지식재산연구 19권 1호, 2024년 3월, 159-177면
The Journal of Intellectual Property Volume 19, Issue 1, March 2024, Pages 159-177.

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

A Study on the Performance Improvement of Automatic Intellectual Property Counseling Classification: Using the Transformer-based AI Model BERT

Dong-Hun Noh[1,2*], Jae-Ok Min[1,3], So-Youn Woo[4,5]

[1]Doctoral Candidate, Department of Intellectual Property Convergence, Chungnam National University, Republic of Korea
[2]Manager, R&D Planning Team, Korea Institute of Patent Information, Republic of Korea
[3]Manager, R&D Team, Korea Institute of Patent Information, Republic of Korea
[4]Graduate Student, Department of Intellectual Property Convergence, Chungnam National University, Republic of Korea
[5]Korea Invention Promotion Association, Republic of Korea

*Corresponding Author: Dong-Hun Noh (laborh@kipi.or.kr)

ABSTRACT

Intellectual property customer counseling is an important public service that supports the creation of intellectual property rights and protection of the rights and interests of applicants and rights holders. To effectively support customers and secure the use of counseling content as a policy, counseling contents are classified according to certain criteria. Until 2020, it was professional counselors who directly classified these contents, but 2021 saw a shift toward automatic classification based on text analysis (TA) of the consultation texts. However, an investigation as to the distribution of counseling case classification over the past five years showed some differences between the 2018–2020 distribution, classified by professional counselors, and the 2021–2022 distribution, automatically classified by TA. Therefore, this study investigated how to improve the performance of the automatic classification system using BERT, a transformer-based AI model. After fine-tuning the BERT model, which was pre-trained using patent counseling text data and professional counselor classification values data, it was observed that the BERT's automatic classification distribution was more similar to that of professional counselors than the classification distribution of the existing TA. These results show that the future application of the "Patent Consultation Classification BERT," a tentative name for the model, to automatic patent consultation classification may yield a better performance than the current TA method. Furthermore, if the automatic classification results become more reliable through the use of this AI model, the purpose behind the policy for the automation of this procedure—namely easing the burden and improving the efficiency of professional counselors—may be achieved with improved continuity and stability. This may then enable a more accurate identification of the current status of patent customer counseling services and customer needs.

KEYWORDS

Intellectual property consultation, automatic classification, artificial intelligence model, BERT model, model learning

Check for updates

Open Access

Citation: Noh DH et al. 2024. A Study on the Performance Improvement of Automatic Intellectual Property Counseling Classification: Using the Transformer-based AI Model BERT. The Journal of Intellectual Property 19(1), 159-177.

DOI: https://doi.org/10.34122/jip.2024.19.1.7

Received: December 21, 2023
Revised: January 30, 2024
Accepted: February 29, 2024
Published: March 30, 2024

Funding: The author received manuscript fees for this article from Korea Institute of Intellectual Property.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

원저

특허상담 자동분류의 성능 향상 방안 연구: 트랜스포머 기반 인공지능 모델 버트(BERT)를 활용*

노동훈[1,2*], 민재옥[1,3], 우소연[4,5]

[1]충남대학교 지식재산융합학과 박사과정, [2]한국특허정보원 연구기획팀장, [3]한국특허정보원 연구실증팀장, [4]충남대학교 지식재산융합학과 석사과정, [5]한국발명진흥회

*교신저자: 노동훈(laborh@kipi.or.kr)

차례

국문초록

특허고객상담은 지식재산권의 창출 및 출원인·권리자의 권익 보호 등을 지원하는 중요한 공공서비스이다. 모든 특허상담 내용은 특허고객을 효과적으로 지원하고 정책적으로 활용하고자 일정한 기준에 따라 분류되고 있다. 2020년까지는 전문상담사가 직접 분류해 왔으나, 2021년부터는 상담 텍스트를 활용하여 TA(Text Analysis)로 자동으로 분류하고 있다.

최근 5년간 상담건 분류 분포를 살펴보면 전문상담사가 분류한 2018~2020년과 TA로 자동분류한 2021~2022년의 분포도가 다소 차이가 나는 것이 관찰되었기에 본 연구에서는 트랜스포머 기반 AI 모델인 버트를 활용한 자동분류의 성능 향상방안을 연구하였다. 특허상담 텍스트 데이터와 전문상담사 분류값을 학습데이터로 활용하여 버트를 사전 훈련시키고 특허상담에 맞도록 파인튜닝한 AI 모델을 활용하여 자동분류한 결과 기존 TA보다 분류 분포가 더 유사하게 나타났다. 이를 근거로 추후 특허상담분류버트(가칭)를 특허상담 자동분류에 적용할 때 보다 나은 성능을 기대할 수 있을 것이라 생각된다. 자동분류 결과가 보다 신뢰성 있게 도출되면 전문상담사의 업무부담 완화 및 효율성을 제고하고자 하는 정책적인 목적을 지속적.안정적으로 달성하고 특허고객상담 서비스 현황 및 고객의 니즈를 보다 정확하게 파악할 수 있어 특허고객 서비스 향상에 더 도움이 될 수 있을 것으로 기대된다.

주제어

특허상담, 자동분류, 인공지능모델, 버트모델, 모델학습

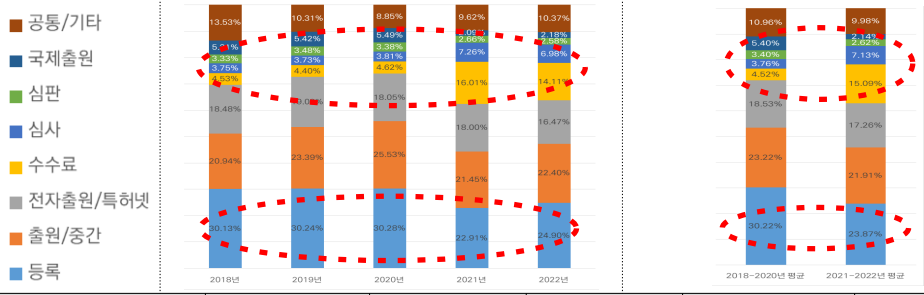
1. 서론

1.1. 연구의 배경

특허고객상담은 지재권 제도·절차, 특허넷 및 전자출원 사용법 등 특허행정 전반에 대한 민원인의 문의사항에 대한 답변을 전화(1544-8080), 채팅, PC원격, 챗봇 등을 통하여 제공하는 공공서비스이다. 상담서비스를 제공하는 상담사들은 모두 특허청 산하 공공기관인 한국특허정보원 소속 직원으로 구성되어 있으며 교육강사, 상담품질관리자, 리더급 상담사 등을 통해 지속적으로 교육 및 코칭을 받고 있어 특허상담에 최적화된 전문성을 보유하고 있다. 특허고객상담은 연평균 71만 건(최근 5년, 2018~2022년)이 특허고객상담센터 전문상담사 등을 통해 제공되고 있으며 지식재산권의 창출 및 출원인·권리자의 권익 보호 등을 지원하는 중요한 공공서비스이다. 모든 특허상담 내용은 특허고객을 효과적으로 지원(특허청 고객서비스 정책 반영, 상담서비스 개선 등)하고 정책적으로 활용하고자 일정한 기준에 따라 분류되고 있다. 2020년까지는 상담센터의 전문상담사가 고객과의 상담이 종료된 이후 전체 상담 내용을 토대로 직접 분류해 왔으나, 2021년부터는 상담사의 업무부담 완화 및 효율성 제고를 위하여 STT(Speech To Text, 음성인식의 한 분야로서 사람의 음성언어를 컴퓨터의 해석으로 문자데이터로 변환하는 처리를 의미[1])하며, 특허고객상담센터는 2018.11월 해당시스템을 도입하여 활용하고 있음)를 통하여 자동으로 변환된 상담 내용 텍스트 데이터를 활용하여 TA(Text Analysis, 텍스트 데이터를 수집하고 분석하여 의미 있는 정보를 도출하는 기술[2])이며, 특허고객상담센터는 2021년부터 활용하고 있음)로 자동분류를 해오고 있다.

