

# 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류\*

## Automatic Classification of Academic Articles Using BERT Model Based on Deep Learning

김인후 (In hu Kim)\*\*

김성희 (Seong hee Kim)\*\*\*

### 초 록

본 연구에서는 한국어 데이터로 학습된 BERT 모델을 기반으로 문헌정보학 분야의 문서를 자동으로 분류하여 성능을 분석하였다. 이를 위해 문헌정보학 분야의 7개 학술지의 5,357개 논문의 초록 데이터를 학습된 데이터의 크기에 따라서 자동분류의 성능에 어떠한 차이가 있는지를 분석, 평가하였다. 성능 평가척도는 정확률(Precision), 재현율(Recall), F 척도를 사용하였다. 평가결과 데이터의 양이 많고 품질이 높은 주제 분야들은 F 척도가 90% 이상으로 높은 수준의 성능을 보였다. 반면에 데이터 품질이 낮고 내용적으로 다른 주제 분야들과 유사도가 높고 주제적으로 확실히 구별되는 자질이 적을 경우 유의미한 높은 수준의 성능 평가가 도출되지 못하였다. 이러한 연구는 미래 학술 문헌에서 지속적으로 활용할 수 있는 사전학습모델의 활용 가능성을 제시하기 위한 기초자료로 활용될 수 있을 것으로 기대한다.

### ABSTRACT

In this study, we analyzed the performance of the BERT-based document classification model by automatically classifying documents in the field of library and information science based on the KoBERT. For this purpose, abstract data of 5,357 papers in 7 journals in the field of library and information science were analyzed and evaluated for any difference in the performance of automatic classification according to the size of the learned data. As performance evaluation scales, precision, recall, and F scale were used. As a result of the evaluation, subject areas with large amounts of data and high quality showed a high level of performance with an F scale of 90% or more. On the other hand, if the data quality was low, the similarity with other subject areas was high, and there were few features that were clearly distinguished thematically, a meaningful high-level performance evaluation could not be derived. This study is expected to be used as basic data to suggest the possibility of using a pre-trained learning model to automatically classify the academic documents.

키워드: 자동분류, 딥러닝, BERT 모델, 트랜스포머, 자연어처리  
automatic classification, deep learning, BERT model, transformer, natural language processing

\* 본 연구는 석사학위논문을 수정·요약한 것임.

\*\* 중앙대학교 문헌정보학과 대학원 석사과정(markaki@cau.ac.kr) (제1저자)

\*\*\* 중앙대학교 문헌정보학과 교수(seonghee@cau.ac.kr) (교신저자)

- 논문접수일자: 2022년 8월 21일 ■ 최초심사일자: 2022년 9월 7일 ■ 게재확정일자: 2022년 9월 13일
- 정보관리학회지, 39(3), 293-310, 2022. <http://dx.doi.org/10.3743/KOSIM.2022.39.3.293>

※ Copyright © 2022 Korean Society for Information Management

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

최근 들어 텍스트를 비롯한 다양한 유형의 데이터가 기하급수적으로 증가하고 빅데이터 기술과 기계학습 기술의 발전으로 문서의 자동 분류에 관한 연구도 활발히 진행되어 왔다.

4차산업혁명의 중심에 있는 인공지능 기술 중에서 기계학습을 비롯한 다양한 분류 알고리즘이 문서의 자동분류에 적용되어왔다. 김판준(2016; 2018; 2019)은 로지오 알고리즘, 나이브 베이즈, SVM, 랜덤포레스트 등의 다양한 알고리즘을 이용하여 분류의 성능을 높이고자 시도하였다. 또한 LDA 알고리즘 및 로지스틱 회귀, SVM, 의사결정 트리의 방법 등을 자동 분류에 활용하여 비교하는 연구들도 진행되어 왔다(배성호 외, 2020; 이상우 외, 2020). 특히 자연어 처리 분야에서 많은 연구가 이루어졌고 인상적인 성과들을 보여주고 있는데 문서에 내재된 의미를 이해하고, 요약하는 자연어 처리 분야에서 딥러닝 기반의 모델들이 괄목할 만한 성과를 보이고 있다(최윤수, 최성필, 2019; 박규현, 정영섭, 2021; 육지희, 송민, 2018; 황상흠, 김도현, 2020). 국외에서는 해당 연구자의 모국어를 사용한 다양한 연구결과가 있거나 새로운 알고리즘을 전형적인 영어 텍스트 데이터셋으로 평가하는 연구들이 진행되어왔다(Hikmah, Adi, & Sulistiyono, 2020; Okur & Sertbas, 2021; Asim et al., 2021; El-Alami, Alaoui, & Nahahi, 2021).

인공지능의 다양한 딥러닝 모델 중에서 특히 2018년에 발표된 BERT(Bidirectional Encoder Representations from Transformers) 모델은 트랜스포머 구조를 뼈대로 하는 언어모델로 자

연어 처리의 전반적인 영역에서 높은 성능을 보여준 모델이다(Devlin et al., 2019). 문서의 자동분류에 BERT 모델을 활용한 연구들을 살펴보면 황상흠, 김도현(2020)은 딥러닝 기반의 사전학습된 BERT모델을 이용하여 기술문서들로부터 자동으로 문서의 특징들을 추출한 후, 이를 문서 분류에 직접 활용하여 자동분류의 성능을 검증한 결과 어느 정도 의미 있는 예측 및 성능의 효과를 보여주었다. 심재권(2021)은 BERT 모델을 활용하여 초등학교 4-6학년의 욕설문장을 자동분류한 실험을 한 결과 온라인 플랫폼에서 문장의 자동분류를 통해 BERT 모델의 적용 가능성을 보여주었다. 이수빈 외(2021)는 공항장에 말뭉치 구축을 한 후 이를 활용하여 공항장에 경향 문서 자동분류 모델을 제안하였다. 김해찬술 외(2017)는 인공지능 기술 중 기계학습을 이용한 기록 텍스트의 자동분류 테스트를 수행한 결과를 통하여 기록물 분류 업무에의 인공지능 기술 적용 가능성을 확인하였다. Devlin et al.(2019)은 BERT 모델을 두 문장의 관계를 예측하는 NLI와 자연어 기반 질문에 대한 해당 부분을 찾는 것에 관한 벤치마크 데이터로 fine-tuning을 활용함으로써 효율성과 성능 면에서 매우 우수함을 보였다. 이처럼 선행연구들은 BERT 언어모델이 문서의 자동분류에 활용 가능성 및 성능에 있어서 우수함을 보여준 반면 학술 문헌의 데이터 크기에 따른 성능분석이라든지 문헌정보학을 대상으로 적용한 연구는 없었다. 이에 본 연구에서는 한국어 데이터로 학습된 BERT 모델을 기반으로 문헌정보학 분야의 문서를 자동으로 분류하는 딥러닝 모델의 예측성능을 확인하고 이를 바탕으로 한국어 BERT 모델의 적용 가능성을

살펴보았다. 이를 위해 문헌정보학 분야의 KCI 등재 논문 전체를 수집하여 7개 학술지의 5357개 논문의 초록을 13개의 주제로 중복 없이 단일 항목으로 분류하였다. 분류 항목은 한국연구재단의 학술연구분야분류표의 13가지 소분류명을 사용하였으며 KoBERT 모델을 사용하여 수집한 초록 데이터를 학습시켜 자동분류를 실험하였다. 수집된 데이터를 학습된 데이터의 크기에 따라서 자동분류의 정확도에 어떠한 차이가 있는지를 분석, 평가하였다. 이러한 연구는 추후 학술 문헌에서 지속적으로 활용할 수 있는 사전학습모델의 활용 가능성을 제시하기 위한 기초자료로 활용될 수 있을 것으로 기대한다.

