Journal of Korea Society of Industrial Information Systems Vol.29 No.3, Jun. 2024 :79-87
http://dx.doi.org/10.9723/jksiis.2024.29.3.079 ISSN:1229-3741

# 금융권에 적용 가능한 금융특화언어모델 구축방안에 관한 연구

# (A Study on the Construction of Financial-Specific Language Model Applicable to the Financial Institutions)

**배 재 권**[1)*]
(Jae Kwon Bae)

**요 약** 최근 텍스트분류, 감성분석, 질의응답 등의 자연어 처리를 위해서 사전학습언어모델(Pre-trained Language Model, PLM)의 중요성은 날로 강조되고 있다. 한국어 PLM은 범용적인 도메인의 자연어 처리에서 높은 성능을 보이나 금융, 제조, 법률, 의료 등의 특화된 도메인에서는 성능이 미약하다. 본 연구는 금융도메인 뿐만 아니라 범용도메인에서도 우수한 성능을 보이는 금융특화언어모델의 구축을 위해 언어모델의 학습과정과 미세조정 방법을 제안하는 것이 주요 목표이다. 금융도메인 특화언어모델을 구축하는 과정은 (1) 금융데이터 수집 및 전처리, (2) PLM 또는 파운데이션 모델 등 모델 아키텍처 선정, (3) 도메인 데이터 학습과 인스트럭션 튜닝, (4) 모델 검증 및 평가, (5) 모델 배포 및 활용 등으로 구성된다. 이를 통해 금융도메인의 특성을 살린 사전학습 데이터 구축방안과 효율적인 LLM 훈련방법인 적응학습과 인스트럭션 튜닝기법을 제안하였다.

**핵심주제어**: 사전학습언어모델, 금융도메인, 금융특화언어모델, 적응학습, 인스트럭션 튜닝

**Abstract** Recently, the importance of pre-trained language models (PLM) has been emphasized for natural language processing (NLP) such as text classification, sentiment analysis, and question answering. Korean PLM shows high performance in NLP in general-purpose domains, but is weak in domains such as finance, medicine, and law. The main goal of this study is to propose a language model learning process and method to build a financial-specific language model that shows good performance not only in the financial domain but also in general-purpose domains. The five steps of the financial-specific language model are (1) financial data collection and preprocessing, (2) selection of model architecture such as PLM or foundation model, (3) domain data learning and instruction tuning, (4) model verification and evaluation, and (5) model deployment and utilization. Through this, a method for constructing pre-learning data that takes advantage of the characteristics of the financial domain and an efficient LLM training method, adaptive learning and instruction tuning techniques, were presented.

**Keywords:** Pre-trained language models(PLM), Natural language processing(NLP), Financial-specific language model, Adaptive learning, Instruction tuning

* Corresponding Author: jkbae99@kmu.ac.kr
Manuscript received June 04, 2024 / revised June 13, 2024 / accepted June 13, 2024
1) 계명대학교 경영정보학과, 제1저자, 교신저자