<표1 최근 5년간 특허상담 분류(대분류) 분포(단위: %)>

구분	연도별 분류 분포			평균 분류 분포	
	전문상담사 분류		TA 자동분류	전문상담사 분류	TA 자동분류



구분	2018년	2019년	2020년	2021년	2022년
등록	30.13%	30.24%	30.28%	22.91%	24.90%
출원/중간	20.94%	23.39%	25.53%	21.45%	22.40%
특허넷/전자출원	18.48%	19.03%	18.05%	18.00%	16.47%
수수료	4.53%	4.40%	4.62%	16.01%	14.11%
심사	3.75%	3.73%	3.81%	7.26%	6.98%
심판	3.33%	3.48%	3.38%	2.66%	2.58%
국제출원	5.31%	5.42%	5.49%	2.09%	2.18%
공통/기타	13.53%	10.31%	8.85%	9.62%	10.37%
비고	상담사가 직접 분류			TA가 자동으로 분류	

* 특허청에서 주최하고 한국지식재산연구원에서 주관한 제18회 대학(원)생 지식재산 우수논문공모전 수상작을 수정·보완하여 작성하였습니다.

1) 민소연 외 3인, "한국어 특성 기반의 STT 엔진 정확도를 위한 정량적 평가방법 연구", 「한국산학기술학회논문지」, 제21권 7호(2020), 699-707면.

2) 포지큐브, "TA(Text Analysis)는 무엇이고 어떻게 활용할 수 있나요?", 포지큐브, <https://www.posicube.com/586335f6-7163-458a-8caf-0daccf2488bf>, 검색일: 2023. 8. 1.

최근 5년간 상담건 분류 분포를 살펴보면 전문상담사가 분류한 2018~2020년과 TA로 자동분류한 2021~2022년의 분포도가 다소 차이가 나는 것을 관찰할 수 있다. 특히 등록, 수수료, 심사, 국제출원 분야 등에서 그 차이가 크게 나타나는 것을 확인할 수 있다. 분류 주체(전문상담사 → TA 자동분류)가 달라진 이후부터 분류 분포가 달라져서 이를 정책적으로 활용하기에는 TA 자동분류의 정확도 및 신뢰성 측면에서 의구심이 들 수 있다. 그렇다고 다시금 전문상담사가 직접 분류하는 과거의 방법으로 회귀한다면 본래 프로세스를 변경한 취지(상담사의 업무부담 완화 및 효율성 제고 등)에 부합하지 않는다. TA는 형태소분석→구문분석→어휘중요도분석→연관어분석 등의 방식으로 텍스트를 분석한다. 이러한 방식은 각 단계마다 동의어/유의어 사전, 감성사전 등 고품질의 사전 데이터가 필요하고 이를 위해 계속적인 투자비용이 요구된다. 오타, 신규어 등에 대해 정확한 의미파악이 어려운 한계도 있다.

최근 한국은 인공지능(AI)을 중심으로 한 지능정보기술 분야에 대한 정보화투자를 확대하고 있다. 2022년도 지능정보기술 투자 규모는 5조 4,813억원으로 전체 국가정보화 투자규모의 47.5% 비중을 차지[3]하고 있으며, 2023년도 지능정보기술 투자 규모는 5조 6,058억원으로 전체 국가정보화 투자규모의 53.5%를 차지[4]하고 있다.(전체 국가정보화 투자규모는 10조 4,741억원으로 전년대비 9.23% 감소) 특허청도 “AI·빅데이터 기술 활용 특허행정 혁신”을 정부 국정과제(22번 수요자 지향 산업기술 R&D 혁신 및 지식재산 보호 강화에 포함)로 삼고 디지털 전환을 통한 구조적인 혁신을 추진하고 있다. 국내외 인공지능 기술을 업무에 활용한 다양한 연구가 진행되어 왔으며 AI 모델을 활용한 기술분류 등도 의미 있는 선행 연구가 있어왔고 소정의 성과들을 거두었다. 본 연구에서는 분류에 많이 활용되고 있는 구글의 트랜스포머 기반 AI 모델인 버트(BERT)를 활용하여 특허상담 자동분류의 성능을 향상시키는 방안을 연구하려 한다.

1.2. 연구의 목적

본 연구는 특허상담 내용 자동분류의 성능 향상을 목적으로 한다. 여기서 성능 향상이란 특허상담 내용 자동분류의 결과와 전문상담사가 직접 분류한 결과와의 유사도를 높이는 것이다. 본 연구를 통하여 특허 전문상담사들의 업무부담 완화 및 효율성을 제고하고자 하는 정책적인 목적을 지속적.안정적으로 달성하고자 하며, 상담분류통계정보가 상담센터 운영방향 설정 및 고객서비스 정책 수립시(예. FAQ, 상담이슈 등에 대한 제공정보 선정 등) 중요한 참고자료 활용되고 있음을 고려할 때 특허고객상담서비스를 이용하는 고객의 V.O.C.를 보다 정확하고 효과적으로 관리(특허고객상담 서비스 현황 파악, 특허고객의 니즈 파악 등)하여 특허고객 서비스 향상에 기여하고자 한다.

2. 이론적 배경

2.1. 트랜스포머와 이에 기반한 AI 언어모델의 개념 및 특징

2.1.1. 트랜스포머와 이에 기반한 AI 언어모델

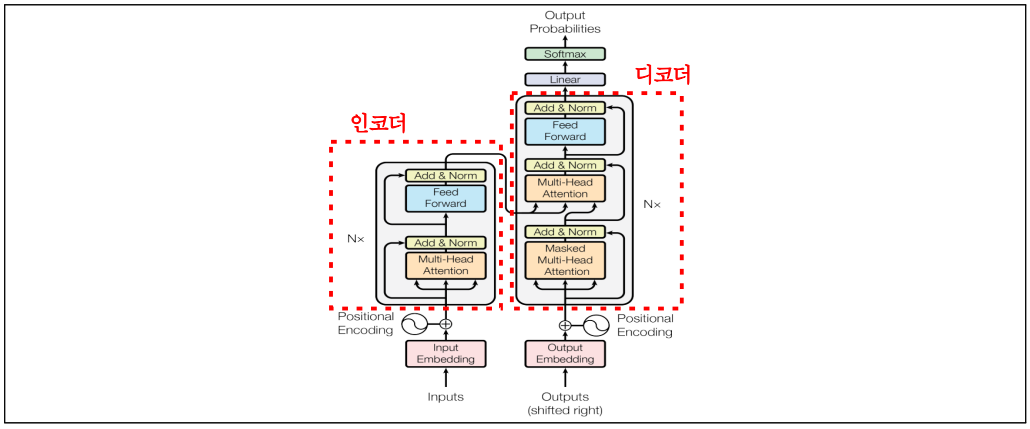
트랜스포머(Transformer)는 2017년 12월 구글이 발표한 인공지능(AI) 언어모델로, 기존 자연어 처리에 기본적으로 사용되어 왔던 RNN(Recurrent Neural Network, 순환 신경망)의 단점(문장의 단어를 순차적으로 처리하여 문장이 길어질수록 성능이 떨어짐)을 보완하며 자연

3) 박미영, “2022년 국가정보화 예산 약 11조, 올해보다 15% 증가 전망”, 보안뉴스, <https://www.boannews.com/media/view.asp?idx=103327>, 2021. 12. 15자.

4) 최아름, “지역 맞춤형 디지털 특화사업 추진, 대한민국 디지털 전략 실현 지름길”, 정보통신신문, <https://www.koit.co.kr/news/articleView.html?idxno=106622>, 2022. 12. 6자.

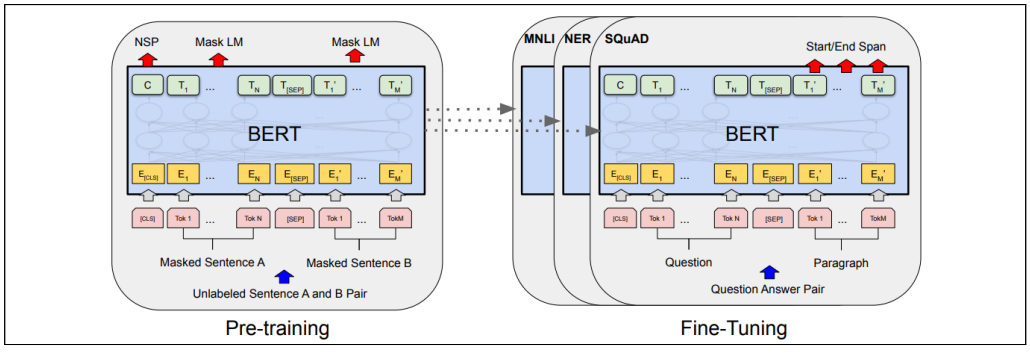
어처리 분야에 큰 변화를 가져왔다. 트랜스포머는 "Attention is all you need" 논문에서 최초로 소개되었으며, <그림 1>와 같이 인코더(Encorder)와 디코더(Decorder)로 이루어진 구조를 갖추고 있다. 트랜스포머 아키텍처에서 핵심적인 역할을 하는 어텐션 기능(Attention Mechanism, 주의 집중 매커니즘)은 각 단어의 중요성을 독립적으로 계산하며, 모든 위치의 단어에 대한 정보를 종합적으로 고려하는 방식으로 자연어처리 성능을 크게 개선하였다.[5)6)]