## 2. 이론적 배경

### 2.1 트랜스포머

트랜스포머(transformer)란 어텐션(attention) 함수를 활용한 자연어 처리모델로, 단어 간의 유사도를 계산하여 텍스트의 맥락을 벡터로 생성하는 방식으로 데이터의 학습을 진행한다(윤희승, 정재은, 2021). 그동안 기계 번역 영역에서 우수한 성능을 보였던 딥러닝 모델들은 순환 신경망(recurrent neural network) 기반의 방법이었는 데 순환 신경망 계열의 모델들은 필연적으로 입력 데이터를 순차적으로 처리해야 하기 때문에 길이가 긴 시퀀스의 경우 학습 속도가 크게 떨어지는 문제가 있었다. 반면에 트랜스포머는 주어진 시퀀스를 한 번에 입력으로 받아 자연어 처리를 하도록 설계되어 이러한 단

점을 해결하였으며 번역 성능에 있어서도 기존의 순환 신경망 계열의 모델들보다 우수하였다.

트랜스포머는 셀프-어텐션 모델을 활용하여 입력을 한 번에 받아 입력된 단어 간의 관계를 분석한다. 이를 통해 문장 안에서 어떤 한 단어의 의미를 다른 단어들과의 관련성 정도를 활용하여 판단할 수 있다. 트랜스포머 모델은 인코더-디코더 구조이다. 인코더-디코더 구조는 시퀀스 투 시퀀스(sequence-to-sequence) 모델로 분류되는데, 문장을 입력 데이터로 받은 후 출력데이터로써 문장을 받기 때문이다(윤희승, 정재은, 2021). 영어를 한국어로 번역하는 경우를 예로 들자면, 인코더는 번역을 해야 되는 대상이 되는 영문 문장을 분석한다. 이후 여러 어텐션 함수로 묶인 인코더 층(encoder layer)을 통해 나온 기울기 값을 디코더(decoder)에 전달한다. 이때 디코더는 그 값을 받아서 영어로 된 문장에 해당하는 한국어 문장을 출력하게 한다. 이러한 구조의 트랜스포머는 기계 번역이나 자연어생성 등의 다양한 분야에서 적용되었다. 그리고 인코더나 디코더만 사용하는 방식으로 모델 연구가 진행되고 있는데, 이 중 각각 BERT나 GPT(Generative Pre-Training)가 있다. 이 중 BERT는 자연어 이해 분야의 실험 데이터 세트에 대해 높은 수준의 성능을 보여주면서, 다양한 자연어 처리(Processing)에 BERT 모델을 적용하고 있다(송의석, 김남규, 2021).

### 2.2 BERT 모델

최근 들어 자연어처리분야(Natural Language Processing)에서 트랜스포머 구조를 활용해 더욱 성능을 향상 시킨 새로운 모델들이 나타나고

있으며 대표적인 모델이 구글의 BERT 모델이다. BERT는 여러 개의 어텐션 헤드를 가진 수십 개의 레이어로 이루어진 트랜스포머 기반 언어모델로 사전학습, 전이학습을 통해 다양한 자연어 처리 태스크에 활용되고 있다(이치훈, 이연지, 이동희, 2020).

BERT는 대량의 텍스트로 사전학습(pre-training)시킨 뒤 적은 치수의 학습 데이터로 해당 모델을 파인튜닝(fine-tuning)하는 방식으로 학습된다. 특히, BERT의 사전훈련 기법은 주석처리가 되어있지 않은 대규모 텍스트를 적용하여 우수한 단어 표현(word representation)을 획득할 수 있는 장점이 있다. 이에 따라 파인튜닝 단계에서 별도의 모델 구조를 추가하지 않더라도 적은 학습 데이터로 높은 수준의 성능을 달성할 수 있다(이수빈 외, 2021). BERT의 학습원리를 살펴보면 사전학습 단계에서 입력내용 중 일부를 생략하고(Masking), 생략된 단어를 모델이 예측하는 방식으로 학습시킴으로써 단어가 포함된 문맥에 따라 그 단어의 임베딩 값을 변화시키도록 만들었다. 이러한 모델을 스탠포드 대학의 NLP 그룹에서 클라우드 소싱을 통해 만든 536개의 위키피디아 아티클에 대한 107,785개의 질문-대답 데이터셋인 SQuAD(SQuAD(The Stanford Question Answering Dataset) 데이터셋을 통해 학습시킨 결과, 최초로 자연어 처리 분야에서 인간을 뛰어넘는 성능을 보여주었다(성소윤, 최재용, 김경철, 2019).

BERT와 비슷한 언어 표현 모델로는 ELMo(Embeddings from Language Models)와 GPT(Generative Pre-Training)가 있는데 이들의 일반적인 언어모델은 앞의 N개의 단어를

가지고 N+1번째의 단어를 예측하는 모델이다. ELMo는 사전학습에 있어서 BERT와 유사하게 많은 양의 문장을 사용하였지만, 성능 면에서는 BERT 모델이 더 우수한 것으로 나타났다(Peters et al., 2018). BERT 이전의 또 다른 모델로 OpenAI에서 만든 GPT(Generative Pre-Training) 모델이 있는데 GPT 모델은 트랜스포머를 이용해 만들어졌지만 ELMo와 유사하게 사전학습 방식이 단순해서 BERT와 동일한 성능을 만들어내지 못하였다(성소윤, 최재용, 김경철, 2019). ELMo와 GPT와 같은 언어모델들은 단방향으로 학습이 이루어졌지만 BERT는 양방향 문맥을 전부 고려할 수 있는 트랜스포머의 인코더 부분을 사용하였다. BERT를 사전 훈련하기 위해 사용한 방법은 마스킹 된 언어 모델(Masked Language Model, MLM)과 다음 문장 예측(Next Sentence Prediction)이다. 마스킹 된 언어모델은 입력에 대해 무작위로 임의의 개수의 토큰들을 마스킹 시킨다. 그리고 이를 Transformer의 인코더에 입력으로 넣어주어 주변 단어의 문맥만을 보고 마스킹 된 단어를 예측하도록 한다(최용석, 이공주, 2020). 다음 문장 예측은 연속된 두 문장과 연속되지 않은 두 문장을 각각 입력에 넣어주어 두 문장이 서로 연속적인지 아닌지에 대해 예측한다. BERT 모델은 Transformer의 인코더를 여러 개 쌓아 구성된 모델로 주어진 맥락에 맞게 특정 단어의 임베딩 벡터를 출력하도록 학습된 모델이다. 대량의 문서 문치로 효과적인 임베딩이 학습되도록 사전 학습한 후 이렇게 학습된 모델을 다양한 하위 과제에 fine-tuning하여 좋은 성능을 달성했다. BERT 모델의 중요한 의미는 사전 학습한 모델의 구조를 유지한

채 다양한 하위 과제들에 대해 좋은 성능을 보였다(황상흠, 김도현, 2020).

