

서지요소 보강에 따른 KDC 자동분류 성능 변화에 대한 연구*

A Study on the Performance Changes in Automatic KDC Classification through Bibliographic Element Augmentation

정 철 (Chul Jung)**
이 수 상 (Soo-Sang Lee)***
노 지 현 (Jee-Hyun Rho)****

목 차

- | | |
|-----------|-----------------------|
| 1. 서론 | 4. 자동분류 성능 평가 및 결과 분석 |
| 2. 이론적 배경 | 5. 결론 |
| 3. 연구 설계 | |

초 록

본 연구의 목적은 서지요소의 보강이 KDC 자동분류 모델의 성능 변화에 미치는 영향을 실증적으로 분석하는 데 있다. 이를 위해 출판유통통합전산망(BNK)에서 도서정보를 수집하고, 서지요소의 구성 수준에 따라 세 가지 서지요소 세트(A:표제, 저자명, B:표제, 저자명, 키워드, C:표제, 저자명, 키워드, 책소개, 목차)를 구성하였다. 이후 각 세트에 대해 KLUE-BERT 기반 자동분류 모델을 적용하여 KDC 강목 수준의 분류 성능을 비교 분석하였다. 분석 결과, A모델과 비교하여 C모델에서 모든 주류에서 분류 성능이 향상되었다. 특히 예술(6XX)은 F1 점수가 124.24% 향상되며 가장 높은 성능 향상률을 보였다. 한편, 서지요소의 보강 정도에 따라 일부 주류에서는 성능이 저하되거나 변화가 없는 경우도 확인되었다. 이에 강목 수준에서 성능 변화 양상과 원인을 실증적으로 분석한 결과, 성능 변화 유형은 지속적 성능 향상형, 제한적 성능 향상형, 성능 저하형, 성능 불변형의 네 가지 유형으로 구분되었다. 이러한 결과는 KDC 자동분류 성능 향상을 위해 서지요소의 구조적 특성과 강목의 주제적 특성을 복합적으로 고려할 필요가 있음을 시사한다.

ABSTRACT

The purpose of this study is to empirically analyze the effects of bibliographic element augmentation on the performance of automatic KDC classification model. To this end, bibliographic data were collected from the BookNet Korea (BNK) system, and three bibliographic element sets were constructed according to the level of bibliographic augmentation: Set A (title, author), Set B (title, author, keywords), and Set C (title, author, keywords, book summaries, tables of contents). KLUE-BERT-based automatic classification models were then applied to each set, and classification performance at the KDC division level was comparatively analyzed. The results showed that, compared with Model A, classification performance improved across all KDC main classes in Model C. In particular, the Arts (6XX) exhibited the highest rate of improvement, with a 124.24% increase in F1-score. However, depending on the level of bibliographic augmentation, some main classes showed performance degradation and no significant change. Additional analysis at the division level identified four patterns of performance change: consistently improved, conditionally improved, decreased, and unchanged. These findings suggest that improving the performance of automatic KDC classification requires a combined consideration of the structural characteristics of bibliographic elements and the subject characteristics of individual divisions.

키워드: 자동분류, 한국십진분류법(KDC), 딥러닝, 서지요소 보강, BERT 모델

Automatic Classification, Korean Decimal Classification(KDC), Deep Learning, Bibliographic Element Augmentation, BERT Model

* 본 연구는 제89회 IFLA WLIC 위성회의(Artificial Intelligence, Bibliographic Control and Legal Matters: Navigating New Horizons) 발표 자료의 일부를 수정·보완한 것임.

** 부산대학교 문헌정보학과 문헌정보학전공 박사수료

(feday0000@pusan.ac.kr / ISNI 0000 0005 0660 9447) (제1저자)

*** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr / ISNI 0000 0000 6434 9851) (공동저자)

**** 부산대학교 문헌정보학과 교수(jhrho@pusan.ac.kr / ISNI 0000 0004 6484 8385) (교신저자)

논문접수일자: 2026년 4월 20일 최초심사일자: 2026년 5월 4일 게재확정일자: 2026년 5월 13일

한국문헌정보학회지, 60(2): 387-408, 2026. <http://dx.doi.org/10.4275/KSLIS.2026.60.2.387>

※ Copyright © 2026 Korean Society for Library and Information Science

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서론

도서관에서 정보조직은 도서관 기능이 정상적으로 작동하기 위한 핵심 업무이며, 그 기반은 목록을 작성하는 과정에 있다. 이 과정에서는 자료의 서지적 특성을 기술하는 기술 편목(descriptive cataloging)뿐만 아니라, 자료의 주제를 분석하여 이에 적합한 주제명이나 분류기호로 표현하는 주제 편목(subject cataloging)이 함께 이루어진다. 특히 주제 편목은 이용자가 주제별로 자료를 탐색하고 접근할 수 있도록 지원한다는 점에서 도서관 정보서비스의 중요한 토대를 형성한다. 그러나 분류기호의 생성에 있어, 전통적인 분류 방식은 많은 인력과 시간을 요구하며, 분류자의 주관적 판단에 의존한다는 점에서 일관성 유지에 한계를 지닌다(노지현 외, 2023; Suominen, 2019).

이러한 한계를 보완하기 위하여 분류기호 생성의 자동화에 대한 연구가 지속적으로 이루어져 왔으며(김정현, 2011), 최근에는 인공지능(Artificial Intelligence, 이하 AI) 기반 자동분류가 주요 연구 흐름으로 자리매김하고 있다. 특히 트랜스포머 모델의 등장 이후(Vaswani et al., 2017), 자연어 처리(Natural Language Processing, 이하 NLP) 기술의 발전에 따라 대규모 텍스트 데이터를 활용한 의미 기반 주제 분류가 가능해졌다.

이와 같은 기술의 발전에 따라 전 세계적으로 도서관 자료의 DDC 및 KDC 분류기호 생성 업무에 기계학습(Machine Learning, 이하 ML) 및 딥러닝(Deep Learning, 이하 DL) 기반 자동분류를 적용하는 연구가 수행되고 있다(강우진 외, 2026; 광철완, 2021; Golub et al., 2018;

2024). 이러한 연구는 각국의 분류체계와 결합하여 발전하고 있으며, 분류모델의 고도화와 서지요소의 보강을 바탕으로 자동분류 성능이 지속적으로 향상되고 있다.

특히 도서관계의 자동분류에서 모델의 학습에 활용된 서지요소는 초기의 표제 중심 접근에서 출발하여, KORMARC 기반의 다양한 메타데이터(표제, 저자명 등)로 확장되었고, 더 나아가 목차, 원문, 책소개, 도서 상세 정보와 같은 데이터를 보강하는 방향으로 발전하였다(강우진 외, 2026; 광철완, 2021; 이용구, 2020; 2022; Golub et al., 2018; 2024). 이러한 서지요소의 보강은 도서의 주제 정보를 보다 풍부하게 반영함으로써 분류 성능 향상에 유의미한 영향을 미치는 것으로 나타난다.

그러나 기존 연구에서 활용한 서지요소는 주로 도서관 내부 데이터에 국한되며, 외부 데이터의 활용은 상대적으로 제한적으로 이루어져 왔다. 그중에서도 출판계에서 제공하는 정보는 도서의 주제를 보다 상세하고 풍부하게 설명하고 있음에도 불구하고, 이를 모델의 학습에 활용한 연구는 부재한 실정이다.

한편, 서지요소 보강에 따른 성능 향상이 보고되고 있으나, 주제 영역에 따라 그 효과에 편차를 보이고 있다(강우진 외, 2026; 이용구, 2022). 이는 서지요소의 보강이 곧 자동분류의 효과로 이어지는 것은 아님을 시사하며, 따라서 전체 주제 영역에서 분류 성능을 안정적으로 확보하기 위해 서지요소의 보강이 미치는 영향을 보다 체계적으로 분석할 필요가 있음을 의미한다.

이에 본 연구는 기존 연구에서 가장 우수한 성능을 보인 BERT 계열 모델 중 한국어 특화

모델인 KLUE-BERT를 기반으로 자동분류 모델을 구축하고, 출판계 데이터를 활용한 서지요소의 보강이 KDC 자동분류 성능에 미치는 영향을 분석하는 것을 목적으로 한다. 이를 통해 서지요소 보강에 따른 주류 및 강목 수준의 분류 성능 변화와 효과를 실증적으로 검토하고, KDC 자동분류의 성능 향상을 위한 시사점을 도출하고자 한다.

2. 이론적 배경

2.1 도서관계의 자동분류

도서관계에서 자동분류에 관한 연구는 지속적으로 이루어져 왔는데, 국내에서는 1980년대 초반부터 관련 연구가 수행된 것으로 확인된다(김정현, 2011). 이러한 연구는 주제 편목 업무의 효율성과 일관성을 제고하기 위한 노력의 일환으로, 크게 주제명 기반 연구와 분류기호 기반 연구로 구분된다.¹⁾

먼저, 주제명 기반 자동분류 연구에서는 초기 단계에서 ML 기법이 주요하게 활용되었다. 기존의 자동분류 연구는 SVM(Support Vector Machine)이나 MNB(Multinomial Naïve Bayes)와 같은 단일 모델을 기반으로 한 지도학습 기법을 중심으로 발전해 왔다(Golub et al., 2018). 그러나 이러한 단일 모델을 바탕으로 한 자동분류는 어휘 기반 접근(lexical approaches) 또

는 연관성 기반 접근(associative approaches)과 같이 단일한 방식에 의존하기 때문에 명시적으로 나타난 용어에 기반한 정보나 단어와 주제 간의 통계적 연관성 중 일부만을 반영할 수 있어, 다른 유형의 정보는 고려되지 못하는 한계가 존재하였다.