## 1. 서 론

ChatGPT로 대표되는 대화형 인공지능(Conversational AI)은 디바이스, 챗봇 등의 애플리케이션에서 자연어를 이용하여 인간들과 대화할 수 있는 인공지능(AI) 시스템이다. 고객상담 및 서비스 지원이 필수적인 금융산업에서 인간 상담원을 대신하여 대화형 AI를 활용하는 사례가 급증하고 있다. 글로벌 시장조사업체 '마켓리서치퓨처(MRF)'에 따르면 대화형 AI 시장 규모는 연평균 22.6% 성장률을 기록하여 2030년에는 325억달러(약 43조원)의 거대 시장을 형성할 것으로 전망하고 있다. 2022년 11월, 미국의 오픈AI(OpenAI)가 출시한 ChatGPT가 세계적인 주목을 받음에 따라 디지털 혁신을 추진중인 글로벌 금융권의 관심도 증대되고 있다. 금융권은 다양한 업무에서 ChatGPT에 탑재된 기능을 활용해 조직의 생산성 및 효율성 향상, 고객경험 및 충성도 개선, 보안 및 리스크 관리 강화 등을 모색할 수 있을 것으로 기대하고 있다(Kim, 2023). 현재 금융권의 AI 도입 수요는 ChatGPT와 더불어 LLM(Large Language Model, 거대언어모델)이 주목받으며 늘어나고 있다. LLM은 수많은 파라미터를 보유한 인공신경망(neural networks) 기반의 언어모델로 대용량의 인간 언어를 이해하고 생성할 수 있도록 훈련된 인공지능이다(Han et al., 2021). 이중에서 GPT(Generative Pretrained Transformer)는 대표적인 LLM이며 이것은 대량의 학습용 데이터를 기반으로 사전학습을 수행한 후 미세조정을 통해 성능을 개선한다. 텍스트분류, 감성분석, 질의응답 등 자연어 처리를 위해서 사전학습언어모델(Pretrained Language Model, PLM)의 중요성은 날로 강조되고 있다(Chen et al., 2024). 국내는 AI의 한국어 능력을 빠르게 높일 수 있는 말뭉치 구축 정책이 활발하게 논의됨과 동시에 LLM 오픈소스에서 한국어 PLM이 공개되고 있다. 한국어 PLM은 범용적인 도메인의 자연어 처리에는 높은 성능을 보이나 금융, 제조, 법률, 의료 등의 특화된 도메인에서는 성능이 미약하다. 그 이유는 이들 모델들이 범용적이고 일반적인 말뭉치를 기반으로 학습되어 특정 도메인에 특화된 업무영역에는 적합하지 않기 때문이다. 이러한 한계점을 극복하기 위해 금융, 제조, 법률, 의료분야 등에서 특화된 개별 PLM을 개발(활용)하기 위한 연구를 시도하고 있다. 특히 금융도메인 지식에 전문화된 언어모델이 부족한 상황에서 금융기관들이 범용적 언어모델을 적용하기에는 어려움이 많은 실정이다(Lee, 2023). 즉, 금융권에서 특화언어모델을 구성하기 위해서는 금융용어, 금융데이터분석, 자연어기술, 머신러닝 및 딥러닝 기술, 금융 규제 및 컴플라이언스, 금융서비스 및 고객경험 등에 대한 이해가 필요하다.

LLM과 PLM 등 초거대 AI 모델은 방대한 양의 데이터를 학습하여 높은 성능을 보이고 있으나 때로는 '환각(hallucination)' 문제가 발생한다. AI 모델이 질문에 답변과 텍스트를 생성하는 경우 학습데이터에 없는 잘못된 정보를 마치 진실인 것처럼 만들어내는 현상으로 인해 금융 업무에 바로 적용하기 어렵다. 이를 위해 경제전문언론사 블룸버그(Bloomberg)는 금융 정보 데이터셋을 대규모로 학습한 '블룸버그GPT'를 초기 사전학습단계부터 직접 개발한 바 있다. 이러한 맞춤형 모델은 특정 분야에 깊은 전문성을 가질 수 있어 금융권에 큰 이점을 제공한다. 다만 사전학습단계는 엄청난 규모의 학습데이터와 GPU 자원이 필요하다. '블룸버그GPT'는 금융 분야의 텍스트 데이터 3,630억 토큰(비금융데이터 3,450억 토큰 포함)을 학습한 바 있다. 개발 비용과 예산문제로 일반 기업은 '블룸버그GPT' 사례처럼 모델을 직접 개발하지 않고 상업적으로 활용 가능한 기존 AI 모델을 선호한다. 메타(Meta)의 '라마(LLaMA)'와 같은 오픈소스 모델을 산업 특화 데이터로 추가학습(파인튜닝)하는 방법이다. 파인튜닝 과정은 상대적으로 적은 양의 학습데이터와 GPU 자원이 필요하다.