이처럼 BERT는 사전학습 과정에서 대규모 텍스트로부터 단어 표현을 학습하며, 이후 파인튜닝을 통해서 얼마나 높은 품질의 단어 표현이 획득되는지가 결정된다. 이러한 차원에서 언어모델을 다양한 데이터로 사전학습하는 것이 최근 중요한 연구영역으로 주목받고 있다(이수빈 외, 2021). 이들 중 한글 데이터만을 사용하여 BERT를 사전학습한 다양한 유형의 모델들도 공개되고 있는데 대표적으로 KoBERT는 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치(corpus)를 기반으로 훈련된 모델이며, 한국전지통신연구원에서 출시한 KorBERT는 한글 백과사전과 뉴스 기사를 사용한 모델이다(한국전지통신연구원, 2019).

본 연구에서는 SKTBrain에서 공개한 한국어 BERT(KoBERT) 모델을 이용하여 학습하였다. 이 BERT 모델은 구글에서 개발한 원래의 BERT 모델과 동일한 구조로 이루어져 있다. 사실상 구글에서 공개한 다국어 지원 BERT 모델도 활용할 수 있지만, 한국어에 특화되어 있지 않기 때문에 한국어 데이터에 대해 최적의 성능을 보이지 않고 있다. 반면에 KoBERT 모델은 한국어 위키 데이터를 사용하여 500만 개의 문장과 5,400만 개의 단어를 학습 데이터로 사전학습되었다. 한국어 텍스트를 토큰화하기 위해 한글 위키와 뉴스 텍스트를 기반으로 문장단위 토큰나이저(Sentence Piece Tokenizer)를 따로 학습하여 제공하므로 본 연구에서 사용하였다.

### 3. 연구설계

본 연구에서는 딥러닝 BERT 모델을 사전학습모델로써 활용하여 학술 문헌 텍스트 데이터에서 자동분류의 성능을 분석하고자 하였다.

#### 3.1 데이터 수집 및 전처리

딥러닝 기반의 BERT 모델을 활용하여 문헌정보학 분야의 KCI 등재 논문 전체를 KCI 등재시점부터 2020년 12월까지 발행된 논문 7,187개 중 초록이 없거나 영문초록만 있는 경우를 제외한 7개의 학술지 5,357개의 논문이 자동분류 성능에 사용되었다. 데이터 분류기준은 한국연구재단(2016)의 KCI 논문 분류인 학술연구분야 분류표의 소분류명 13가지 범주를 사용하여 저자가 제목과 국문 초록, 키워드를 참고하여 사전분류를 수행하였으며 대상 데이터가 5,000개 이상이라 전문가의 검증을 따로 수행하지는 않았다. 자동분류 모델 학습을 위한 데이터의 형태는 한글 초록 데이터만을 사용하였는데 제목을 활용하지 않은 이유는 초록과 합칠 경우 문맥이 이어지지 않으며 제목만 활용하는 경우는 학습을 위한 데이터양으로는 부족하기 때문이다. 또한, 입력 길이 하이퍼 파라미터 조정을 통해서 데이터양의 변화를 위해서는 단일한 초록 데이터가 가장 적합한 형태를 가지고 있기 때문이다. 논문에 대한 13개의 분류명은 모두 숫자로 바꿔 기입한 데이터로 <표 1>과 같이 구성하였다. <표 1>에 보는 바와 같이 각 13개의 주제에 번호가 부여되었다. 예를 들면 '문헌정보학 일반'의 주제번호는 1이며 논문 수는 406개이고 이 중 103개의 논문이 평가데이터로 사용되었다.

〈표 1〉 분류명별 학습 및 평가데이터

	문헌정보학 일반 (0)*	기록관리/ 보존 (1)	서지학 (2)	도서관/ 정보센터경영 (3)	정보 서비스 (4)	정보자료/ 미디어 (5)	정보조직 (6)
논문 수	406 (103)**	648 (200)	795 (229)	949 (289)	702 (209)	175 (62)	488 (137)
	정보검색 (7)	디지털도서관 (8)	정보공학 (9)	계량정보학 (10)	정보교육 (11)	기타 문헌정보학 (12)	
논문 수	235 (66)	42 (11)	123 (29)	166 (60)	562 (189)	66 (24)	

\* ( )는 각 주제에 부여한 번호임

\*\* ( )는 평가데이터임

사전학습모델로는 딥러닝 모델인 BERT 모델을 사용하여 워드 임베딩을 학습시켰다. SK의 TBrain에서 오픈소스로 배포한 KoBERT<sup>1)</sup> 모델을 사용하여 수집한 데이터의 초록을 분류명과 같이 학습시켰다. 데이터의 70%는 학습데이터로 30%는 평가데이터로 주제의 키워드나 동향이 시간이 지남에 따라 변화가 있을 수 있기 때문에 특정 시점을 기준으로 정하지 않고 랜덤하게 나눠 학습시켰다.

초록 데이터를 학습시킬 때, KoBERT 모델의 감정 분석 작업 예제<sup>2)</sup>에서 사용한 64개의 입력 길이에서 시작하여 2배수로 늘려 하나의 문장의 토큰 개수를 64개(short model)와 128개(middle model), 256개(long model)로 나누어 데이터양의 변화량에 따른 성능을 분석하였다.

### 3.2 학습 모델 환경 및 평가척도

문헌정보학 분야의 학술논문 5,357개의 자동

분류의 성능분석을 위해 70%는 fine-tuning 방식으로 학습되었다. 30%는 평가데이터로 검증 과정을 진행하였다. 학습 환경은 Google colab pro를 사용하였으며 학습 횟수(epoch)는 13번, 학습률(learning rate)은 0.00004로 설정하였고 배치사이즈(batch size)는 64로 설정 하였다. 본 연구에서는 문서의 자동분류의 성능을 평가하기 위해 정확률(precision)과 재현율(recall)을 구하고 F 척도를 통해서 평가하였다. 하나의 문서가 자동분류가 실행되었을 때, 해당 문서가 참에 해당하고, 참으로 자동분류를 실행하면 True Positive에 해당하고 거짓으로 잘못 분류를 실행하면 False Negative에 해당한다. 하나의 문서가 실제 거짓에 해당하고, 자동분류가 거짓인 문서면 True Negative에 해당하고 실제 참일 경우에는 False Negative에 해당한다.

1) SKTBrain, KoBERT, [online][cited 2022.6.6.] <<https://github.com/SKTBrain/KoBERT>>2) SKTBrain, naver\_review\_classifications\_pytorch\_kobert.ipynb, [online][cited 2022.5.15.] <[https://colab.research.google.com/github/SKTBrain/KoBERT/blob/master/scripts/NSMC/naver\\_review\\_classifications\\_pytorch\\_kobert.ipynb](https://colab.research.google.com/github/SKTBrain/KoBERT/blob/master/scripts/NSMC/naver_review_classifications_pytorch_kobert.ipynb)>

〈표 2〉 성능평가척도

	실제 참	실제 거짓
예측 참	True Positive	False Positive
예측 거짓	False Negative	True Negative

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F\text{-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

정확률은 모델이 참이라고 예측한 모든 값 중에서 실제 참을 맞힌 경우의 비율로 참으로 바르게 맞힌 경우인 True Positive와 False를 잘못 예측한 False Positive인 경우의 합을 실제 맞힌 경우에 나눈 것이다. 재현율은 실제 정답을 True로 예측한 경우에 실제 True Positive와 True를 False로 잘못 분류한 False Negative인 경우를 더한 값을 나눈 것이다.

정확률은 모델의 예측값을 중심으로 평가한 것이고 재현율은 실제 값을 중심으로 평가한 값이다. F 척도는 양쪽의 평가지표의 조화평균으로 불균형한 데이터를 하나의 수치로 평가할 수 있다. 본 연구에서는 학술 문헌 집단의 정확률과 재현율, 그 둘의 조화평균인 F 척도를 사용하여 성능을 평가하였다.