<그림1 트랜스포머 모델 아키텍처>[7)]



트랜스포머가 발표된 이후 이에 기반하여 버트(BERT)[8)], GPT[9)] 등과 같은 언어모델이 등장하였다. 버트(BERT, Bidirectional Encoder Representations form Transformers)는 2018년 10월 트랜스포머 이후에 구글이 발표하였으며 트랜스포머의 인코더만으로 구현한 언어모델이다. 대규모 데이터에서의 사전 훈련(Pre-training)을 통해 풍부한 문맥 정보를 습득하며, 인코더에서 양방향 어텐션 메커니즘을 적용하여 문맥 정보를 양쪽 방향으로 고려함으로써 단어 간 상호작용과 문맥적 특성을 더욱 정교하게 이해하는 특성을 가진다. 이러한 특징으로 버트는 다양한 자연어 처리 작업에서 뛰어난 성능을 보이며, 사안에 따라 파인튜닝(Fine-tuning) 또는 전이학습(Transfer Learning)을 통해 문제를 해결하는 방식이 효과적인 것으로 나타나고 있으며 분류, 기계독해, 기계번역, 검색 등 다양한 분야에서 연구가 진행되어 왔다.

<그림2 버트(BERT)의 사전 훈련 및 파인튜닝>[10)]



5) 한규동, 「AI 상식사전」, 길벗, 2022, 317면, 351-352면.
6) Ashish Vaswani et al., "Attention is all you need", Edited by I. Guyon et al., Advances in Neural Information Processing Systems 30, 2017, pp. 6000-6010.
7) 위의 글 자료에 텍스트 삽입.
8) Jacob Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv*, 1810.04805(2018).
9) Alec Radford et al., "Improving language understanding by generative pre-training", Preprint, <https://paperswithcode.com/paper/improving-language-understanding-by>, (2018).
10) Jacob Devlin et al., 앞의 논문에서 그림 발췌.

GPT(Generative Pre-trained Transformer)는 자연어 텍스트 생성 능력을 향상시키기 위한 모델로 버트와는 반대로 트랜스포머 아키텍처에서 디코더(단방향 방식)만을 활용한다. 디코더는 이전 단어들을 기반으로 다음 단어를 예측하는 방식으로, 대화, 문서 생성, 질의응답과 같은 작업에서 높은 성능을 나타내고 있다. OpenAI社에서 디코더를 사용하여 2018년 6월 'GPT-1'을 만들었으며 이후 계속적인 성능 개선을 통해 GPT-3.5를 기반으로 한 LLM(Large Language Model, 초거대 언어모델인 챗GPT를 발표하며 사회적 이슈를 일으켰다. 생성형 AI인 GPT는 논문, 보고서, 문학, 코딩 등 다양한 분야에서 활용되어 왔다.

<표2 챗GPT 출현 이전의 인공지능(AI) 분야 주요 기술 등장 경과>[11)]

2012	이미지 처리 딥러닝 본격화 합성곱 신경망(CNN)	'12년 이미지넷 챌린지에서 CNN기반 딥러닝 알고리즘 AlexNet이 우승을 차지하며 딥러닝 부흥의 계기가 됨
2014	생성적 AI 분야의 새로운 혁신 적대적 생성 신경망(GAN)	생성자와 판별자가 서로 대립하며 데이터를 생성하는 모델(기존 생성 AI 대비 성능 우수)
2016	인공지능 대중화 알파고(AlphaGo)	구글 딥마인드가 개발한 인공지능 바둑 프로그램
2017	언어모델의 혁신적 돌파구 마련 트랜스포머(Transformer)	기존 RNN 구조의 단점(병렬처리 어려워 속도 느림)을 극복하여 여러 모델 파생 (BERT, GPT 등의 기반이 됨)
	연합학습(Federated Learning)	데이터를 중앙 취합하지 않고, 사용자 기기에서 학습한 모델의 가중치만 중앙으로 취합
2018	자기지도학습 부각 (Self-supervised Learning)	비지도학습의 한 방법으로 얀 르쿤 교수가 중요성 강조
	BERT 언어모델	언어모델의 새로운 표준 역할
2019	GPT-2 언어모델	15억 개 파라미터, 웹문서 800만 개 학습
2020	초거대 모델의 범용성 부각 GPT-3	초거대 언어모델의 시작 (1,750억 개 파라미터, 3,000억 개 학습데이터)
	비주얼트랜스포머	이미지 분야에 트랜스포머를 적용하여 최고수준 경신
	SimCLRv2	이미지 분류 자기지도학습이 지도학습 성능에 근접
	알파폴드2(AlphaFold2)	단백질 구조와 기능 예측(사람수준까지 성능 개선)

CNN(Convolutional Neural Network) : 이미지에 다양한 필터를 적용하여 특징 추출, 압축하여 일반화하는 알고리즘
GAN(Generative Adversarial Network) : 생성자, 판별자를 활용하여 가상모델같이 실제같은 사진, 영상 등을 만드는 알고리즘
RNN(Recurrent neural network) : 유닛 간의 연결이 순환적 구조를 같는 특징을 보유하는 알고리즘
자기지도학습 : 라벨링 없이 입력 값의 일정 부분으로 입력 값의 다른 부분을 예측하며, 입력 값 자체로 지도학습
파라미터 : 딥러닝 모델의 노드 사이를 연결하는 가중치
알파폴드2 : 알파고를 만든 구글 딥마인드가 개발한 모델, 유선정보만으로 단백질의 3차원 입체구조를 예측

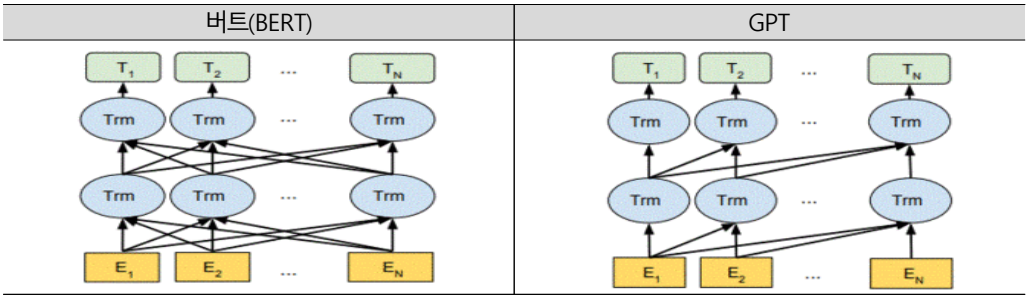
11) 한국지능정보사회진흥원, "IT & Future Strategy 2021년 보고서", 한국지능정보사회진흥원, 2022, 6면, 내용 일부를 재구성.

2.1.2. 버트 모델의 특징

버트의 기본 구조는 트랜스포머의 인코더를 쌓아올린 구조로서, 자연어 처리 분야의 범용 모델로 사용되고 있다. 다양한 자연어 처리 임무를 동일한 구조의 학습된 모델로 해결할 수 있는 특징을 가지고 있다.[12] 버트는 사전 훈련한 임베딩을 바탕으로, 적은 데이터셋으로 파인튜닝(미세 조정)한 후 다른 과제에 적용하여 좋은 성능을 낼 수 있다는 것이 큰 장점이다.[13]

사전 훈련 단계에서 라벨링이 되지 않은 대규모의 데이터를 학습시켜 임베딩이 만들어지면, 파인튜닝 단계에서는 그것을 기반으로 라벨링 된 작은 규모의 데이터를 학습시켜 구체적인 과제를 수행한다. 버트와 기존 모델과의 차이점은 왼쪽에서 오른쪽으로, 오른쪽에서 왼쪽으로 동시에 고려하여 문장을 읽는 양방향의 학습을 통해 사전 훈련이 이루어진다는 것이다. 아래 그림에서 보이는 것과 같이 GPT는 한 방향으로만 문장을 학습하지만, 버트는 네트워크가 양쪽방향을 다 고려하는 양방향 학습을 하여 벡터에 문맥 정보를 잘 반영할 수 있게 된다. ELMo에서 양방향성을 가지는 구조가 제안되기도 하였으나 단방향과 단방향의 단순한 결합이라는 점에서 얕은 양방향성(shallow bidirectional)을 가지는 반면 버트는 깊은 양방향성(deep bidirectional)을 가진다.[14]

<그림3 버트(BERT) 및 GPT의 사전 훈련 모델 구조 차이>[15]



2.2. 버트를 활용한 자동분류 연구

버트는 양방향으로 문맥을 이해하는 특징이 있다. 입력문장의 모든 단어를 고려하여 문맥 정보를 파악하므로 단어들 간의 상호작용을 더욱 잘 이해할 수 있는 장점이 있다. 버트는 사안에 따라 파인튜닝 또는 전이학습을 통해 분류를 효과적으로 구현할 수 있어 이와 관련한 다양한 선행연구들이 진행되어 왔다. 최근의 선행연구들은 다음과 같다.

