석사학위논문

BERT와 Llama를 활용한 국내 학술지 논문의 자동분류 성능 비교

Comparison of automatic classification performance of Korean journal articles using BERT and Llama

지도교수 홍아름

경희대학교 테크노경영대학원

AI기술경영학과

강 광 선

2024년 2월

BERT와 Llama를 활용한 국내 학술지 논문의 자동분류 성능 비교

Comparison of automatic classification performance of Korean journal articles using BERT and Llama

지도교수 홍아름

이 논문을 석사학위논문으로 제출함.

경희대학교 테크노경영대학원

AI기술경영학과

강 광 선

2024년 2월

강광선의 경영학 석사학위
논문을 인준함

주심교수 마 민 철 ㊞

부심교수 강 송 희 ㊞

부심교수 홍 아 름 ㊞

경희대학교 테크노경영대학원

2024년 2월

< 목 차 >

< 표 차 례 >

< 그 림 차 례 >

국문초록

BERT와 Llama를 활용한 국내 학술지 논문의 자동분류 성능 비교

경희대학교 테크노경영대학원

AI기술경영학과

강 광 선

초거대 인공지능 오픈 AI사의 ChatGPT의 열풍으로 다양한 LLM 모델이 발표되었다. 2023년 2월 발표한 메타의 Llama 모델은 연구 커뮤니케이션에 오픈 하면서 거대 언어 모델의 생태계를 활성화하였다. Llama2는 SFT, RLHF를 반복 학습하여 ChatGPT 3.5와 유사한 성능을 구현 하면서 상업적으로도 이용한 모델이다. 문서 자동분류 분야에 많이 이용되고 있는 Bert 모델과 최신 LLM 모델인 Llama2 모델을 비교하여 Llama2 모델이 Bert 모델에 대비 문서 자동분류에서 성능이 향상되었는지 검증하려고 한다. 학습데이터는 AI-HUB에 '논문자료 요약' 데이터셋 사용하였다. 학습데이터는 1995년부터 2020년까지 데이터 16만건이며 대상 분류는 한국연구재단의 연구 분야 분류기준으로 8개 분류로 정의되어 있다. 본 연구를 위한 python 프로그램을 작성하였으며 Bert, Llama2의 학습 및 자동분류 성능 평가를 실행하였다. 본 실험의 결과는 학습데이터의 오차의 경우 Short model의 경우 Bert가 더 낮았고 middle, long model의 경우 Llama2가 더 낮았다. 학습데이터 정확도의 경우 short, middle, long model에서 Llama2가 Bert 보다 높은 정확도를 보였다. 행렬 분석한 결과 Bert의 경우 사회과학, 공학, 농수해양에서 높았으며 Llama2는 인문학, 자연과학, 의약학, 예술체육, 복합학에서 빈도가 높게 나왔다. 분류 평가에서 short 모델의 경우 Bert가 Llama2보다 정밀도, 재현율, F1 스코어에서 우세한 결과가 나

왔다. middle, long 모델의 경우 Llama2가 Bert 보다 정밀도, 재현율, F1 스코어에서 우세한 결과가 나왔다.

두 모델의 유의수준 5%의 쌍체 비교 t-검정을 실시하였다. short 모델은 성능 차이가 없었고 middle 모델의 경우 정밀도는 성능 차이가 있고 재현율, F1 스코어는 차이가 없는 것으로 나왔다. long 모델의 경우 재현율은 성능 차이가 없고 정밀도, F1 스코어가 성능 차이가 있는 것으로 나왔다. 문서 자동 분류 모델 선택시 입력 길이가 Short 텍스트일 경우 Bert 모델, Long 텍스트일 경우 Llama2 모델의 사용을 고려할 필요가 있다. 자동분류 모델 선택시 입력 데이터 길이에 따라 지표 판단의 기준이 되는 실증분석 결과를 제시 하였다.

향후 연구에서는 다양한 LLM 모델의 활용해 보고 제로샷(Zero-shot) 및 퓨샷(Few-shot) 학습을 이용한 문서 자동분류를 영역으로 연구하고자 한다.

주제어 : 인공지능, 자동분류, BERT, Llama, LLM

제1장 서 론

제1절 연구의 배경 및 목적

1. 연구의 배경

세계적으로는 '챗GPT(ChatGPT)'[1] 열풍이 생성형AI(Generative AI)[2]를 촉발시키면서 문서 내의 정보를 추출하고, 문서 간의 관계를 파악하고, 문서의 내용을 요약하고, 문서의 품질을 평가하는 등의 작업이 인공지능 기술을 활용하여 자동화될 수 있다(정의석, 2023). 많은 데이터가 기하급수적으로 증가하고 해당 문서의 자동분류에 관한 연구도 활발히 진행되어져 왔다. 2018년 11월에 BERT 모델이 발표되면서 자동분류 분야에서 많이 사용되고 있다. 황상흠, 김도형(2020)은 한국기술문서를 전처리하고 SKT BERT 모델을 사용하여 문서의 특징 토큰들을 추출한 다음에 문서분류에 적용하여 기술문서의 자동분류에 BERT가 의미 있는 성능의 효과를 보여주었다. 소현지, 이종태(2021)은 BERT-Triplet 모델 기반으로 데이터의 클래스 정보를 활용하는 word embedding 모델 학습을 수행하고, 자동분류의 성능 향상된 것을 확인하였다. 김인후, 김성희(2022)는 BERT 모델을 문헌정보학에 적용하여 문서 자동분류를 실행하여 데이터의 양과 품질에 따른 의미 있는 결과를 확인하였다. ChatGPT에 대항하기 위한 다양한 구글의 PaLM 모델 기반 바드(Bard), 메타의 Llama (Large Language Model Meta AI), 네이버의 HyperClovaX 등 LLM 모델을 출시 하였다(신성필. 2023). Llama는 GPT-3(175B)에 비하면 10%도 안 되는 크기의 모델에 데이터를 4배 이상 사용하여 학습한 결과 Llama-13B가 GPT-3(175B)보다 더 좋은 성능을 보여주었다(Touvron et al., 2023). Llama 2-Chat이라고 불리는

1) 2022년 11월 OpenAI가 개발한 인공지능 대화형 챗봇(위키피디아)
2) 이용자의 특정 요구에 따라 결과를 생성해내는 인공지능을 말한다(나무위키)

모델을 미세 조정된 Llama2는 상업적 이용 가능한 오픈소스로 2023년 8월에 배포했다. Llama 모델이 공개적으로 개방 하면서 LLM 모델 연구의 OpenSource화 및 대중화가 촉발하였다(주하영 외, 2023). Llama 모델을 파생된 다양한 언어 모델이 Github[3], Huggingface[4] 와 같은 공개 플랫폼에 공유하면서 LLM 모델의 오픈소스 환경이 조성되었으며, KoAlpaca, Ployglot 등 뛰어난 성능과 경량화 된 모델들이 등장하며 LLM 모델 생태계가 활성화하였다

공개된 Llama의 최신 모델인 Llama2의 모델을 문서 자동분류 분야에서 이용가능한지 확인하고 Bert 모델과 성능을 비교하여 유의미한 성능 향상이 있는지 파악하려고 한다. LLM Llama 모델의 향후 가능성을 확인하고 인공지능 모델을 이용한 다양한 서비스 구현하는데 도움이 되는 연구로 제안한다.

2. 연구의 목적

본 연구는 문서 자동분류 분야에 많이 이용되고 있는 인공지능 Bert 모델과 최신 LLM 모델인 Llama2 모델을 비교하여 Llama2 모델이 Bert 모델에 대비 문서 자동분류에서 성능이 향상되었는지 검증하려고 한다.

학습 및 검증 데이터는 AI 허브의 논문요약 데이터를 활용하여 논문 초록 데이터를 이용하였다. Bert 모델과 Llama2 모델을 미세조정(Fine-tuning) 학습하였으며 테스트 데이터를 이용하여 성능평가를 진행하였다. 분류 모델 성능 측정 지표인 정밀도(precision), 재현율(recall), F1 스코어(F1-score)를 측정하고 Bert 모델과 Llama2 모델의 성능 차이가 있는지 쌍체 비교 t-검증을 하였다.

3) 분산 버전 관리 툴인 깃 저장소 호스팅을 지원하는 웹 서비스 . 2011년 오픈 소스 소프트웨어 인터넷 호스팅 서비스(위키피디아).
4) 트랜스포머나 데이터셋 같은 머신러닝 프레임워크를 제공하는 세계 최대의 인공지능 플랫폼 중 하나이다(나무위키).

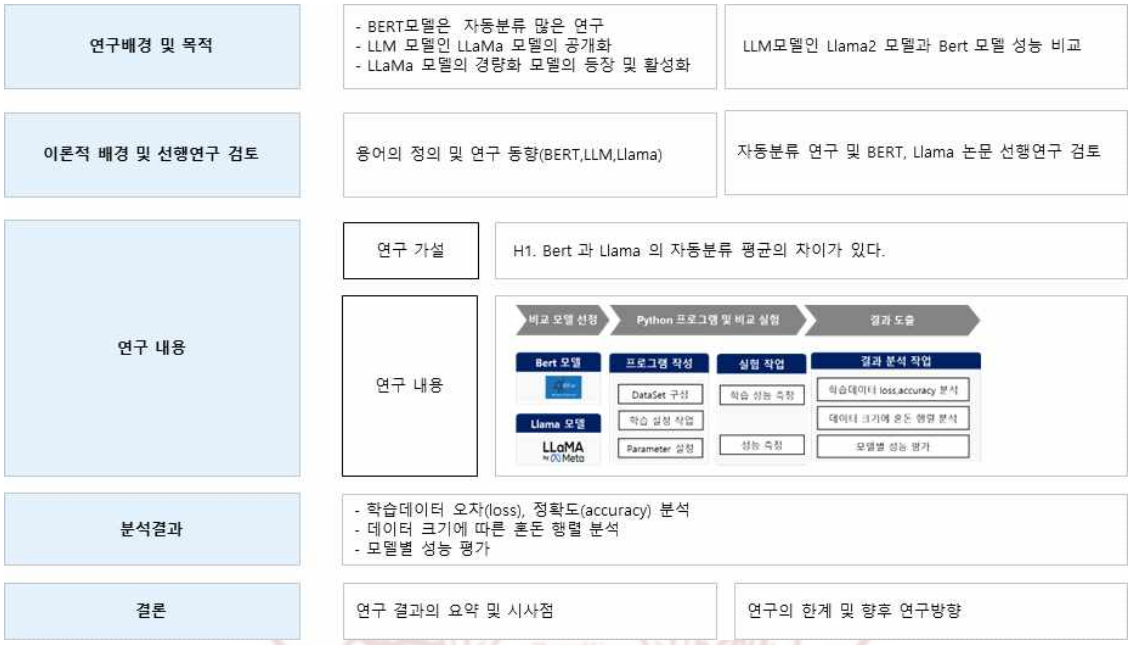
제2절 연구의 범위 및 방법

본 연구는 인공지능 LLM 모델인 Llama의 최신 모델인 Llama2와 기존 자동분류 분야에서 많이 사용 하는 Bert 모델을 비교하여 Llama2 모델이 Bert 모델보다 문서 자동분류에서 성능 향상이 있는지 확인하는 목적으로 하며 구체적인 연구를 위하여 Bert, LLM, Llama등 이론적 배경과 자동분류 선행연구를 검토하였다. 기존 연구에서는 Bert를 이용한 자동분류에 대한 논문과 Llama 한글과 관련된 논문을 검토하였다. 자동분류 논문인 한국어 기술문서 분석을 위한 Bert 기반의 분류모델(황상흠, 김도영, 2020)에서는 기술문서에 Bert를 이용한 자동분류의 성능의 의미성을 확인 하였다. Bert와 FastText를 활용한 온라인 진로상담 문서 분류(김인후, 김성희, 2022)에서는 FastText 보다 Bert가 문서분류에서 우수함을 입증하였다. 딥러닝 기반의 Bert 모델을 활용한 학술 문헌 자동분류 (김인후, 김성희, 2022)에서는 문헌정보학 분야의 13개 분류에서 자동분류를 실행하여 입력 길이가 64인 모델만 9개 분류에서 128, 256인 경우 10개 분류에서 유의미한 결과를 확인하였다. Llama 한글 논문인 인간과 ChatGPT의 대화내용을 이용한 공개 대형 언어모델 Llama 한국어 대화 능력 개선 (주호택 외, 2023)에서는 Llama 에 LoRA 방법의 파인튜닝을 하여 Llama 모델은 한국어 능력이 부족하지만 한국어 의사소통 능력을 개선효과가 있다는 것을 증명하였다.

본 연구에 사용한 데이터는 AI-HUB[5]에 '논문자료 요약' 데이터셋 16만건을 이용하여 실험하였다. 본 연구를 위한 python 프로그램을 작성하였으며 Bert, Llama2의 학습 및 자동분류 성능 평가를 실행하였다. 분석 결과는 학습데이터 오차(loss), 정확도(accuracy) 분석 및 데이터 크기에 따른 혼돈 행렬 분석을 하였다. 분류척도는 정밀도, 재현율, F1-Score의 데이터를 분석하였으며 이 결과를 이용하여 T-test을 실행하여 자동분류의 성능 차이가 유의미한지 평가하였다.

5) AI-Hub는 한국지능정보사회진흥원이 운영하는 AI 통합 플랫폼(위키피디아).

<그림 1> 논문의 구성



본 논문은 <그림 1>과 같이 구성하였다. 제1장의 서론에서는 연구의 배경과 목적을 설명하며, 제2장에서는 이론적 배경 및 선행연구 검토를 하였으며. 제3장에서는 연구 가설, 연구 내용을 이용하여 실험을 진행 하였다. 제4장에 분서결과에서는 모델별 성능평가를 진행 하였다. 제 5장에서는 연구 결과의 요약 및 시사점, 연구의 한계 및 향후 연구방향에 대해 기술하였다.

제2장 이론적 배경과 선행연구 검토

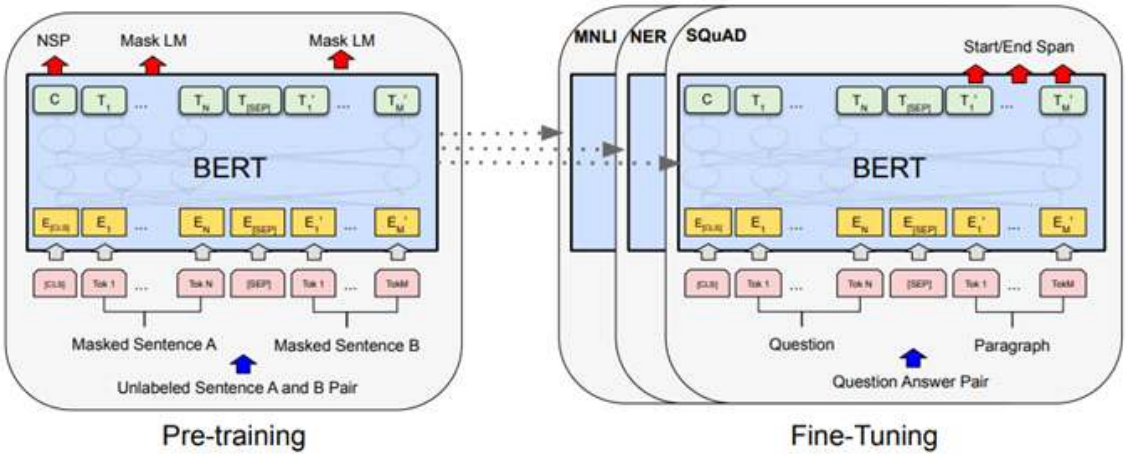
제1절 이론적 배경

1. BERT

최근 인공지능기술의 발달에 따라 다양한 문제 해결을 위한 연구가 많아지고 있으며, 텍스트에 대한 자연어처리 분야에 활발한 연구가 진행되고 있다. 자연어처리 분야의 대표적인 문제 유형으로는 문헌 분류(Classification), 기계독해(Machine Reading Comprehension), 기계번역(Neural Machine Translation), 의미 검색(Semantic Search) 등이 있다. 기존에는 각 문제의 특성에 따른 데이터와 딥러닝 모델을 별도 학습하여 문제를 해결하는 방식이 주된 흐름이었다. 2018년 11월 Google BERT 언어모델이 공개된 이후로는 각 분야에 존재하는 방대한 자연어데이터를 1차적으로 사전학습(Pre-training)하여 언어모델을 구성한 뒤, 이를 활용하여 2차적으로 각 개별적인 문제에 맞게끔 미세조정(Fine-tuning) 또는 전이학습(Transfer Learning)을 통하여 문제를 해결하는 방식이 효과적인 것으로 검증되었다. 그래서 최근까지 이러한 방식이 자연어처리 문제를 해결하는 데에 활용되고 있다(박진우 외, 2022).