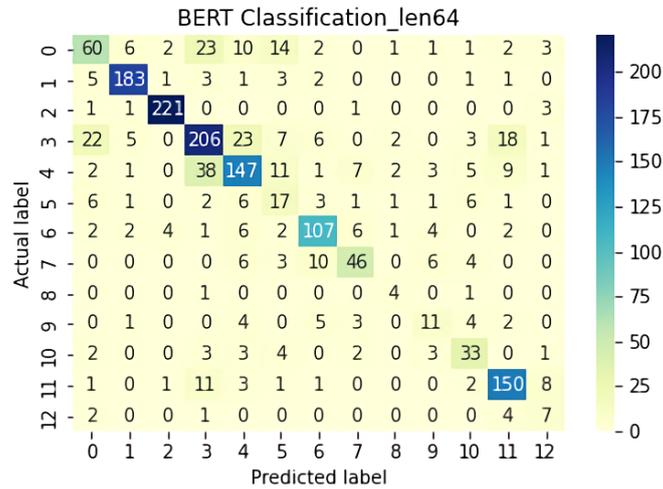
## 4. 데이터 분석

### 4.1 데이터 크기에 따른 혼돈 행렬 분석 (Confusion Matrix Analysis)

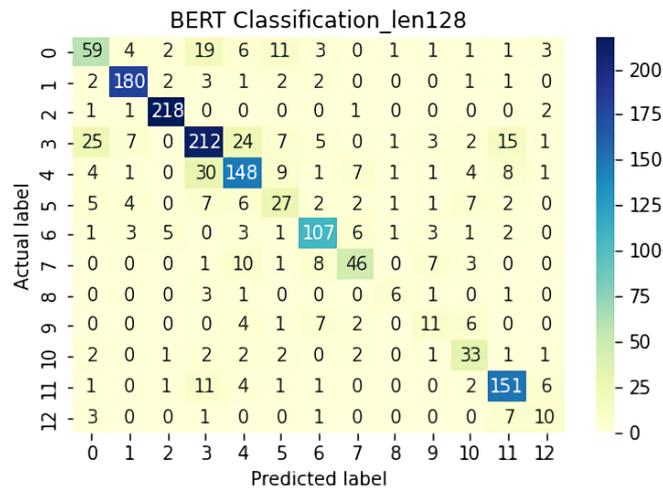
본 연구에서는 학술논문 초록의 입력 길이에

따라 데이터의 크기를 64개(short model)와 128개(middle model), 256개(long model) 등 3가지 유형으로 나누어 자동분류 실험을 시행하였다. 그 결과를 살펴보면 먼저 short model의 경우 〈그림 1〉에 보이는 바와 같이 명암이 짙을수록 해당 행렬 공간에 값이 높은 것을 보여준다. y축은 실제 분류에 해당하는 경우고 x축은 예측하여 분류를 실행한 경우의 값이다. 대각선에 위치한 숫자들이 모두 올바른 분류명을 맞춘 경우로 True Positive에 해당한다. 각 열에서 올바르게 분류한 경우를 제외한 경우는 모두 False Positive에 해당하고 각 행에서 올바르게 분류한 경우를 제외한 경우를 제외한 나머지 열에 해당하는 값들이 모두 False Negative에 해당한다.

13개의 주제 분야 중 3번 분류명인 ‘도서관/정보센터경영’ 주제에 해당하는 총 논문 수가 949개이고 이 중 289의 논문이 평가에 사용된 경우를 보면 〈그림 1〉에 보이는 바와 같이 논문에서 열을 중심으로 해당 분류명으로 분류하였을 경우에 같은 3번 행에 겹치는 206건의 값이 해당 논문을 올바르게 분류한 경우이고 나머지 행에 위치한 3번 열의 값들은 모두 거짓을 참이라고 판단한 경우로 False Positive에 해당한다. 좀 더 구체적으로 살펴보면 23개의 논문은 ‘문헌정보학 일반’에 분류되었고 38개 논문은 ‘정보서비스’에 분류되었다. 3번 주제와 4번 주제인 ‘도서관/정보센터경영’ 논문과 ‘정보서비스’ 주제 논문은 공통되게 잘못 분류한 빈도가 상대적으로 높은 것으로 나타났다. 이는 정보센터의 경영적인 측면과 정보를 서비스하는 내용의 유사성으로 해당 초록 내용이 유사하게 학습되었기 때문인 것으로 보인다. 0번에 해당



〈그림 1〉 Short Model의 혼동행렬



〈그림 2〉 Middle Model의 혼동행렬

하는 '문헌정보학일반'과도 유사성으로 인한 오답 분류가 상대적으로 높게 나타나는데 정보/도서관 정책 항목이 일반문헌정보학 분류에 들어가기 때문에 정책과 경영의 연관성으로 인해서 오답으로 분류된 것으로 판단된다. 11번에 해당하는 '정보교육'의 경우 3번인 '도서관/정보센터경영' 사이에서 많은 오분류가 있는데 이

는 도서관의 경영과 교육 관련 논문에서 도서관 프로그램 운영에 관한 내용들이 유사하기 때문에 이와 같은 결과가 나타난 것으로 판단된다. 12번에 해당하는 '기타문헌정보학'의 경우 7개를 정답으로 분류했지만 '정보교육'에 해당하는 8개 문서를 '기타문헌정보학'으로 잘못 분류하는 결과가 나오는 것을 보아 정보교육과 높은

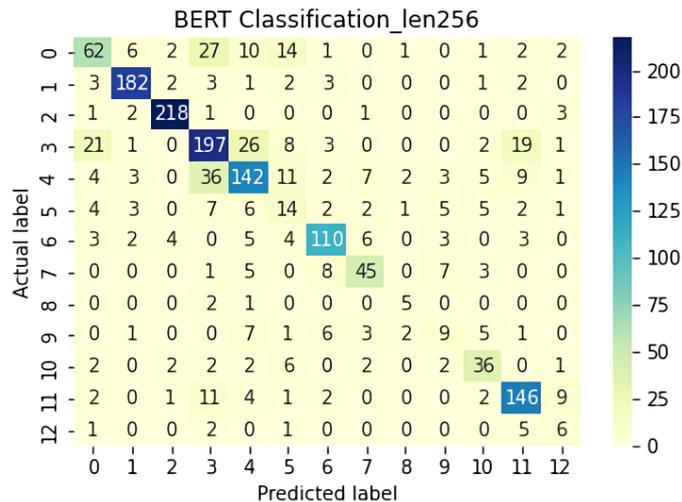
연관성을 가지며 데이터도 적기 때문에 데이터가 많은 정보교육으로 분류가 몰려 성능 저하가 일어난 것으로 판단된다.

데이터 치수가 middle model의 전반적인 혼돈 행렬(Confusion Matrix Analysis)의 모습은 short model과 큰 차이가 없는 것으로 나타났다. '도서관/정보센터경영'과 '정보서비스'의 유사성으로 인한 잘못된 분류도 동일하게 나타나고 '문헌정보학 일반'에 해당하는 분야와 '도서관/정보센터경영'에 해당하는 분야의 잘못된 분류도 동일하게 나타났다.