버트를 활용한 온라인 진로상담 문서 자동분류 연구가 있었다. 학생들이 올린 온라인 진로상담 텍스트 데이터 4,400여건을 활용하여 4개의 유형으로 정리하였다. 이 연구에서 버트를 활용한 자동분류에서 좋은 결과를 확인하였으며 사례 수가 적은 범주에서 더 우수한 성능이 확인되었다.[16]

버트를 활용한 한국어 특허 문장 기반의 CPC 자동분류 연구도 있었다. CPC는 2012년 유럽 심사관들의 주도로 미국과 유럽특허청이 공동으로 개발한 특허분류체계이며 특허분류체계 중

12) 권순보, “BERT를 활용한 진로상담 텍스트데이터 분석”, 한국교원대학교대학원, 박사, 2022, 15-16면.
13) 서울대학교 AI연구원, “AI용어사전”, <https://aiis.snu.ac.kr/>, 검색일: 2023. 8. 1.
14) Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv, 1810.04805(2018).
15) 위의 논문에서 그림 발췌.
16) 권순보·유진은, “BERT와 FastText를 활용한 온라인 진로상담 문서 분류”, 한국데이터정보과학회지, 제33권 제6호(2022), 991-1006면.

가장 세분화된 분류로 선진 5개국 특허청을 중심으로 CPC 분류를 위해 많은 인력과 예산을 투입하고 있다. 이 연구에서 특허문헌을 버트로 학습시켜 기존 모델 대비 우수환 KorPatBERT를 생성하였으며 분류 학습데이터셋 구축방안도 제안하였다. 이 연구의 결과로 실제 서비스가 가능한 수준의 CPC 자동분류 모델을 생성하였다.[17)]

버트를 활용한 법률상담 데이터 자동분류 연구도 있었다. 이 연구에서 버트 사전학습 데이터를 활용하여 법률상담에 맞는 분류 모델을 구현하였으며 대한법률구조공단에서 운영하는 홈페이지에 게시되어있는 법률상담사례 데이터로 연구를 수행하였다. 분류는 상담사례가 많은 9개(가사, 손해배상, 민사일반, 노동, 민사소송, 형사범죄, 주택임대차, 강제집행, 부동산일반)로 설정하여 진행하였으며 약 9천여개의 법률상담 데이터를 분석하여 버트 사전학습을 통해 약 69%의 정확도로 자동분류 결과를 도출하였다.[18)]

버트를 활용한 1:1 게시판 상담문의에 대한 자동분류 연구도 있었다. 논문 전문 데이터 구축을 수행하면서 1,380명의 청년들이 문의한 약 36,000건의 상담 데이터를 활용하여 질문 유형을 분류하여 자동분류 연구를 수행하였다.

<표3 논문 전문 데이터 구축 1:1 상담 질문 분류 유형 및 정의>

구분	분류명	정의
1	구축불가	원문 텍스트 복사 오류가 폭넓게 발생하는 논문
2	신고	작업지침을 따르지 않고 구축/검토한 참여자 신고
3	오류문의	데이터 구축 도구에 대한 오류 문의
4	근태문의	목표량 미달, 휴가, 자리 비움에 대한 처리 문의
5	작업지침	논문 전문 편집 대상 및 방법에 대한 문의
6	의사소통	공지사항, 1:1 상담, 카카오 채널 등에 대한 문의
7	기타	위 내용 이외의 문의

상담 질문 텍스트에 대해서 데이터 전처리(분류되지 않은 상담 질문 제거, 형태소 분석을 통한 토큰화, 불용어 제거, 정수 인코딩)를 수행한 후 모델 학습을 위해 훈련, 검증 및 평가 데이터셋으로 분할하여 연구를 수행하였으며 비교적 의미있는 수준의 분류모델을 개발하였다.[19)]

버트를 활용한 주제명 자동분류 연구도 있었다. 국립중앙도서관이 소장한 장서에 대해 분류와 편목 업무를 통해 생산한 국가서지 데이터의 일부를 사용하여 연구를 진행하였다. 주제명의 부여 횟수에 따라 6개의 실험 데이터셋을 구축하여 분류 실험을 진행하였으며 실험 데이터셋은 25개부터 최대 3,506개 주제명 개수로 구성하여 자동분류 연구를 수행하였다. 주제명의 자동분류와 주제명이 부여된 KDC 분류체계와 주제명의 범주 유형에 따른 성능도 분석하였으며 '식물', '법률명', '상품명'에서 높은 성능을 보였다.[20)]

버트를 활용한 학술문헌 자동분류 연구도 있었다. 문헌정보학 분야의 KCI 등재 논문 전체를 수집하여 7개 학술지 5,357개의 논문의 초록 데이터로 문헌을 13개의 단일 항목으로 분류하여 학습데이터를 구축하였다. 분류항목은 한국연구재단의 학술연구분야분류표의 13가지 소분류명을 사용하였다. 데이터양과 데이터품질에 따라 정확도가 차이가 있었으며 50% 이상부터 90% 이상까지의 정확도를 도출하였다.[21)]

17) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022), 209-256면.
18) 전영호, "BERT Transformer와 Deep Learning을 활용한 전이학습 효과 검증 연구: 법률상담데이터 분류 문제 적용", 「한국경영공학회지」, 제24권 제4호(2019), 77-89면.
19) 신진섭 외 5인, "양방향 LSTM 기반의 기계학습 데이터 구축 상담 자동 분류", 「한국정보과학회 2021 한국컴퓨터종합학술대회 논문집」, (2021), 328-330면.
20) 이용구, "BERT 모형을 이용한 주제명 자동 분류 연구", 「한국문헌정보학회지」, 제57권 제2호(2023), 435-452면.

버트를 활용한 기술문서의 한국표준산업분류 자동분류 연구도 있었다. 기술과 산업간 분류 활용성을 높이고자 논문, R&D 과제정보 등 특허와 문서의 성격이 유사한 기술문서에 한국표준산업분류코드를 자동부여 할 수 있는 자동분류 버트 모델을 연구하였으며 특허문헌에 한국표준산업분류를 자동분류하는데 79.77%의 분류 정확도를 도출하였다.[22]

버트를 활용한 초등학교 고학년의 욕설문장 자동분류 연구도 있었다. 초등학생이 작성한 문장을 학습시켜 자동으로 욕설문장을 필터링하는 실험을 진행했으며 온라인 학습 플랫폼에서 초등학교 4~6학년의 채팅내역을 수집하고 채팅 내역 중에 욕설로 신고 되어 판정된 욕설문장 데이터를 함께 수집하고 학습시켜 모델을 구성하였다. 실험결과 욕설문장 자동분류에서 약 75%의 정확도를 보이는 것으로 분석되었다.[23]

버트를 활용한 코로나-19 감염병 다국어 기사 자동 색인 및 분류 연구도 있었다. 뉴스 중 감염병 관련 기사 데이터를 수집하기 위해 보건복지부에서 정의한 법정 감염병 주요 증상 50개를 검색 키워드 및 챗GPT 색인으로 사용하였으며 장기적인 관점에서 해당 국가의 감염병 발생 및 전파 위험도를 표현하는 INFORM 감염병 위험 지수 및 정부정책분류인 OxCGRT 분류(봉쇄 정책 C, 경제 정책 E, 의료보건 정책 H, 백신 정책 V, 그 밖의 정책 M)와 10가지 사건 분류(확진자수, 완치자수, 사망여부, 집단감염, 백신관련, 방역지침, 경제지원, 마스크, 국제기구, 병원관련)를 제안하여 연구를 진행하였다. 약 46천건의 학습데이터를 활용하여 90% 이상의 높은 정확도를 구현하였다.[24]