이에 Suominen(2019)은 어휘 기반 모델(Maui, STWFSA, MLLM 등)과 연관성 기반 모델(fastText, Omikujii 등)을 결합하는 앙상블(Ensemble) 기법을 제안하였다. 해당 연구에서 제안한 Annif는 앙상블 기법을 바탕으로 주제어가 생성된 문서 데이터를 학습하여 주제명을 자동으로 분류하는 도구로, 단일 모델에 비해 일관되고 우수한 성능을 보였다. 이후 Suominen et al.(2022)은 Annif를 기반으로 한 주제명 자동 분류 기법을 실제 서비스 환경으로 확장하여 Finto AI로 구현하고, 이를 도서관 색인 실무에 적용한 사례를 제시함으로써 유용성을 확인하였다.

이후 AI의 발전이 가속화되면서, BERT(Bidirectional Encoder Representations from Transformers)와 같은 트랜스포머 기반 모델을 활용한 자동분류 연구가 수행되었다. 대표적인 사례로, 이용구(2022)는 국가서지 데이터를 대상으로 BERT 기반 주제명 자동분류를 수행하였다. 이 연구에서는 표제와 저자명 뿐만 아니라 목차와 원문 데이터를 포함한 서지요소 세트를 구성하였으며, 서지요소의 다양성이 증가할수록 분류 성능이 향상되는 경향을

1) ChatGPT와 Gemini와 같은 생성형 AI를 바탕으로 DDC 분류기호나 LCSH 주제명 등을 생성하는 연구가 최근 들어 수행되고 있지만(Bodenhamer, 2023; Brzustowicz, 2023; Noruzi, 2024), 해당 방식은 현 시점에서 실효성이 낮은 한계가 지적되고 있다(강우진 외, 2026; Noruzi, 2024). 이에 본 연구에서는 생성형 AI 기반의 접근은 고려하지 않았다.

확인하였다. 그러나 주제 영역 및 데이터 분포에 따라 성능 편차가 나타났으며, 일부 분야에서는 상대적으로 낮은 성능을 보이는 한계가 존재하였다.

다음으로, 분류기호 기반 자동분류 연구의 양상도 주제명 기반 자동분류와 유사하게 전개되어 왔다. 초기 연구에서는 SVM, MNB, MLP (Multi-Layer Perceptron)와 같은 단일 모델을 바탕으로 KDC 및 DDC의 자동분류를 실험하였다(곽철완, 2021; Golub et al., 2018). 이후, 단일 모델 기반 접근의 한계를 보완하고자 Golub et al.(2024)은 Annif의 앙상블 기법을 스웨덴 국가종합목록(LIBRIS)에 적용하여 DDC 자동분류를 수행하였으며, 단일 모델 대비 안정적이고 향상된 성능을 확인하였다. 특히 모델의 학습에 활용되는 서지요소에 주제가 포함된 경우 분류 정확도가 더욱 향상되는 경향을 보였다.

이어서 모델의 발전에 따라 BERT 기반 분류모델이 도입되면서 기존의 ML 기반 모델에 비해 전반적인 분류 성능 향상이 확인되었다. 강우진 외(2026)는 DDC 기술과학(600) 분야 문헌을 대상으로 Omikujii, FastText, BERT 기반 모델을 비교한 결과, BERT 기반 분류모델이 가장 우수한 성능을 보였으며, 특히 한국문화정보원의 문화빅데이터 플랫폼에서 국립중앙도서관이 제공하는 '도서별 상세정보'와 같은 문헌 설명 정보를 서지요소로 보강할 경우 성능 개선 효과가 더욱 두드러지는 것으로 나타났다. 또한 해당 모델은 요목 수준에서 약 79.52%의 정확도를 기록하여, Golub et al.(2024)에서 보고된 분류 성능을 상회하는 결과를 보였다.

이상의 선행연구들을 종합하면, 도서관계의

자동분류 연구는 모델의 발전에 따라 성능이 지속적으로 향상되어 왔으며, 최근에는 BERT 기반 모델이 가장 우수한 성능을 보이는 것으로 확인된다(강우진 외, 2026; 이용구, 2022). 그리고 표제와 같은 단일의 서지요소보다 복합적인 서지요소를 활용하는 것이 더 높은 분류 성능을 보이며, 특히 문헌 설명 정보, 원문, 주제어, 목차 등 주제나 내용과 관련된 서지요소는 분류 성능 향상에 중요한 역할을 하는 것으로 보고되고 있다(강우진 외, 2026; 이용구, 2020; 2022; Golub et al., 2018; 2024).

2.2 KLUE-BERT

NLP는 인간의 언어를 컴퓨터가 이해하고 처리할 수 있도록 하는 기술 분야를 의미하며, 최근 DL 모델의 발전과 함께 성능이 크게 향상되어 왔다. 이는 NLP의 영역에서 기계가 단어와 문장이 어떤 의미를 갖는지 학습하고 이해하도록 하는 데에 DL 모델이 적용되기 때문이다(고영수 외, 2021). 특히 트랜스포머 구조는 대규모 데이터 기반의 언어 모델 학습을 가능하게 하며 NLP 발전의 전환점으로 작용하였다.

트랜스포머 기반 모델 중 BERT는 MLM (Masked Language Model)과 NSP (Next Sentence Prediction)라는 두 가지 사전학습 기법을 사용하여 양방향 문맥 정보를 반영하는 구조를 통해 문장 내 단어 간 관계를 정교하게 학습할 수 있도록 설계된 언어 모델이다(Rogers et al., 2020). BERT는 사전학습(pre-training)과 파인튜닝(fine-tuning) 구조를 기반으로 다양한 NLP 과제에서 우수한 성능을 보였으며, GLUE, SQuAD 등 다양한 벤치마크에서 기존

모델 대비 성능 향상을 달성한 것으로 보고된다(Devlin et al., 2019).

이후, BERT를 기반으로 한 다양한 변형 모델이 제안되었으며, 한국어 특화 사전학습 언어모델도 활발히 개발되고 있다. 대표적인 예로 KLUE-BERT, KoBERT, KorBERT 등이 있으며, 이러한 모델들은 한국어 말뭉치를 기반으로 학습되어 한국어 NLP 과제에서 높은 성능을 보이도록 설계되었다. 이 중 KLUE-BERT는 KLUE(Korean Language Understanding Evaluation) 벤치마크를 기반으로 구축된 대표적인 한국어 사전학습 언어모델로, 타 한국어 사전학습 모델 대비 NLP 과제에서 우수한 성능을 보이고 있다(Park et al., 2021).

Park et al.(2021)이 제안한 KLUE-BERT는 약 62GB 규모의 한국어 말뭉치를 바탕으로 사전학습되었으며, 이후 토픽 분류(TC), 문장 의미 유사도(STS), 자연어 추론(NLI), 개체명 인식(NER), 관계 추출(RE) 등 8개 주요 NLP 과제에 대해 NLP 과제별 데이터 세트를 활용하여 파인튜닝 되었다. 이 과정에서 사용된 데이터는 뉴스 기사, 위키백과, 사용자 리뷰, 질의문 등 다양한 도메인의 텍스트를 포함하고 있어, 형식적인 문체뿐 아니라 구어체와 사용자 생성 텍스트까지 반영할 수 있도록 설계되었다. 이러한 설계를 통해 KLUE-BERT는 실제 한국어 텍스트 환경을 효과적으로 반영할 수 있도록 하며, 다양한 주제와 문체를 포함하는 텍스트 처리에 적합한 구조를 갖는다.

3. 연구 설계

본 연구에서는 도서의 표제와 저자명으로 구성된 기본 서지정보에 출판계에서 수집한 외부 데이터를 보강함으로써 자동분류 성능이 어떻게 변화하는지를 분석하고자 한다. 이를 위해서 다른 구성 수준의 세 가지 서지요소 세트를 설계하였다.

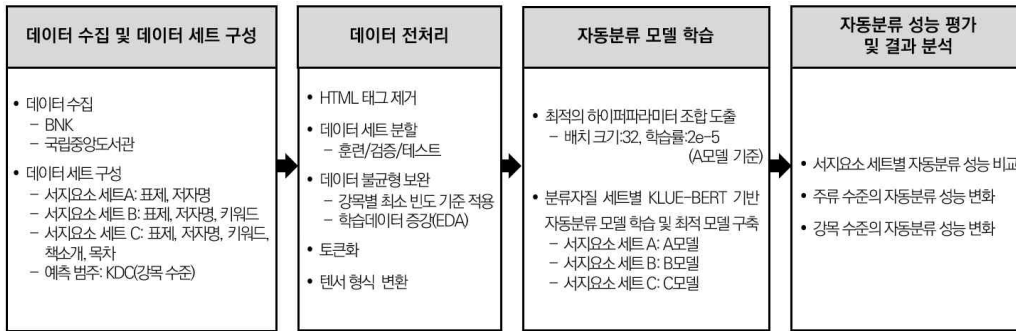
다음으로 각 서지요소 세트를 바탕으로 자동분류 모델을 각각 학습하였으며, 서지요소 세트의 보강 정도에 따른 분류 성능의 변화를 분석하였다. 이때 자동분류 모델은 KDC 강목 수준의 클래스를 예측하도록 학습되었다.