5번 항목인 '정보자료/미디어'는 short model보다 좀 더 높은 정확도를 보이고 있다. 올바르게 분류한 경우가 약 10건 증가하였고 정답으로 분류한 경우는 short model과 비슷한 수치를 보이고 오답 분류의 경우 연관성이 높아 상관관계를 많이 보이던 분류 항목들 사이의 오답 분류 수치는 약간 감소하였다. 12번 항목인 '기타문헌정보학'을 올바르게 분류한 경우가 10

건으로 늘었으며, 정보교육으로 잘못 분류한 경우가 줄어든 모습을 볼 수 있었다. '정보자료/미디어'가 short model보다 20건 정도 분류빈도가 상승하였고 나머지 항목들은 근소한 수치 내에서 변화가 있는 것으로 나타났다.

입력 길이를 256으로 가장 길게 잡은 long 모델은 BERT 모델의 학습에 가장 많은 데이터를 입력한 모델이지만 성능은 <그림 3>에서 보는 바와 같이 3가지 모델 중에서 가장 낮게 나타났다. short model보다 오답 분류의 경우가 근소하게 늘어난 경우가 있으며 정답 분류의 경우 short model, middle model 양쪽보다 낮게 나오는 경우도 있다. '도서관/정보센터경영'의 경우 short model보다 9건 더 줄어들었고 middle model보다 15건 줄어들었다. '기타문헌정보학' 부분에서는 short model에서 나온 것처럼 오답 분류가 정답 분류보다 많은 경우가 다시 발생했는데 더 심하게 나타난 모습을 보였다. Long model의 경우에도 전반적인 성능



<그림 3> Long Model의 혼동행렬

저하는 많은 데이터가 입력되어도 초록 데이터를 정제하지 않아 품질이 낮은 데이터이기 때문에 데이터의 양이 늘어도 분류성능의 개선으로 이어지지 않는 못했다는 점이다. 상대적으로 많은 양의 데이터가 학습이 진행될수록 쉽게 과적합이 일어나 학습 데이터 세트에만 최적화가 이루어지고 평가데이터에서는 성능 저하로 이어지게 되었다.

#### 4.2 데이터 크기에 따른 성능분석

데이터의 크기에 따라 Short model, Middle model, Long model 등 3가지 모델로 구분하여 분류에 대한 성능을 정확률, 재현율, F 척도를 측정하여 평가한 결과 <표 3>과 같이 나타났다. 일반적으로 F 척도와 평균적인 정확률 값이 0.5 이상이면 분류성능에 의미 있다고 판단을 한다(황상흠, 김도현, 2020). 먼저 Short model에서는 '정보자료/미디어', '디지털도서관', '정

보공학', '기타문헌정보학' 4개의 주제 분야에서 자동분류의 유의미한 성능을 보이지 못했다. F 척도가 0.5를 넘기는 경우는 정확률이나 재현율이 0.7 이상인 경우가 대부분으로 70%의 확률로 분류의 성능이 효과적으로 나타났다. 데이터의 입력 길이가 긴 경우와 짧은 경우의 분류성능의 차이는 10% 가까이 상승하는 경우도 있으며 반대로 성능 저하가 일어나는 경우도 있는 것으로 나타났다. '기록관리'와 '서지학'의 경우 정확률과 재현율이 모두 90%를 넘기며 F 척도도 90%를 넘기고 입력 길이에 상관없이 똑같은 성능을 보였다. 이는 주제의 성격이 다른 주제와 구분이 상대적으로 명확하여 식별능력이 우수하기 때문인 것으로 판단된다. '정보조직'은 middle model이 short model보다 재현율이 2% 상승하여 1%의 성능향상이 나타났고 long model에서는 정확률이 2% 상승하여 1%의 성능향상을 이루었다. Middle model에서는 다른 분류 항목들이 '정보조직' 주제의 논

<표 3> 3가지 모델에 따른 성능 평가

	Short model			Middle model			Long model		
	Precision	recall	F 척도	Precision	recall	F 척도	Precision	recall	F 척도
문헌정보학일반	0.582	0.48	0.526	0.572	0.531	0.551	0.601	0.484	0.536
기록관리/보존	0.915	0.915	0.915	0.9	0.927	0.913	0.91	0.914	0.912
서지학	0.965	0.973	0.969	0.951	0.977	0.964	0.951	0.964	0.958
도서관/정보센터경영	0.712	0.703	0.707	0.733	0.701	0.717	0.681	0.708	0.694
정보서비스	0.703	0.647	0.674	0.708	0.688	0.698	0.679	0.631	0.654
정보자료/미디어	0.274	0.377	0.317	0.435	0.421	0.428	0.225	0.269	0.245
정보조직	0.781	0.781	0.781	0.781	0.804	0.792	0.802	0.785	0.794
정보검색	0.696	0.613	0.652	0.696	0.605	0.647	0.681	0.652	0.666
디지털도서관	0.363	0.666	0.47	0.545	0.5	0.521	0.454	0.625	0.526
정보공학	0.379	0.366	0.372	0.379	0.354	0.366	0.31	0.257	0.281
계량정보학	0.55	0.647	0.594	0.55	0.702	0.616	0.6	0.654	0.626
정보교육	0.793	0.842	0.817	0.798	0.848	0.822	0.772	0.82	0.795
기타문헌정보학	0.291	0.5	0.368	0.416	0.454	0.434	0.25	0.4	0.307

문을 오답 분류하지 않아서 분류성능이 올라갔기 때문에 재현율이 상승하여 성능향상이 이루어졌지만, long model에서는 정확률이 향상되는 것으로 해당 항목에서 모델 자체의 분류 정확도가 향상된 것으로 판단된다.

‘계량정보학’에서는 입력 길이가 길어질수록 성능이 향상되는 것으로 나타났다. 정확률에서는 short model과 middle model이 동일한 수치를 가지고 Long model에서는 5% 상승하였다. 재현율에서는 short model과 long model이 65%의 수치를 가지며 middle model이 70%의 성능 수치를 가진다. short model에서 59%의 성능이 middle model에서는 61%, long model에서는 62%로 상승하였다. 이는 데이터의 양이 적어도 데이터 품질이 좋은 경우로 해당 항목의 논문들은 계량정보학에 관계된 명확한 키워드를 사용한 경우로 분류가 명확하여 입력한 데이터의 양이 많아질수록 성능향상이 나타난 것으로 보인다. ‘도서관/정보센터경영’에서 재현율의 경우 세 가지 모델 모두 비슷한 수치를 가지고 있지만, 정확률은 오히려 long model에서 약간 감소하는데 이는 명확하게 분류되어 학습에 좋은 품질이 아닌 데이터가 들어가서 오히려 모델의 예측을 떨어뜨린 것으로 볼 수 있다. ‘정보서비스’ 분야의 경우 정확률이 short model, middle model에서는 변화가 없었지만, long model에서는 하락하는 것으로 데이터가 많아져 오히려 예측성능의 하락을 불렀고 재현율은 middle model에서 4% 향상되며 해당 모델이 ‘정보서비스’ 항목에서 69%의 성능을 보였다.

‘정보자료/미디어’는 short model과 middle model 사이에서는 10%의 성능향상을 보여 정

확률이 16% 상승, 재현율이 5% 상승하였다. 입력되는 데이터가 늘어나서 일시적인 성능향상이 보였는데 반대로 long model에서는 정확률과 재현율이 모두 22%, 26%로 크게 떨어져 해당 분류 항목의 학습 데이터들의 품질이 좋지 못하다는 것을 알 수 있었다.

‘정보공학’은 short model과 middle model 사이에서는 재현율이 1% 줄어서 middle model의 성능이 0.5% 떨어졌고 long model은 정확률이 6%, 재현율이 10% 하락하여 9% 가까이 성능이 하락하였다. 123개의 데이터 숫자로 학습 데이터의 양은 성능 대비 부족한 편은 아니며 long model에서 큰 폭의 성능 하락은 데이터의 품질이 낮기 때문인 것으로 판단된다.