버트를 활용한 기업 공시문서의 자동분류 연구도 있었다. 기업 공시자료는 기업간 거래 및 투자 결정에 있어 필수적인 정보이며 기업 및 산업에 대한 중요 연구자료이다. 기계학습에 기반한 자연처처리 기법을 활용하여 미국 수시공시 회계문서를 자동으로 분류하는 연구를 수행하였으며 3가지 도구로 자동분류결과를 비교분석하였으며 그 중 버트 모형도 포함되어 있다. 비록 이 연구에서는 버트 모형이 가장 높은 정확도를 나타내진 않았지만 버트를 기업공시자료에 활용한 의미있는 사례라고 볼 수 있다.[25]

버트를 활용한 내비게이션 장소 자동분류에 대한 연구도 있었다. 행정안전부에서 제공하는 '지방행정 인허가 데이터'를 활용하였으며, 이 연구를 통하여 자동차 내비게이션의 검색 결과에 해당하는 장소 데이터 구축시 딥러닝을 활용하여 장소의 종별을 자동으로 분류하기 위한 모델을 개발하였다. 버트 모델로 별도의 색인 DB 사용 없이 딥러닝만으로 자동분류의 정확도를 71.8%까지 도출(연구의 정확도가 비교적 낮은 이유는 방대한 양의 데이터와 한정된 자원으로 연구를 진행하다보니 epoch를 3회 밖에 진행하지 못했으며 인허가 데이터 특성상 학습데이터의 정답이 여러 개일 수 있는 한계점 때문)하였다.[26]

---

21) 김인후·김성희, "딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류", 「정보관리학회지」, 39권 3호(2022), 293-310면.

22) 김명선·한동희, "기술문서의 한국표준산업분류 자동분류를 위한 특허기반 BERT 모델", 「한국정보과학회 2022 한국소프트웨어종합학술대회 논문집」, (2022), 455-457면.

23) 심재권, "BERT 를 활용한 초등학교 고학년의 욕설문장 자동 분류방안 연구", 「창의정보문화연구」, 제7권 제2호(2021), 91-98면.

24) 강승태·장길진, "ChatGPT와 다국어 BERT를 이용한 코로나-19 감염병 다국어 기사 자동 색인 및 분류", 「전자공학회논문지」, 제60권 제7호(2023), 20-29면.

25) 이경란·강창묵, "자연어처리 기계학습 기법을 이용한 공시문서의 자동분류: Confidential treatment를 가진 8-K 문서를 중심으로", 「한국전자거래학회지」, 제28권 제2호(2023), 21-36면.

26) 강유라, "KoBERT를 활용한 내비게이션 장소 자동 분류 시스템", 고려대학교 컴퓨터정보통신대학원, 석사, 2023, 6-9면.

<표4 논문 전문 데버트를 활용한 자동분류 선행연구 사례>

구분	연구개요	연구연도
1	법률상담 데이터 자동분류	2019
2	논문 전문 데이터 구축 관련 1:1 게시판 상담문의 자동분류	2021
3	초등학교 고학년의 욕설문장 자동분류	2021
4	온라인 진로상담 문서 자동분류 연구	2022
5	한국어 특허문장 기반 CPC 자동분류 연구	2022
6	문헌정보학 분야 학술문헌 자동분류	2022
7	특허/기술문서에 대한 한국표준산업분류 자동분류	2022
8	국가서지에 대한 주제명 자동분류	2023
9	코로나-19 감염병 다국어 기사 자동분류	2023
10	기업 공시문서 자동분류	2023
11	네비게이션 장소 자동분류	2023

앞서 언급한 바와 같이 트랜스포머 기반 AI 모델인 버트는 법률상담, 게시판상담, 진로상담, 학술문헌, 산업분류 등 상담영역을 포함한 다양한 데이터의 자동분류에 연구모델로 사용되어 왔으며, 자동분류에서의 효과가 어느 정도 검증되어왔다. 그래서 특허고객상담 데이터의 자동분류에도 좋은 성능을 도출할 수 있을 것으로 판단하였다. 버트 모델은 분류 작업에서 뛰어난 성능을 보이지만, 모델의 크기와 복잡성 때문에 학습 과정에서 많은 계산 리소스가 소모된다. 이로 인해 전산환경 등이 충분히 뒷받침되지 않을 경우 속도에서 상대적인 제약이 있을 수 있다. 본 특허상담분류 연구에서는 약 112만 건의 방대한 데이터를 확보하고 있으며, 일정 주기로 분류작업을 수행해도 정보활용에 지장이 없기 때문에 모델의 처리 속도보다는 분류 정확도 향상을 중심으로 연구를 진행하였다.

3. 연구 방법 및 결과

3.1. 개요

앞서 연구의 범위 및 방법에서 설정했듯이 먼저 약 112만 건의 2019 ~ 2020년도 특허상담 텍스트 데이터와 전문상담사가 직접 분류한 분류값을 활용하여 버트를 학습시키고 특허상담에 맞게 파인튜닝하여 만들 특허상담분류버트(가칭) 모델을 활용하여 2021 ~ 2022년도 특허상담 텍스트 데이터를 자동 분류하였다. 특허상담분류버트로 자동 분류한 2021 ~ 2022년도 분류 분포와 기존 TA가 자동 분류한 2021 ~ 2022년도 분류 분포 데이터를 우선 비교하고, 실제 전문상담사가 분류한 2018 ~ 2020년의 분류 분포의 유사성을 비교하여 자동분류의 성능을 확인하였다.

3.2. 연구의 범위 및 방법

본 연구에서는 특허고객상담센터에서 보유했던 특허상담 관련 데이터를 활용하였다. 먼저 데이터 현황(전체 상담 텍스트 및 분류값)을 파악하고 비교 분석이 가능한 새로운 데이터 생성이 가능한지 여부를 분석하여 연구의 범위 및 방법을 결정하였다.

<표5 기보유한 데이터 현황>

데이터 구분	2018년	2019년	2020년	2021년	2022년	비고
상담 텍스트	X	○	○	○	○	STT 자동변환
전문상담사 분류	○	○	○	X	X	
TA 분류	X	X	X	○	○	

<표5>의 기보유한 데이터 현황에 따라 2019~2022년까지의 전체 상담 텍스트 데이터, 2018~2020년 전문상담사 분류 데이터(특허고객상담센터 전문상담사가 상담을 종료한 후 직접 분류한 분류 데이터), 2021~2022년 TA 분류 데이터(상담시스템 내 TA로 자동 분류된 분류 데이터) 등을 활용하여 가정을 세우고 몇 가지 연구의 방향을 설정하였다.

가정	① 전문상담사가 고객과의 상담을 통해 직접 분류한 분류값이 가장 정확하며, ② 자동분류값이 전문상담사가 직접 분류한 분류값에 근접할 때 그 자동분류 도구의 성능이 우수하다고 판단한다.
※ 원활한 연구를 위해 가정에 대한 검증은 별도로 진행하지 않았음	

위의 가정을 바탕으로 첫번째 방법(1안)은, 2020년도 전체 상담 텍스트 데이터를 활용하여 ①상담시스템에 적용된 TA로 자동분류하고, ②버트로 자동분류(2019년도 데이터로 학습시켜 모델을 구성하고 2020년도 데이터를 자동분류)하여 ③전문상담사가 직접 분류한 2020년도의 데이터와 TA 자동분류, 버트 자동분류 데이터 분포를 비교하여 성능을 판단하는 방법이다.

<표6 1안에 대한 검토>

구분	검토내용	검토결과
① 상담시스템에 적용된 TA로 '20년도 상담 텍스트 데이터 자동분류	• 1일 단위 자동분류하는 방식으로 구현되어 있으며, 대량의 데이터를 TA로 자동분류하려면 별도의 용역 필요 • 자체적으로 실행 불가	TA 분류가 안되어 불가
② 버트로 '20년도 상담 텍스트 데이터 자동분류	• '19년도 상담 텍스트 데이터를 학습데이터로 활용하여 자체적으로 학습 실행 가능	
③ '20년도 전문상담사 분류, TA 자동분류, 버트 자동분류 데이터 분포 비교		

1안의 경우 2020년도의 동일한 데이터로 직접 비교가 가능하여 높은 신뢰성·타당성을 확보할 수 있는 방법이지만, TA가 1일 단위로 자동분류 테스크를 실행하는 방식으로 시스템에 구현되어 있어 2020년도 전체 특허상담 텍스트 데이터를 자동분류하려면 별도의 용역이 필요하므로 시간, 비용 등의 한계로 연구를 진행하기 불가하다.

두 번째 방법(2안)은, 2021~2022년도 전체 상담 텍스트 데이터를 활용하여 ①버트로 자동분류(2019~2020년도 데이터로 학습시켜 모델을 구성하고 2021~2022년도 데이터를 자동분류)하고, ②전문상담사가 2021~2022년도 상담 데이터를 직접 분류하여 ③TA가 자동분류한 2021~2022년도 데이터와 버트 자동분류, 전문상담사 분류 데이터 분포를 비교하여 성능을 판단하는 방법이다.