BERT는 <그림 2>와 같이 대량의 텍스트를 사전학습 된 모델을 만들고 비교적 작은 학습 데이터로 파인튜닝(fine-tuning)하는 방식으로 재학습 하게 된다. BERT의 사전훈련 기법은 대규모 텍스트를 적용하여 우수한 단어 표현(word representation)을 획득할 수 있는 장점이 있다. 파인튜닝 단계에서 모델 자체를 재구조화 하지 않고 적은 학습 데이터로 높은 성능을 달성할 수 있다(이수빈 외, 2021).

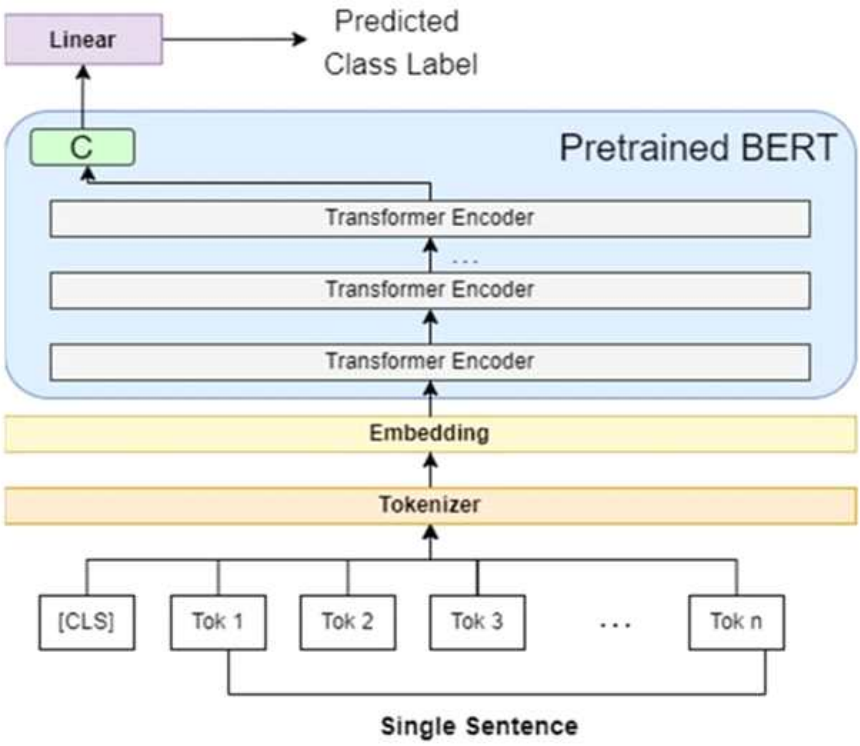
<그림 2> BERT 전이학습 개념



*자료: (Devlin et al., 2018)

BERT 학습 원리는 데이터 학습시 입력내용 중 일부 단어를 생략하고, 이 단어를 모델이 예측 학습 하도록 하였으며 단어가 포함된 문장에 따라 그 단어의 임베딩 값을 변화되도록 하였다. BERT 모델은 536개의 위키피디아 문장에 대한 107,785개의 질문-대답 데이터세트인 SQuAD 데이터세트를 통해 학습한 결과, 인간을 뛰어넘는 성능을 보여주었다(성소윤 외, 2019). 자동분류에서 BERT는 <그림 3>과 같이 사전 훈련된 모든 매개변수를 미세 조정하며 분류의 경우 최종 은닉 벡터 C만 사용하여 레이블을 예측한다(Lee, E et al., 2022).

<그림 3> BERT 분류 개념도



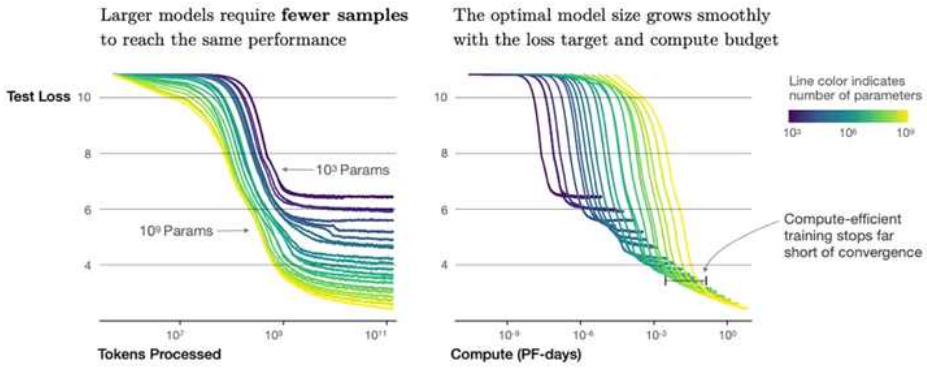
*자료: (Lee, E et al., 2022)

2. LLM (Large Language Model)

다음 단어 예측 및 생성과 관련 범용 인공지능 언어모델이 경쟁이 치열하게 발전하였다. <그림4>과 같이 모델의 구조보다는 언어모델의 성능은 매개변수, 데이터, 컴퓨팅 파워에 좌우된다는 연구가 발표되어 초거대 모델에 대한 관심이 집중되었다(Hu, E et al., 2021). 초거대 인공지능은 오픈 AI사의 ChatGPT라는 게임체인저의 등장으로 인하여 사회 전반에 다양한 산업적, 경제적 패러다임 변화를 일으키고 있다. AI 챗봇 시스템으로, 간단한 질의응답 기능뿐 아니라 다양한 도메인에서 전문가 수준의 답변을 제공하는 등 강력한 성능을 제공하여 세계적인 흥행에 성공하였다(신성필, 2023). ChatGPT에 대항하기 위한 다양한 구글의 바드(Bard, 구글의 PaLM

모델 기반), 메타의 Llama (Large Language Model Meta AI), UAE에 팰컨 40B가 있으며 국내에서는 LG의 EXAONE, 네이버의 HyperClovaX 등 LLM 모델을 출시되었다.

<그림 4> 언어모델 성능의 요소



* 자료 : (Hu, E et al., 2021)

<표 1>은 지금까지 소개한 초거대 AI 기반모델의 특징을 보여준다.

<표 1> 국내외 초거대 AI 기반모델 동향

기반모델	출시년월	기업	파라미터	AI 서비스	기타
GPT-3	2020년 6월	오픈AI / MS	1,750억	ChatGPT 등	약 800GB의 신경망 규모
GPT-4	2023년 3월	오픈AI / MS	미공개	ChatGPT, MS Bing,등	이미지 처리/해석 등의 멀티모달 기능 지원
PaLM	2022년 4월	구글	5,400억	Bard 등	영어와 다국어 데이터 학습
Llama	2023년 2월	META	650억	Alpaca 등	비상업 라이센스로 가중치 공개
Llama2	2023년 8월	META	700억	추후 공개	상업적 이용 가능한 오픈소스
HyperClovaX	2023년 8월	네이버	2,040억	ncloud 서비스 등	ChatGPT에 비해서도 한국어 데이터를 6,500배 더 학습

*자료: (신성필, 2023)

대규모 언어 모델(LLM)의 사전 지식을 활용한 임베딩 하는 능력에서 벗어나 언어 자동 번역, 자동 프로그램 코딩까지 매우 다양한 능력을 보여주고 있다. 대규모 언어 모델들이 가진 폭넓은 상식과 추론 능력, 자동 작업 계획 능력들도 다양한 분야들에서 효과적으로 활용될 수 있다고 알려지고 있다(백호준, 김인철, 2023).

3. Llama

2023년 2월 페이스북의 메타(Meta)는 대규모언어 모델인 Llama를 배포했다. AI연구자들의 연구 발전을 돕기 위해 Llama를 출시했다고 언급하며, 텍스트 생성, 대화, 자료 요약, 수학 풀이, 단백질 구조 예측, 코드 도움 등이 가능한 이 모델을 AI 연구 커뮤니티에 제공하였다(Touvron, H et al. 2023). 메타가 공개한 이 대형모델은 파라미터가 70억 개부터 650억 개에 이르는 모델 4종을 출시하였으며, 비상업적 라이센스에 따라 대학, 기관 등에서 무료로 사용할 수 있다(주호택 외, 2023). 이 거대한 GPT계열 디코더 언어 모델이 발표되고, 더 좋은 성능을 내기 위해 더 큰 모델을 학습하는 흐름 속에서 Llama는 GPT-3(175B)에 비하면 10%도 안 되는 크기의 모델에 데이터를 4배 이상 사용하여 학습한 결과 Llama-13B가 GPT-3(175B)보다 더 좋은 성능을 보여주었다(Touvron, H. et al., 2023). Llama 2-Chat이라고 불리는 모델을 미세 조정된 Llama2 는 기존의 LLM 언어 모델보다 우수한 성능이 나왔으며 상업적으로도 이용 가능하며, 가장 최근에 공개된 모델임에도 연구가 활성화되고 있다(주하영 외, 2023). <표 2>와 같이 ChatGPT-0301 (gpt-3.5-turbo)버전과 성능 유사하다(Touvron, H. et al., 2023).

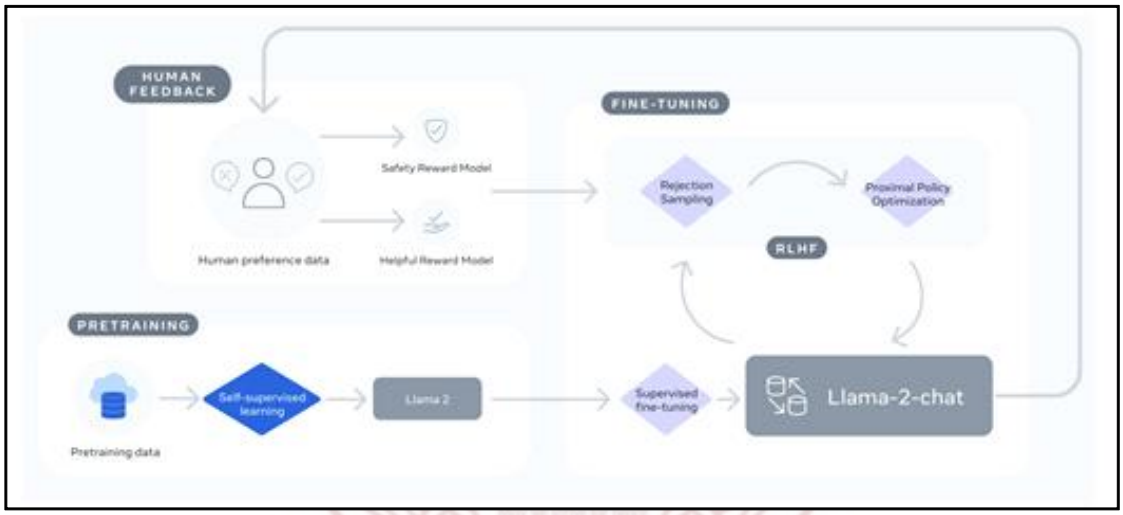
<표 2> Llama2의 벤치마크 비교

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	Llama 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	-	-	81.4	**86.1**	85.0
Natural Questions(1-shot)	-	-	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	-	29.9
BIG-Bench Hard (3-shot)	-	-	52.3	**65.7**	51.2

*자료: (Touvron, H. et al., 2023)

<그림5>와 같이 Llama2는 공개된 데이터셋을 자기지도학습을 하여 생성된 Llama2를 SFT[6]하여 Llama2-Chat 모델을 생성하게 한 후, Llama2-Chat은 RLHF[7]을 사용하여 반복적인 튜닝을 하였다.

<그림 5> Llama2 학습도



*자료: (Touvron, H. et al., 2023)

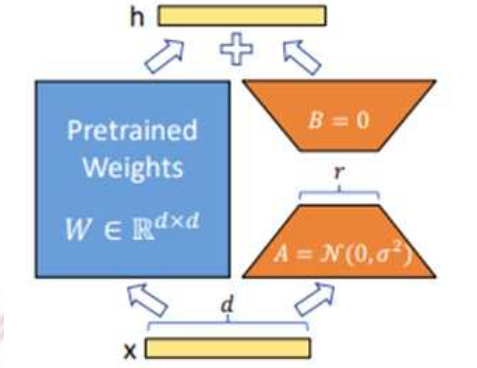
4. Low-Rank Adaptation

LLM 모델들을 파라미터 수가 1,000억 개가 넘어가면서 Fine-tuning 하려면 GPU 연산이 많은 시간이 소요된다. 거대한 파라미터를 효율적으로 다루기 위한 방법(Parameter Efficient Fine tuning, PEFT)으로 최근 Low-Rank Adaptation(LoRA)이 주목을 받고 있다(정단호 외, 2023). Low Rank Adaptation는 pretrained model의 모든 weight를 파인튜닝 하는 방법 대신 pretrained model weight를 모두 freeze하고 각 layer마다 rank decomposition matrice을 추가하는 방식으로 parameter 효율적으로 훈련하는 방법이다. <그림 6>과 같이 사전 학습한 모델의

6) Supervised Fine-Tuning
7) 인간 피드백을 통한 강화 학습(Reinforcement Learning from Human Feedback)

weight 들은 업데이트 하지 않고, LoRA 의 rank decomposition matrices 의 weight 들만 업데이트 하는 것이다.

<그림 6> LoRA 파라미터 개념도



*자료 : (Hu, E et al., 2021)

<표 3>와 BART와 T5 모델을 미세조정 (FT)과 LoRA를 적용하여 질문 생성 태스크에 대한 성능을 보여준 표이며 해당 LoRA를 이용시 파라미터가 100배, 80배 정도 차이나는 것을 볼 수 있으며 SQuAD-QG[8] 성능에서 비슷한 성능을 보여주었다 (박규민 외, 2023).

<표 3> SQuAD-QG의 벤치마크 비교

Model & Method	# Trainable Parameters	rank r	SQuAD-QG		
			SacreBLEU[9]	METEOR[10]	ROUGE-L[11]
$BART_{Large}$ (FT)	406M	-	20.22	46.82	45.30
$BART_{Large}$ (LoRA)	4M	32	20.59	47.26	45.15
$T5_{Large}$ (FT)	737M	-	21.47	48.44	46.07
$T5_{Large}$ (LoRA)	9M	64	20.80	47.25	45.20

*자료: (박규민 외, 2023)

8) SQuAD 1.0 데이터셋은 위키피디아 기반으로 생성된 질의응답 데이터셋이다. SQuAD-QG은 <passage, answer, question> 묶음의 데이터셋임
9) 학습된 모델의 생성 sequence 결과가 실제 정답과 얼마 유사한지 측정하는 metric (Kim, A., Kim, J, 2022)
10) METEOR는 Unigram들의 precision과 recall의 조화 평균으로 정의(위키피디아)
11) LCS 기법을 이용해 최장 길이로 매칭되는 문자열을 측정(위키피디아)

Adam Optimizer으로 파인튜닝 한 GPT-3 175B의 경우 약 1000배 가량의 파라미터 수를 줄였으며, 학습 중 VRAM 소비를 1.2TB에서 350GB로 3배 정도 줄였다 (Hu, E et al., 2021).

5. 학술연구분야분류표

한국연구재단의 연구 분야 분류기준표이다. (인문사회)학술연구지원사업의 효율적인 추진 및 관리 운영 등을 목적으로 활용되고 있으며 연구자정보관리, 학술연구지원의 관리 통계, 대학의 연구 활동 실태 등의 조사, 인문사회분야 연구과제의 접수와 심사 및 평가자의 선정 등에 활용되고 있다. <표 4>와 같이 재단 연구 분야는 대분류-중분류-소분류-세분류 등 4개 단계로 구성되어 있다. 대분류 체계는 인문학, 사회과학, 자연과학, 공학, 의약학, 농수해양, 예술체육, 복합학 등 8개로 구성되어 있다.

<표 4> 학술연구분야분류표

대분류	중분류	소분류	세분류
인문학	23	167	298
사회과학	22	269	479
자연과학	13	135	371
공 학	28	310	457
의약학	39	409	648
농수해양학	7	64	132
예술체육학	12	104	61
복합학	8	93	22
합계	152	1,551	2,468

제2절 선행연구 검토

<표 5> 선행연구 논문

논문명	주요 내용
Text classification에 특화 시킨 개선된 BERT 활용 방법론 제안 (소현지, 이종태, 2021)	- BERT-Triplet 모델은 데이터의 문맥만이 아닌 클래스 정보를 추가로 반영하는 분류 반영 - BERT-Triplet 모델을 적용 한 결과, 기존 방법론에 비하여 텍스트 분류의 정확도가 향상 확인
한국어 기술문서 분석을 위한 BERT 기반의 분류모델 (황상흠, 김도형, 2020)	- 총 33 개의 중분류기술명로 한국어 기술문서 데이터 는 7,108로 테스트 - BERT를 이용한 문서 분류에서 F-score 및 평균 정밀도 값이 의미가 있음을 확인. - 문서단위의 분류성능이 의미가 있음.
딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류 (김인후, 김성희, 2022)	- 문헌정보학 분야의 KCI 등재 논문 5357개 논문의 초록을 13개의 주제로 분류 - 성능 평가척도는 정확률,재현율,F 척도를 사용 - 입력 길이가 64인 모델만 9개, 128, 256인 경우 10개를 유의미하게 분류
BERT와 FastText를 활용한 온라인 진로상담 문서 분류 (권순보, 유진은, 2022)	- 커리어넷 진로상담 게시판의 데이터를 4,412건을 학습 평가 - 정확도가 BERT 0.768, FastText는 0.624로 BERT가 더 좋은 결과를 보였으며 정밀도, 재현율, F1-score에서 BERT가 우수하였다.
특허문서 분류를 위한 딥러닝 개별 모델 분류기 성능비교 (김인후, 김성희, 2022)	- US특허의 CPC 분류데이터를 사용하도록 한다. MLP, CNN, LSTM, Attention, Transformer 모델을 자동분류에 성능 비교를 하였다. - 데이터셋 1은 Attention이 데이터셋2는 LSTM이 , 데이터셋3는 Attetion이 좋은 성능이 나옴.