‘정보교육’은 short model과 middle model 사이에서는 0.5%의 성능향상이 있는데 정확률과 재현율에서 0.5%씩 상승이 있어 81%, 82%의 성능 평가를 기록하였다. Long model에서는 2%씩 성능의 하락이 있어 79%의 성능을 기록하였는데 이미 유의미한 성능을 보이는 상황이어서 데이터의 품질에 문제가 있는 것으로 보이지 않지만, long model로 토큰 집합의 입력 길이를 256개로 늘리는 순간 성능이 하락하였다. ‘정보교육’에 해당하는 논문의 초록 데이터들의 평균적인 토큰 집합의 크기가 213개이기 때문에 뒷부분에 위치한 단어 정보와 문맥 정보가 분류의 성능에 악영향을 미쳤다고 볼 수 있다.

‘기타문헌정보학’은 short model과 middle model 사이에서 약 8%의 성능향상이 일어났는데 정확률이 12% 상승하고 재현율이 5% 하락했기 때문에 43%의 성능 평가를 보인다. Long model에서는 정확률은 short model보다 낮고

재현율은 middle model보다 낮아 전반적으로 제일 낮은 수치를 기록하였다. 적은 학습 데이터가 확장되면서 성능이 향상되었지만 데이터의 품질이 떨어지기 때문에 오히려 long model에서 분류의 성능이 낮아진 것을 볼 수 있다.

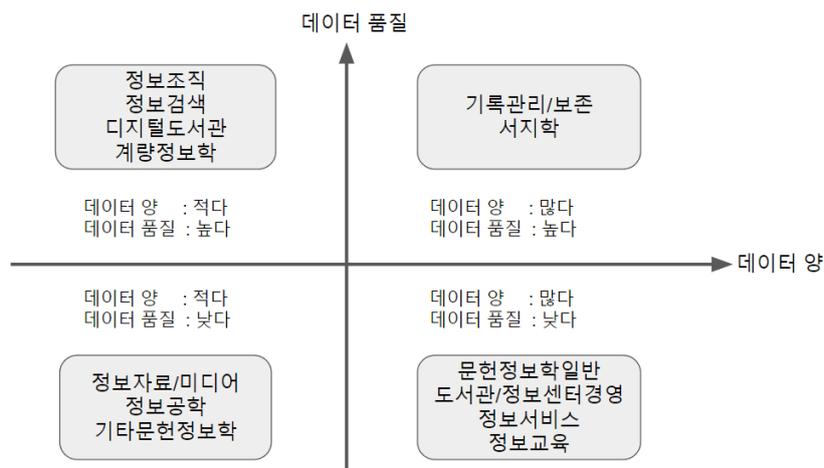
이상의 내용을 요약하면 데이터 크기와 F 척도의 상관관계는 short model에서 middle model로 변화량은 상승 한 반면 long model에서는 대부분 성능이 감소하는 것으로 나타났다. 이는 각 논문의 초록을 토큰화시키면 앞에서부터 model의 필요한 토큰 개수에 맞추어 잘리는데 초록의 앞부분은 논문의 주제가 잘 표현하는 문장들로 구성이 되어있어서 데이터의 크기가 어느 정도까지는 학습 데이터가 많아 성능이 향상되어 middle model에서는 F 척도가 향상된 것으로 판단된다. 반면에, long model에서는 대부분 성능이 하락하는데 토큰이 초록 전체의 내용 중 뒤에 나오는 문장들이 그 논문을 대표하기보다는 미래에 관한 내용 및 의의 등이 포함되어서 문헌정보학의 다양한 분야를 포

괄하기 때문으로 추측된다. 이 부분은 앞으로 좀 더 연구해야 할 필요가 있다.

### 4.3 데이터의 크기와 품질에 따른 비교 분석

이상에서 분석한 13가지 항목 전체를 분석한 결과 학습 데이터의 크기와 품질 두 가지 측면을 중심으로 성능 평가의 변화를 분석한 결과 <그림 4>와 같이 나타났다. 여기서 데이터의 품질은 초록 내용이 해당 분류명에 적합하고 정확하게 부여되었는지를 통해서 데이터 품질의 높고 낮음을 설정하였다. 모든 데이터에서 한글 초록이 없는 데이터는 삭제하였기 때문에 데이터의 완전성 측면에서는 동일한 수준을 유지하였다.

먼저 1 사분면에 위치한 ‘기록관리/보존’, ‘서지학’의 경우 90%를 넘기는 높은 수치와 동시에 입력 길이와 상관없이 일관된 성능 평가를 보여 데이터의 양과 품질이 모두 보장된 가장



<그림 4> 데이터 크기 및 품질에 따른 분류

이상적인 형태를 보여주었다. 2 사분면에 해당하는 경우는 '정보조직', '디지털도서관', '계량정보학' 주제의 경우 입력 길이가 길어질수록 성능 평가가 상승하였다. 다만 '정보검색'의 경우 middle model에서 다소 떨어지는 것으로 나타났으나 3 사분면에 해당하는 주제들에 비해 0.5 이상의 성능을 나타내서 자동분류가 가능하므로 2 사분면에 배치했다. 2 사분면에 해당하는 항목들은 모두 특정 전문용어가 자주 등장하여 데이터의 품질이 높다고 판단할 수 있는 항목들이라 할 수 있다. 데이터의 양에서는 적은 숫자를 가지고 있지만, F 척도가 계속 상승하는 형태를 띠고 있어 short model에서 long model까지 입력 길이를 지속적으로 늘려도 성능이 계속 상승하였다. 이는 해당 분류 항목 수준의 데이터 품질이 전제된다면 단순한 데이터의 확장으로 높은 성능을 기대할 수 있는 것으로 판단된다.

3 사분면에 위치한 '정보자료/미디어', '정보공학', '기타문헌정보학' 등 3가지 항목은 50% 미만의 성능 평가를 보여 유의미한 자동분류에 도달하지 못했다고 판단된 항목들로 '정보자료/미디어'와 '기타문헌정보학'은 middle model로의 입력 길이 증가로 10%에 가까운 성능향상을 보였지만 long model에서는 short model에서 10% 가까이 하락한 성능을 보였다. '정보공학'에서는 길이가 증가할수록 성능이 하락하였다. 이들의 공통점은 데이터의 품질이 좋지 못하다는 것으로 내용적으로 다른 항목들과 유사도가 높고 확실히 구별되는 자질이 적어 오답 분류가 많이 발생할 수밖에 없으며 데이터의 숫자 또한 많지 않기 때문에 유의미한 성능 평가가 도출되지 못한 것으로 판단된다.

데이터 품질에 문제가 있으면 유의미한 자동분류 성능을 나타낸 항목에서도 문제가 발생할 수 있는데 4 사분면에 있는 항목들이 여기에 해당한다. 80%의 높은 성능을 보인 '정보교육'도 long model을 사용하면서 오히려 79%로 성능이 근소하게 하락하는 모습을 보였다. '도서관/정보센터경영', '정보서비스' 또한 70%에 가까운 높은 성능 평가를 보였지만 long model에서는 2% 가까이 하락한 성능이 나타났다. '문헌정보학일반' 항목도 middle model에서는 성능향상이 이루어졌지만, long model에서 다시 성능이 하락하였는데 그래도 short model보다는 높은 성능을 기록하고 있다.