PLM은 사전학습단계에서 활용된 학습 말뭉치의 도메인에 따라 해당 분야에 특화된 지식을 습득한다. 이는 PLM의 성능과 활용가능성에 직접적인 영향을 미친다. 이와 같은 이유로 금융분야에서 PLM을 최적화된 방식으로 활용하기 위해 각 도메인에 특화된 PLM을 학습시킬 수

있는 방법론 연구가 필요한 시점이다(Noh, 2024). 그러나 금융권의 LLM 도입에 대한 연구 및 금융특화언어모델의 적용가능성과 구현방법에 대한 실증연구는 부족한 실정이다. 이와 같은 논의를 바탕으로 본 연구의 목적은 다음과 같다. 본 연구는 금융도메인 뿐만 아니라 범용 도메인에서도 우수한 성능을 보이는 금융특화언어모델의 구축을 위해 언어모델의 학습과정과 응용 관점에서 금융특화언어모델의 성능향상 및 활용방안을 제안한다. 또한 적응학습(Adaptation Learning)과 인스트럭션 튜닝(Instruction Tuning) 기법을 활용한 금융특화 PLM의 기본구조 및 구축과정을 제시한다.

## 2. PLM 개념과 LLM에 관한 선행연구

PLM은 대규모 코퍼스(corpus) 기반의 미리 학습된 언어모델로 다양한 자연어 처리 태스크에 활용된다. 대표적인 PLM인 BERT(Bidirectional Encoder Representations from Transformers)는 구글(Google)에서 개발한 트랜스포머 기반의 자연어 처리 모델이다. BERT는 기존의 LSTM(Long Short-Term Memory) 기반 언어모델과 비교하여 양방향 문맥 학습, 셀프 어텐션(Self-Attention) 구조, 대규모 사전학습, 마스크드 언어 모델(Masked Language Model) 등의 특징을 지닌다. BERT는 문장의 양방향 문맥을 고려하여 단어의 표현을 학습하나 LSTM 기반 언어모델은 단방향으로 문맥을 학습하므로, 문맥 정보를 충분히 활용하지 못한다. BERT는 트랜스포머 아키텍처의 핵심인 셀프 어텐션 구조를 사용한다. 셀프 어텐션은 문장 내 단어 간의 관계와 중요도를 직접적으로 계산하여, 단어의 문맥 표현을 효과적으로 학습한다. BERT는 대규모 말뭉치(Wikipedia, BookCorpus 등)로 사전학습을 진행하며, 이를 통해 광범위한 도메인 지식을 습득하고, 다양한 자연어 처리 태스크에 활용될 수 있는 강력한 언어 표현을 학습한다. 또한 BERT는 입력 문장에서 무작위로 일부 단어를 마스크(mask)하고, 주변 문맥을 통해 마스크된 단어를 예측하는 방식으로 학습한다. 이는 단어의 문맥 표현을 강력하게 만들어주며, 모델이 문맥을 깊이 이해하도록 도움을 준다(Lee et al., 2020). 이렇게 학습된 BERT 모델은 다양한 자연어 처리 응용 태스크에 활용되며, PLM에 목표 태스크의 학습데이터를 이용해 미세조정으로 추가 학습을 진행한다(Kim, 2023). PLM은 특정 태스크나 도메인의 성능 향상을 목표로 대규모 코퍼스로 학습된 언어 표현을 특정 태스크에 전이하여 활용한다. 이처럼 PLM은 주로 특정 태스크를 위한 파인튜닝(fine-tuning) 또는 특징 추출(feature extraction)에 사용된다.

최근까지 진행된 LLM에 관한 선행연구는 LLM 분야 최신 연구(기술)동향과 LLM 구축 핵심전략, LLM 품질평가기준, 프롬프트 엔지니어링 작성가이드, LLM 모델의 데이터 평가기준에 대한 연구 등이 수행되었다. 그러나 LLM 구축 및 활용가능성(적용가능성)에 관한 연구와 금융특화언어모델 구현방안에 관한 연구는 미흡한 실정이다. Kim et al.(2023)은 LLM의 기술동향을 소개하면서 학습데이터 및 평가데이터의 역할과 평가요소(언어생성능력, 지식활용능력, 추론능력)를 제시하였다. 또한 디코더(decoder) 사전학습모델에서 튜닝과정을 거쳐 강화학습 기반의 사용자 피드백을 반영한 LLM 구축과정을 제시하였다. Han et al.(2023)은 LLaMA, GPT-4, Alpaca, Vicuna 등의 LLM을 데이터 공개 및 코드 공개 여부, 평가방식 공개 여부 등으로 비교·분석하였다. 또한 이들은 LLM 품질평가기준으로 신뢰성(추론, 언어이해), 기능성, 유효성을 제시하였다. Park and Kang(2023)은 생성형 AI에서 프롬프트 엔지니어링에 대한 중요성을 강조하면서 프롬프트 엔지니어링 기법들을 비교·분석하였다. 이들은 퓨샷, 지식생성, 제로샷, 액티브 등 프롬프트 기법들의 발전과정과 상호 간의 관계를 도식화하였고, 이를 통해 추론 성능 향상을 위한 방안과 주요 벤츠마크 데이터를 제안하였다. Park and Lee(2023)는 감정 분류 작업에 사용된 데이터셋을 활용한 LLM보다 ChatGPT에서 추출한 추론 데이터셋 기반의 감정 분류 모델 성능이 우수하다는 연구결과를 제시하였다. Han et al.(2022)은 기존 LLM이 도메인 특성을 이해하지 못한다는 한계점을 지적하