1. 한국어 기술문서 분석을 위한 BERT 기반의 분류모델 (황상흠, 김도형, 2020)

한국어 기술문서 데이터는 국가과학기술 지식정보서비스에 등록된 자료를 활용하였다. 인공지능 및 지능형로봇 분과 데이터를 사용하였다. 데이터 7,108개를 33개의 중분류 기술명으로 분류하였다. 테이터는 학습 데이터 4,976개, 테스트 데이터 2,132개를 이용하였다. BERT 모델은 파인튜닝 학습하였으며 Adam optimizer를 사용했고 Learning rate는 0.00005로 했다. 총 50 epoch 학습을 수행했고, mini batch의 32 사이즈로 실험 하였다.

<표 6>와 같이 개별요소 단위 성능 지표를 보면 기술문서 데이터의 분류별 불균형이 심하였으며, 개별 기술문서가 속하는 기술 분류의 수가 매우 적은 점 등을 고려하면, 우수한 예측 성능으로 판단할 수 있다

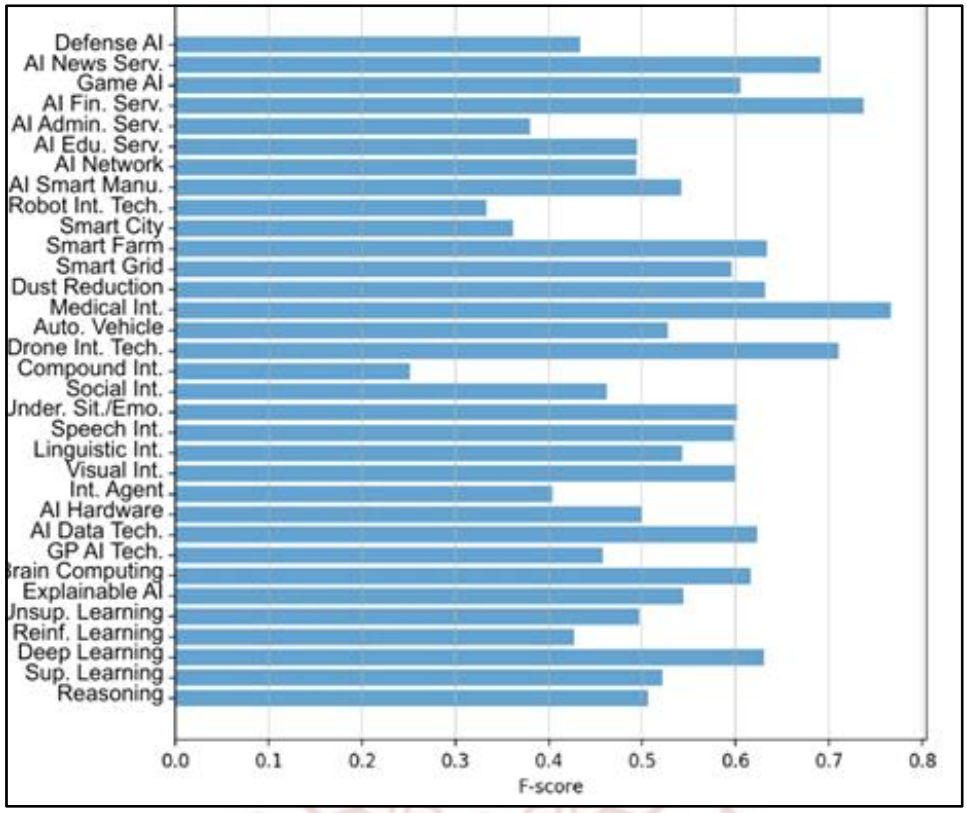
<표 6> 개별요소 단위 성능 지표

Performance measures			
Accuracy	F-score	Precision	Recall
0.9368	0.556	0.5591	0.5531

*자료: (황상흠, 김도형, 2020)

<그림 7>과 같이 기술 분류 단위별 성능 지표를 보면 ”정밀의료”, “AI 기반금융” 등 예측 성능이 가장 높은 경우에는 해당 기술 분야들의 일부 단어들이 분류 특징짓는 것이 쉽기 때문이다. “복합지능” 기술에 대해서는 예측 성능이 낮은데 기술 분야 자체가 명확하게 특징짓기 어렵기 때문이다.

<그림 7> 기술 카테고리 F-score



*자료: (황상흠, 김도형, 2020)

<표 7>와 같이 문서 단위 성능 지표의 결과, 주어진 과제의 특성과 데이터를 고려해보면 예측력 있음을 확인했다. 학습된 모델이 평균 0.5 이상의 F-score 및 정밀도 결과가 나왔으며, 문서단위의 분류에서는 문서 단위 성능이 유의미함을 확인했다.

<표 7> 문서 단위 성능 지표

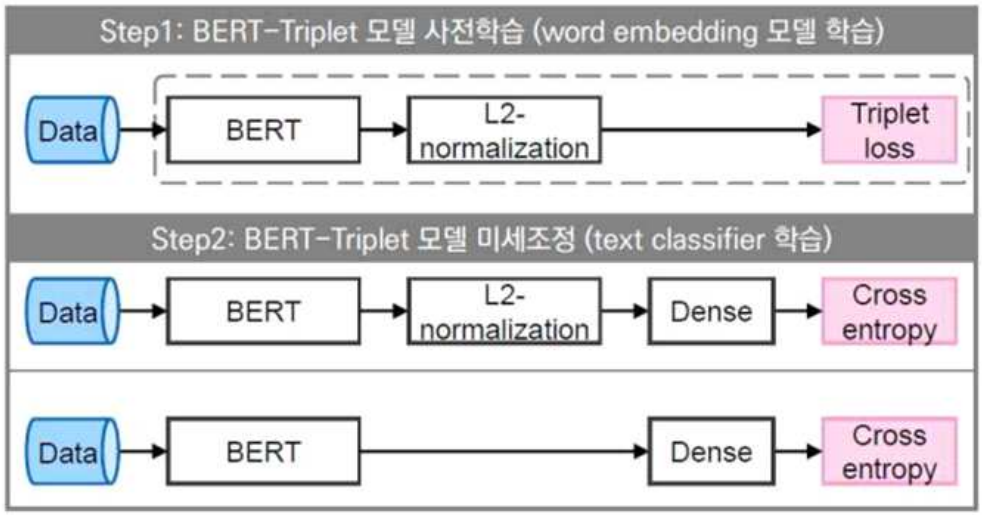
Performance measures		
F-score	Precision	Recall
0.5336	0.5625	0.5655

*자료: (황상흠, 김도형, 2020)

2. Text classification에 특화 시킨 개선된 BERT 활용 방법론 제안(소현지, 이종태, 2021)

<그림 8>과 같이 BERT-Triplet 모델은 단어를 임베딩 벡터로 변환하는 과정에서 데이터의 문맥만이 아닌 클래스 정보를 추가로 반영하는 것을 목적으로 한다. TREC 6가지 대분류를 기준으로 분류한 데이터 세트 및 50가지 소분류 중 테스트 데이터가 존재하는 42가지의 클래스를 기준으로 분류한 데이터 세트를 사용하였다. 대분류는 TREC-6, TREC-42로 명명하였다.

<그림 8> BERT-Triplet 모델 프레임워크



*자료: (소현지, 이종태, 2021)

<표 8>, <표 9>과 같이 사전 학습 word embedding 모델 BERT를 자연어처리의 대표 작업인 text classification에 특화 시켜 활용할 수 있는 방법론을 제안하였다. 실제 제안된 BERT-Triplet 모델을 두 가지 데이터세트에 적용 한 결과, 기존 방법론에 비하여 텍스트 분류의 정확도가 향상되었다. BERT-TRI-L2-FT가 가장 높은 정확도 및 가장 낮은 손실을 가짐을 확인하였다.

<표 8> TREC-6 text classification 테스트 결과

분류	모델명	Accuracy	Loss
기존 방법론	BERT-FT	0.9720	0.1561
	BERT-FUR-FT	0.8480	0.5131
제안 방법론	BERT-TRI-L2-FT	0.9680	1.2285
	BERT-TRI-FT	0.9780	0.1186

*자료: (소현지, 이종태, 2021)

<표 9> TREC-42 text classification 테스트 결과

분류	모델명	Accuracy	Loss
기존 방법론	BERT-FT	0.9720	0.1561
	BERT-FUR-FT	0.8480	0.5131
제안 방법론	BERT-TRI-L2-FT	0.9680	1.2285
	BERT-TRI-FT	0.9780	0.1186

*자료: (소현지, 이종태, 2021)

3. 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류 (김인후, 김성희, 2022)

문헌정보학 분야의 KCI 등재된 논문을 7 개 학술지의 5357 개 논문의 초록을 수집하였으며 중복 없이 13 개의 주제로 단일 항목으로 분류하여 분석하였다. 수집된 데이터를 이용하여 학습된 데이터의 크기에 따른 자동분류의 정확도에 변화가 있는지를 분석, 평가하였다. 입력크기를 Short Model(64), Middle Model(128), Long Model 3 가지로 구분하였으며 정확률(Precision), 재현율(Recall), F 척도를 이용하여 성능 평가하였다.

<표 10>과 같이 3 가지 모델에 따른 분류 결과는 입력 길이가 short(64) 모델은 9 개를 유의미하게 분류하였다. middle(128), long(256) 모델들은 10 개를 유의미하게 분류하였다. 데이터의 품질에 따라 결과는 분류 품질이 좋은 경우 입력 길이가 커지면서 성능이 향상되었고 반대로 데이터 품질이 나쁜 경우 입력 길이가 커지면서 성능이 하락 되었다.

<표 10> BERT 3가지 모델에 따른 성능 평가

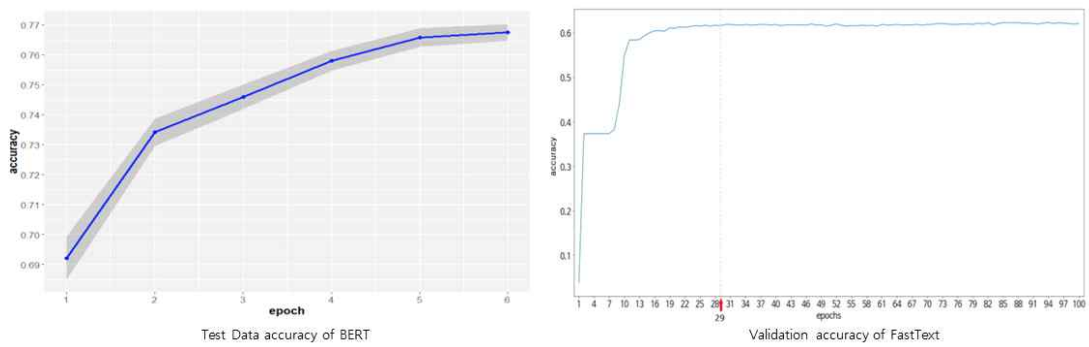
	Short model			Middle model			Long model		
	Precision	recall	F 척도	Precision	recall	F 척도	Precision	recall	F 척도
문헌정보학일반	0.582	0.48	0.526	0.572	0.531	0.551	0.601	0.484	0.536
기록관리/보존	0.915	0.915	0.915	0.9	0.927	0.913	0.91	0.914	0.912
서지학	0.965	0.973	0.969	0.951	0.977	0.964	0.951	0.964	0.958
도서관/정보센터경영	0.712	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

*자료: (김인후, 김성희, 2022)

4. BERT와 FastText를 활용한 온라인 진로상담 문서 분류(권순보, 유진은, 2022)

중학생과 고등학생의 상담 내용이 있는 커리어넷 진로상담 게시판의 데이터 4,689건을 수집 및 전처리하여 4,412건을 네 가지유형으로 분류하였으며 훈련 데이터넷(3,529 개)과 시험 데이터(883 개)로 나누어 진행하였다. 온라인 진로상담 자료 분류 문제에 있어 FastText와 BERT의 예측 성능을 비교 실험하였다. <그림 9>와 같이 정확도가 BERT 0.768, FastText는 0.624로 BERT가 더 좋은 결과를 보였다.

<그림 9> BERT FastText 학습시 정확도



*자료: (권순보, 유진은, 2022)

<표 11>와 같이 정밀도, 재현율, F1-score 에서 BERT 가 우수하였다. BERT 는 사례 수가 적은 범주의 예측에 특히 두각을 나타내었다.

<표 11> BERT와 FastText 분류 결과

Measures \ Categories		1		2		3		4	
		Bert	FastText	Bert	FastText	Bert	FastText	Bert	FastText
Precision	mean	0.805	0.626	0.764	0.640	0.703	0.547	0.724	0.606
	sd	0.019	0.005	0.028	0.006	0.079	0.063	0.032	0.005
Recall	mean	0.784	0.697	0.787	0.638	0.677	0.109	0.736	0.590
	sd	0.027	0.005	0.029	0.008	0.084	0.009	0.031	0.006
F1-score	mean	0.794	0.660	0.775	0.639	0.684	0.181	0.730	0.598
	sd	0.016	0.004	0.019	0.006	0.060	0.014	0.020	0.005

*자료: (권순보, 유진은, 2022)

5. 특허문서 분류를 위한 딥러닝 개별 모델 분류기 성능비교(김성훈, 김승천, 2021)

US특허의 CPC 분류데이터를 사용하며 <표 12>과 같이 CPC C01 계열 데이터를 훈련데이터, 검증데이터 그리고 테스트 데이터로 분리한다. MLP, CNN, LSTM, Attention, Transformer 모델을 자동분류에 성능 비교를 하였다.

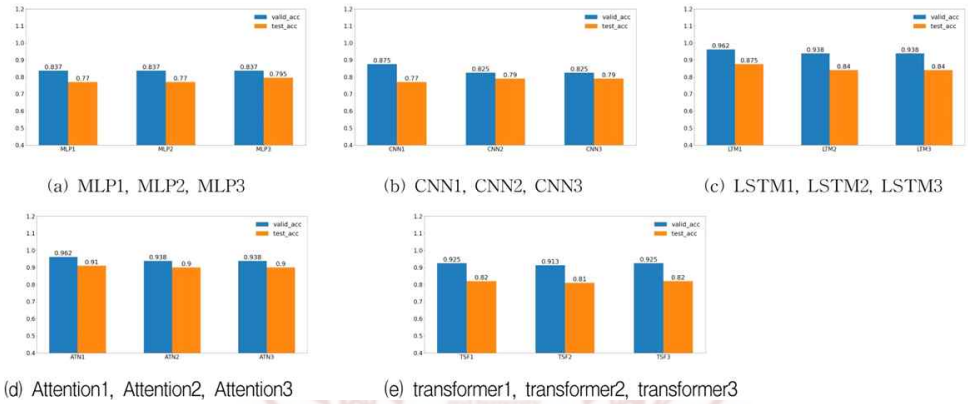
<표 12> CPC C01 계열 학습 테스트 데이터

	class	Train	Valid	test	total
데이터셋 1	5	720	80	200	1,000
데이터셋 2	14	2,016	224	560	2,800
데이터셋 3	19	2,736	304	760	3,800

*자료: (김성훈, 김승천, 2021)

<그림 10>과 같이 MLP 은 0.795,CNN 의 최대 정확도는 0.790, LSTM 의 최대 정확도는 0.875,Attention 의 최대 정확도는 0.91 그리고 Transformer 의 최대 정확도는 0.820 이다. 데이터셋 1 에서는 개별 모델중 Attention 이 가장 좋았고, 데이터셋 2 에서는 LSTM 이 가장 좋았다. 그리고 데이터셋 3 에서는 Attention 의 최대정확도가 좋았던 것으로 실험 결과 확인이 된다.