이들의 공통점은 유의미한 자동분류가 가능하며 middle model에서는 성능향상이 나타나지만, long model에서 성능 하락이 나타난다는 것이다. 이는 데이터 품질에 문제가 있는 것으로 추가적인 데이터양의 증가로 인해서 오히려 학습에 방해(noise)가 발생하여 자동분류의 성능이 하락한 것으로 판단된다. 입력 길이가 긴 모델이기 때문에 초록 데이터에서 뒷부분에 위치한 문장 정보들이 포함되어 학습이 진행되는데 데이터 품질이 낮은 경우, 논문 초록의 뒤쪽에 위치한 문장 정보가 성능 하락의 원인이 된 것으로 보인다.

이들 항목은 다른 분류 항목과 내용적으로 많은 상관관계를 보인 항목들인 동시에 3가지 항목 모두 long model에서 정확률이 약 3% 감소한 모습을 보여주었다. 초록의 뒷부분에 위치한 문장들에서 상관관계가 있는 다른 분류 항목의 내용이 등장하여 해당 분류 항목의 정답 분류성능이 하락하였다고 볼 수 있다. 데이터의 품질이 뒤에 위치한 문장이 추가되면서

더욱 하락한 것이다. 결국, 올바른 분류의 성능을 높이려면 데이터의 품질이 전제되어야 할 것으로 보인다.

## 5. 결론

빅데이터 기술과 기계학습 기술의 발전으로 현대 사회에서 다루는 문헌의 크기는 점차 크고 복잡해졌다. 학술 문헌 또한 점차 데이터의 양이 증가하고 있지만, 한국어 학술 문헌을 대상으로 딥러닝을 사용한 자동분류 연구는 거의 이루어지지 않았다.

이에 본 연구에서는 학술 문헌을 자동 분류하기 위한 자질 추출을 사전학습모델인 BERT 모델을 사용하였을 때 자동분류의 가능성을 확인하고 BERT 모델의 활용 가능성을 탐색하고자 하였다. 본 연구는 문헌정보학 분야의 KCI 등재 논문 전체를 수집하여 7개 학술지의 5357개 논문의 초록 데이터로 문헌을 13개의 단일 항목으로 분류하여 학습 데이터를 구성하였다. 분류 항목은 한국연구재단의 학술연구분야분류표의 13가지 소분류명을 사용하였다. KoBERT 모델을 사용하여 수집한 초록 데이터를 학습시켜 자동분류를 실험하였다. 분류작업에 사용된 기법은 단순한 단층 퍼셉트론을 사용하여 BERT 모델의 성능에 집중하였다. 학습 과정에서 하나의 문서에 입력되는 토큰의 길이에 구분을 주어 데이터의 양과 품질에 따라서 분류 정확도의 변화를 분석하였다. 토큰의 입력 길이를

64, 128, 256 세 가지로 나누어 학습시킨 모델에서 동일한 데이터를 검증하여 각각의 성능을 평가하였다.

성능 평가척도는 정확률(Precision), 재현율(Recall), F 척도를 사용하였고 분류 결과는 세 가지 모델 중에서 입력 길이가 64인 모델만 9개를 유의미하게 분류하였고 나머지 모델들은 10개를 유의미하게 분류하였다. 각 모델의 성능 평가를 분석하여 데이터의 품질과 양에 따라서 성능이 변화하였는데 4가지 유형으로 나타났다. 데이터의 양이 많고 품질이 높은 항목들은 F 척도가 90% 이상으로 높은 수준의 성능을 보였다. 데이터양이 적어도 데이터 품질이 좋거나 데이터 품질이 낮아도 데이터양이 많으면 F 척도가 50% 이상으로 자동분류가 유의미하게 나타났다. 데이터의 품질이 좋은 경우 입력 길이가 늘어나면 성능이 향상하고 데이터의 품질이 낮을 때는 입력 길이가 늘어나면 성능이 약간 하락하였다.

결론적으로 데이터의 양과 품질에 따라서 BERT 모델을 사용한 자동분류에서 유의미한 성능이 나타났으며 5,000개 이상의 논문 초록 집합이라는 빅데이터 형태의 학술 분야 텍스트 데이터에서 사전학습모델의 활용 가능성을 확인하였다. 문헌정보학 안에서 문헌을 분류하거나 서지정보의 생성 혹은 분류체계를 정립할 때, 해당 분야의 이상적인 단어 벡터가 존재하면 빅데이터 형태의 데이터 처리에서 효과적인 도구로써 활용이 가능할 것으로 기대한다.

## 참 고 문 헌

- 김관준 (2016). 기계학습에 기초한 자동분류의 성능 요소에 관한 연구. 정보관리학회지, 33(2), 33-59.  
<https://doi.org/10.3743/KOSIM.2016.33.2.033>
- 김관준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 35(2), 37-62. <https://doi.org/10.3743/KOSIM.2018.35.2.037>
- 김관준 (2019). 랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 36(2), 57-77. <https://doi.org/10.3743/KOSIM.2019.36.2.057>
- 김해찬술, 안대진, 임진희, 이해영 (2017). 기계학습을 이용한 기록 텍스트 자동분류 사례 연구. 정보관리학회지, 34(4), 321-344. <https://doi.org/10.3743/KOSIM.2017.34.4.321>
- 박규환, 정영섭 (2021). KoBERT를 사용한 한국어 일상 주제 분류, 2021년 한국컴퓨터종합학술대회 논문집, 1735-1737.
- 배성호, 구자일, 박찬봉, 김정수 (2020). 국내 군사학 학술논문 주제 분류를 위한 잠재 토픽 모델링. 한국군사학논집, 76(2), 181-216. <http://doi.org/10.31066/kjmas.2020.76.2.008>
- 성소윤, 최재용, 김경철 (2019). 트랜스포머를 이용한 향상된 댓글 생성에 관한 연구. 한국게임학회 논문지, 19(5), 103-113. <https://doi.org/10.7583/JKGS.2019.19.5.103>
- 송의석, 김남규 (2021). 사전학습 언어모델을 활용한 트랜스포머 기반 텍스트 요약. 경영과 정보연구, 40(4), 31-47. <http://doi.org/10.29214/damis.2021.40.4.002>
- 심재권 (2021). BERT를 활용한 초등학교 고학년의 욕설 문장 자동 분류방안 연구. 창의정보문화 연구, 7(2), 91-98. <http://www.doi.org/10.32823/jcic.7.2.202105.91>
- 육지희, 송민 (2018). 토픽모델링과 딥러닝을 활용한 생의학 문헌 자동분류 기법 연구. 정보관리학회지, 35(2), 63-88. <https://doi.org/10.3743/KOSIM.2018.35.2.063>
- 윤희승, 정재은 (2021). 효율적인 트랜스포머를 이용한 팩트체크 자동화 모델. 한국정보통신학회논문지, 25(9), 1275-1278. <https://doi.org/10.6109/jkiice.2021.25.9.1275>
- 이상우, 권정혁, 김남, 최형도, 김의직 (2020). 전자과 인체영향 연구논문에 대한 연구형태 자동분류 연구. 한국전자과학회논문지, 31(10), 839-842.  
<https://doi.org/10.5515/KJKIEES.2020.31.10.839>
- 이수빈, 김성덕, 이주희, 고영수, 송민 (2021). 딥러닝 자동분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석. 정보관리학회지, 38(2), 153-172.  
<https://doi.org/10.3743/KOSIM.2021.38.2.153>
- 이치훈, 이연지, 이동희 (2020). 사전 학습된 한국어 BERT의 전이학습을 통한 한국어 기계 독해 성능 개선에 관한 연구. 한국IT서비스학회지, 19(5), 83-91.  
<https://doi.org/10.9716/KITS.2020.19.5.083>