자동분류 모델은 분류기호의 자동분류를 다룬 기존 연구에서 BERT 계열 분류 모델이 우수한 성능을 보인다는 결과를 바탕으로(강우진 외, 2026), BERT 계열 모델을 기반으로 구성하였다. 기존 연구에서는 다국어 BERT 사전학습 모델을 기반으로 하였으나, 본 연구에서는 한국어 처리 성능을 고려하여 KLUE-BERT를 활용하였다.²⁾ 모든 과정은 Google Colab 환경에서 Python을 사용하여 NVIDIA A100 GPU에서 수행되었다. 연구의 절차를 도식화하면 <그림 1>과 같다.

3.1 데이터의 수집 및 구조

도서를 대상으로 한 자동분류에서 단일의 서지요소보다 복합적인 서지요소를 활용하는 것이 더 높은 분류 성능을 보이며, 특히 주제와 관련된 데이터는 분류 성능 향상에 중요한 역할

2) 허깅 페이스의 klue/bert-base, <https://huggingface.co/klue/bert-base> [최종접근 25.01.05]



〈그림 1〉 연구 절차 도식화

을 하는 것으로 확인되었다(강우진 외, 2026; 이용구, 2022; Golub et al., 2018; 2024). 이에 본 연구에서는 도서의 주제와 관련된 다양한 정보를 반영한 서지요소를 구성하기 위하여, 한국출판문화산업진흥원의 출판유통통합전산망(BookNet Korea, 이하 BNK)³⁾에서 제공하는 도서정보를 수집하였다.

데이터 수집은 2023년 12월 정보공개청구를 통해 수행되었으며, 2023년 1월 1일부터 12월 3일까지 발행된 도서를 대상으로 BNK 도서정보 10,106건을 확보하였다. 이후, 020 필드(ISBN)를 기준으로 국립중앙도서관의 국가서지과로부터 대한민국 국가서지 데이터를 제공받아 동일 도서의 KORMARC 056 필드를 추출하고, 이를 통해 예측 범주인 KDC 6판 분류기호를 강목 수준까지 수집하였다.

수집된 BNK 데이터의 경우, '서평'과 '본문 인용' 데이터는 기입률이 낮아 분석에 활용하기에 제한이 있어 제외하였다. 대신, 기입률이 높은 '키워드', '책소개', '목차'를 연구에 활용하고자 하였다. 이에 따라 해당 정보가 모두 확보된 5,902건을 분석 대상으로 선정하였다.

BNK의 도서정보는 출판사가 도서의 내용을 설명하기 위해 직접 입력한 정보를 기반으로 구성되며, 도서의 주제는 국제 표준 주제분류 체계인 테마(Thema 1.4)에 기반한 주제어(메인 주제어)와 수식자(추가 주제어)의 조합으로 이루어져 있다. '키워드'는 이러한 주제어와 수식자에 더해 출판사가 자유롭게 기술한 정보(태그어)를 포함하여, 도서의 주제와 맥락이 풍부하게 표현된다는 특징을 지닌다. 이에 본 연구에서는 BNK의 '키워드'가 KORMARC 주제명 필드(6XX 필드)에 비해 도서의 주제를 보다 구체적으로 반영하고 있다고 판단하여 이를 서지요소로 활용하였다.

다음으로 서지요소의 보강 수준에 따른 KDC 자동분류 성능 변화를 살펴보기 위하여, 세 가지 세트를 정의하였다. 〈표 1〉은 하나의 도서에 대해 수집된 데이터의 예시를 제시한 것이다. 구체적으로 서지요소 세트 A는 '표제'와 '저자명'으로 구성되며, 서지요소 세트 B는 여기에 '키워드'를 보강하였다. 서지요소 세트 C는 서지요소 세트 B에 '책소개'와 '목차'를 보강한 것으로, 가장 많은 내용 정보를 반영하고 있다.

3) 한국출판문화산업진흥원 출판유통통합전산망(BNK) 홈페이지, <https://bnk.kpipa.or.kr/> [최종 접근: 2026.04.10.]

〈표 1〉 서지요소 세트의 구성 예시

서지요소	예시	포함 여부		
		A	B	C
표제	사랑과 상실의 뇌과학	○	○	○
저자명	메리프랜시스 오코너	○	○	○
키워드	뇌과학, 뉴런, 신경생물학, 신경전달물질, 심리학, 상실, 애, 비애, 슬픔, 행동, 감정, 인지심리, 수학, 과학, 의학, 간호학, 건강, 인간관계, 개인발전, 철학, 종교, 사회, 사회과학	X	○	○
책소개	사랑하는 가족 친구를 잃고 그 상실로 고통 받을 때 우리 뇌에선 무슨 일이 (하략)	X	X	○
목차	차례 들어가는 말 사랑하는 이를 잃었을 때 뇌에선 무슨 일이 일어날까 1부 (하략)	X	X	○
예측 범주 (KDC)	18X	○	○	○

3.2 데이터의 전처리

일반적으로 BERT 계열 모델은 입력된 텍스트를 토큰 단위로 분할한 후 이를 벡터로 변환하여 처리한다. 이 과정은 워드피스(wordpiece) 기반 토큰라이저를 통해 수행되며, 기본적인 텍스트 정제 기능을 포함하고 있다. 다만, 불필요한 토큰이 생성될 경우 분류모델의 성능에 악영향을 미칠 수도 있다. 이에 데이터에 존재하는 HTML 태그를 제거하였다.

그러나 본 데이터 세트는 KDC 강목별 도서 수의 불균형이 존재하여, 모델 학습 시 클래스 간 데이터 불균형으로 인한 학습 편향이 발생할 가능성이 있다. 이에 본 연구에서는 이러한 문제를 보완하기 위해 두 가지 방안을 적용하였다.

먼저, 강목별 최소 빈도 기준을 적용하여 데이터 불균형을 사전에 조정하였다. KDC 강목별 도서 수에 큰 편차가 존재하여 일부 강목은 매우 적은 수의 데이터만 포함하고 있었으며, 이에 따라 최소 빈도가 10건 미만인 20개 강목

을 제외하였다. 그 결과 최종적으로 61개 강목에 대한 5,882건의 데이터를 분석 대상으로 확정하였다.

다음으로, 데이터 분할 이후 훈련 데이터에 대한 증강을 통해 클래스 불균형을 보완하였다. 데이터는 자동분류 모델의 학습과 평가를 위해 강목별 비율을 유지한 채, 훈련(train), 검증(validation), 테스트(test) 세트로 분할되었다. 분할 비율은 지도학습 기반 자동분류 연구에서 일반적으로 활용되는 데이터 세트 구성 비율에 따라(이용구, 2023), 전체 데이터의 80%를 훈련 데이터, 20%를 테스트 데이터로 설정한 뒤, 훈련 데이터의 20%를 검증 데이터로 재분할하였다.

한편, 선행된 BERT 계열 모델 기반 DDC 자동분류 연구에서 보고된 바와 같이(강우진 외, 2026), DL 기반 모델은 훈련 데이터의 규모와 분포에 크게 의존하는 특성을 갖는다. 데이터 규모가 제한적이고 강목 간 분포가 불균형한 경우, 자동분류 모델이 훈련 데이터에 포함된 특정 표현이나 분류 패턴에 과도하게 의존하게

되어 검증 데이터에 대한 일반화 성능이 저하될 가능성이 있다. 본 연구에서도 데이터 분할 후 훈련 데이터가 3,764건으로 축소되면서 과적합의 위험이 존재하였기 때문에, 이를 방지하고자 Wei와 Zou(2019)가 제안한 EDA(Easy Data Augmentation) 기법 중 RD(Random Deletion)와 RS(Random Swap)를 활용하여 훈련 데이터를 증강하였다. RD는 각 단어를 일정 확률로 무작위로 삭제하는 방식으로 데이터를 증강하며, RS는 문장 내의 두 단어를 무작위로 선택하여 서로 위치를 바꾸는 방식으로 데이터를 증강한다. 이와 같은 방식들은 문장의 의미를 크게 훼손하지 않는 범위에서 다양한 문장 패턴을 생성함으로써 모델의 과적합을 완화하며, 특히 소규모 데이터셋 환경에서 효과적이라고 보고된 바 있다(Wei & Zou, 2019). 본 연구에서는 이러한 데이터 증강을 통해 훈련 데이터 2,401건이 추가되어 6,165건으로 확장되었다.

이후, KLUE-BERT에서 BertTokenizerFast를 이용하여 입력 텍스트를 토큰화하고, [CLS]와 [SEP] 토큰이 포함된 입력 시퀀스를 구성하였다. 이 과정에서 최대 입력 길이인 토큰 수(512)를 기준으로 패딩(padding)과 트렁케이션(truncation)을 적용하고, 어텐션 마스크(attention mask)를 생성하였다. 최종적으로 전처리된 데이터는 텐서(tensor) 형식으로 변환하여 모델 학습에 사용하였다.