<그림 10> 데이터 크기 및 품질에 따른 분류



(a) MLP1, MLP2, MLP3 (b) CNN1, CNN2, CNN3 (c) LSTM1, LSTM2, LSTM3

(d) Attention1, Attention2, Attention3 (e) transformer1, transformer2, transformer3

*자료: (김성훈, 김승천, 2021)

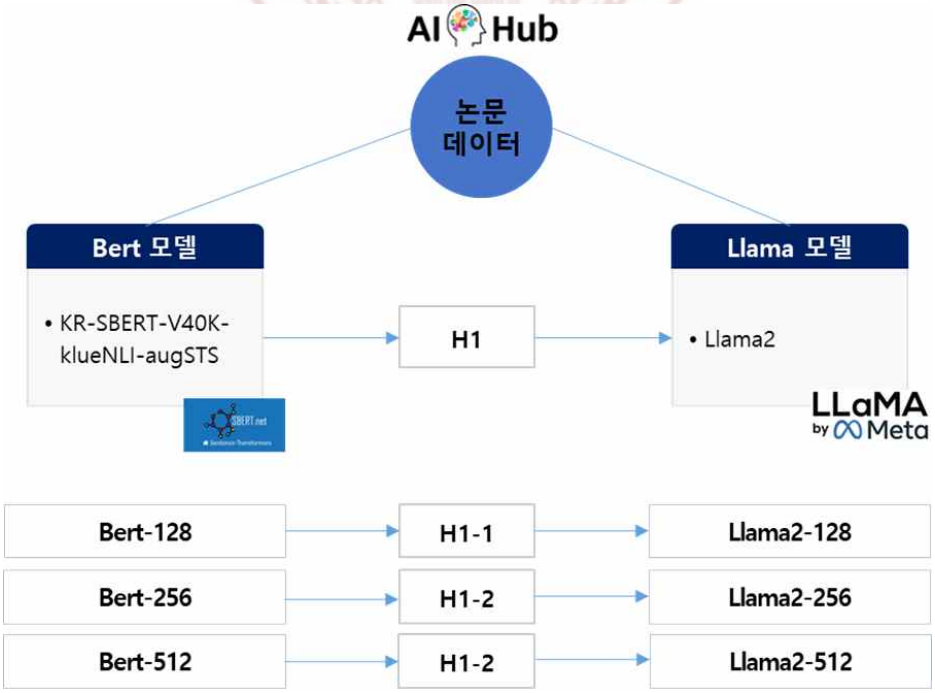
제3장 연구내용

제1절 연구 모형 및 연구 가설

1. 연구 모형

본 연구는 인공지능은 언어 모델을 활용한 논문자료를 범주별로 분류한다. 추론기반의 BERT와 LLM의 하나인 Llama2의 분류의 성능평가를 비교하는 목적을 가지고 있다. 본 연구의 목적을 달성하기 위해 AI-HUB의 논문데이터를 활용하여 Bert 모델과 Llama2의 자동분류 결과를 측정하였고 모형은 <그림 11>과 같다.

<그림 11> 연구 모형



2. 연구 가설

인공지능은 언어 모델인 추론기반의 BERT와 LLM의 모델인 Llama2의 분류의 성능을 비교하기 위해 다음과 같이 가설을 설정하였다. 입력 크기별로 Short(128), Middle(256), Long(512) 3가지를 모델을 가지고 각각 설정하였다.

H1. Bert 과 Llama 의 자동분류 성능이 차이가 있다.

H1-1. 모델 입력 크기가 Short 모델일 경우 Bert 와 Llama 의 자동분류 성능이 차이가 있다.

H1-2. 모델 입력 크기가 Middle 모델일 경우 Bert 와 Llama 의 자동분류 성능이 차이가 있다.

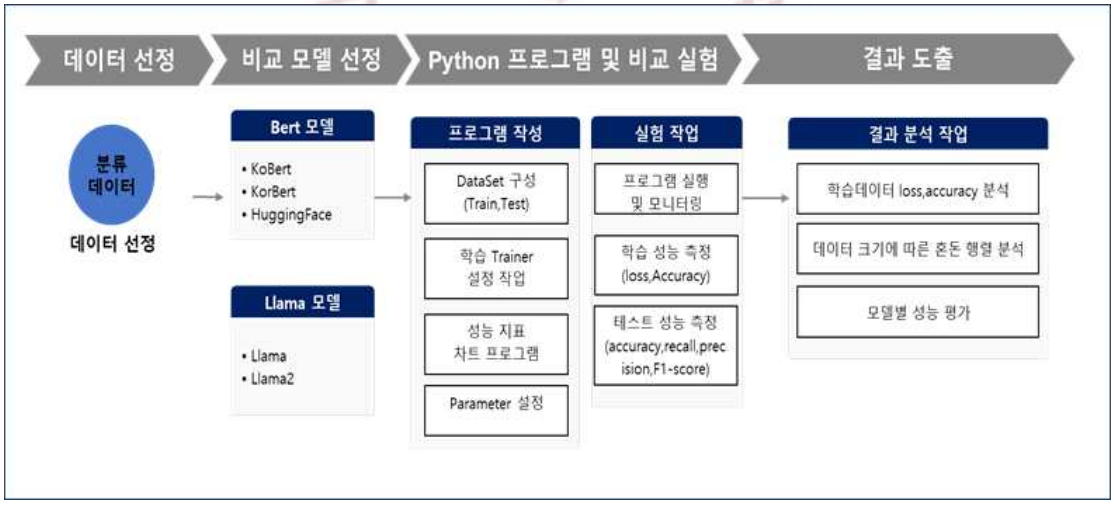
H1-3. 모델 입력 크기가 Long 모델일 경우 Bert 와 Llama 의 자동분류 성능이 차이가 있다.

제2절 연구 내용

1. 연구 실험 방법

본 연구는 <그림 12>와 같은 연구 흐름도를 가지고 있다. 연구를 수행하기 위하여 먼저 분류에 적합한 데이터를 찾아서 선정 작업을 진행하였다. 데이터셋의 수집은 신뢰성 있는 출처의 데이터 선정이 중요하며 데이터의 정확성이 확보되어야 한다. 모델 훈련 및 예측 테스트를 실험하기 위해서는 일정이상 데이터 건수를 확보해야 한다. BERT와 Llama 모델의 경우 Hugging Face에서 공유된 모델을 선정하였다. BERT 및 Llama의 경우 한글지원이 되는 모델을 고려 대상으로 하였다. 실험하기 위한 DataSet 구성 작업 및 Trainer 설정 작업, Parameter를 고려하여 Python 프로그램을 작성하였으며 GPU 메모리 용량에 맞게 튜닝 설정을 하였다. 결과 도출할 수 있는 학습성능 Loss, Accuracy 및 분류모델척도 Recall, Precision, F1-Score를 Python 프로그램에서 출력하게 하였다. 결과 분석 작업에서는 모델별 학습 결과 및 데이터 크기에 따른 혼돈 행렬 분석과 성능 평가를 진행하였다.

<그림 12> 연구 흐름도

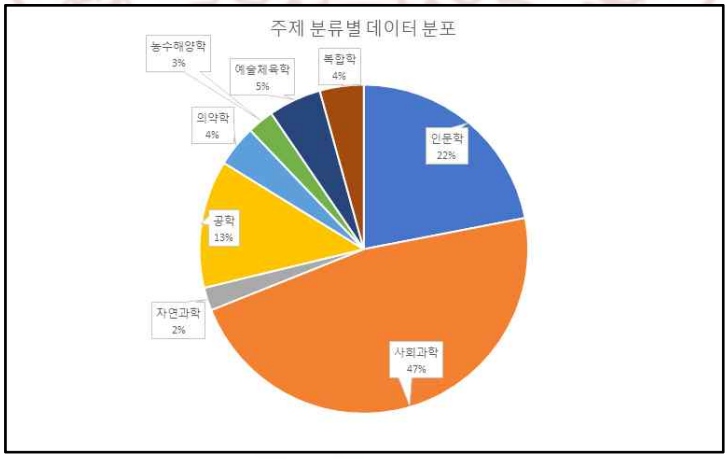


2. 연구 실험 내용

2.1 데이터 선정

실험에 사용한 데이터의 경우 AI-HUB[12]에 '논문자료 요약' 데이터셋을 이용하였다. 데이터는 학국학술지인용색인(KCI), 국립중앙도서관, KISTI 등의 관련 연구기관에서 제공하는 Open Access 한국어 한술 논문을 대상으로 수집하였다. 수집, 정제 대상 원천데이터의 대상 분류는 한국연구재단의 학술연구분야분류표을 따르고 있다. <그림13>, <표 13>에서 사회과학 47%, 인문학 22%, 공학 13% 등 상위 3개 분류에 82% 데이터가 분포되어 있다.

<그림 13> 주제 분류별 데이터 분포 현황



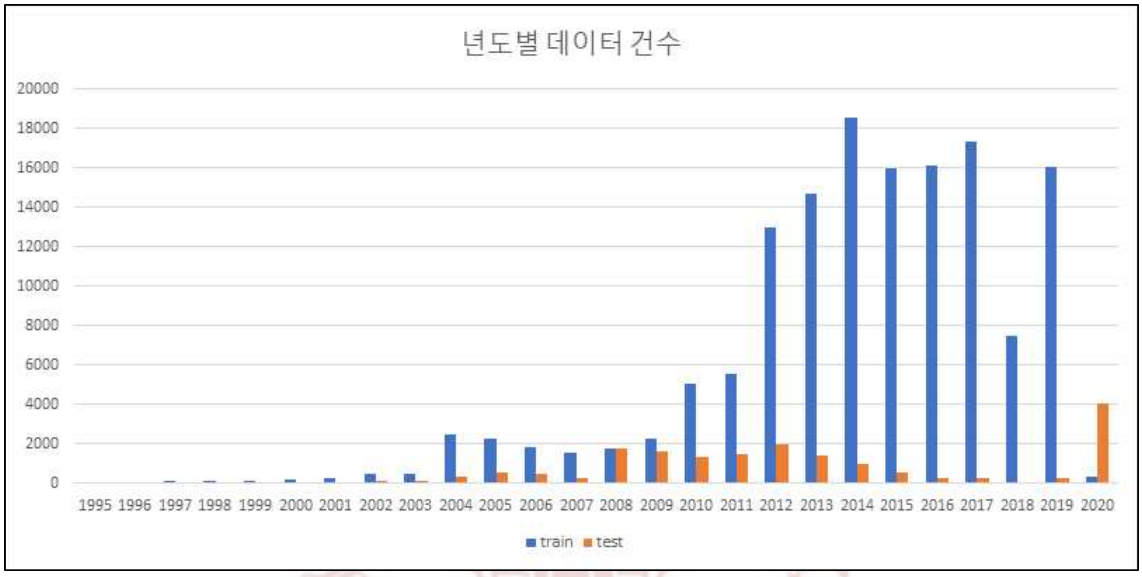
<표 13> 주제별 학습 데이터 및 테스트 데이터 현황

	인문학	사회과학	자연과학	공학	의약학	농수해양학	예술체육학	복합학	계
학습 데이터	27,981	70,569	3,180	19,304	5,686	3,720	7,757	6,063	144,280
테스트 데이터	7,660	5,736	499	1,092	937	505	705	927	18,061
합계	35,641	76,305	3,679	20,396	6,623	4,225	8,462	7,010	162,341

12) AI-Hub는 한국지능정보사회진흥원이 운영하는 AI 통합 플랫폼이다(위키피디아)

<그림14>, <표 14>과 1995년부터 2020년까지 데이터 중 2010~2020년 데이터가 학습데이터는 90%이며 테스트 데이터는 70%이다. 2010년 이후에 데이터의 경우 2018년을 데이터의 경우가 적은 편이고 테스트 데이터의 경우 2020년 데이터가 많이 존재했다.

<그림 14> 년도 데이터 건수



<표 14> 년도별 데이터 건수

년도	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2020
학습	380	16016	7481	17327	16104	15956	18566	14707	12984	5581	380
테스트	4066	314	4	243	297	553	1020	1437	1964	1488	4066

년도	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000~
학습	5082	2269	1752	1531	1838	2270	2522	492	462	246	5082
테스트	1332	1663	1747	297	481	537	365	128	118	0	1332

<표 15>과 2000년 이전 데이터는 자연과학, 공학, 의약화, 인문학이 있었으며 농수해양학, 예술체육학, 복합학의 경우 2002년 이후 데이터로 구성되었다. 사회과학, 인문학, 공학을 제외한 데이터는 경우 연도별로 고르게 분포 되어 있다.

<표 15> 년도/카테고리별 데이터 건수

년도	인문학	사회과학	자연과학	공학	의약학	농수해양학	예술체육학	복합학	계
1995	2	4							6
1996		2		78					80
1997		20		88	26				134
1998	2	37		82	42				163
1999	4	48		54	38				144
2000		26		114	54				194
2001		100		120	26				246
2002	87	171		175	92	49	2	4	580
2003	128	268	4	94	70	48	4	4	620
2004	528	1,587	94	358	211	62	22	25	2,887
2005	501	1,628	76	168	249	149	19	17	2,807
2006	524	1,319	42	147	149	58	18	62	2,319
2007	651	903	19	90	119	9	18	19	1,828
2008	1,626	1,215	110	223	145	23	61	96	3,499
2009	1,948	1,064	124	303	236	57	92	108	3,932
2010	3,050	698	72	1,478	726	287	85	18	6,414
2011	2,680	313	404	1,760	1,086	730	78	18	7,069
2012	3,721	6,112	404	2,654	710	658	666	23	14,948
2013	4,018	6,827	596	2,668	493	404	958	180	16,144
2014	4,222	9,519	448	2,539	637	336	1,350	535	19,586
2015	689	11,152	224	2,253	197	213	1,181	600	16,509
2016	1,183	9,987	213	1,705	444	293	1,160	1,416	16,401
2017	3,251	9,743	314	1,189	322	339	985	1,427	17,570
2018	1,560	3,671	104	544	201	177	570	658	7,485
2019	4,111	7,769	334	1,268	287	241	915	1,405	16,330
2020	1,155	2,122	97	244	63	92	278	395	4,446
합계	35,641	76,305	3,679	20,396	6,623	4,225	8,462	7,010	162,341

2.2 비교 모델 선정

BERT는 기존 BERT 구조에서 문장 임베딩의 성능을 우수하게 개선한 Sentence BERT를 이용하였다. KUE-nli와 KorSTS의 우수한 데이터로 파인튜닝한 snunlp/KR-SBERT-V40K-klueNLI-augSTS 모델을 사용하였다. Llama2 모델의 경우 <표 16>와 같이 0.06%만 한국어 학습데이터로 사용되었으며 한국어 소통 능력이 떨어지는 것을 볼 수 있으며, 대화 내용의 질적 측면에서도 많이 부족하다(주호택 외, 2023). Llama-2 7b 모델의 한국어 미세튜닝을 한 모델인 beomi/Llama-2-ko-7b에 학습 데이터 nlpai-lab/kullm-v2를 통해 학습한 kfkas/Llama-2-ko-7b-Chat 모델을 사용하였다.

<표 16> 사전 학습 데이터의 언어 분포

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

*자료: (Touvron, H. et al., 2023)

<표 17>와 같이 BERT와 Llama2의 모델을 선정하여 실험을 진행하였다.

<표 17> 시험 사용된 모델

모델	내용	파라미터
BERT	snunlp/KR-SBERT-V40K-klueNLI-augSTS[13]	117백만건
Llama2	kfkas/Llama-2-ko-7b-Chat[14]	6,674백만건

13) 자료 : https://huggingface.co/snunlp/KR-SBERT-V40K-klueNLI-augSTS
14) 자료 : https://huggingface.co/kfkas/Llama-2-ko-7b-Chat

2.3 Python 프로그램 및 실험

모델평가 서버는 <표 18>과 같이 32 core, 125G ram, SSD 3.0TB, RTX A6000(48G)을 사용하였다. Deep 러닝의 모델의 학습을 하기 위해서는 gpu가 필수이며 Llama2의 파인튜닝을 하기 위해 Peft LoRA를 사용하여 1 gpu에서 가능하도록 하였다. Application의 Python 모듈은 <표 19>과 같이 Python 버전 python3.8, Torch 버전 2.0.1+cuda 117, Transformers 라이브러리 버전 4.31.0, Peft 라이브러리 버전 0.4.0을 사용하였다. Sentence 학습에 사용되는 파라미터는 <표 20>와 같이 epoch(학습 횟수)는 5를 기준으로 사용하였으며 learning_rate는 1e-4를 사용하였다. Batch 크기는 BERT에 경우 128(short,middle), 64(long)을 사용하였으나 Llama2의 12(short), 6(middle), 3(long)으로 Llama2가 gpu 메모리 사용이 많았다.