면서 도메인 특화 추가 LLM의 중요성을 강조하였다. 이들은 정치, 경제, 법률, 의료 등의 데이터를 이용하여 도메인 특수성 지표 산출과 분류 태스크의 성능 개선 정도를 측정하고 PLM 개발과정을 제시하였다. 최근에는 소형특화언어모델(sLLM)에 관한 연구가 수행되고 있다(Heo et al., 2024; Jung et al., 2023). sLLM은 특정한 작업이나 도메인에 특화된 작은 규모의 언어 모델(통상 파라미터가 1000억 개 이하)을 의미한다. 파라미터 개수가 많을수록 성능향상에 유리하나 최근 생성형 AI 기술 발달로 sLLM도 LLM 또는 PLM보다 우수한 성능을 내며 비용도 절약할 수 있다는 점에서 다수의 기업들이 주목하고 있다. 주로 산업에 특화된 모델을 만들고, 연산 작업이 적어 스마트폰과 같은 개인용 기기에서도 작동해 수요가 많을 것으로 전망된다.

## 3. 적응학습과 인스트럭션 튜닝을 활용한 금융특화언어모델

### 3.1 금융특화언어모델의 기본구조

본 장에서는 금융특화언어모델의 기본구조 및 학습단계(과정)를 제시한다. 본 연구에서 제안하는 금융특화언어모델은 적응학습과 인스트럭션 튜닝기법을 활용하고, 이를 위해 데이터 수집 및 구조화 과정이 필수이다. 수집 단계는 금융시장 데이터, 고객 거래 기록, 규제 관련 문서 등 다양한 원천으로부터 필요한 금융빅데이터를 수집하고, 가공 단계에서 데이터를 정제·분류하며, 적절한 형태로 변환한다. 구조화된 데이터는 모델이 답변에 참조할 데이터를 쉽게 이해하고 처리할 수 있도록 돕는다. 금융특화언어모델은 기존 LLM(챗GPT, 바드, 하이퍼클로바 등)보다 파라미터 수를 10분의 1로 줄여 GPU 비용은 낮추되 금융회사에 필요한 분야만 집중 학습한다. 이것은 BERT와 동일한 트랜스포머 신경망으로 구성되며, 약 200M개의 파라미터로 구성된다. 금융특화 문서로 구성된 말뭉치를 구축하고, 이를 사전학습에 활용한다. 금융특화언어모델의 학습을 위해서 금융도메인에 특화된 대규모 말뭉치가 필요하다. 여기에는 금융 뉴스 기사, 금융 보고서 및 분석 자료, 금융 관련 학술논문 및 서적, 금융 챗봇 및 고객 상담 데이터, 금융 용어 사전, 공시데이터 및 금융계약서 등의 문서들을 포함한 말뭉치를 구성하여 학습에 사용한다. 대부분의 자료는 DB화가 되어있지 않고, 파일 또는 스킨 형태로 존재한다. Document Intelligence (DI) 기술을 통해 아날로그 및 디지털 문서의 정보를 AI가 이해할 수 있는 형태로 벡터(vector)화하여 저장하는 작업을 수행한다.

### 3.2 금융도메인 특화언어모델 구축과정

금융도메인 특화언어모델을 구축하는 과정은 Table 1과 같이 (1) 금융 데이터 수집 및 전처리, (2) PLM 또는 파운데이션 모델 등 모델 아키텍처 선정, (3) 도메인 데이터 학습과 인스트럭션 튜닝, (4) 모델 검증 및 평가, (5) 모델 배포 및 활용 등의 5단계로 요약할 수 있다.