3.3 자동분류 모델의 학습

사전학습 모델인 KLUE-BERT는 입력 시퀀스를 바탕으로 양방향 문맥 정보를 반영한 표현(representations)을 생성하는 모델이다(Park

et al., 2021). BERT 계열 모델은 입력 시퀀스로부터 문맥 정보를 반영한 표현을 생성하며, 이러한 표현은 최종 은닉 상태(hidden state)로 처리된다(Devlin et al., 2019). 이때 입력 시퀀스의 첫 토큰인 [CLS]에 해당하는 최종 은닉 상태는 입력 시퀀스 전체의 정보를 반영한 집계 표현(aggregate representation)으로 활용되며(Devlin et al., 2019; Park et al., 2021), 이는 본 연구의 KDC 자동분류 모델에서 범주 예측을 위한 핵심 자질(feature)로 사용된다. 분류 과제(classification task)에서는 해당 자질이 선형 변환을 거쳐 분류 계층에 전달되는 방식으로 분류가 수행된다(Park et al., 2021; Devlin et al., 2019). 본 연구에서는 복수의 서지요소를 하나의 입력으로 결합하여 처리하는 단일 문장 분류(single sentence classification) 과제로 모델을 구성하고, KDC 강목 수준의 클래스에 대한 확률을 예측하도록 하였다.

이와 같이 구성된 모델의 학습 과정에서는 안정성과 효율성을 높이기 위해 Adam 계열의 최적화 기법을 적용하였으며, 다중 클래스 분류에 적합한 크로스 엔트로피(cross entropy) 기반 손실 함수를 사용하였다. 또한 데이터 불균형으로 인한 학습 편향을 완화하기 위해 클래스별 가중치를 반영하였다.

이후, 모델의 성능을 최적화하기 위해 하이퍼파라미터 탐색을 수행하였다. 하이퍼파라미터는 모델의 학습 과정과 구조를 제어하는 변수로, BERT 계열 모델에서는 적절한 설정이 필요하다. 이에 Devlin et al.(2019)의 실험 설정에서 활용된 하이퍼파라미터 범위를 참고하여 배치 크기(16, 32)와 학습률($2e-5$, $5e-5$)의 조합을 구성하였다.

이러한 설정을 바탕으로 서지요소 세트 A에 대해 에포크를 최대 10까지 확장하여 학습을 수행한 결과, <표 2>와 같이 최적의 하이퍼파라미터 조합은 학습률 $2e-5$, 배치 크기 32로 확인되었다.

자동분류 모델의 성능 평가를 위한 지표는 정확도(accuracy), 정확률(precision), 재현율(recall), F1 점수(F1-score), 그리고 ROC-AUC를 활용하였다. 정확도는 전체 예측 중 올바르게 분류된 비율을 의미하며, 정확률은 특정 클래스로 예측한 결과 중 실제로 해당 클래스에 속하는 비율을, 재현율은 실제 해당 클래스에 속하는 데이터 중 모델이 올바르게 분류한 비율을 의미한다(강우진 외, 2026). F1 점수는 정확률과 재현율의 조화 평균으로 산출되는 지표이며(이용구, 2023), ROC-AUC는 무작위로 선택된 양성 사례가 음성 사례보다 더 높은 점수를 받을 확률로 해석된다(Li, 2024).

이진 분류에서는 TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)의 개념이 명확하여 계산이 단순하므로 해당 단일 지표들로 성능 평가가 가능하다. 그러나 다중 분류에서는 클래스가 두 가지

이상 존재하기에 각 클래스별로 TP, FP, FN이 다르게 정의된다. 따라서 다중 자동분류 모델의 성능 평가에는 마이크로 평균(micro average)과 매크로 평균(macro average) 같은 방법으로 확장된 지표가 요구된다.

이때 마이크로 평균은 전체 데이터를 기준으로 성능을 산출하여 자동분류 모델의 전반적인 분류 성능을 평가하는 방식이며, 매크로 평균은 각 클래스별 성능을 동일한 비중으로 평균하여 클래스 간 성능 차이를 균형 있게 반영하는 방식이다(Raschka & Mirjalili, 2019). 본 연구에서는 클래스 간 데이터 불균형이 존재하는 상황에서 각 KDC 강목에 대한 분류 성능을 균형 있게 평가하기 위해 정확도는 단일 지표로 활용하고, 정확률, 재현율, F1 점수, ROC-AUC는 매크로 평균 방식으로 평가하였다. 이때 정확률과 재현율은 상충 관계에 있으므로, 두 지표의 조화 평균인 F1-점수를 주요하게 고려하였다(이용구, 2023). 또한 ROC-AUC는 다중 분류 환경에 적용하기 위해 OvR(One-vs-Rest) 방식으로 산출하였다.

앞서 정의한 성능 지표를 바탕으로, <표 2>와 같이 학습률 $2e-5$ 와 배치 크기 32에서 최적

<표 2> 하이퍼파라미터 조합에 따른 모델 성능(서지요소 세트 A 기준)

배치 크기	학습률	데이터 유형	매크로 평균		
			F1	정확률	재현율
16	$2e-5$	검증	0.4953	0.6359	0.4852
		훈련	0.9990	0.9987	0.9994
	$5e-5$	검증	0.4771	0.6025	0.4816
		훈련	0.9848	0.9813	0.9900
32	$2e-5$	검증	0.5114	0.6581	0.4977
		훈련	0.9960	0.9939	0.9982
	$5e-5$	검증	0.5106	0.6123	0.5094
		훈련	0.9993	0.9993	0.9993

의 성능을 확인하였다. 이에 동일한 설정을 적용하여 각 서지요소 데이터 세트에 대한 KDC 자동분류 모델(A모델, B모델, C모델)의 학습을 수행하였다. 에포크는 10회로 설정하였으며, 검증 데이터에서 가장 높은 성능을 보이는 시점의 자동분류 모델을 성능 평가에 활용하고자 하였다.

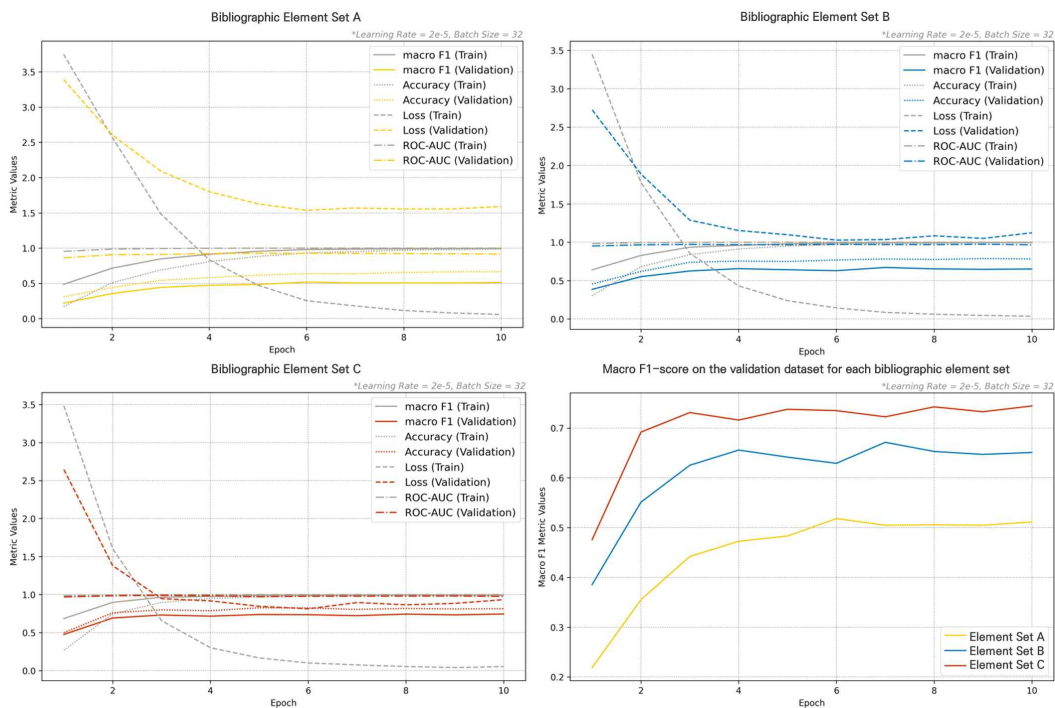
〈그림 2〉는 서지요소 세트별 에포크에 따른 훈련 데이터와 검증 데이터의 성능 변화를 나타낸 것이다. 분석 결과에 따르면, 모든 서지요소 세트에서 6 에포크 시점에 검증 데이터의 성능이 최고점에 도달한 이후 정체되거나 감소하는 경향이 나타났다. 반면, 훈련 데이터에서는 성능이 지속적으로 향상되어 과적합이 발생하는 양상이 확인되었다. 이에 따라 각 서지요소

세트에 대해 6 에포크 시점의 모델을 최적 모델로 선정하여 이후 분석에 활용하였다.