<표 18> 실험 서버 환경

OS	CPU	Memory	HDD	GPU
CentOS Linux release 7.7.1908 X64"	Intel(R) Xeon(R) CPU E5-2698 v3 @ 2.30GHz	125G	SSD: 3.0TB HDD: 4.0TB	RTX A6000 48G

<표 19> 실험 프로그램 환경

Application	내용
Python	python3.8
Torch	2.0.1+cuda 117
Transformers	4.31.0
Peft	0.4.0

<표 20> 실험 학습 환경

<table>
<tr><th>파라미터 정보</th><th>내용</th></tr>
<tr><td>Epoch</td><td>5</td></tr>
<tr><td>learning_rate</td><td>1e-4</td></tr>
<tr><td>Max length</td><td>128(short), 256(middle), 512(long)</td></tr>
<tr><td>Batch</td><td>
<table>
<tr><th></th><th>128
(short)</th><th>256
(middle)</th><th>512
(long)</th></tr>
<tr><td>BERT</td><td>128</td><td>128</td><td>64</td></tr>
<tr><td>Llama2</td><td>12</td><td>6</td><td>3</td></tr>
</table>
</td></tr>
</table>

Transformers의 Trainer 클래스를 이용하여 학습하였으며 미리 학습된 모델을 미세 조정(fine-tuning) 할 수 있다. <표 21>와 같이 144,280건의 학습 데이터 중 8:2 비율로 115,424는 train_dataset을 이용하고 28,856는 eval_dataset을 이용하여 epoch 5의 학습을 실행한다.

<표 21> 학습에 사용된 데이터 건수

	인문학	사회과학	자연과학	공학	의약학	농수 해양학	예술체육학	복합학	계
학습 데이터	22,436	56,441	2,522	15,392	4,539	2,999	6,214	4,881	115,424
검증 데이터	5,545	14,128	658	3,912	1,147	721	1,543	1,202	28,856
합계	27,981	70,569	3,180	19,304	5,686	3,720	7,757	6,083	144,280

모델별로 입력 크기(short-128, middle-256, long-512)별 실험을 진행하고 학습 데이터의 학습 loss, acuuracy, 테스트 데이터의 대한 혼돈 행렬 분석, 성능 평가를 진행한다.

2.4 실험 평가 척도

<표 22>와 같이 Confusion Matrix를 보면 실제 값의 경우 True, False인지와 예측 클래스 값이 positive, negative인지를 나태내고 있다. T는 True를 의미이며, F는 False를 이다. P는 Positive를 이며, N은 Negative이다.

<표 22> Confusion Matrix

		예측 클래스 (Predicted Class)	
		Positive	Negative
실제 클래스 (Actual Class)	Positive	TP (Tru Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

가) 정밀도(Precision)

예측 값이 Positive 인 것 중에 정확하게 실제 값도 Positive값 것의 비율.

$$\text{정밀도} = \frac{\text{정답 실제 개수}}{\text{정답 예측 전체 개수}} = \frac{TP}{TP + FP}$$

나) 재현율(Recall)

실제 정답 값이 Positive 인 것 중 실제 값이 True Positive를 맞춘 것의 비율.

$$\text{재현율} = \frac{\text{정답 실제 개수}}{\text{실제 정답 전체 개수}} = \frac{TP}{TP + FN}$$

다) F1 스코어(F1-score)

정밀도와 재현율의 조화 평균

$$\text{F1 스코어} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

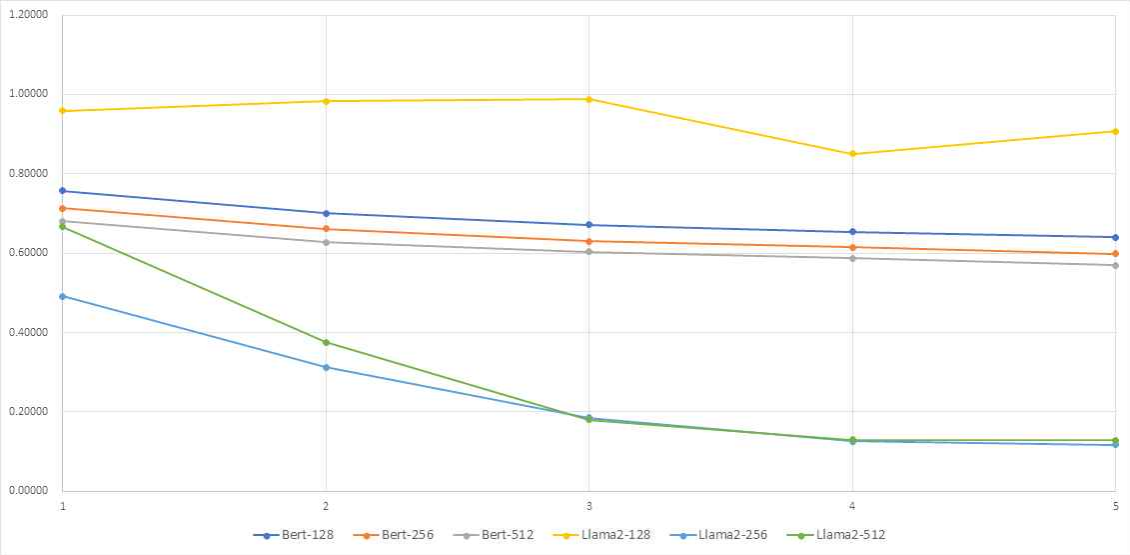
제4장 연구결과

제1절 학습데이터 분석 결과

1. 학습데이터 오차(Loss) 비교

<그림 15>, <표 23>와 같이 Bert, Llama2 모델은 epoch 횟수가 증가할수록 더 낮은 loss를 결과를 보이고 있다. Bert 모델의 경우 epoch 1에서 0.75716(short), 0.71267(middle), 0.67982(long) 으로 시작하여 epoch 5에서는 0.63983(short), 0.59789(middle), 0.56898(long) 으로 끝났다. 입력크기가 커지수록 오차가 작았으며 epoch를 올라갈수록 미세하게 낮아졌다. Llama2 모델의 경우 epoch 1에서 0.95907(short), 0.49080(middle), 0.66660(long) 으로 시작하여 epoch 5에서는 0.90675(short), 0.11661(middle), 0.12796(long) 이였다. middle 모델의 경우가 가장 낮은 수치를 보여 주었으며 middle, long 모델의 경우 epoch 3부터 오차가 급격히 낮아졌다.

<그림 15> 학습모델별 epoch별 오차(loss) 비교



<표 23> epooh별 모델 오차(loss) 비교

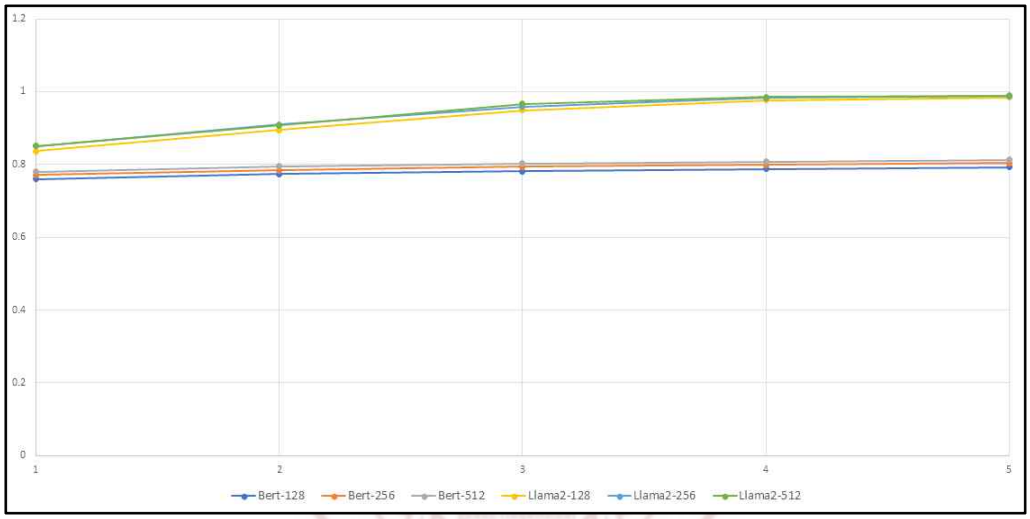
epoch	Bert-128	Bert-256	Bert-512	Llama2-128	Llama2-256	Llama2-512
1	0.75716	0.71267	0.67982	0.95907	0.49080	0.66660
2	0.70011	0.66114	0.62712	0.98307	0.31231	0.37551
3	0.67143	0.62946	0.60341	0.98777	0.18517	0.17977
4	0.65398	0.61448	0.58722	0.85039	0.12649	0.12949
5	0.63983	0.59789	0.56898	0.90675	0.11661	0.12796

2. 학습데이터 정확도(Accuracy) 비교

<그림 16>, <표 24>과 같이 Bert, Llama2 모델은 epoch 횟수가 증가할수록 더 높은 정확도를 보인다. Bert 모델의 경우 epoch 1에서 0.75932(short), 0.777129(middle), 0.77998(long)으로 시작하여 epoch 5에서는 0.79199(short), 0.80351(middle), 0.81303(long)으로 끝났다. Llama2 모델의 경우 epoch 1에서

0.83686(short), 0.84987(middle), 0.58039(long)으로 시작하여 epoch 5에서는 0.98413(short), 0.98777(middle), 0.98913(long)으로 끝났다. Llama2가 Bert 보다 모든 크기에서 더 높은 정확도를 보였다. Llama2의 epoch 3이상 정확도가 94.8%, 95.9%, 96.6%등 높은 정확성을 나타냈다.

<그림 16> epooh별 학습모델 정확도(accuracy) 비교



<표 24> epooh별 모델 정확도(accuracy) 비교

epoch	Bert-128	Bert-256	Bert-512	Llama2-128	Llama2-256	Llama2-512
1	0.75932	0.77129	0.77998	0.83686	0.84987	0.85039
2	0.77334	0.78363	0.79479	0.89459	0.90907	0.90675
3	0.78121	0.79429	0.80273	0.94878	0.95907	0.96661
4	0.78699	0.79877	0.80740	0.97643	0.98307	0.98558
5	0.79199	0.80351	0.81303	0.98413	0.98777	0.98913

제2절 데이터 크기에 따른 혼돈 행렬 분석 (Confusion Matrix Analysis)

1. Short 모델 데이터 행렬 분석

<그림 17>, <그림 18>과 같이 Bert 와 Llama2 를 비교하면 Bert 의 경우 사회과학, 공학, 농수해양, 예술체육 분야에서 높았으며 Llama2 는 인문학, 자연과학, 의약학, 복합학 분야에서 빈도가 높게 나타났다. Bert 의 경우 공학, Llama2 의 경우 자연과학, 복합학 분야가 상대적으로 높게 나왔다. 다른 분야에서는 미세한 차이가 있었다. 복합학, 자연과학 분야에서는 Bert 나 Llama 두 모델에서 낮게 빈도가 나타났다.

<그림 17> Bert Short 모델 행렬 분석

<표 25> Bert Short 모델 정확도 분석

	precision	recall	f1-score
인문학	0.93683	0.80535	0.86613
사회과학	0.72045	0.92015	0.80815
자연과학	0.70556	0.25451	0.37408
공학	0.62548	0.89927	0.73779
의약학	0.84926	0.85379	0.85152
농수해양	0.76231	0.70495	0.73251
예술체육	0.54659	0.68227	0.60694
복합학	0.36937	0.04423	0.07900
합계	0.78911	0.78811	0.77135

<그림 18> Llama2 Short 모델 행렬 분석

<표 26> Llama2 Short 모델 정확도 분석

	precision	recall	f1-score
인문학	0.89479	0.85274	0.87326
사회과학	0.73208	0.87796	0.79841
자연과학	0.64873	0.41082	0.50307
공학	0.63305	0.47711	0.54413
의약학	0.81079	0.91462	0.85958
농수해양	0.68193	0.56040	0.61522
예술체육	0.47353	0.68511	0.56000
복합학	0.35857	0.09709	0.15280
합계	0.76622	0.77554	0.76223

2. Middle 모델 데이터 행렬 분석

<그림 19>, <그림 20>와 같이 Bert 와 Llama2 를 비교하면 Bert 의 경우 사회과학, 공학, 농수해양 분야에서 높았으며 Llama2 는 인문학, 자연과학, 의약학, 예술체육, 복합학 분야에서 빈도가 높게 나타났다. Bert 의 경우 공학, Llama2 의 경우 인문학, 자연과학, 복합학 분야가 상대적으로 높게 나왔다. 의약학 및 농수해양, 예술체육 분야는 비슷한 빈도를 나타냈다. Short 모델과 비교시 Bert 는 사회과학, 공학, 의약학 분야에서 빈도가 높아졌지만 인문학, 자연과학, 농수해양, 예술체육, 복합학 분야에서는 빈도 내려갔다.

Llama2 의 경우는 Middle 입력 모델이 Short 입력 모델보다 빈도가 높아졌다.

<그림 19> Bert Middle 모델 행렬 분석

<표 27> Bert Middle 모델 정확도 분석

	precision	recall	f1-score
인문학	0.94468	0.79360	0.86258
사회과학	0.71482	0.92817	0.80765
자연과학	0.61268	0.17435	0.27145
공학	0.57776	0.91850	0.70934
의약학	0.85016	0.87193	0.86091
농수해양	0.75599	0.68713	0.71992
예술체육	0.57759	0.66525	0.61833
복합학	0.35294	0.02589	0.04824
합계	0.78544	0.78346	0.76413

<그림 20> Llama2 Middle 모델 행렬 분석

<표 28> Llama2 Middle 모델 정확도 분석

	precision	recall	f1-score
인문학	0.91527	0.87428	0.89430
사회과학	0.75955	0.90812	0.82722
자연과학	0.79720	0.45691	0.58089
공학	0.75214	0.72527	0.73846
의약학	0.85507	0.94450	0.89757
농수해양	0.81546	0.64752	0.72185
예술체육	0.58333	0.71489	0.64245
복합학	0.48583	0.12945	0.20443
합계	0.81178	0.81734	0.80502

3. Long 모델 데이터 행렬 분석

<그림 21>, <그림 22>와 같이 Bert 와 Llama2 를 비교하면 Bert 의 경우 공학, 농수해양 분야에서 높았으며 Llama2 는 인문학, 자연과학, 의약학, 예술체육, 복합학 분야에서 빈도가 높게 나타났다. Bert 의 경우 공학, Llama2 의 경우 인문학, 자연과학, 복합학 분야에서 상대적으로 높게 나왔다. Middle 모델과 비교시 Bert 는 사회과학, 공학, 의약학 분야에서 빈도가 높았졌지만 사회과학, 공학, 예축체육 분야에서는 빈도가 내려갔다. Llama2 의 경우는 Long 입력 모델이 Middle 입력 모델보다 전체적으로 높아졌다.

<그림 21> Bert Long 모델 행렬 분석

<표 29> Bert Long 모델 정확도 분석

	precision	recall	f1-score
인문학	0.94799	0.80196	0.86888
사회과학	0.73185	0.92451	0.81698
자연과학	0.68898	0.3507	0.46481
공학	0.62531	0.90934	0.74104
의약학	0.86084	0.85165	0.85622
농수해양	0.79332	0.75248	0.77236
예술체육	0.55043	0.72766	0.62676
복합학	0.43226	0.07228	0.12384
합계	0.80185	0.79575	0.78246

<그림 22> Llama2 Long 모델 행렬 분석

<표 30> Llama2 Long 모델 정확도 분석

	precision	recall	f1-score
인문학	0.93389	0.90000	0.91663
사회과학	0.79585	0.92294	0.85470
자연과학	0.84375	0.59519	0.69800
공학	0.76705	0.84432	0.80384
의약학	0.90688	0.95624	0.93091
농수해양	0.86480	0.73465	0.79443
예술체육	0.71508	0.73333	0.72409
복합학	0.52252	0.18770	0.27619
합계	0.84449	0.85073	0.84104

4. 전체 데이터 행렬 분석

<그림 23>, <그림 24>와 같이 Bert 와 Llama2 의 short, middle, long 의 행렬값을 더한 데이터를 비교하면 Bert 의 경우 사회과학, 공학, 농수해양 분야에서 높았으며 Llama2 는 인문학, 자연과학, 의약학, 예술체육, 복합학 분야에서 빈도가 높게 나왔다. Bert 는 공학, 농수해양 분야의 경우 Llama2 보다 10%이상 우수한 결과를 보였다. Llama2 는 자연과학, 복합학 분야의 경우 2 배 정도 우수한 결과가 확인되었다.