<표7 2안에 대한 검토>

구분	검토내용	검토결과
① 버트로 '21~'22년도 상담 텍스트 데이터 자동분류	• '19~'20년도 상담 텍스트 데이터를 학습 데이터로 활용하여 자체적으로 학습 실행 가능	전문상담사 분류가 안되어 불가
② 전문상담사가 '21~'22년도 상담 데이터를 직접 분류	• 데이터 분량이 많아, 다수의 전문상담사가 일정기간 투입되어 분류작업을 수행해야 하므로 자체적으로 실행 불가	
③ '21~'22년도 버트 자동분류, 전문상담사 분류, TA 자동분류 데이터 분포 비교		

2안의 경우 2021~2022년도의 동일한 데이터로 직접 비교가 가능하여 높은 신뢰성·타당성을 확보할 수 있지만, 2021~2022년도 특허상담 텍스트 데이터 분량이 많아 다수의 전문상담사가 투입되어야 하므로 시간, 비용 등의 한계로 연구를 진행하기 불가하다.

세 번째 방법(3안, 2안의 수정안)은, 2021 ~ 2022년도 전체 상담 텍스트 데이터를 활용하여 ①버트로 자동분류(2019 ~ 2020년도 데이터로 학습시켜 모델을 구성하고 2021 ~ 2022년도 데이터를 자동분류)하여, ②TA가 자동분류한 2021 ~ 2022년도 데이터와 2021 ~ 2022년도 버트 자동분류 데이터를 전문상담사가 분류한 2020년까지의 데이터 분포와 유사도를 비교하여 성능을 판단하는 방법이다.

<표8 3안(2안의 수정안)에 대한 검토>

구분	검토내용	검토결과
① 버트로 '21 ~ '22년도 상담 텍스트 데이터 자동분류	• '19 ~ '20년도 상담 텍스트 데이터를 학습 데이터로 활용하여 자체적으로 학습 실행 가능	동일년도에 대한 비교는 아니지만 과거 전문상담사 분류 분포가 일정하기에 가능할 것으로 사료
② '21 ~ '22년도 버트 자동분류, TA 자동분류 데이터와 '20년까지의 전문상담사 분류 데이터 분포의 유사도를 비교		

3안의 경우 동일년도에 대한 직접 비교는 아니지만 2018 ~ 2020년도 전문상담사 분류 분포가 일정한 분포를 유지하고 있어 대략적인 유사성 비교가 가능할 것으로 사료된다. 그러므로 시간, 비용의 한계상 현재 상황에서 연구 가능한 방안인 3안으로 연구의 범위 및 방법을 설정하고 연구를 진행하였으며 분류결과의 범위는 대분류(1차 분류)에 한정하여 진행하였다.

<표9 특허상담 대분류(1차 분류)>

①	②	③	④	⑤	⑥	⑦	⑧
등록	출원/중간	특허넷/전자출원	수수료	심사	심판	국제출원	공통/기타

3.3. 연구 결과

3.3.1. 특허상담 자동분류 모델 구현을 위한 데이터 학습

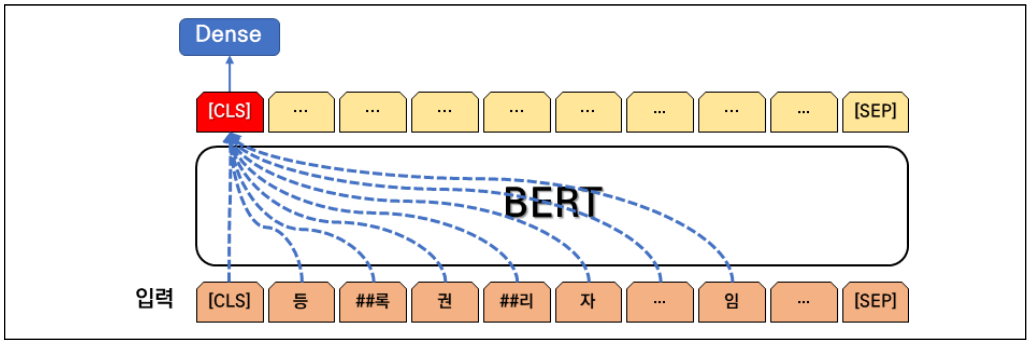
특허상담 자동분류 모델을 구현하기 위해 2019 ~ 2020년도 특허상담 데이터 약 112만 건을 데이터 전처리(data cleaning)를 통해 학습데이터셋(train dataset)를 구축하였고, NVIDIA社 Tesla V100 32GB GPU 4개가 설치되어 있는 Ubuntu기반의 Linux 서버 1대로 모델 학습을 진행하였다.

특허상담 자동분류 모델을 구현하기 위한 특허상담 텍스트 데이터 전처리
① 인바운드 및 아웃바운드 상담 텍스트 중 아웃바운드 데이터 삭제 ② 의미없는 특수문자 제거 / ③ 의미없는 한 글자 제거 (예: 아, 네 등) ④ 인사말 등 불용어 제거 (예: 감사합니다, 고맙습니다, 안녕하세요, 안녕하십니까, 수고하십니다. 수고하세요, 여보세요, 상담사, 특허청 등) ⑤ 글자 수가 너무 적으면 문장의 특징 추출이 어려우므로 50글자 이하 제거 ⑥ 문장의 앞뒤 공백, tab으로 벌어진 긴 공백 등 제거

전처리된 학습데이터셋을 9:1의 비율로 나누어 검증데이터셋(validation dataset)을 구성하였으며 이를 활용하여 특허상담 분류에 맞도록 딥러닝 학습을 진행하였다. 분류에 대한 학습을 실행하기 위해서 학습데이터셋은 문장(sentence)와 라벨(label) 컬럼으로 구분하였으며 전체 상담내용 문장들이 등록, 출원/중간, 특허넷/전자출원, 수수료, 심판, 국제출원, 공통/기타으로 분류될 수 있도록 라벨 상에 분류코드를 지정하였다. 학습에 사용한 파라미터는 토큰(token) 수(sequence length) 256으로 적용하여 연구를 진행하였으며 7epoch 학습에서 가장

낮은 로스(loss : 0.5250)와 높은 정확도(정확도 80.05%)가 달성되어 예측을 위해 7epoch까지 학습한 모델을 사용하였다. 학습데이터에서 문장 컬럼에 해당하는 텍스트를 버트의 토크나이저(tokenizer)로 토큰화하여 인코딩을 하여 입력으로 들어가고, 임베딩(embedding) 레이어를 통과하여 입력 데이터 코튼의 수만큼의 형태를 갖는 출력을 내보내었다. 입력 필드의 대표적인 의미가 담겨 있는 버트의 컨텍스트(context) 벡터인 [CLS] 토큰(Special Classification token으로 모든 문장의 가장 첫 번째 토큰으로 삽입되며, 분류작업에 주요하게 사용됨[27])의 벡터를 분류하고자 하는 클래스 개 수 만큼의 덴스(dense) 레이어로 전달하고 활성화하여 결과값을 출력하였다.

<그림4 버트 컨텍스트 벡터인 [CLS] 토큰을 활용한 특허상담 텍스트 데이터 임베딩 예시>



3.3.2. 테스트 데이터에 대한 자동분류 실행 및 데이터 비교

위와 같이 학습한 모델 '특허상담분류버트'로 2021년, 2022년 테스트 데이터를 자동분류를 실행해 본 결과 다음과 같은 분류 분포가 도출되었다.