4. 자동분류 성능 평가 및 결과 분석

4.1 서지요소 세트별 자동분류 성능 비교

서지요소 세트별 자동분류 모델의 테스트 데이터를 평가한 결과, 〈표 3〉과 같이 서지요소가 보강됨에 따라 KDC 강목에 대한 분류 성능이 전반적으로 향상되는 경향이 확인되었다. F1 점수, 정확률, 재현율, 정확도는 모두 일관된 증가 패턴을 보였으며, ROC-AUC 역시 동일한 양상으로 상승하여 판별력 향상이 확인되었다.



〈그림 2〉 에포크에 따른 서지요소 세트별 모델 성능(훈련 및 검증 데이터)

〈표 3〉 서지요소 세트별 자동분류 모델 성능 비교(테스트 데이터)

평가 지표		A모델	B모델	C모델
매크로 평균	F1	0.5286	0.6568	0.7009
	정확률	0.6253	0.7194	0.7662
	재현율	0.5330	0.6654	0.6998
	ROC_AUC	0.9252	0.9623	0.9673
정확도		0.6525	0.7723	0.8199
손실률		1.5219	0.9492	0.8308

*학습률: 2e-5, 배치 크기:32, 에포크: 6

반면 손실률은 점차 감소하여 서지요소 보강이 학습 안정성에도 긍정적인 영향을 미친 것으로 나타났다.

4.2 주류 수준의 자동분류 성능 변화

강목 수준에서 산출된 F1 점수를 상위 주류 기준으로 매핑한 후 주류별 평균값을 계산하여 주류 수준의 분류 성능 변화를 분석한 결과, 대부분의 주류에서 분류 성능이 점진적으로 향상된 것으로 나타났다. 〈표 4〉와 같이 모든 주류에서 C모델은 A모델보다 높은 F1 점수를 보였으

며, 특히 예술(6XX) 분야에서는 124.24%의 높은 향상률을 보였다. 이러한 결과는 서지요소의 보강이 KDC 자동분류 성능 개선에 전반적으로 긍정적인 영향을 미친다는 점을 시사한다.

한편 일부 주류에서는 서지요소를 보강하는 것이 항상 분류 성능의 향상으로 이어지지 않는 것으로 나타났다. 예를 들어 총류(0XX)와 같이 키워드가 보강되는 과정에서 성능이 일시적으로 저하되거나, 종교(2XX)와 기술과학(5XX)과 같이 B모델에서 최고 성능을 보인 이후 책소개와 목차의 보강으로 성능이 감소하는 경우도 확인되었다. 또한 언어(7XX)의 경

〈표 4〉 주류 수준의 서지요소 세트별 자동분류 모델 성능과 향상률

주류	샘플 수	매크로 평균 F1 점수			성능 향상률(%)			
		A모델	B모델	C모델	A → B	B → C	A → C	
0XX	총류	81	0.68	0.65	0.70	-4.41	7.69	2.94
1XX	철학	85	0.49	0.53	0.58	8.16	9.43	18.37
2XX	종교	32	0.79	0.95	0.92	20.25	-3.16	16.46
3XX	사회과학	239	0.45	0.57	0.63	26.67	10.53	40
4XX	자연과학	46	0.59	0.64	0.74	8.47	15.63	25.42
5XX	기술과학	105	0.42	0.67	0.66	59.52	-1.49	57.14
6XX	예술	45	0.33	0.61	0.74	84.85	21.31	124.24
7XX	언어	35	0.57	0.62	0.62	8.77	0	8.77
8XX	문학	463	0.67	0.77	0.84	14.93	9.09	25.37
9XX	역사	46	0.47	0.72	0.63	53.19	-12.5	34.04

*성능 향상률(%) = (이후 단계 모델 성능 - 이전 단계 모델 성능)/이전 단계 모델 성능 * 100

우, B모델에서 성능이 향상되었으나 C모델에서 성능의 변화는 확인되지 않았다.

이에 따라 다음 절에서는 강목 수준에서 성능 변화를 실제 분류 사례를 통해 세부적으로 분석함으로써, 서지요소 보강 이후 나타나는 성능 변화의 주요 양상과 그 원인을 살펴보고자 한다.

4.3 강목 수준의 자동분류 성능 변화

본 연구에서는 주류 수준 분석에서 확인된 성능 변화를 보다 정밀하게 검토하기 위해, KDC 강목 61개를 대상으로 서지요소의 보강에 따른 성능 변화를 기준으로 유형을 구분하였다. 이때 A모델과 C모델 간의 성능 차이를 중심으로 유형을 구분하되, 서지요소가 단계적으로 보강되는 과정에서 나타나는 B모델의 성능 변화 양상을 함께 고려하였다. 그 결과, 성능 변화 유형

은 지속적 성능 향상형, 제한적 성능 향상형, 성능저하형, 성능 불변형의 네 가지 유형으로 구분되었다.

4.3.1 지속적 성능 향상형

지속적 성능 향상형은 서지요소가 단계적으로 보강됨에 따라 성능이 지속적으로 향상된 경우로, 전체 61개 강목 중 29개 강목이 이에 해당하였다(〈표 5〉 참조). 이 유형은 각 서지요소가 핵심 주제를 비교적 명확하게 반영하고, 보강되는 서지요소가 상호 보완적으로 작용하면서 성능 향상으로 이어졌다.

이 유형은 주로 예술, 기술과학, 문학 분야에서 주로 나타났다. 특히 음악(67X), 전기공학, 통신공학, 전자공학(56X), 중국문학(82X) 등의 강목에서는 단계별 서지요소 보강에 따라 모델의 성능이 향상되며, C모델에서는 F1 점수가 1.00에 도달하며 완전한 분류 성능을 보였다.

〈표 5〉 서지요소 보강에 따른 분류 성능 변화(지속적 성능 향상)

강목	매크로 평균 F1 점수			빈도	
	A모델	B모델	C모델		
00X	총류	0.80	0.87	0.90	65
15X	동양철학, 동양사상	0.43	0.67	0.80	5
16X	서양철학	0.50	0.69	0.74	11
18X	심리학	0.53	0.64	0.73	44
19X	윤리학, 도덕철학	0.18	0.32	0.43	15
21X	비교종교	0.86	1.00	1.00	3
32X	경제학	0.73	0.79	0.85	104
33X	사회학, 사회문제	0.44	0.57	0.67	45
36X	법률, 법학	0.65	0.76	0.81	17
38X	풍습, 예절, 민속학	0.29	0.50	0.67	4
47X	생명과학	0.29	0.67	0.80	6
49X	동물학	0.18	0.42	0.53	8
51X	의학	0.67	0.80	0.82	39
52X	농업, 농학	0.42	0.62	0.65	11

강목	강목명	매크로 평균 F1 점수			빈도
		A모델	B모델	C모델	
53X	공학, 공업일반, 토목공학, 환경공학	0.42	0.53	0.59	14
56X	전기공학, 통신공학, 전자공학	0.40	0.67	1.00	3
59X	생활과학	0.46	0.64	0.72	24
60X	예술	0.25	0.62	0.67	7
63X	공예	0.00	0.50	0.67	2
65X	회화, 도화 디자인	0.48	0.55	0.67	14
67X	음악	0.78	0.96	1.00	11
68X	공연예술, 매체예술	0.46	0.50	0.91	6
80X	문학	0.33	0.51	0.71	17
81X	한국문학	0.72	0.89	0.91	254
82X	중국문학	0.78	0.90	1.00	11
83X	일본문학 및 기타 아시아 제문학	0.92	0.93	0.95	79
84X	영미문학	0.66	0.86	0.88	53
91X	아시아	0.49	0.77	0.78	21
98X	지리	0.67	0.69	0.72	16

구체적인 분류 사례를 살펴보면, 음악(67X)의 「안드라스 쉬프」는 A모델에서는 사회문제(33X)로 오분류되었으나, ‘음악’, ‘예술에세이’, ‘문학에세이’와 같은 주제 관련 정보가 보강되면서 음악(67X)으로 정확히 분류되었다. 이와 같은 양상은 문학 분야에서도 확인되었는데, 한국문학(81X)의 「우주먼지에 관한 명상」, 「수상한 교장실」, 「초당 연애 비법서」 등의 도서에서도 서지요소가 단계적으로 보강됨에 따라 분류 성능이 향상되었다.

한편, 일부 강목에서는 키워드가 보강된 시점에 이미 최고 성능에 도달한 이후, 추가적인 서지요소 보강에도 불구하고 성능 변화가 나타나지 않는 경우가 관측되었다. 예를 들어 비교종교(21X)는 키워드의 보강만으로 F1 점수가 1.00에 도달하였는데, 이는 특정 강목의 경우에 키워드와 같은 핵심적인 서지요소만으로도 주제 판별에 효과적임을 보여준다.

지속적 성능 향상형은 표제만으로는 주제를

충분히 파악하기 어려운 예술, 기술과학, 문학 분야에서 주로 나타났으며, 보강된 서지요소가 주제 판단에 필요한 정보를 보완함에 따라 두드러진 성능 향상을 보인 것으로 해석된다.

4.3.2 제한적 성능 향상형

제한적 성능 향상형은 A모델에 비하여 C모델의 성능 향상이 나타났으나, 단계적인 보강 과정에서 일관된 성능 개선이 이루어지지 않은 경우로, 전체 61개 강목 중 19개 강목이 이에 해당하였다(〈표 6〉 참고).