<그림 23> Bert 전체 모델 행렬 분석

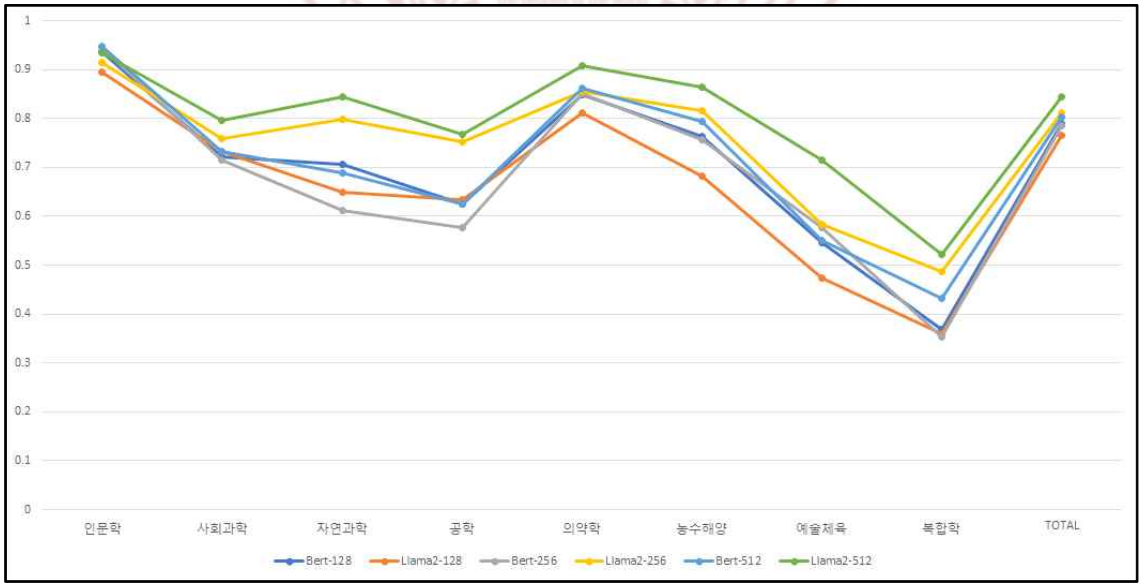
<그림 24> Llama2 전체 모델 행렬 분석

제3절 모델별 성능 평가

1. 정밀도(Precision)

<그림 25>, <표 31>과 같이 정밀도를 분석한 결과 Bert 와 Llama2 에서 인문학, 의약학, 농수해양 분야에서는 높은 정밀도가 나왔다. 복학합, 예술체육 분야는 낮은 결과가 나왔다. 인문학, 사회과학, 의약학 분야에서는 입력 크기별 차이가 미미하게 나왔다. Bert 모델의 경우 자연과학, 복합학 분야에서는 입력 크기별 차이가 났으며 나머지 분야에서는 거의 차이가 없었다. Llama2 의 경우 자연과학, 공학, 농수해양, 의과학, 예술체육, 복합학 분야에서는 모델의 입력 크기가 커질수록 의미 있는 성능 향상이 있었다.

<그림 25> 모델별 정밀도 비교

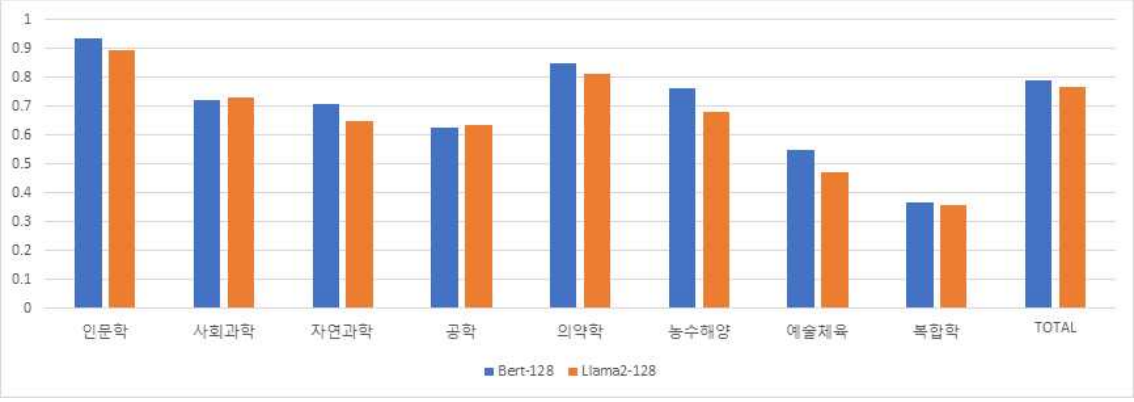


<표 31> 모델별 정밀도 비교

	Short-model(128)		Middle-model(256)		Long-model(512)	
	BERT	Llama2	BERT	Llama2	BERT	Llama2
인문학	0.93683	0.89479	0.94468	0.91527	0.94799	0.93389
사회과학	0.72045	0.73208	0.71482	0.75955	0.73185	0.79585
자연과학	0.70556	0.64873	0.61268	0.79720	0.68898	0.84375
공학	0.62548	0.63305	0.57776	0.75214	0.62531	0.76705
의약학	0.84926	0.81079	0.85016	0.85507	0.86084	0.90688
농수해양	0.76231	0.68193	0.75599	0.81546	0.79332	0.86480
예술체육	0.54659	0.47353	0.57759	0.58333	0.55043	0.71508
복합학	0.36937	0.35857	0.35294	0.48583	0.43226	0.52252
TOTAL	0.78911	0.76622	0.78544	0.81178	0.80185	0.84449

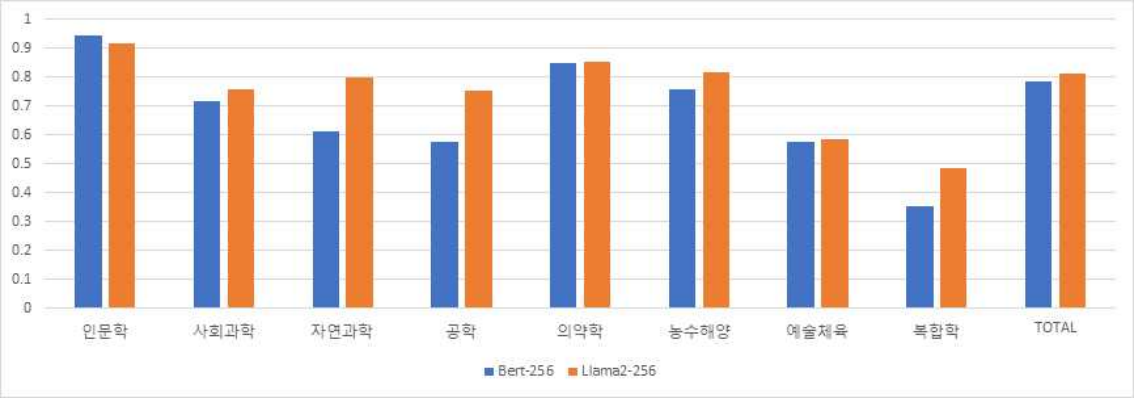
<그림 26>과 같이 short model 일 경우 Bert 가 인문학, 자연과학, 의약학, 농수해양, 예술체육, 복합학 분야에서 Llama2 가 사회과학, 공학 분야에서 우수한 성능이 나왔다. 전체적으로 Bert 가 Llama2 보다 더 좋은 결과가 나왔다.

<그림 26> Short 모델별 정밀도 비교



<그림 27>와 같이 middle model 의 경우 Bert 가 인문학, 의약학, 예술체육에서 Llama2 가 사회과학, 자연과학, 공학, 농수해양, 복합학에서 우수한 성능이 나왔다. 전체적으로 Llama2 가 Bert 보다 약간 좋은 결과가 나왔다.

<그림 27> Middle 모델별 정밀도 비교



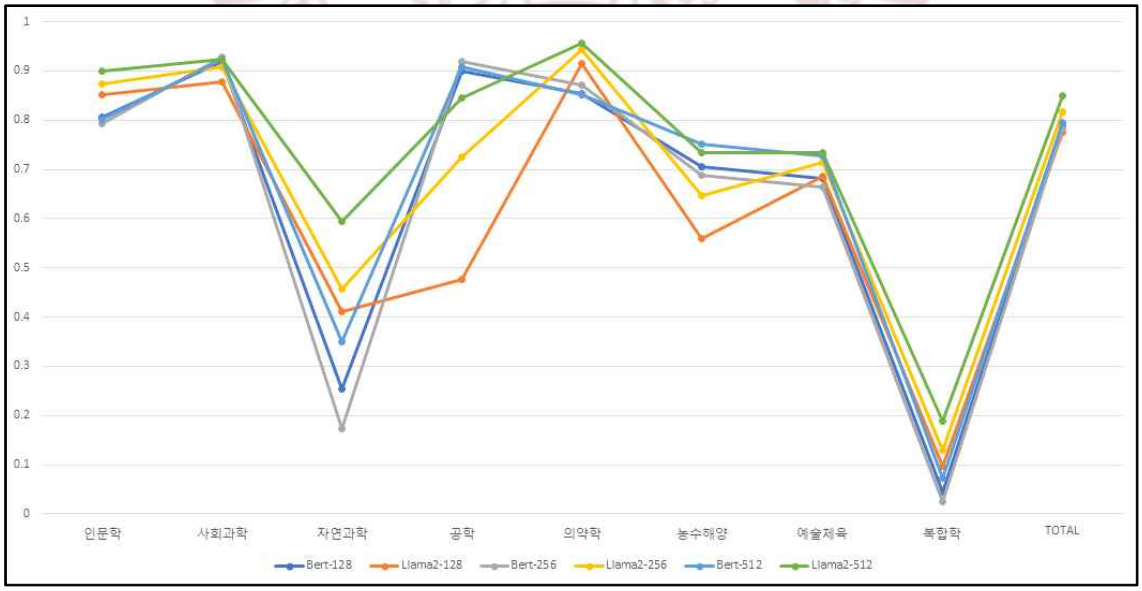
<그림 28>과 같이 long model의 경우 Bert는 인문학에서 Llama2의 경우 사회과학, 자연과학, 공학, 의약학, 농수해양, 예술체육, 복합학에서 우수한 성능이 나왔다. 자연과학, 공학, 예술체육, 복합학의 경우 10%이상 Llama2가 우수하게 평가되었다. long model의 경우 Llama2가 더 우수한 정밀도가 나왔다.

<그림 28> Long 모델별 정밀도 비교

2. 재현율(Recall)

<그림 29>, <표 32>과 같이 재현율을 분석한 결과 인문학, 사회과학, 의약학 분야에서 높은 성능이 나왔으며 자연과학 및 복합학 분야에서는 낮은 성능이 나왔다. Bert 의 경우 인문학, 사회과학, 공학, 의약학 분야는 우수한 성능을 보이며 농수해양, 예술체육은 준수한 성능을 보였다. Llama2 의 경우 인문학, 사회과학, 의약학 분야에서는 우수한 성능을 나왔으며 예술체육 분야는 준수한 성능이 나왔다. 자연과학, 복합학 분야에서는 낮은 성능이 나왔다. Llama2 에서는 자연과학, 공학, 농수해양 분야는 모델의 입력 사이즈가 높아질수록 성능 차이가 크게 높아졌다.

<그림 29> 모델별 재현율 비교

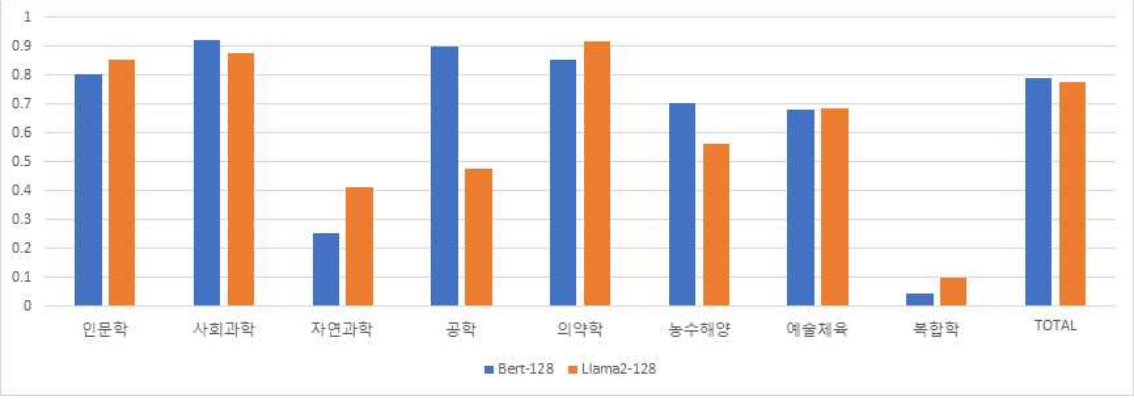


<표 32> 모델별 재현율 비교

	Short-model(128)		Middle-model(256)		Long-model(512)	
	BERT	Llama2	BERT	Llama2	BERT	Llama2
인문학	0.80535	0.85274	0.79360	0.87428	0.80196	0.90000
사회과학	0.92015	0.87796	0.92817	0.90812	0.92451	0.92294
자연과학	0.25451	0.41082	0.17435	0.45691	0.35070	0.59519
공학	0.89927	0.47711	0.91850	0.72527	0.90934	0.84432
의약학	0.85379	0.91462	0.87193	0.94450	0.85165	0.95624
농수해양	0.70495	0.56040	0.68713	0.64752	0.75248	0.73465
예술체육	0.68227	0.68511	0.66525	0.71489	0.72766	0.73333
복합학	0.04423	0.09709	0.02589	0.12945	0.07228	0.18770
TOTAL	0.78811	0.77554	0.78346	0.81734	0.79575	0.85073

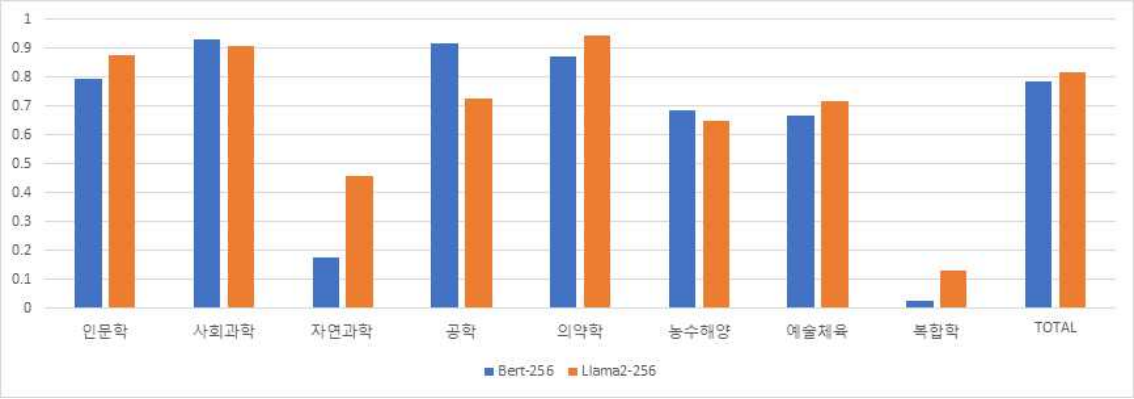
<그림 30>에서 short 모델의 경우 Bert가 사회과학, 공학, 농수해양 분야에서 점수가 높았으며 Llama2는 인문학, 자연과학, 의약학, 예술체육, 복합학 분야에서 점수가 높았다. 전체적으로 Bert가 Llama2보다 더 좋은 결과가 나왔다.

<그림 30> Short 모델별 재현율 비교



<그림 31>에서 middle 모델의 경우 Bert 의 경우 공학, 농수해양 분야에서 높았으며, Llama2 는 인문학, 의약학, 예술체육 분야에서 높았다. 전체적으로 Llama2 가 Bert 보다 더 좋은 결과가 나왔다.

<그림 31> Middle 모델별 재현율 비교



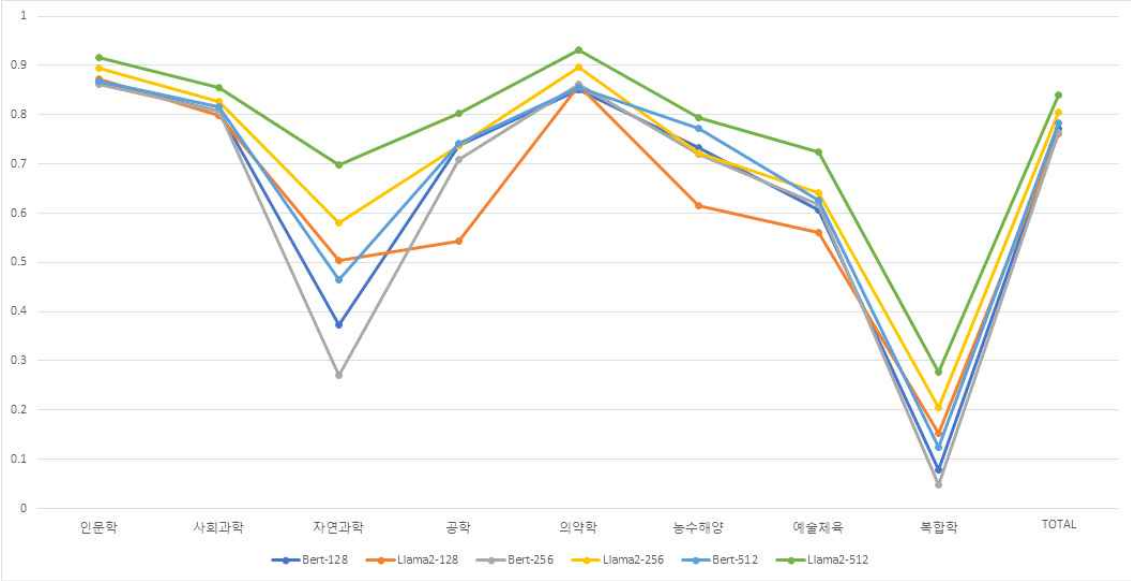
<그림 32>에서 Long mode의 경우 Bert가 사회과학, 공학, 농수해양 분야에서 우수하고 Llama2의 경우 인문학, 자연과학, 의약학, 복합학 분야에서 우수하였다. 전체적으로 Llama2가 Bert보다 더 좋은 결과가 나왔다. middle 모델과 long 모델에서 자연과학, 복합학 분야의 경우 Bert와 10% 이상 재현율 차이가 났다.