<표10 2021 ~2022년 특허상담분류버트 자동분류와 TA 자동분류 분포(단위: %)>

구분	2021년	2022년	2021년	2022년
등록	33.17%	36.46%	22.91%	24.90%
출원/중간	24.29%	22.96%	21.45%	22.40%
특허넷/전자출원	14.74%	13.35%	18.00%	16.47%
수수료	3.43%	3.49%	16.01%	14.11%
심사	4.22%	4.05%	7.26%	6.98%
심판	3.25%	3.08%	2.66%	2.58%
국제출원	5.20%	5.09%	2.09%	2.18%
공통/기타	11.70%	11.51%	9.62%	10.37%
분류주체	특허상담분류버트		기존 TA	

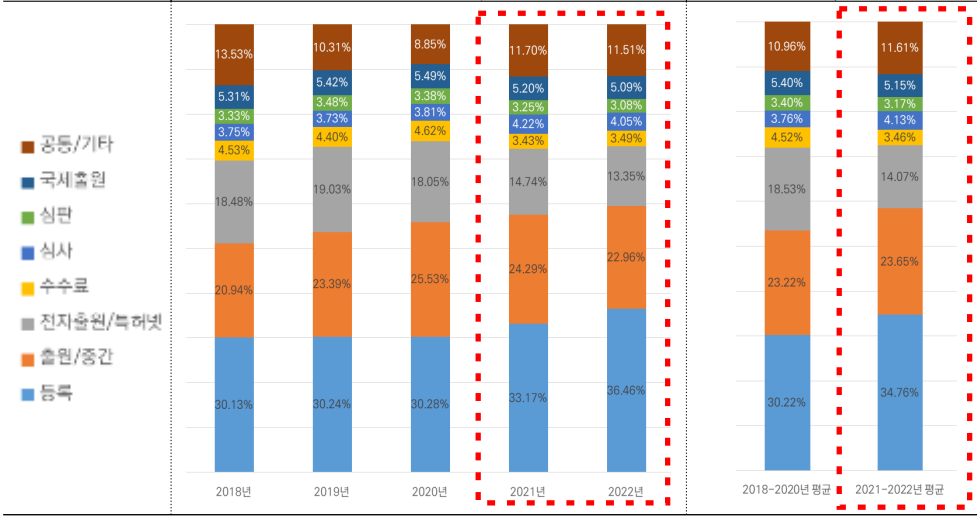
기존 TA가 분류한 자동분류 분포와 비교할 때 대부분의 분류에서 다소 큰 차이(특히 등록, 수수료 등에서 분포값이 크게 차이남)를 보였다. 2018 ~ 2020년 전문상담사가 직접 분류한 분포와 2021 ~ 2022년 특허상담분류버트 자동 분류 분포 결과로 최근 5년간 특허상담 분류 분포를 살펴보니 기존 TA 분류가 대부분의 분류에서 큰 차이를 보인 반면 특허상담분류버트로 자동 분류한 분류 분포는 기존 전문상담사가 직접 분류한 분포와 비슷한 양상을 보였다. 2018 ~ 2020년 전문상담사 분류 분포 평균값과 2021 ~ 2022년 특허상담분류버트 자동분류 평균값도 비슷한 양상을 보였다.

27) Jacob Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv*, 1810.04805(2018).

<표11 최근 5년간 특허상담 분류(대분류) 분포(단위: %)>

구분	2018년	2019년	2020년	2021년	2022년
등록	30.13%	30.24%	30.28%	33.17%	36.46%
출원/중간	20.94%	23.39%	25.53%	24.29%	22.96%
특허넷/전자출원	18.48%	19.03%	18.05%	14.74%	13.35%
수수료	4.53%	4.40%	4.62%	3.43%	3.49%
심사	3.75%	3.73%	3.81%	4.22%	4.05%
심판	3.33%	3.48%	3.38%	3.25%	3.08%
국제출원	5.31%	5.42%	5.49%	5.20%	5.09%
공통/기타	13.53%	10.31%	8.85%	11.70%	11.51%
분류주체	전문상담사			특허상담분류버트	

구분	연도별 분류 분포		평균 분류 분포	
	전문상담사 분류	특허상담분류버트 자동분류	전문상담사 분류	특허상담분류버트 자동분류

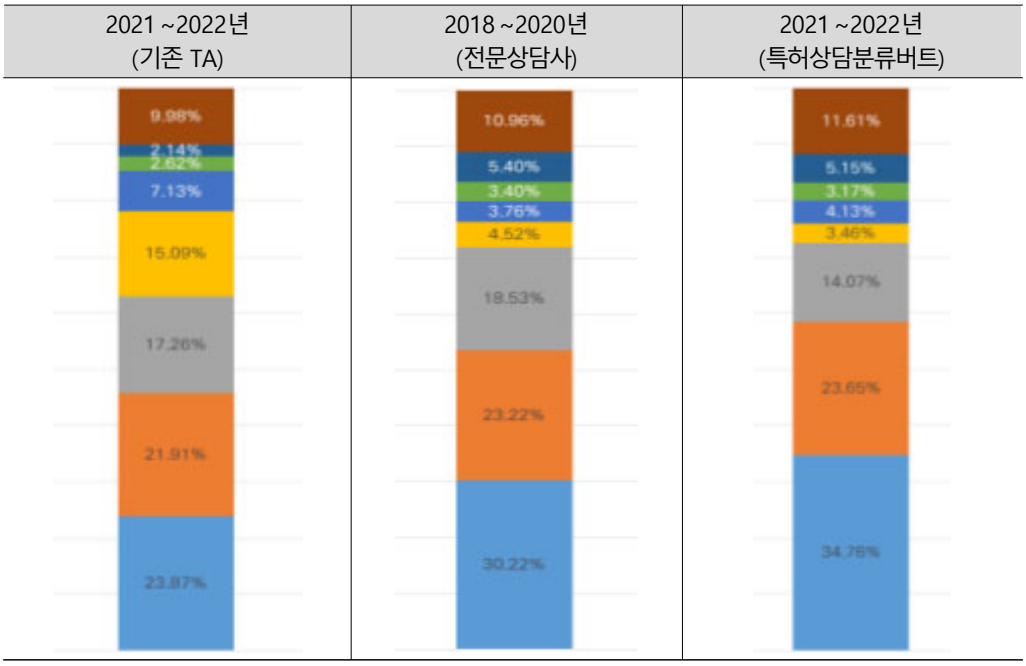


기존 전문상담사가 직접 분류한 2018~2020년 특허상담 분류 분포와 2021~2022년 TA 자동분류, 특허상담분류버트 자동분류 이 3개의 분류 분포의 평균값을 함께 놓고 비교해 보았다. 수치 및 그래프에서 TA 자동분류 분포보다 특허상담분류버트 자동분류 분포가 전문상담사가 직접 분류한 분류 분포 대비 차이가 적음(차이값의 절대값 합산이 15.90% 적음)을 확인했다. 또한, 전문상담사가 직접 분류한 분포값과 특허상담분류버트 자동 분류 분포값 중 가장 큰 차이를 나타내는 분류가 '등록'인데 이는 2018~2020년 산업재산권 연평균 등록건수(298,949건) 대비 2021~2022년 산업재산권 연평균 등록건수(334,307건)가 약 11.8% 증가한 것[28]으로 볼 때 자연스러운 증가 현상이라 해석할 수도 있을 것 같다.

28) 특허청, "2022년 지식재산 통계연보", 특허청, 2023, 7면.

<표12 2018~2020년 대비 2021~2022년 분류분포 평균값 및 차이 비교>

구분	2018~2020년 (전문상담사)	2021~2022년 (기존 TA)		2021~2022년 (특허상담분류버트)	
	분포값(a)	분포값(b)	| a-b |	분포값(c)	| a-c |
등록	30.22%	23.87%	6.34%	34.76%	4.54%
출원/중간	23.22%	21.91%	1.31%	23.65%	0.43%
특허넷/전자출원	18.53%	17.26%	1.27%	14.07%	4.45%
수수료	4.52%	15.09%	10.57%	3.46%	1.06%
심사	3.76%	7.13%	3.37%	4.13%	0.37%
심판	3.40%	2.62%	0.77%	3.17%	0.22%
국제출원	5.40%	2.14%	3.26%	5.15%	0.25%
공통/기타	10.96%	9.98%	0.98%	11.61%	0.65%
계	100.00%	100.00%	27.87%	100.00%	11.97%



3.4. 문제점

본 연구에서 기존 전문상담사가 직접 분류한 2018~2020년 특허상담 분류 분포와 2021~2022년 TA 자동분류, 특허상담분류버트 자동분류 이 3개의 분류 분포를 비교하여 특허상담분류버트가 전문상담사가 분류한 분포와 좀 더 유사함을 발견할 수 있었고, 이에 따라 특허상담 자동분류에서 기존의 TA 보다는 버트를 활용하여 자동분류를 수행하는 것이 더 성능이 우수할 것으로 사료된다. 그러나 연구의 여건상 전문상담사가 직접 분류한 데이터와 동일한 데이터로 3개의 분류 분포를 분류하면(1안, 2안) 보다 명확하고 신뢰할 수 있는 연구결과를 도출할 수 있을 것으로 예상되나, 시간 및 비용적인 한계로 3안으로 연구를 진행하여 1안, 2안의 연구방법만큼의 신뢰성은 확보하지 못했다. 다만 과거의 분류분포 데이터가 큰 변화 없이 비슷한 분포를 보이고 있는 것을 감안할 때 어느 정도의 성능 향상 효과가 있었다고 할 수 있겠다.