모델별 분류 결과를 바탕으로 이러한 성능 변화의 양상을 살펴보면, 서지요소가 단계적으로 보강됨에 따라 성능이 서로 다른 방향으로 변동하는 두 가지 양상이 확인된다.

먼저, 문헌정보학(02X), 자연과학(40X), 기계공학(55X), 화학공학(57X) 등의 강목에서는 B모델에서 성능이 향상되었으나, C모델에서는 성능이 유지되거나 오히려 저하되는 양상이 나

〈표 6〉 서지요소 보강에 따른 분류 성능 변화(제한적 성능 향상)

강목	매크로 평균 F1 점수			빈도	
	A모델	B모델	C모델		
02X	문헌정보학	0.56	0.67	0.67	7
40X	자연과학	0.67	0.82	0.82	9
69X	오락, 스포츠	0.00	0.50	0.50	5
71X	한국어	0.76	0.90	0.90	13
85X	독일문학	0.62	0.79	0.79	21
87X	스페인문학 및 포르투갈문학	0.57	0.67	0.67	2
22X	불교	0.80	0.88	0.80	8
23X	기독교	0.72	0.98	0.95	21
35X	행정학	0.00	0.80	0.67	3
37X	교육학	0.72	0.90	0.87	40
55X	기계공학	0.29	0.86	0.57	4
57X	화학공학	0.00	0.55	0.25	5
90X	역사	0.40	0.73	0.62	5
92X	유럽	0.33	0.67	0.40	4
89X	기타 제문학	0.71	0.71	0.80	8
41X	수학	0.89	0.80	0.95	10
42X	물리학	0.67	0.60	0.73	5
45X	지학	0.50	0.33	0.80	3
86X	프랑스문학	0.73	0.71	0.86	18

타났다.

이러한 양상이 나타난 주요 원인은 책소개와 목차에 포함된 정보가 특정 강목의 핵심 주제를 보완하기보다, 핵심 주제와 직접적으로 관련되지 않은 내용까지 함께 반영되면서 분류 판단에 필요한 정보가 상대적으로 희석되었기 때문으로 보인다. 특히 책소개는 도서의 내용을 설명하는 과정에서 지나치게 포괄적인 정보나 다양한 주제와 관련된 내용을 함께 포함하는 경우가 많으므로, 핵심 주제와 직접적으로 관련되지 않은 정보가 분류 판단에 영향을 줄 수 있다. 예를 들어 기계공학(55X)의 「2024 이기적 운전면허 필기 문제은행 1」의 경우, 키워드가 보강되면서 분류 성능이 크게 향상되었으나, 책소개에 포함된 '도로교통법'과 같은 비주

제적 정보가 반영되면서 법률, 법학(36X)으로 오분류되었다. 또한 초기에는 영미문학(84X)으로 오분류된 유럽(92X)의 「고양이 대학살」은 키워드가 보강되면서 정확하게 분류되었다. 그러나 C모델에서는 책소개에 포함된 '단턴의 역사 방법론', '역사학 분야의 필독서'와 같은 비핵심적인 정보가 반영되면서 역사(90X)로 오분류된 것으로 나타났다.

또한 목차는 도서의 주제를 압축적으로 제시하기보다 세부 내용을 나열하는 경우가 많기 때문에, 문제집이나 수험서와 같이 여러 단원과 개념이 병렬적으로 제시되는 특징을 갖는 도서에서는 인접 분야의 용어가 분류 판단에 과도하게 반영될 가능성이 있다. 구체적인 사례로, 화학공학(57X)의 「2024 정나나의 화공

기사 실기 필답형+작업형」에서는 목차에 포함된 정보가 오분류에 영향을 미친 것으로 나타났다. 해당 도서의 목차에는 '증류', '촉매반응'과 같이 화학 분야와 관련성이 높은 용어가 포함되어 있었으며, 이로 인해 C모델에서는 화학 공학이 아닌 화학(43X)으로 오분류되었다.

다음으로, 수학(41X)과 프랑수문학(86X)와 같은 일부 강목은 B모델에서는 성능이 저하되거나 변화가 나타나지 않다가, C모델에서 성능이 향상되는 양상을 보였다. 이는 키워드만으로는 주제 판별에 필요한 정보가 충분하지 않았지만, 책소개와 목차의 보강을 통해 주제 파악에 필요한 단서가 보완되면서 성능이 향상된 결과로 해석된다.

따라서 제한적 성능 향상형은 서지요소 보강이 성능 향상에 기여할 수 있더라도, 보강되는 정보의 내용에 따라 그 효과가 달라질 수 있음을 보여준다. 특히 보강된 서지요소가 핵심 주제를 보완하는 경우에는 성능 향상으로 이어질 수 있으나, 핵심 주제와 직접적으로 관련되지 않은 정보가 함께 반영되는 경우에는 오히려 분류 판단을 혼란스럽게 하여 성능 저하를 초래할 수 있다.

4.3.3 성능 저하형

성능 저하형은 A모델과 비교하여 C모델의 성능이 낮아진 경우로, 전체 61개 강목 중 7개 강목이 이에 해당하였다(〈표 7〉 참조). 이 유형은 서지요소의 보강이 오히려 분류 성능을 저해한 경우에 해당한다.

특히 본 연구에서 활용된 출판계의 키워드는 도서의 주제를 설명하기 위해 다수 부여된 경우가 많아, 앞서 〈표 1〉의 사례에서 본 바와 같이 핵심 주제뿐 아니라 관련 주제, 상위 개념, 인접 분야의 표현까지 함께 포함되어 있다. 이로 인해 일부 강목에서는 키워드가 주제 정보를 보완하기보다 분류 판단에 혼동을 유발하여, 표제만을 활용한 경우보다 오히려 낮은 성능을 보인 것으로 판단된다.

정치학(34X), 신문, 저널리즘(07X)의 분류 사례는 키워드가 성능 저하에 대한 주요한 원인으로 나타난 대표적인 사례에 해당한다. 예를 들어, 신문, 저널리즘(07X)의 「가짜뉴스 판독법」은 A모델에서 정확하게 분류되었으나, 키워드에 포함된 '사회'와 '사회과학'과 같은 정보로 인하여 이후 모델에서 사회학, 사회문제(33X)로 오분류되었다. 정치학(34X)의 「좌파

〈표 7〉 서지요소 보강에 따른 분류 성능 변화(성능 저하)

강목	매크로 평균 F1 점수			빈도	
	A	B	C		
14X	경학	0.80	0.80	0.67	3
43X	화학	1.00	1.00	0.80	2
34X	정치학	0.55	0.48	0.52	17
07X	신문, 저널리즘	0.80	0.55	0.67	4
30X	사회과학	0.63	0.36	0.57	7
10X	철학	0.33	0.00	0.00	3
54X	건축, 건축학	0.73	0.67	0.67	5

의 길, 역시 A모델에서는 정확하게 분류되었지만, 키워드의 '기후문제', '환경운동'과 같은 정보가 반영되면서 사회학, 사회문제(33X)로 분류된 것이 확인된다. 한편, 경학(14X)과 화학(43X)과 같이 책소개에 포함된 정보가 지나치게 포괄적으로 반영되면서 성능이 저하된 강목도 확인되었다.

4.3.4 성능 불변형

성능 불변형은 A모델과 비교하여 C모델의 성능 변화가 없는 유형으로, 전체 61개 강목 중 6개 강목이 이에 해당하였다(〈표 8〉 참조). 이 유형은 서지요소 보강이 성능 변화로 이어지지 않은 경우로, 주제 식별에 필요한 정보가 이미 충분하거나 학습이 충분히 이루어지지 않은 강목에서 나타났다.

구체적인 사례를 살펴보면, 영어(74X)는 모든 모델에서 F1 점수 0.95를 유지하며 성능 변화가 나타나지 않았으며, A모델에서 이미 안정적인 분류 성능이 확보된 사례로 해석된다.

반면 식물학(48X), 도서학, 서지학(01X), 철학의 체계(13X) 또한 서지요소의 보강에도 불구하고 동일하거나 유사한 수준의 성능을 유지하였으나, 해당 강목들의 데이터 규모가 제한적

이기 때문에 이를 초기 단계에서 충분한 분류 성능이 확보된 결과로 단정하기는 어렵다. 또한 국방, 군사학(39X)과 언어(70X)은 모든 단계에서 F1 점수 0을 기록하였으며, 훈련 데이터의 불균형 문제로 인해 해당 강목에 대한 분류 성능이 제한적으로 나타난 것으로 파악된다.

4.4 논의

분석 결과에 따르면 서지요소의 보강은 전반적으로 주류 수준의 분류 성능 향상에 기여하는 것으로 나타났다. 표제만을 서지요소로 활용한 주제명 자동분류 연구에서 타 주류에 비해 비교적 낮은 성능을 보였던 예술(6XX) 분야(이용구, 2023)에서 124.24%의 높은 성능 향상률이 관측되었으며, 기술과학(5XX), 사회과학(3XX), 역사(9XX), 자연과학(4XX), 문학(8XX)에서 20% 이상의 유의미한 성능 개선이 확인되었다. 이는 출판계 서지요소의 보강을 통해 표제와 저자명만으로는 파악하기 어려운 주제 정보가 효과적으로 보완될 수 있으며, 그 효과가 특정 주제 영역에 한정되지 않고 KDC 분류체계 전반의 자동분류 성능 개선으로 이어질 수 있음을 시사한다.