첫째, 금융 데이터 수집 및 전처리 단계이다. 금융도메인의 텍스트 데이터(금융도메인의 뉴스기사, 보고서, 소셜 미디어 게시물 등)를 대량으로 원시데이터(음성포함)를 수집한다. 수집한 텍스트 데이터를 언어모델 학습에 적합한 형태로 정제한다. 토큰화, 불용어 제거, 맞춤법 교정, 형태소 분석 등의 작업이 이루어진다. 다음으로 데이터 분포의 적합성을 검토하고, 데이터형식 요건 검사 및 원시데이터 품질검사를 수행한다.

둘째, PLM 또는 파운데이션 모델 등 모델 아키텍처 선정이다. 파운데이션 모델은 PLM을 대규모의 데이터로 학습시킨 모델이다. PaLM, SOLAR, LLaMA3(라마 3) 등이 대표적인 파운데이션 모델이다. 파운데이션 모델은 광범위한 태스크와 도메인에 대한 일반적인 지식 습득을 목표로 한다. 즉, 텍스트 분류, 감성 분석, 개체명 인식, 질의응답 등의 태스크에 적용 가능하다. 파운데이션 모델은 제로샷 러닝(zero-shot learning), 원샷 러닝(one-shot learning), 퓨샷 러닝(few-shot learning) 등을 통해 적은 양의 데이터나 태스크 설명만으로 새로운 태스크를 수행할 수 있다는 장점이 있다(Brown et al.,

2020). 그러나 PLM은 특정 태스크나 도메인에 특화되어 새로운 태스크나 도메인으로의 확장이 제한적이다. 금융도메인 내 특정 태스크나 하위 도메인에 특화된 모델이 필요한 경우 해당 분야의 데이터로 사전학습된 모델을 활용하는 것이 효과적이다. 금융도메인 전반에 걸친 다양한 태스크와 하위 도메인에 대한 확장성과 일반화 능력이 중요한 경우 파운데이션 모델을 활용하는 것이 유리하다. 즉, 금융도메인에 특화된 언어모델을 개발하는 경우 해당 도메인의 특성과 요구사항을 고려하여 적합한 모델 구조 또는 아키텍처를 선택해야 한다. PLM을 활용해 효율적인 도메인 적응을 하고, 파운데이션 모델을 활용해 고차원적인 금융 태스크를 수행하는 것이 효과적인 전략이 될 수 있다.

금융도메인에서 사용되는 모델 구조와 아키텍처에는 트랜스포머 모델, CNN(Convolutional Neural Network) 모델, 그리고 금융특화 PLM 등이 사용된다. 트랜스포머 기반 모델은 사전학습된 BERT나 GPT 모델을 금융도메인 데이터로 미세 조정하여 사용한다. 텍스트의 길이가 길고 복잡한 구조를 가진 경우 트랜스포머 모델이 적합하며 셀프 어텐션 구조를 통해 문맥 정보를 효과적으로 포착할 수 있다. CNN 기반 모델은 금융 텍스트에서 중요한 패턴이나 키워드를 추출(텍스트 분류, 감성분석)하는 태스크에 적합하다. 최근에는 CNN과 트랜스포머를 결합한 CNN-Transformer가 사용된다. 최근에 등장한 금융 특화 PLM(FinBERT, BERT-Finance 등)은 금융 특화 어휘 및 문맥을 이해할 수 있어 다양한 금융 태스크에 적용 가능하다. 금융도메인에 적합한 모델 구조와 아키텍처는 태스크의 유형, 데이터 특성, 가용한 자원 등에 따라 선택되어야 한다. 또한 도메인 전문가와의 협업을 통해 금융 분야의 요구사항과 제약 조건을 고려하는 것이 중요하다. 최적의 모델을 선정하기 위해서는 다양한 구조와 아키텍처를 지속적으로 실험하고 평가하는 과정이 필요하다.