- 최용석, 이공주 (2020). 트랜스포머와 BERT로 구현한 한국어 형태소 분석기의 성능분석. 정보과학회 논문지, 47(8), 730-741. <http://doi.org/10.5626/JOK.2020.47.8.730>
- 최윤수, 최성필 (2019). 기술용어 분산표현을 활용한 특허 문헌 분류에 관한 연구. 한국문헌정보학회지, 53(2), 179-199. <https://doi.org/10.4275/KSLIS.2019.53.2.179>
- 한국연구재단 (2016). 학술연구분야분류표.  
출처: [https://www.nrf.re.kr/biz/doc/class/view?menu\\_no=323](https://www.nrf.re.kr/biz/doc/class/view?menu_no=323)
- 한국전자통신연구원 (2019). KorBERT. 출처: [https://aiopen.etri.re.kr/service\\_dataset.php](https://aiopen.etri.re.kr/service_dataset.php)
- 황상흠, 김도현 (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214. <https://doi.org/10.7838/jsebs.2020.25.1.203>
- Asim, M. N., Ghani, M. U., Ibrahim, M. A., Mahmood, W., Dengel, A., & Ahmen, S. (2021). Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification. *Neural Computing and Applications*, 33, 5437-5469. <https://doi.org/10.1007/s00521-020-05321-8>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805>
- El-Alami, F., El Alaoui, S. O., & Nahnahi, N. E. (2021). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences*, 2021, 1-7. <http://doi.org/10.1016/j.jksuci.2021.02.005>
- Hikmah, A., Adi, S., & Sulistiyono, M. (2020). The best parameter tuning on RNN layers for inonesian text classification. *Proceedings 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems*, 94-99. <https://doi.org/10.1109/ISRITI51436.2020.9315425>
- Okur, H. I. & Sertbas, A. (2021). Pretrained neural models for turkish text classification. *Proceeding of 2021 6th International Conference on Computer Science and Engineering*, 174-179. <https://doi.org/10.1109/UBMK52708.2021.9558878>
- Peters, M. E., Neumann M., Iyyer M., & Gardner M. (2018). Deep Contextualized Word Representations. <https://arxiv.org/abs/1802.05365>

• 국문 참고문헌에 대한 영문 표기  
(English translation of references written in Korean)

Bae, Seongho, Ku, Xyle, Park, Chanbong, & Kim, Jungsu (2020). A latent topic modeling

- approach for subject summarization of research on the military art and science in South Korea. *Korean Journal of Military Art and Science*, 76(2), 181-216.  
<http://doi.org/10.31066/kjmas.2020.76.2.008>
- Choi, Yongseok & Lee, Kong Joo (2020). Performance analysis of Korean morphological analyzer based on transformer and BERT. *Journal of Korean Institute of Information Scientists and Engineers*, 47(8), 730-741. <http://doi.org/10.5626/JOK.2020.47.8.730>
- Choi, Yunsoo & Choi, Sung-Pil (2019). A study on patent literature classification using distributed representation of technical terms. *Journal of the Korean Society for Library and Information Science*, 53(2), 179-199. <https://doi.org/10.4275/KSLIS.2019.53.2.179>
- Electronics and Telecommunications Research Institute (2019). KorBERT. Available:  
[https://aiopen.etri.re.kr/service\\_dataset.php](https://aiopen.etri.re.kr/service_dataset.php)
- Hwang, Sangheum & Kim, Dohyun (2020). BERT-based classification model for Korean documents. *Journal of Society for e-Business Studies*, 25(1), 203-214.  
<https://doi.org/10.7838/jsebs.2020.25.1.203>
- Kim, Hae-Chan-Sol, An, Dae-Jin, Yim, Jin-Hee, & Lieh, Hae-Young (2017). A study on automatic classification of record text using machine learning. *Journal of the Korean Society for Information Management*, 34(4), 321-344. <https://doi.org/10.3743/KOSIM.2017.34.4.321>
- Kim, Pan-Jun (2016). An analytical study on performance factors of automatic classification based on machine learning. *Journal of the Korean Society for Information Management*, 33(2), 33-59. <https://doi.org/10.3743/KOSIM.2016.33.2.033>
- Kim, Pan-Jun (2018). An analytical study on automatic classification of domestic journal articles based on machine learning. *Journal of the Korean Society for Information Management*, 35(2), 37-62. <https://doi.org/10.3743/KOSIM.2018.35.2.037>
- Kim, Pan-Jun (2019). An analytical study on automatic classification of domestic journal articles using random forest. *Journal of the Korean Society for Information Management*, 36(2), 57-77. <https://doi.org/10.3743/KOSIM.2019.36.2.057>
- Lee, Chi-Hoon, Lee, Yeon-Ji, & Lee, Dong-Hee (2020). A study of fine tuning pre-trained Korean bert for question answering performance development. *Journal of Information Technology Services*, 19(5), 83-91. <https://doi.org/10.9716/KITS.2020.19.5.083>
- Lee, Sang-Woo, Kwon, Jung-Hyok, Kim, Nam, Choi, Hyung-Do, & Kim, Eui-Jik (2020). Research category classification for scientific literature on human health risk of electromagnetic fields. *The Journal of Korean Institute of Electromagnetic Engineering and Science*, 31(10), 839-842. <https://doi.org/10.5515/KJKIEES.2020.31.10.839>

- Lee, Soobin, Kim, Seongdeok, Lee, Juhee, Ko, Youngsoo, & Song, Min (2021). Building and analyzing panic disorder social media corpus for automatic deep learning classification model. *Journal of the Korean Society for Information Management*, 38(2), 153-172.  
<https://doi.org/10.3743/KOSIM.2021.38.2.153>
- National Research Foundation of Korea (2016). the classification table of academic research fields. Available: [https://www.nrf.re.kr/biz/doc/class/view?menu\\_no=323](https://www.nrf.re.kr/biz/doc/class/view?menu_no=323)
- Park, Kyu Hwon & Jeong, Young-Seob (2021). Korean daily conversation topics classification using KoBERT. *Proceedings of Korea Computer Congress 2021*, 1735-1737.
- Seong, So-yun, Choi, Jae-yong, & Kim, Kyoung-chul (2019). A study on improved comments generation using transformer. *Journal of Korea Game Society*, 19(5), 103-113.  
<https://doi.org/10.7583/JKGS.2019.19.5.103>
- Shim, Jaekwoun (2021). A study on automatic classification of profanity sentences of elementary school students using BERT. *Journal of Creative Information Culture*, 7(2), 91-98.  
<http://www.doi.org/10.32823/jcic.7.2.202105.91>
- Song, Euseok & Kim, Namgyu (2021). Transformer-based text summarization using pre-trained language model. *Management & Information Systems Review*, 40(4), 31-47.  
<https://doi.org/10.29214/DAMIS.2021.40.4.002>
- Yuk, Jee Hee & Song, Min (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. *Journal of the Korean Society for Information Management*, 35(2), 63-88.  
<https://doi.org/10.3743/KOSIM.2018.35.2.063>
- Yun, Hee Seung & Jung, Jason J. (2021). Automated fact checking model using efficient transformer. *Journal of the Korea Institute of Information and Communication Engineering*, 25(9), 1275-1278. <https://doi.org/10.6109/jkiice.2021.25.9.1275>