〈표 8〉 서지요소 보강에 따른 분류 성능 변화(성능 불변)

강목	매크로 평균 F1 점수			빈도	
	A	B	C		
01X	도서학, 서지학	0.57	0.50	0.57	5
13X	철학의 체계	0.67	0.57	0.67	4
39X	국방, 군사학	0.00	0.00	0.00	2
48X	식물학	0.50	0.50	0.50	3
70X	언어	0.00	0.00	0.00	2
74X	영어	0.95	0.95	0.95	20

그러나 강목 수준의 분석 결과, 서지요소 보강의 효과가 모든 경우에 동일하게 나타나지는 않았다. 기존 연구에서 서지요소 보강 이후에 일부 주류 및 요목 수준에서 성능 저하가 보고된 바와 같이(강우진 외, 2026; 이용구, 2022), 일부 강목에서는 서지요소가 보강된 이후 성능이 정체되거나 오히려 저하되었음이 확인되었다. 이는 KDC의 자동분류에서 서지요소 보강의 효과가 단순한 주제 정보의 양적 확대에 의해 결정되는 것이 아니라, 서지요소의 구조적 특성과 강목의 주제적 특성에 따라 달라질 수 있음을 보여준다. 따라서 KDC 자동분류의 성능을 안정적으로 향상시키기 위해서는 다음의 세 가지 측면을 고려할 필요가 있다.

먼저, 서지요소의 구조적 특성을 고려하여 서지요소의 구성을 정교화할 필요가 있다. 본 연구에서 활용된 출판계의 키워드, 책소개, 목차는 모두 도서의 주제 식별에 기여할 수 있는 정보를 포함하지만, 주제 정보를 제공하는 범위와 구조는 서지요소별로 다르게 나타난다. 키워드는 도서의 핵심 주제를 직접적으로 포함하지만, 관련 주제, 상위 개념, 인접 분야의 표현까지 함께 포함될 수 있다. 책소개는 도서의 내용을 포괄적으로 설명하는 과정에서 핵심 주제와 직접적으로 관련되지 않은 정보를 포함할 수 있으며, 목차는 세부 개념이 병렬적으로 제시됨에 따라 인접 분야의 용어가 함께 나타날 수 있다는 점에서 구별된다.

이와 같은 서지요소의 구조적 특성은 주제 판단에 필요한 정보를 보완하는 데 기여할 수 있지만, 경우에 따라 핵심 주제 정보가 비핵심 정보나 인접 분야의 표현에 의해 희석되는 원인이 될 수 있다. 따라서 서지요소의 선별은 각

서지요소에 포함된 정보의 범위와 핵심 주제와의 관련성을 함께 고려하여 이루어질 필요가 있다. 키워드는 많은 용어를 활용하기보다 도서의 핵심 주제를 명확하게 드러내는 소수의 용어를 선별하는 것이 중요하며, 책소개와 목차 역시 핵심 주제 정보가 희석되지 않도록 자료 유형과 주제 분야에 따라 선택적으로 활용할 필요가 있다. 특히 문제집이나 수험서와 같은 도서에서는 목차 정보가 주제 판단을 보완하기보다 인접 분야의 용어를 과도하게 반영하게 하여 오분류를 유발할 수 있으므로, 목차의 활용은 제한적으로 이루어질 필요가 있다.

다음으로, 자동분류 성능 향상을 위해 추가적인 서지요소의 활용 가능성을 모색할 필요가 있다. 서지요소 보강의 효과가 각 서지요소가 지닌 구조적 특성에 따라 달라질 수 있음을 고려할 때, 키워드, 책소개, 목차 외에도 분류 성능 향상에 기여할 수 있는 추가 서지요소를 모색할 필요가 있다. 특히 본 연구에서는 서지요소로 '저자명'을 포함하였으나, 저자명 자체는 도서의 주제를 직접적으로 드러내는 정보로 작용하는 데 한계가 존재하였던 것으로 판단된다. 그러나 저자의 전공, 활동 분야, 주요 저작 등을 포함하는 '저자 소개'가 함께 보강될 경우, 저자명만으로는 파악하기 어려운 도서의 주제적 맥락을 보완할 수 있을 것으로 예상된다. 또한 '서평'은 도서의 주제, 대상, 내용적 특징을 외부 평가의 형태로 제시할 수 있으므로, 책소개나 목차와는 다른 방식으로 주제 판단에 필요한 정보를 제공할 가능성이 존재한다.

마지막으로, 강목의 주제적 특성을 반영한 분류 전략이 필요하다. 연구 결과에서 나타난 바와 같이, 표제만으로 주제를 파악하기 어려운

예술, 문학 분야에서는 보강된 서지요소가 주제 판단에 필요한 단서로 작용하여 성능 향상에 기여하였다. 반면 인접 분야와 용어가 중첩되거나 주제 범위가 넓고 경계가 모호한 강목에서는 보강된 정보가 핵심 주제와 인접 주제의 구분을 어렵게 하여 오분류를 유발할 수 있는 것으로 나타났다.

이러한 양상은 앞서 살펴본, 화학공학(57X)의 「2024 정나나의 화공기사 실기 필답형+작업형」이 화학(43X)으로 오분류된 사례에서 구체적으로 확인된다. 화학(43X)은 A모델과 B모델에서 F1 점수 1.00을 유지하였으나, C모델에서는 F1 점수가 0.80으로 하락하였다. 이는 실제 화학 도서를 올바르게 분류하지 못한 결과가 아니라, 화학공학 도서가 화학 범주로 잘못 분류되면서 정확률이 낮아졌기 때문으로 해석된다. 동시에 화학공학(57X)에서는 실제 화학공학 도서가 다른 강목으로 분류되면서 재현율 저하가 발생한 것으로 볼 수 있다.

따라서 강목의 주제적 특성을 고려할 때, 개별 강목을 서로 독립적인 범주로만 처리하는 방식에는 한계가 있다. 특히 KDC 분류체계와 같이 주제 간 경계가 모호한 경우에는 단순히 서지요소를 보강하는 것만으로는 자동분류 성능을 안정적으로 향상시키기 어렵다. 이러한 점에서, 기존 연구에서 제안한 바와 같이 주제와 하위 주제를 구분하는 계층형 분류 기반 모델이 효과적인 것으로 예상된다(곽철완, 2021; Golub et al., 2018). 이는 도서관 분류체계의 위계적 구조와도 부합하는 형태로(정연경, 2017), KDC의 주류-강목-요목 구조를 반영하여 주류 수준에서 분류 범위를 먼저 설정한 뒤, 강목 및 요목 수준에서 세부 주제를 단계적으로 구분할 수 있는

구조이다. 특히 유사하거나 관련성이 높은 강목 간에 용어가 중첩되는 경우에도 상위-하위 주제 관계를 함께 고려할 수 있어, 개별 강목을 독립적으로 분류하는 방식보다 분류 판단의 일관성을 높이는 데 기여할 수 있을 것으로 판단된다.

5. 결론

최근 들어, AI 기술의 발전과 함께 도서관계 자동분류의 활용 가능성이 점차 확대되고 있다. 그러나 KDC를 대상으로 한 자동분류 연구는 제한적으로 이루어져 왔으며, 표제와 같은 한정된 서지요소를 활용하거나 사회과학(3XX) 및 예술(6XX)과 같은 일부 주제 영역에 국한되어 수행되어 왔다(곽철완, 2021). 이에 본 연구는 KDC 전체 분류체계를 대상으로 자동분류의 범위를 확장하고, 출판계에서 수집한 데이터를 활용하여 서지요소를 보강함으로써 KDC 자동분류 성능 변화에 미치는 영향을 실증적으로 분석하고자 하였다. 특히 기존 연구에서 충분히 논의되지 않았던 서지요소 보강에 따른 강목별 성능 변화의 차이와 그 원인을 실제 분류 사례를 바탕으로 살펴봄으로써, KDC 자동분류의 성능 향상을 위한 시사점을 도출하고자 하였다.

이를 위해 본 연구에서는 '표제'와 '저자명'으로 구성된 기본 서지요소 세트 A를 기반으로, '키워드', '책소개', '목차'를 단계적으로 보강하여 서지요소 세트 B와 C를 구성하고, 각 세트에 KLUE-BERT 기반의 KDC 자동분류 모델을 적용하여 성능 변화를 분석하였다. 분석 결과에 따르면 서지요소의 보강은 편차가 존재하

였지만, 전반적으로 주류 수준의 분류 성능 향상에 기여하는 것으로 나타났다. 특히 모든 서지요소가 보강된 C모델에서 F1 점수가 최저 0.62에서 최대 0.92로 나타나, KDC 자동분류 모델이 분류자의 판단을 지원하는 보조 도구로 활용될 가능성이 있는 것으로 판단된다.