Table 1 The Construction Process of Financial-Specific Language Model

| Construction Process | Description |
|---|---|
| Financial data collection and preprocessing | Collect raw data, including text data from the financial domain, and perform preprocessing |
| Model architecture selection | Establish a model structure and architecture suitable for the financial domain using PLM and foundation models. |
| Domain data learning and instruction tuning | Fine-tuning is performed to fit areas specialized in the financial domain using adaptation learning and instruction tuning |
| Model verification and evaluation | Measures task-based evaluation indicators such as text classification, entity name recognition, and text summarization. |
| Model deployment and utilization | Consists of model optimization and light-weighting, API design and implementation, security and authentication system construction, distribution automation, and user feedback. |

셋째, 도메인 데이터 학습과 튜닝과정이다. 선정된 PLM 또는 파운데이션 모델을 금융도메인 데이터로 미세 조정하는 과정으로 여기에서 모델이 도메인 특화 지식과 어휘를 습득하게 된다. 일반적인 튜닝 과정은 태스크 수행에 필요한 입력-출력 쌍(예: 질문-답변, 문서-요약)으로 구성된 데이터셋을 사용하며 특정 태스크(예: 감성 분석, 개체명 인식, 문서 분류 등)에 대한 성능 향상에 초점을 맞춘다(Seo et al., 2022). 최근에는 '적응학습(Adaptation Learning)'과 '인스트럭션 튜닝(Instruction Tuning)'기법이 활용되고 있다. 적응학습이란 PLM을 특정 도메인이나 태스크에 맞게 조정하는 과정이다. 대규모 코퍼스로 학습된 언어모델은 일반적인 언어이해 능력을 가지고 있으나 특정 분야나 태스크에 최

적화되어 있지 않다(Yu and Kim, 2023). 적응 학습을 통해 금융도메인에 맞게 언어모델을 조정하여 해당 분야에서 성능을 향상시킬 수 있다. 또한 특정 태스크에 맞게 언어모델을 조정하는 인스트럭션 튜닝을 통해 해당 태스크에 대한 성능을 향상시킨다. 인스트럭션 튜닝은 NLP 분야에서 PLM을 질의응답, 요약, 번역업무 등에 맞게 미세 조정하는 기술이다. 인스트럭션 튜닝은 자연어로 기술된 명령어(instruction, 지시사항)와 그에 따른 출력 쌍으로 구성된 데이터셋을 사용하는데 자연어 명령어를 이해하고 따르는 능력 자체를 향상시키는 것이 주된 목표이다. 금융 업무를 수행하기 위한 지시사항과 관련 입출력 예시 데이터를 준비하고, 금융 특화 데이터를 사용하여 PLM을 추가로 학습시킨다. 이 과정에서 모델은 특정 태스크를 수행하는 방법을 학습한다. 태스크에 맞게 미세 조정된 모델을 사용하여 해당 태스크를 수행한다. 이처럼 인스트럭션 튜닝은 PLM을 활용하여 다양한 NLP 태스크에 대한 고성능 모델을 빠르게 개발할 수 있는 방법으로 주목받고 있다.

넷째, 모델 검증 및 평가이다. 기본적으로 데이터는 학습용(training set), 검증용(validation set), 평가용(test set)으로 분리되며 이 단계는 학습된 모델의 성능을 평가하기 위해 평가용 데이터셋을 활용하며 텍스트 분류, 개체명 인식, 텍스트 요약 등 태스크 기반 평가지표를 측정한다. 전문가 피드백을 통해 모델의 타당성을 정성적으로 검토하기도 한다. 금융 특화 언어모델의 평가지표는 Table 2와 같이 퍼플렉서티(Perplexity), BLEU 스코어, ROUGE 스코어, F1 스코어, 사전훈련 태스크평가, 생성 태스크 평가, 편향성 평가, 효율성 평가 등이 있다(Chang et al., 2024). 퍼플렉서티는 언어모델이 실제 텍스트 데이터를 얼마나 잘 예측하는지 측정하는 지표이다. 퍼플렉서티가 낮을수록 예측 성능이 우수한 것이다. BLEU(Bilingual Evaluation Understudy) 스코어는 기계번역 태스크에서 모델이 생성한 번역문과 참조 번역문 간의 유사도를 측정하는 지표이다. ROUGE(Recall Oriented Understudy for Gisting Evaluation) 스코어는 텍스트 요약 태스크에서 모델이 생성한 요약문과 참조 요약문 간의 유사도를 측정하는 지표이다. F1 스코어는 질의응답, 개체명 인식 등의 태스크에서 모델정확도를 측정하는 지표이다. 이것은 정밀도(Precision)와 재현율(Recall)의 조화평균으로 계산된다. 사전훈련 태스크평가는 GLUE(General Language Understanding Evaluation), SuperGLUE 등의 벤치마크 태스크 세트를 통해 모델의 언어이해능력을 평가한다. 텍스트 분류, 자연어 추론, 의미론적 유사도 판단 등 다양한 하위 태스크로 구성된다.