<그림 32> Long 모델별 재현율 비교

3. F1 스코어(F1-score)

<그림 33>, <표 33>과 같이 F1 스코어를 분석한 결과를 확인하면 인문학, 사회과학, 의약학 분야에서 높은 점수가 나왔으며 자연과학, 예술체육, 복합학 분야에서는 낮은 점수가 나왔다. 자연과학, 공학, 농수해양 분야는 입력 크기가 커질수록 크게 영향을 받았으며 Llama2 가 Bert 보다 상대적으로 높게 평가되었다. 인문학, 사회과학, 의약학, 예술체육, 복합학은 크기가 커질수록 좋아졌지만 미세하게 영향을 받았다.

<그림 33> 모델별 F1 스코어 비교

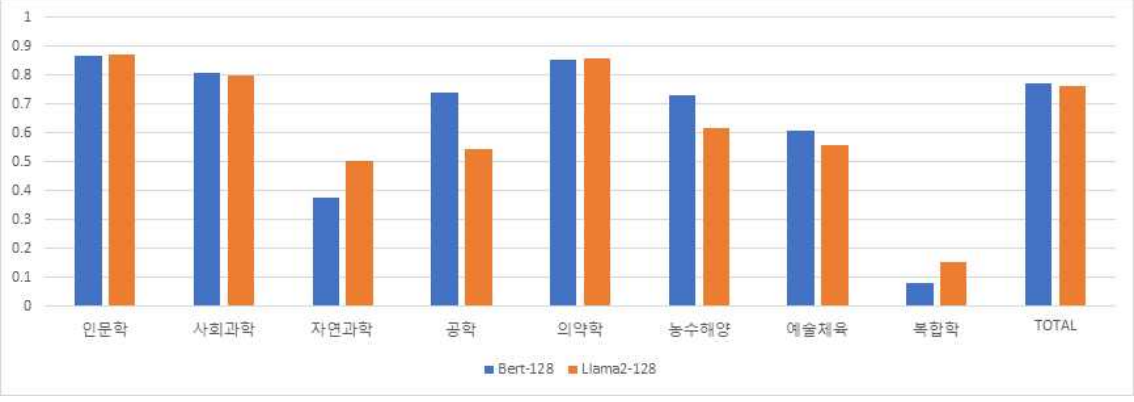


<표 33> 모델별 F1 스코어(F1-score) 비교

	Short-model(128)		Middle-model(256)		Long-model(512)	
	BERT	Llama2	BERT	Llama2	BERT	Llama2
인문학	0.86613	0.87326	0.86258	0.89430	0.86888	0.91663
사회과학	0.80815	0.79841	0.80765	0.82722	0.81698	0.85470
자연과학	0.37408	0.50307	0.27145	0.58089	0.46481	0.69800
공학	0.73779	0.54413	0.70934	0.73846	0.74104	0.80384
의약학	0.85152	0.85958	0.86091	0.89757	0.85622	0.93091
농수해양	0.73251	0.61522	0.71992	0.72185	0.77236	0.79443
예술체육	0.60694	0.56000	0.61833	0.64245	0.62676	0.72409
복합학	0.07900	0.15280	0.04824	0.20443	0.12384	0.27619
TOTAL	0.77135	0.76223	0.76413	0.80502	0.78246	0.84104

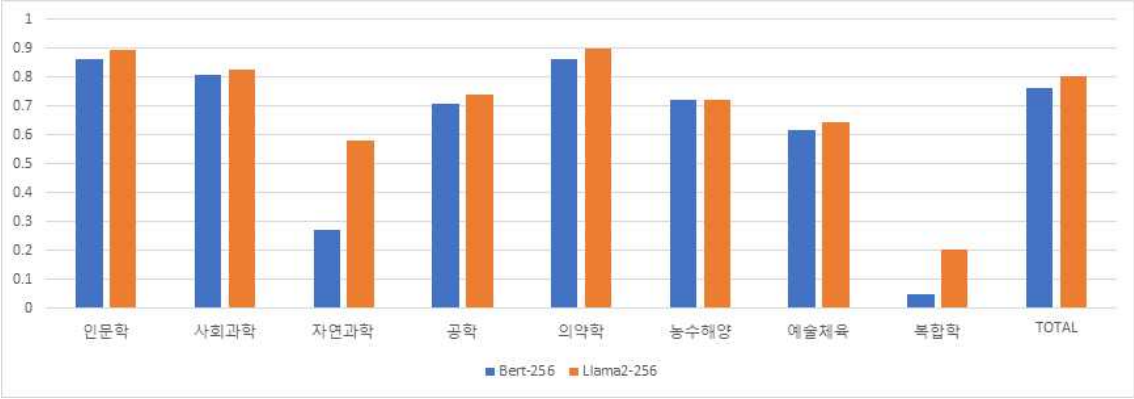
<그림 34>에서 Short model의 경우 Bert는 공학, 농수해양, 예술체육 분야가 우수했으며 Llama2의 경우 자연과학, 복합학 분야가 우수했다. 인문학, 사회과학, 의약학 분야는 유사한 성능을 나타냈다. Bert가 Llama2보다 우수한 성적이 나왔다.

<그림 34> Short 모델별 F1 스코어 비교



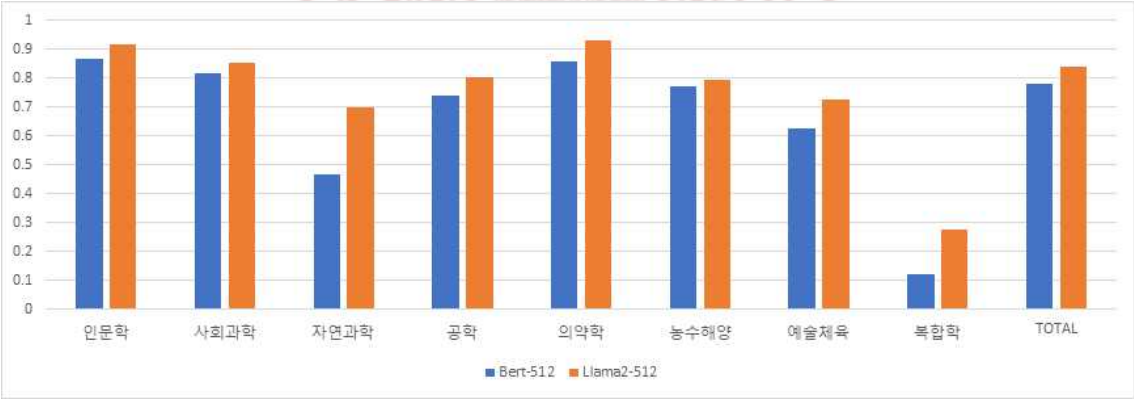
<그림 35>와 같이 Middle model의 경우 Bert는 공학, 농수해양이 Llama2는 인문학, 의약학, 자연과학이 우수했다. Llama2의 경우 자연과학과 복합학이 높은 성능이 나왔다. Llama2가 Bert보다 우수한 성능이 나왔다.

<그림 35> Middle 모델별 F1 스코어 비교



<그림 36>와 같이 Long model의 경우 Llama2가 모든 분류에서 높게 평가되었다. Llama2가 Bert 보다 높게 평가되었다.

<그림 36> Long 모델별 F1 스코어 비교



제4절 모델 성능 비교 검증

1. 쌍체 비교 t-검정

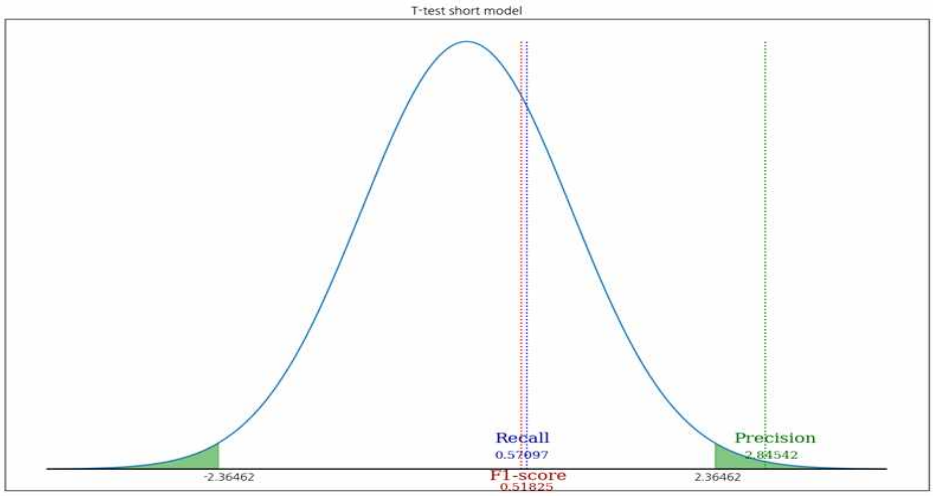
BERT와 Llama의 자동분류 성능 비교를 하기 위해 아래와 같은 조건으로 평가하였다.

- 유의수준(α) : 5%
- 귀무가설 H_0 : 성능의 차이가 없다
- 대립가설 H_1 : 성능의 차이가 있다.

2. Short model t-test 검증

<그림 37>, <표 34>와 같이 정밀도 T 통계량 2.84542이 기각역 2.36462 보다 크고 P 값이 0.02485이므로 해당 대립가설이 채택되어 성능의 차이가 있다. 재현율 T 통계량 0.57097이므로 기각역 2.36462 보다 작고 P 값이 0.58588이므로 해당 대립가설이 기각(귀무가설 채택)되어 성능 차이가 없다. F1 스코어 T 통계량 0.51825이므로 기각역 2.36462 보다 작고 P 값이 0.562026이므로 해당 대립가설이 기각(귀무가설 채택)되어 성능 차이가 없다.

<그림 37> 쌍체 비교 T-검정 short 모델 결과



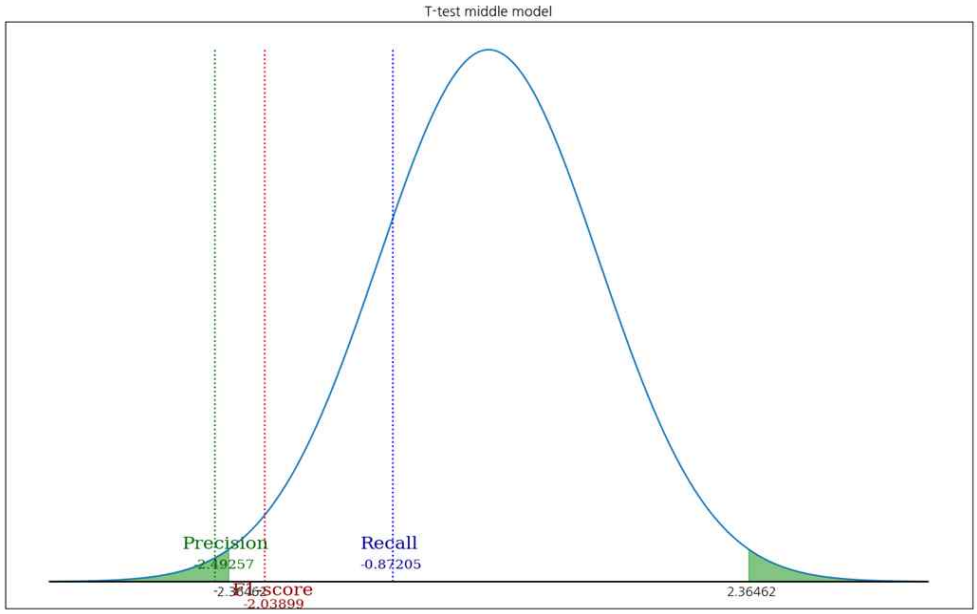
<표 34> 쌍체 비교 T-검정 short 모델 결과

	Precision		Recall		F1-score	
t-test	BERT	Llama2	BERT	BERT	BERT	Llama2
평균	0.68948	0.65418	0.64557	0.60948	0.63202	0.61331
분산	0.03146	0.02996	0.10401	0.07886	0.07547	0.05633
관측수	8	8	8	8	8	8
피어슨 상관 계수	0.98025		0.83319		0.93078	
가설 평균차	0		0		0	
자유도	7		7		7	
t 통계량	2.84542		0.57097		0.51825	
P(T<=t) 단측 검정	0.01243		0.29294		0.31013	
t 기각치 단측 검정	1.89458		1.89458		1.89458	
P(T<=t) 양측 검정	0.02485		0.58588		0.62026	
t 기각치 양측 검정	2.36462		2.36462		2.36462	

3. Middle model t-test 검증

<그림 38>, <표 35>와 같이 정밀도 T 통계량 -2.49257이 기각역 -2.36462 보다 작고 P 값이 0.04144이므로 해당 대립가설이 채택되어 성능의 차이가 있다. 재현율 T 통계량 -0.87205이 기각역 -2.36462 보다 크고 P 값이 0.41209이므로 해당 대립가설이 기각(귀무가설 채택)되어 성능 차이가 없다. F1 스코어 T 통계량 -2.03899이므로 기각역 -2.36462 보다 크고 P 값이 0.08083이므로 해당 대립가설이 기각(귀무가설 채택)되어 성능 차이가 없다.

<그림 38> 쌍체 비교 T-검정 middle 모델 결과



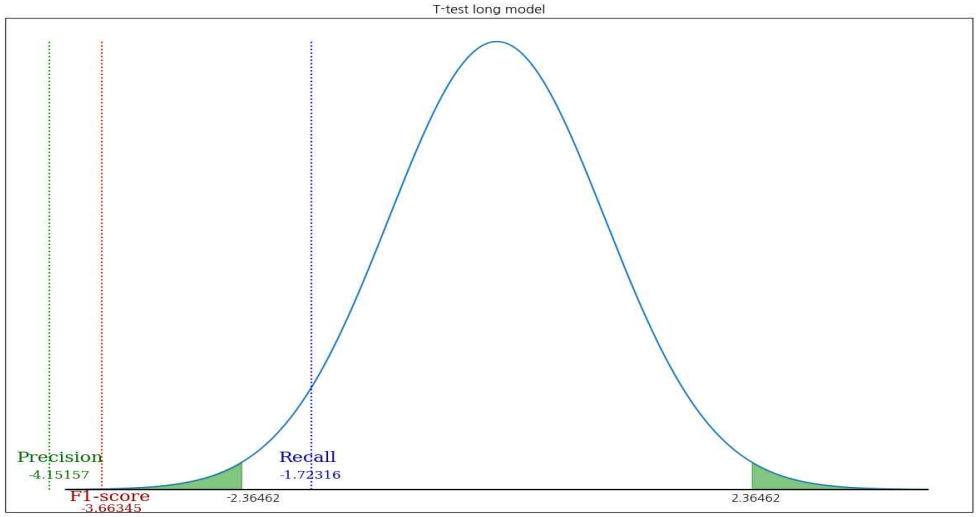
<표 35> 쌍체 비교 T-검정 middle model 결과

	Precision		Recall		F1-score	
t-test	BERT	Llama2	BERT	BERT	BERT	Llama2
평균	0.67333	0.74548	0.63310	0.67512	0.61230	0.68840
분산	0.03401	0.02034	0.11920	0.07382	0.08828	0.05099
관측수	8	8	8	8	8	8
피어슨 상관 계수	0.90579		0.92986		0.95487	
가설 평균차	0		0		0	
자유도	7		7		7	
t 통계량	-2.49257		-0.87205		-2.03899	
P(T<=t) 단측 검정	0.02072		0.20605		0.04042	
t 기각치 단측 검정	1.89458		1.89458		1.89458	
P(T<=t) 양측 검정	0.04144		0.41209		0.08083	
t 기각치 양측 검정	2.36462		2.36462		2.36462	

4. Long model t-test 검증

<그림 39>, <표 36>와 같이 정밀도 T 통계량 -4.15157이 기각역 -2.36462 보다 작고 P 값이 0.00429이므로 해당 대립가설이 채택되어 성능의 차이가 있다. 재현율 T 통계량 -1.72316이 기각역 -2.36462 보다 크고 P 값이 0.12852이므로 해당 대립가설이 기각(귀무가설 채택) 성능 차이가 없다. F1 스코어 T 통계량 -3.66345이므로 기각역 -2.36462 보다 작고 P 값이 0.00803이므로 해당 대립가설이 채택되어 성능 차이가 있다.