서지요소 보강에 따른 성능 향상의 편차를 보다 구체적으로 파악하기 위해 강목 수준에서 성능 변화 양상을 검토한 결과, 강목별 성능 변화는 지속적 성능 향상형, 제한적 성능 향상형, 성능 저하형, 성능 불변형의 네 가지 유형으로 구분되었다. 이를 통해 KDC 자동분류의 성능 향상이 단순히 얼마나 많은 서지요소를 보강하였는지에 따라 결정되는 것이 아니라, 특정 강목을 변별하는 데 필요한 핵심 주제 정보가 얼마나 적절하게 제공되는지에 영향을 받는다는 점을 확인하였다. 나아가 KDC 자동분류 성능 향상을 위해 고려해야 할 사항으로 서지요소 구성의 정교화, 추가적인 서지요소 활용 가능성 모색, 그리고 강목의 주제적 특성을 고려한 계층형 분류의 적용을 제시하였다.

한편, 본 연구는 데이터의 규모에 따른 한계가 존재한다. 본 연구는 5,882건의 소규모 데이터를 바탕으로 수행되었으며, 강목 간 데이터의 불균형이 존재하였다. 이에 전체 강목이 아닌 61개 강목을 대상으로 분석을 한정하여 수행하였으며, 일부 강목에 대한 성능 변화는 반

영되지 못하였다. 요목 수준에서는 분류 범주가 더욱 세분화되므로, 각 범주별 훈련 데이터가 충분히 확보되지 못하였기 때문에 요목 수준에서의 자동 분류 성능 변화 또한 검증하지 못하였다. 그리고 소규모 데이터의 한계로 인한 과적합을 방지하기 위해 EDA 기법을 적용하여 학습 데이터를 증강하였음에도 불구하고, 훈련 데이터와 검증 데이터의 분류 성능 차이가 비교적 크게 나타나 과적합 양상이 확인되었다. 따라서 본 연구 결과를 KDC 전체 분류체계에 일반화하는 데에는 제약이 있다.

따라서 향후 연구에서는 데이터 규모를 확충하여 KDC 자동분류 모델의 일반화 성능을 검증하고, 보다 세분화된 주제 단위에서의 자동 분류 성능을 확인하기 위해 분석 범위를 요목 수준까지 확장할 필요가 있다. 또한 모델 고도화 측면에서는 주류-강목-요목의 구조를 고려한 계층형 분류 방식을 적용하고, 서지요소별 특성을 개별적으로 반영할 수 있는 입력 구성 방식과 모델 구조 설계 방안을 함께 모색할 필요가 있다. 아울러 과적합 완화를 위한 추가적인 정규화 기법도 검토되어야 할 것이다.

본 연구의 결과는 KDC 뿐만 아니라 주제명 분류를 포함한 다양한 지식조직체계 기반 자동 분류 모델 개발을 위한 기초 자료로 활용될 것으로 기대된다.

참 고 문 헌

- 강우진, 나상오, 이종욱 (2026). 기술과학 분야 문헌의 DDC 자동 분류에 관한 연구. *한국문헌정보학회지*, 60(1), 173-194. <https://doi.org/10.4275/KSLIS.2026.60.1.173>

- 고영수, 이주희, 송민 (2021). 딥러닝 및 토픽모델링 기법을 활용한 소셜 미디어의 자살 경향 문헌 판별 및 분석. *한국비블리아학회지*, 32(3), 247-264. <https://doi.org/10.14699/kbiblia.2021.32.3.247>
- 곽철완 (2021). 공공도서관 도서 분류를 위한 머신러닝 적용 가능성 연구: 사회과학과 예술분야를 중심으로. *한국비블리아학회지*, 32(1), 133-150. <https://doi.org/10.14699/kbiblia.2021.32.1.133>
- 김정현 (2011). 한국의 자료조직 분야 연구동향 분석: 1970~2010. *한국도서관·정보학회지*, 42(3), 149-164. <https://doi.org/10.16981/kliss.42.3.201109.149>
- 노지현, 이미화, 이은주 (2023). *목록이론의 이해와 적용*. 서울: 한국도서관협회.
- 이용구 (2020). 목차 정보와 kNN 분류기를 이용한 사회과학 분야 도서 자동 분류에 관한 연구. *정보관리학회지*, 37(1), 1-21. <https://doi.org/10.3743/KOSIM.2020.37.1.001>
- 이용구 (2022). 국가서지를 활용한 주제명 자동분류 적용방안 연구 (11-1371029-000236-01). 국립중앙도서관.
- 이용구 (2023). BERT 모델을 이용한 주제명 자동분류 연구. *한국문헌정보학회지*, 57(2), 435-452. <https://doi.org/10.4275/KSLIS.2023.57.2.435>
- 정연경 (2017). *지식 정보 분류론*. 서울: 이화여자대학교출판문화원.
- Bodenhamer, J. (2023). The reliability and usability of ChatGPT for library metadata. Oklahoma State University. Available: <https://openresearch.okstate.edu/handle/20.500.14446/339626>
- Brzustowicz, R. (2023). From ChatGPT to CatGPT: the implications of artificial intelligence on library cataloging. *Information Technology and Libraries*, 42(3). <https://doi.org/10.5860/ital.v42i3.16295>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- Golub, K., Hagelbäck, J., & Ardö, A. (2018). Automatic classification using DDC on the swedish union catalogue. *Proceedings of the 18th European Networked Knowledge Organization Systems (NKOS 2018) Workshop*, 4-16.
- Golub, K., Suominen, O., Mohammed, A. T., Aagaard, H., & Osterman, O. (2024). Automated dewey decimal classification of swedish library metadata using Annif software. *Journal of Documentation*, 80(5), 1057-1079. <https://doi.org/10.1108/JD-01-2022-0026>
- Li, J. (2024) Area under the ROC Curve has the most consistent evaluation for binary classification. *PLOS ONE* 19(12): e0316019. <https://doi.org/10.1371/journal.pone.0316019>
- Noruzi, A. (2024). The use of artificial intelligence in knowledge organization and subject indexing. *Informology*, 3(1), 1-8.

- Park, S. J., Moon, J. H., Kim, S. D., Cho, W. I., Han, J. Y., Park, J., Song, C. S., Kim, J. S., Song, Y., Oh, T., Lee, J. H., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S. K., Lim, K. T., Lee, J., Park, K., Shin, J., Kim, S. H., Park, L., Oh, A., Ha, J. W., & Cho, K. (2021). KLUE: Korean language understanding evaluation. arXiv:2105.09680. <https://doi.org/10.48550/arXiv.2105.09680>
- Raschka, S. & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2 (3rd. ed.). Birmingham: Packt.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics* 2020, 8, 842-866. https://doi.org/10.1162/tacl_a_00349
- Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *Liber Quarterly: The Journal of European Research Libraries*, 29(1), 1-25. <https://doi.org/10.18352/lq.10285>
- Suominen, O., Juho, I., & Mona, L. (2022). Annif and Finto AI: Developing and implementing automated subject indexing. *JLIS.It*, 13(1), 265-282. <https://doi.org/10.4403/jlis.it-12740>
- Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 17)*. Curran Associates Inc., Red Hook, NY, USA, 6000-6010.
- Wei, J. & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv:1901.11196. <https://doi.org/10.48550/arXiv.1901.11196>

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- Chung, Yeon Kyoung (2017). *Introduction to Knowledge & Information Classification*. Seoul: Ewha Womens University Press.
- Kang, Woo Jin, Na, Sang o, & Lee, Jong Wook (2026). A study on automatic DDC classification of documents in technology. *Journal of the Korean Society for Library and Information Science*, 60(1), 173-194. <https://doi.org/10.4275/KSLIS.2026.60.1.173>
- Kim, Jeong Hyen (2011). An analytical study on research trends of resource organization in Korea: 1970~2010. *Journal of Korean Library and Information Science Society*, 42(3), 149-164. <https://doi.org/10.16981/kliss.42.3.201109.149>

- Ko, Young Soo, Lee, Ju Hee, & Song, Min (2021). Examining suicide tendency social media texts by deep learning and topic modeling techniques. *Journal of the Korean Biblia Society for Library and Information Science*, 32(3), 246-264.
<https://doi.org/10.14699/kbiblia.2021.32.3.247>
- Kwak, Chul Wan (2021). A study on applicability of machine learning for book classification of public libraries: Focusing on social science and arts. *Journal of the Korean Biblia Society for Library and Information Science*, 32(1), 133-150.
<https://doi.org/10.14699/kbiblia.2021.32.1.133>
- Lee, Yong Gu (2022). Study on the application of automatic subject classification using national bibliography (11-1371029-000236-01). National Library of Korea.
- Lee, Yong Gu (2023). A study on automatic classification of subject headings using BERT model. *Journal of the Korean Society for Library and Information Science*, 57(2), 435-452.
<https://doi.org/10.4275/KSLIS.2023.57.2.435>
- Lee, Yong-Gu (2020). A study on book categorization in social sciences using kNN classifiers and table of contents text. *Journal of the Korean Society for Information Management*, 37(1), 1-21. <https://doi.org/10.3743/KOSIM.2020.37.1.001>
- Rho, Jee Hyun, Lee, Mi Hwa, & Lee, Eun Ju (2023). *Cataloging Theory and Practice*. Seoul: Korean Library Association.