Table 2 Model Verification and Evaluation Indicators

| Evaluation indicators | Description |
|---|---|
| Perplexity | Measuring the prediction performance of a language model |
| BLEU score | Measure similarity between machine translation results and reference translation |
| ROUGE score | Measure the similarity between text summary results and reference summaries |
| F1 score | Measuring accuracy in question answering, entity name recognition, etc. |
| Pre-training task evaluation | Assess language comprehension skills with task sets such as GLUE and SuperGLUE |
| Generation task evaluation | Assessing the quality of text produced in tasks such as sentence generation and dialogue generation; |
| Bias assessment | Verification of bias regarding gender, race, etc. in the generated text, |
| Efficiency evaluation | Evaluation of performance, inference speed, energy efficiency, etc. compared to model size |

생성 태스크 평가는 문장 생성, 대화 생성, 스토리 생성 등의 태스크에서 모델이 생성한 텍스트의 품질을 평가하고, 편향성 평가는 모델이

생성한 텍스트에서 성별, 인종 등에 대한 편향이 나타나는지 평가한다. 효율성 평가는 모델 크기 대비 성능, 추론 속도, 에너지 효율성 등을 평가한다. 실제 서비스 환경에 적용 가능한 수준인지 판단하는 기준이 된다.

마지막 단계는 모델 배포 및 활용이다. 이 단계는 세부적으로 모델 최적화 및 경량화, 모델 서빙(Serving) 환경 구축, API 설계 및 구현, 보안 및 인증 체계구축, 배포 자동화 및 사용자 피드백으로 구성된다. 모델 최적화 및 경량화는 모델의 크기를 줄이고 추론 속도를 높이기 위해 압축 기술을 적용한다. 양자화(Quantization), 가지치기(Pruning), 지식 증류(Knowledge Distillation) 등의 기술을 활용한다. 모델 경량화를 통해 배포 및 실행 환경에 맞는 모델을 준비한다. 모델 서빙 환경 구축은 모델을 서비스로 제공하기 위한 서버 환경을 구축한다. Flask, FastAPI 등의 웹 프레임워크를 사용하여 API 서버를 구현한다. 서버 환경에서 모델을 로드하고 추론을 수행하도록 설정한다. API 설계 및 구현은 모델을 활용할 수 있는 API를 설계한다. 입력 데이터의 형식, 출력 결과의 형식, 요청 방식 등을 정의한다. API 문서화를 통해 개발자들이 모델을 쉽게 활용할 수 있도록 가이드를 제공한다. 보안 및 인증 체계 구축은 모델 API에 대한 접근 제어 및 인증 체계를 구축한다. 데이터 암호화, 접근 로깅 등의 보안 조치를 수행하여 개인정보보호와 사생활 침해를 방지한다. 배포 자동화 및 피드백은 모델의 버전을 관리하고 배포 과정을 자동화하며, 사용자 피드백을 수집하여 최종적으로 반영한다.