<그림 39> 쌍체 비교 T-검정 long model 결과



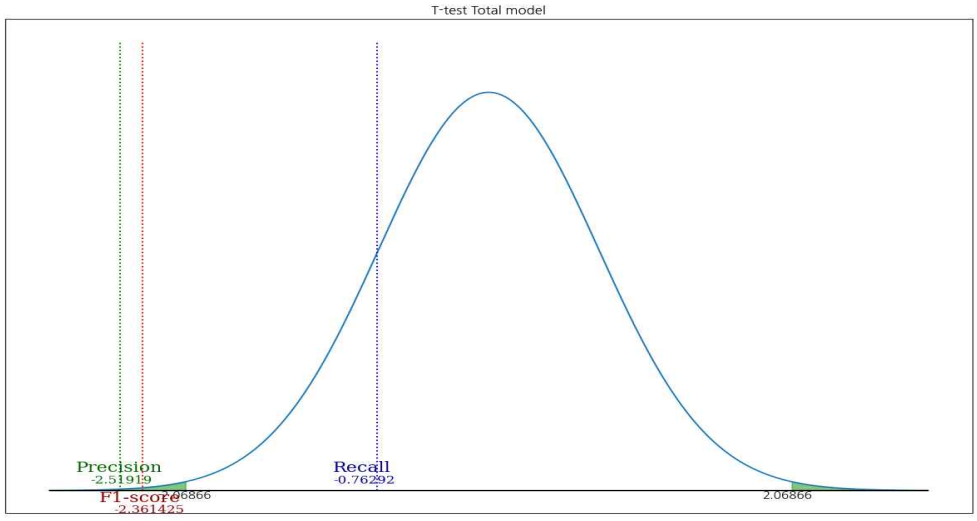
<표 36> 쌍체 비교 T-검정 Long model 결과

	Precision		Recall		F1-score	
t-test	BERT	Llama2	BERT	BERT	BERT	Llama2
평균	0.70387	0.79373	0.67382	0.73430	0.65886	0.74985
분산	0.02810	0.01721	0.09167	0.06322	0.06466	0.04346
관측수	8	8	8	8	8	8
피어슨 상관 계수	0.94502		0.95260		0.97326	
가설 평균차	0		0		0	
자유도	7		7		7	
t 통계량	-4.15157		-1.72316		-3.66345	
P(T<=t) 단측 검정	0.00214		0.06426		0.00402	
t 기각치 단측 검정	1.89458		1.89458		1.89458	
P(T<=t) 양측 검정	0.00429		0.12852		0.00803	
t 기각치 양측 검정	2.36462		2.36462		2.36462	

5. 전체 model t-test 검증

<그림 40>, <표 37>와 같이 정밀도 T 통계량 -2.51919이 기각역 -2.06866 보다 작고 P 값이 0.01916이므로 해당 대립가설이 채택되어 성능의 차이가 있다. 재현율 T 통계량 -0.76293이 기각역 -2.06866 보다 크고 P 값이 0.45326이므로 해당 대립가설이 기각(귀무가설 채택)6 성능 차이가 없다. F1 스코어 T 통계량 -2.36143이므로 기각역 -2.06866 보다 작고 P 값이 0.02705이므로 해당 대립가설이 채택되어 성능 차이가 있다.

<그림 40> 쌍체 비교 T-검정 전체 model 결과



<표 37> 쌍체 비교 T-검정 전체 model 결과

	Precision		Recall		F1-score	
t-test	BERT	Llama2	BERT	BERT	BERT	Llama2
평균	0.68889	0.73113	0.65083	0.67297	0.63439	0.68385
분산	0.02864	0.02404	0.09614	0.06842	0.06990	0.04914
관측수	24	24	24	24	24	24
피어슨 상관 계수	0.87529		0.88995		0.92574	
가설 평균차	0		0		0	
자유도	23		23		23	
t 통계량	-2.51919		-0.76293		-2.36143	
P(T<=t) 단측 검정	0.00958		0.22663		0.01353	
t 기각치 단측 검정	1.71387		1.71387		1.71387	
P(T<=t) 양측 검정	0.01916		0.45326		0.02705	
t 기각치 양측 검정	2.06866		2.06866		2.06866	

제5장 결론

제1절 연구결과 요약

본 연구는 문서 자동분류 분야에 많이 이용되고 있는 Bert 모델과 최근 인공지능 LLM 모델인 Llama2 모델을 AI 허브의 논문 초록데이터를 분석대상으로 하여 실험하였다. 논문 초록데이터는 한국연구재단의 연구 분야 분류기준에 맞추어 인문학, 사회과학, 자연과학, 공학, 의약학, 농수해양학, 예술체육학, 복합학 총 8개의 분류하였다. 학습데이터는 144,280건, 테스트 데이터는 18,061건으로 총 162,341건을 사용하였다. 사회과학 47%, 인문학 22%, 공학 13%등 상위 3개의 카테고리의 데이터가 82% 분류되어 있다. 1995년부터 2020년 까지 데이터 중 2010~2020년 데이터가 학습데이터는 90%이며 테스트 데이터는 70%이다. 학습데이터의 오차의 경우 Short model의 경우 Bert가 더 낮았고 middle, long model의 경우 Llama2가 더 낮았다. 학습데이터 정확도의 경우 Llama2가 Bert에 대해서 높은 정확도를 보였으며 epoch 4 이상에서는 95% 이상 되었다.

Short model 행렬 분석 결과는 Bert의 경우 사회과학, 공학, 농수해양, 예술체육에서 높았으며 Llama2는 인문학, 자연과학, 의약학, 복합학에서 빈도가 높게 나타났다. Middle model 행렬 분석 결과는 Bert의 경우 사회과학, 공학, 농수해양에서 높았으며 Llama2는 인문학, 자연과학, 의약학, 예술체육, 복합학에서 빈도가 높게 나타났다. Long model 행렬 분석 결과는 Bert의 경우 공학, 농수해양에서 높았으며 Llama2는 인문학, 자연과학, 의약학, 예술체육, 복합학에서 빈도가 높게 나타났다. 이 모든 결과를 합쳐 비교해 보면 Bert의 경우 사회과학, 공학, 농수해양에서 높았으며 Llama2는 인문학, 자연과학, 의약학, 예술체육, 복합학에서 빈도가 높게 나왔다.

정밀도 결과는 short model일 경우 Bert가 인문학, 자연과학, 의약학, 농수해양, 예술체육, 복합학에서 Llama2가 사회과학, 공학에서 우수한 성능이 나왔다. middle model의 경우 Bert가 인문학, 의약학, 예술체육에서 Llama2가 사회과학, 자연과학, 공학, 농수해양, 복합학에서 우수한 성능이 나왔다. long model의 경우 Bert는 인문학에서 Llama2의 경우 사회과학, 자연과학, 공학, 의약학, 농수해양, 예술체육, 복합학에서 우수한 성능이 나왔다.

재현율 결과는 Short 모델의 경우 Bert가 인문학, 사회과학, 공학, 농수해양이 점수가 높았으며 Llama2는 인문학, 자연과학, 의약학이 점수가 높았다. Middle 모델의 경우 Bert의 경우 공학, 농수해양이 높았으며, Llama2는 인문학, 의약학, 예술체육이 높았다. Long mode의 경우 Bert가 공학에서 우수하고 Llama2의 경우 인문학, 자연과학, 의약학, 복합학이 우수하였다.

F1 스코어 결과는 Short model의 경우 Bert는 공학, 농수해양, 예술체육이 우수했으며 Llama2의 경우 자연과학, 복합학이 우수했다. 인문학, 사회과학, 의약학은 유사한 성능을 나타냈다. Middle model의 경우 Bert는 공학, 농수해양이 Llama2는 인문학, 의약학, 자연과학이 우수했다. Llama2의 경우 자연과학과 복합학이 높은 성능이 나왔다. Long model의 경우 Llama2가 모든 분류에서 높게 평가되었다.

두 모델을 비교하기 위해 유의수준 5%의 쌍체 비교 t-검정을 실시하였다.

Short model의 경우 정밀도, 재현율, F1 스코어가 대립가설이 채택되어 성능 차이가 없는 것으로 결과가 나왔다. Middle model의 재현율, F1 스코어 경우 대립가설이 기각되어 성능 차이가 없었고 정밀도는 대립가설이 채택되어 성능차이가 있는 것으로 확인되었다. Long model의 경우 재현율은 대립가설이 채택되어 성능 차이가 없었고 정밀도, F1 스코어는 대립가설이 채택되어 성능차이가 있는 것으로 확인되었다. short 모델은 성능 차이가 없었고 long 모델의 경우 정밀도, F1 스코어가 성능 차이가 있는 것으로 보아 입력 크기가 커질수록 성능 차이가 있는 것으로 판단되어 진다.

제2절 연구결과 의미

<표 38>과 같이 입력 데이터의 크기와 분류 품질의 데이터에 따라 모델의 선택 사이에 상호작용을 보여준다. 기업이나 조직에서 문서 자동 분류나 자연어 처리 모델을 선택할 때, 데이터의 특성, 분류의 목표에 따라 적절한 모델을 선택해야 한다. Short 텍스트의 경우 Bert는 Precision 측면에서 우수한 성능을 보이므로, 정확한 예측이 중요한 응용 프로그램에서 Bert를 고려할 가치가 있다. Middle 텍스트의 경우, Bert와 Llama2가 비슷한 성능을 보였다.

Long 텍스트의 경우, Llama2가 Precision 및 F1 스코어에서 우수한 결과를 제공하므로, 정밀한 분류가 필요한 상황에서 Llama2를 고려할 필요가 있다.

<표 38> 분류 품질별 모델 입력 크기에 따른 모델 선택

모델 입력 크기		
Short	Middle	Long
BERT	BERT, Llama	Llama

이와 같이 기술의 경영학적 함의를 요약해보자면, 인공지능 자연어 처리 모델의 선택을 할 경우 다양한 성능 지표와 모델의 입력 데이터 길이에 대한 분석 등 다양한 측면에서 종합적으로 이해하고, 최종적으로 문제에 해결에 있어 가장 적합하고 효율적인 모델을 선택하는 기준을 제시하였다. 본 논문에서 제안한 기준이외에도 문서의 다른 요소랑 결합하여 적합한 모델을 찾는 제언이 가능하다. 향후 인공지능 문서 분류 모델을 선택할 때 본 연구 지표가 시사점을 줄 수 있을 것으로 기대한다.

제3절 연구의 한계 및 향후 연구방향

본 연구는 문서 자동분류에서 Bert와 LLM인 Llama2 성능 비교하였다는 점에서 의의가 있다. 하지만, 다음과 같은 한계점과 추가적인 향후 연구가 필요하다.

LLM 모델들을 파라미터 수가 1,000억 개가 넘어가기 파인튜닝 하려면 GPU 연산이 많이 필요하며 하드웨어적 장비의 제한이 되어 있다. 실험의 사용된 GPU의 경우 48GB의 메모리가 있다. Llama2의 입력 크기가 512이고 Batch size가 3 경우 GPU 메모리가 47GB정도 사용되어 더 큰 사이즈(768,1024)의 입력크기 테스트가 진행하기 어려웠다. Llama2의 경우 한글 학습 자체가 0.06%로 한글 이해도 낮은 편이므로 한글 된 문서 분류의 한계가 있다.

LLM 모델의 하나인 Llama2 7B를 바탕으로 파인 튜닝한 모델을 사용하였다. Llama2의 13B, 70B 더 많은 파라미터가 있는 모델이 존재한다. 해당 모델을 사용하여 문서 자동분류의 성능 비교가 연구할 필요성이 있다. LLM 모델이 계속해서 국내외로 출시되고 있으며 한글지원 LLM 모델을 이용한 연구할 필요성 있다. LLM 모델의 강점 중 하나인 제로샷(Zero-shot) 및 여러 가지의 샘플 답변을 주고 하는 퓨샷(Few-shot)학습 과 파인튜닝을 이용한 학습 결과를 비교하는 연구를 할 필요성이 있다. Llama2의 경우 영어 데이터 이해력이 높은 수준이므로 한글 데이터를 번역을 통해 영문 데이터로 변경 후 자동분류를 하는 연구가 필요한다.

참고문헌

1. 참고문헌

권순보, & 유진은. (2022). BERT 와 FastText 를 활용한 온라인 진로상담 문서 분류. 한국데이터정보과학회지, 33(6), 991-1006.

김성훈, & 김승천. (2021). 특허문서 분류를 위한 딥러닝 개별 모델 분류기 성능 비교. 대한전자공학회 학술대회, 1904-1906.

김인후, & 김성희. (2022). 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동 분류. 정보관리학회지, 39, 293-310.

박규민, 홍충선, & 박성배. (2023). 질문 생성을 위한 대규모 언어 모델에 Low-Rank Adaptation 적용 방법. 한국정보과학회 학술발표논문집, 394-396.

박진우, 심우철, 이상헌, 고봉수, & 노한성. (2022). 한국어 특허 문장 기반 CPC 자동분류 연구-인공지능 언어모델 KorPatBERT 를 활용한 딥러닝 기법 접근. 지식재산연구, 17(3), 209-256.

백호준, & 김인철. (2023). 대규모 언어 모델 (LLM) 의 사전 지식을 활용한 3 차원 장면 그래프 생성. 멀티미디어학회논문지, 26(8), 859-873.

소현지, & 이종태. (2021). Text classification 에 특화 시킨 개선된 BERT 활용 방법론 제안. 한국통신학회 학술대회논문집, 1043-1044.

신성필. (2023). 초거대 AI 의 기반모델 (Foundation Model) 개념 및 표준화 동향. 한국통신학회지 정보와통신, 40(6), 12-21.

이수빈, 김성덕, 이주희, 고영수 and 송민. (2021). 딥러닝 자동 분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석. 정보관리학회지, 38(2), 153-172.

정단호, 김운, & 정유철. (2023). 초거대 인공지능 생성 모델 동향 연구. 한국통신학회지 정보와통신, 40(6), 22-28.

정의석. (2023). 기계가독형 데이터 시맨틱 상호운용성: 디지털 전환 시대의 당면 과제. 한국통신학회지 정보와통신, 40(5), 3-8.

주하영, 오현택, & 양진홍. (2023). 오픈 소스 기반의 거대 언어 모델 연구 동향: 서베이. 한국정보전자통신기술학회 논문지, 16(4), 193-202.

주호택, 이성하, & 김경중. (2023). 인간과 ChatGPT 의 대화내용을 이용한 공개 대형 언어모델 LLaMA 한국어 대화 능력 개선. 한국정보과학회 학술발표논문집, 991-99.

황상흠, & 김도현. (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214.

Kim, A., & Kim, J. (2022, May). Vacillating Human Correlation of SacreBLEU in Unprotected Languages. In Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval) (pp. 1-15).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Abstract

Kang, Kwang Sun
Dept. of AI Technology Management
The Graduate School of Technology Management
Kyung Hee University
Supervised by Professor. Ah Reum, Hong, Ph. D.

Different version of LLM models have been announced in the wake of the ChatGPT craze by AI giant Open AI. Meta's Llama model, announced in February 2023, opened up to research communication and revitalized the commercially while realizing similar performance to ChatGPT 3.5 by iteratively learning SFT and RLHF. By comparing the Bert model, which was widely used in the field of automatic document classification, and the latest LLM model, Llama2, we wanted to verify whether the Llama2 model had improved performance in automatic document classification compared to the Bert model. For training data, we used the 'Article Summary' dataset on AI-HUB. The training data was 160,000 documents from 1995 to 2020, and the target classification was defined as 8 categories based on the research field classification of the National Research Foundation of Korea. This study used a python program for this study and executed learning and automatic classification performance evaluation of Bert and Llama2. The results of this experiment showed that Bert had lower errors for shorter models, whereas Llama2 exhibited lower errors for medium and long models in terms of training data error.

For training data accuracy, Llama2 demonstrated higher accuracy than Bert across short, medium, and long models. The matrix analysis revealed that Bert

was more prevalent in social sciences, engineering, agriculture, and marine fields, while Llama2 was more prevalent in humanities, natural sciences, medicine, arts and sports, and multidisciplinary areas. In the classification evaluation, Bert outperformed Llama2 in precision, recall, and F1-score for the short model. However, in the medium and long models, Llama2 outperformed Bert in precision, recall, and F1-score.

A paired comparison t-test at a significance level of 5% was conducted for two models. The short model showed no performance difference, while the middle model exhibited a difference in precision but no variance in recall or F1 score. As for the long model, there was no difference in recall, but there were performance differences in precision and F1 score. The empirical analysis results provide evidence that when selecting an automatic document classification model, considering the input length using the Bert model for short text and the Llama2 model for long text may be necessary. These empirical analysis results serve as criteria for metric judgment based on the length of input data when selecting an automatic classification model.

In future research, this study plan to utilize various LLM models and investigate automatic document classification using zero-shot and few-shot learning.

keyword : Artificial intelligence, automatic classification, BERT, Llama, LLM