4. 결론

4.1. 연구의 요약

트랜스포머 기반 AI 모델인 버트를 사전 훈련시키고 파인튜닝하여 특허상담 내용에 대한 자동분류를 수행하고 전문상담사가 분류한 분포값과 기존 TA로 분류한 자동분류 분포값과 비교하여 연구를 수행한 결과, 기존 TA보다 특허상담분류버트로 자동 분류했을 때 분류 분포가 더 유사하게 나타났다. 이를 근거로 특허상담분류버트를 특허상담 자동분류에 활용할 때 보다 나은 성능을 기대할 수 있을 것이라 생각된다. 자동분류 결과가 보다 신뢰성 있게 도출되면 전문상담사의 업무부담 완화 및 효율성을 제고하고자 하는 정책적인 목적을 지속적.안정적으로 달성하고 특허고객상담 서비스 현황 및 고객의 니즈를 보다 정확하게 파악할 수 있어 특허고객 서비스 향상에 더 도움이 될 수 있을 것으로 사료된다.

4.2. 정책적 시사점 및 연구의 한계

본 연구를 진행한 결과, 특허상담 자동분류에 기존 특허상담 시스템에 구현된 TA 대신 사전훈련된 AI 모델 버트를 활용하면 더 신뢰성 있는 자동분류가 가능할 것으로 기대하며, 향후 특허상담 시스템 개선시 추가적인 대안이 없는 한 버트를 활용한 자동분류 기술을 도입하는 것에 대한 정책적인 검토가 요구된다. 본 연구에서는 자동분류를 1차 분류(대분류, 8개 분류)에 한정하였으며 정확도 80.05%의 자동분류모델을 구현하여 연구를 수행하였다. 더 면밀한 데이터 전처리 및 사전 훈련을 통해 버트를 활용한 자동분류의 정확도를 높이고 2차 분류(중분류), 3차 분류(소분류) 까지 특허상담분류버트(가칭)를 활용하여 연구를 확대 진행한다면 상담시스템에 실질적으로 적용할 수 있는 특허상담자동분류 시스템을 구현할 수 있을 것이다. 또한 1안, 2안의 연구방법 사용, 분포의 변화를 측정하기 위한 단일 지표 사용(예. national average index[29]), 개별 분류건에 대한 세부적인 비교 등 추가적인 후행 연구가 진행된다면 자동분류 성능 향상 연구의 신뢰성 및 타당성을 더 높일 수 있을 것으로 예상된다. 또한 상담서비스를 제공하고 있는 타 정부·공공기관 및 기업에서 테스크별 사전 훈련된 AI 모델 버트를 활용하면 양질의 자동분류를 구현할 수 있을 것으로 예상된다. 최근에는 버트 외에도 XLNet[30] 또는 DeBERTa[31] 모델 등이 텍스트 분류에서 좋은 성능을 발휘하고 있으므로[32] 이를 활용한 후행 비교 연구가 진행된다면 관련 기여도를 높일 수 있을 것으로 예상된다.

29) Australian Bureau of Statistics, “1381.0.55.001 - Research Paper: A Review of Selected Regional Industrial Diversity Indexes, 2011”, Australian Bureau of Statistics, <https://www.abs.gov.au/ausstats/abs%40.nsf/0/4276628C76F84A10CA257DAF0018B599%3FOpendocument>, 검색일: 2024. 2. 18.

30) Zhilin Yang et al., “XLNet: Generalized autoregressive pretraining for language understanding”, Edited by H. Wallach et al., Advances in Neural Information Processing Systems 32, 2019.

31) Ziyang Luo et al., “DecBERT: Enhancing the Language Understanding of BERT with Causal Attention Masks”, *arXiv*, 2204.08688(2022).

32) Papers with Code, “Text Classification”, Papers with Code, <https://paperswithcode.com/task/text-classification?>, 검색일: 2024. 2. 20.

참고 문헌(References)

단행본(국내 및 동양)

특허청, “2022년 지식재산 통계연보”, 특허청, 2023

한국지능정보사회진흥원, “IT&Future Strategy 2021년 보고서”, 한국지능정보사회진흥원, 2022

한규동, “AI 상식사전”, 길벗, 2022.

단행본(서양)

Ashish Vaswani et al., “Attention is all you need”, Edited by I. Guyon et al., Advances in Neural Information Processing Systems 30, 2017

Zhilin Yang et al., “XLNet: Generalized autoregressive pretraining for language understanding”, Edited by H. Wallach et al., Advances in Neural Information Processing Systems 32, 2019.

학술지(국내 및 동양)

강승태·장길진, “ChatGPT 와 다국어 BERT 를 이용한 코로나-19 감염병 다국어 기사 자동 색인 및 분류”, 「전자공학회논문지」, 제60권 제7호(2023).

권순보·유진은, “BERT 와 FastText 를 활용한 온라인 진로상담 문서 분류”, 「한국데이터정보과학회지」, 제33권 제6호(2022).

김명선·한동희, “기술문서의 한국표준산업분류 자동분류를 위한 특허기반 BERT 모델”, 「한국정보과학회 2022 한국소프트웨어종합학술대회 논문집」, (2022).

김인후·김성희., “딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류”, 「정보관리학회지」, 39권 3호(2022).

민소연 외 3인, “한국어 특성 기반의 STT 엔진 정확도를 위한 정량적 평가방법 연구”, 「한국산학기술학회논문지」, 제21권 7호(2020).

박진우 외 4인, “한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근”, 「지식재산연구」, 제17권 제3호(2022).

신진섭 외 5인, “양방향 LSTM 기반의 기계학습 데이터 구축 상담 자동 분류”, 「한국정보과학회 2021 한국컴퓨터종합학술대회 논문집」, (2021).

심재권, “BERT 를 활용한 초등학교 고학년의 욕설문장 자동 분류방안 연구”, 「창의정보문화연구」, 제7권 제2호(2021).

이경란·강창묵, “자연어처리 기계학습 기법을 이용한 공시문서의 자동분류: Confidential treatment를 가진 8-K 문서를 중심으로”, 「한국전자거래학회지」, 제28권 제2호(2023).

이용구, “BERT 모형을 이용한 주제명 자동 분류 연구”, 「한국문헌정보학회지」, 제57권 제2호(2023).

전영호, “BERT Transformer 와 Deep Learning 을 활용한 전이학습 효과 검증 연구: 법률상담데이터 분류문제 적용”, 「한국경영공학회지」, 제24권 제4호(2019).

학술지(서양)

Alec Radford et al., “Improving language understanding by generative pre-training”, Preprint, <https://paperswithcode.com/paper/improving-language-understanding-by>, (2018).

Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv*, 1810.04805(2018).

Ziyang Luo et al., “DecBERT: Enhancing the Language Understanding of BERT with Causal Attention Masks”, arXiv, 2204.08688(2022).

학위논문(국내 및 동양)

강유라, “KoBERT 를 활용한 내비게이션 장소 자동 분류 시스템”, 고려대학교 컴퓨터정보통신대학원, 석사, 2023.

권순보, “BERT를 활용한 진로상담 텍스트데이터 분석”, 한국교원대학교 대학원, 박사, 2022.

신문기사

박미영, “2022년 국가정보화 예산 약 11조, 올해보다 15% 증가 전망”, 보안뉴스, <https://www.boannews.com/media/view.asp?idx=103327>, 2021. 12. 15자.

최아름, “지역 맞춤형 디지털 특화사업 추진, 대한민국 디지털 전략 실현 지름길”, 정보통신신문, <https://www.koit.co.kr/news/articleView.html?idxno=106622>, 2022. 12. 6자.

인터넷 자료

서울대학교 AI연구원, “AI 용어사전”, <https://aiis.snu.ac.kr/>, 검색일 : 2023. 8. 1.

포지큐브, “TA(Text Analysis)는 무엇이고 어떻게 활용할 수 있나요?”, 포지큐브, <https://www.posicube.com/586335f6-7163-458a-8caf-0daccf2488bf>, 검색일: 2023. 8. 1.

Australian Bureau of Statistics, “1381.0.55.001 - Research Paper: A Review of Selected Regional Industrial Diversity Indexes, 2011”, Australian Bureau of Statistics, <https://www.abs.gov.au/ausstats/abs%40.nsf/0/4276628C76F84A10CA257DAF0018B599%3FOpendocument>, 검색일: 2024. 2. 18.

Papers with Code, “Text Classification”, Papers with Code, <https://paperswithcode.com/task/text-classification?>, 검색일: 2024. 2. 20.