학습은 다양한 경험과 융합적 사고를 함양하고 문제해결력과 협업 능력을 지향하는 21세기 학습자 역량을 위한 교육 방법이다. 이러한 교육 방법에서 학습자 행위를 분석하고 이를 바탕으로 수업을 설계 또는 개선하기 위해 학생들의 프로젝트 결과물인 텍스트를 분석하는 것은 중요하다. 본 연구에서는 이러한 학생들의 다양한 활동 결과물을 해석하기에 적합한 BERTopic 기반 토픽 마이닝 기법을 통해 학습자의 행위를 어떻게 모델링할 수 있는지를 제시하였다. 이런 관점에서 향후 교육 분야에서 AI를 활용한 다양한 평가와 학습자 행위 분석을 통해 미래 교육에서의 교육 방법 및 전략에 대한 시사점을 도출하는 연구가 다각적으로 시도될 필요가 있음을 제언하고자 한다.

참고문헌

- [1] Kang, U., & Lee, J. (2022). Design of an artificial intelligence liberal arts education course for non-computer majors. *Journal of Korea Digital Contents Society*, 23(1), 57-66.
- [2] Hwang, H., & Hwang, Y. (2023). A study on AI literacy conceptualization and detailed competencies by sub-dimensions. *Journal of Cyber Communication*, 40(2), 89-148.
- [3] Ng, D., Su, J., & Chu, S. (2023). Fostering secondary school students' AI literacy through making AI-driven recycling bins. *Education and Information Technologies*. Advance online publication. <https://doi.org/10.1007/s10639-023-12183-9>
- [4] OECD (2025). Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (Review draft). OECD. Paris. <https://ailiteracyframework.org>
- [5] Chee, H., Ahn, S., & Lee, J. (2024). Competency Framework for AI literacy: Variations by different learner groups and an implied learning pathway. *British Journal of Educational Technology*, 56(5), 1-3. <https://doi.org/10.1111/bjet.13556>
- [6] Cha, H. (2025). Designing a AI Literacy Course for Non-IT Major Undergraduates in Higher Education: based on Backward Design. *The Journal of Korean association of computer education*, 28(7), 29-42. <https://doi.org/10.32431/kace.2025.28.7.003>
- [7] Ng, D., Leung, J., Chu, K., & Qiao, M. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, 58, 504-509. <https://doi.org/10.1002/pra2.487>
- [8] Southworth, J., Migliaccio, K., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, 4, Article 100127. <https://doi.org/10.1016/j.caeai.2023.100127>
- [9] Yuan, C., Tsai, H., & Chen, Y. (2024). Charting competence: A holistic scale for measuring proficiency in artificial intelligence literacy. *Journal of Educational Computing Research*, 62(7), 1675-1704. <https://doi.org/10.1177/07356331241261206>
- [10] Kang, K., Yun, S., & Jeon, D. (2025). Comparative analysis of digital and AI literacy education discourse in South Korea, the U.S., and China universities: A BERTopic approach. *The Korean Journal of Literacy Research*, 16(6), 329-354. <https://doi.org/10.37736/KJLR.2025.12.16.6.11>
- [11] Ho, M., Lin, M., Orr, C., Ryoo, J., & Jeon, R. (2025). AI literacy through a project-based learning course. In 2025 IEEE Smart World Congress (SWC) (pp. 283-288). IEEE. <https://doi.org/10.1109/SWC65939.2025.00065>
- [12] Bang, E., & Jeong, H. (2026). Exploratory study on a generative AI literacy education program for college students focusing on PBL. *Journal of Digital Contents Society*, 27(1), 165-177.
- [13] Li, H., Xiao, R., Nieu, H., Tseng, Y., & Liao, G. (2024). From unseen needs to classroom solutions: Exploring AI literacy challenges and opportunities with a project-based learning toolkit in K-12 education (arXiv No. 2412.17243). arXiv. <https://doi.org/10.48550/arXiv.2412.17243>
- [14] Tseng, Y., Xiao, R., Bogart, C., Savelka, J., & Sakr, M. (2024). Assessing the efficacy of goal-based scenarios in scaling AI literacy for non-technical learners. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2 (pp. 1842-1843). Association for Computing Machinery.
- [15] Skowronek, M., Gilberti, R., Petro, M., Sancomb, C., Maddern, S., & Jankovic, J. (2022). Inclusive STEAM education in diverse disciplines of sustainable energy and AI. *Energy and AI*, 7, 100124. <https://doi.org/10.1016/j.egyai.2021.100124>

- [16] Kasinidou, M. (2023). AI literacy for all: A participatory approach. *ITiCSE 2023: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*, 2, 607–608. <https://doi-org-ssl.eproxy.sch.ac.kr/8443/10.1145/3587103.3594135>
- [17] Jeong, Y., Son, N., & Yoo, J. (2020). Development of a project performance competency assessment framework for evaluating project-based learning in higher education. *Journal of Educational Evaluation*, 33(2), 455–486. <https://dx.doi.org/10.31158/JEEV.2020.33.2.455>
- [18] Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- [19] Gil, W., Jang, H., & Shin, I. (2024). An analysis of academic research trends in text mining using KeyBERT and BERTopic. *Journal of Information Technology and Architecture*, 21(2), 159–169. <https://doi.org/10.22865/jita.2024.21.2.159>
- [20] Lee, B., & Lee, J. (2025). Exploring the AI literacy competency framework from the perspectives of knowledge, skills, and attitudes: A text mining analysis of OECD documents. *The Korean Journal of Literacy Research*, 16(4), 141–171.



차현진

- 2001년 한양대학교 컴퓨터교육(이학사)
- 2004년 University College London, HCI with Ergonomics(MSc)
- 2012년 한양대학교 교육공학(교육학 박사)
- 2015년 한국교육학술정보원 연구원
- 2019년~현재 순천향대학교 기초공동교양학부 부교수

✚ 관심분야 : 사용자 경험(UX), UDL, UCD, AI와 학습분석학, 리터러시

✉ lois6934@sch.ac.kr



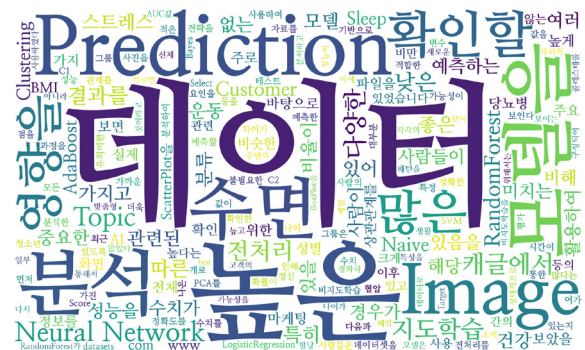
김연성

- 2024년 ~ 현재 순천향대학교 컴퓨터공학과 학사 과정 재학

✚ 관심분야 : 게임 및 인터랙티브 콘텐츠 개발, AI 리터러시 교육, 에듀테크

✉ schkys20244033@sch.ac.kr

부록



〈그림 2〉 Wordclouds in project reports(한글버전)



〈그림 5〉 Wordclouds in students' reflection reports (한글 버전)