## 4. 결 론

PLM은 학습단계에 사용된 자연어 말뭉치의 특성에 영향을 받으며 이후 PLM이 실제 활용되는 응용단계가 적용되는 도메인에 따라 최종 모델 성능에 큰 차이를 보인다. 최근 등장한 한국어 PLM은 일반적이고, 범용적인 도메인의 자연어 처리에는 우수한 성능을 보이나 제조, 금융, 의료, 법률 등의 특화된 영역에서는 답변 완성도와 안정성이 미흡한 것으로 알려져 있다. 이러한 한계점을 극복하기 위해 금융, 의료, 법률분야에서 해당 도메인에 특화된 PLM을 활용하기 위한 연구가 시도되고 있다. 특히 금융용어 관련 도메인 지식에 전문화된 언어모델은 부족하고, 은행을 비롯한 금융기관에서 최신 언어모델을 쓰기에 아직 어려움이 많은 실정이다. 본 연구는 금융도메인 뿐만 아니라 범용도메인에서도 우수한 성능을 보이는 금융특화언어모델 구축을 위해 언어모델의 학습과정과 미세조정 방법을 제안하고, 응용 관점에서 금융특화언어모델의 성능향상 방안을 제안하는 것이 주요 목표이다. 금융도메인 특화언어모델을 구축하는 과정은 (1) 금융 데이터 수집 및 전처리, (2) PLM 또는 파운데이션 모델 등 모델 아키텍처 선정, (3) 도메인 데이터 학습과 인스트럭션 튜닝, (4) 모델 검증 및 평가, (5) 모델 배포 및 활용 등의 5단계로 구성된다. 이를 통해 금융도메인의 특성을 살린 사전학습 데이터 구축방안과 단순하고 효율적인 LLM 훈련방법인 적응학습과 인스트럭션 튜닝과정을 제시하였다.

본 연구는 금융권의 금융특화언어모델에 관한 연구가 전무한 상황에서 금융특화언어모델의 기본구조 및 학습 말뭉치 구성방법, 도메인 데이터 학습과 튜닝과정에 관한 내용을 제시하여 금융권 LLM 연구의 이론적 발전에 기여하였다. 또한 금융권 대화형 AI 서비스 개발자, 한국어 PLM 개발자, LLM 개발자 등에게 금융권의 언어모델 운영 및 관리적인 고려사항, 금융특화언어모델의 활용방안 탐색 등의 기본 지침을 제공할 수 있다. 향후 연구에서는 제조와 법률 분야에 특화된 언어모델 구축과정과 성능개선방안을 탐색하고자 한다.

## References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A. (2020). Language Models are Few-shot Learners, *Advances in Neural Information Processing*

*Systems*, 33(1), 1877-1901.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K. and Xie, X. (2024). A Survey on Evaluation of Large Language Models, *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45.

Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X. and Tang, J. (2024). Exploring the Potential of Large Language Models in Learning on Graphs, ACM SIGKDD Explorations Newsletter, 25(2), 42-61.

Han, M. A., Kim, Y. H. and Kim, N. G. (2022). The Effect of Domain Specificity on the Performance of Domain-specific Pre-trained Language Models, *Journal of Intelligence and Information Systems*, 28(4), 251-273.

Han, N. E., Seo, S. and Um, J. H. (2023). A Proposal of Evaluation of Large Language Models Built Based on Research Data, *Journal of the Korean Society for Information Management*, 40(3), 77-98.

Han, X., Zhang, Z. Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A. and Zhang, L. (2021). Pre-trained Models: Past, Present and Future, AI Open, 2(1), 225-250.

Heng, J., Teo, D. B. and Tan, L. F. (2023). The Impact of Chat Generative Pre-trained Transformer(ChatGPT) on Medical Educati-on, *Postgraduate Medical Journal*, 99(1176), 1125-1127.

Heo, H. D., Kang, D. G., Kim, Y. S. and Chun, S. H. (2024). A Study on the Intelligent Document Processing Platform for Document Data Informatization, *The Journal of The Institute of Internet, Broadcasting and Communication*, 24(1), 89-95.

Jung, J. K., Choi, S. K. and Kwon, H. C. (2023). Combining sLLM and Re-ranking Strategies for an Efficient GEC Model, *Korean Institute of Information Scientists and Eng*, 2023(12), 362-364.

Kim, A. Muhn, M. and Nikolaev, V. (2023). Bloated Disclosures: Can ChatGPT Help Investors Process Financial Information?, *General Economics*, 13, 1-20.

Kim, J. S. (2023). A Study on Fine-Tuning and Transfer Learning to Create a Sentiment Binary Classification Model in Korean Text, *Journal of Korea Society of Industrial Information Systems*, 27(5), 1-11.

Kim, S., Shin, J., Yun, H. G., Lee, J., Choi, J. and Han, J. (2023). Technology Trends of Large Language Models in the Age of Generative AI, *Korean Institute of Information Scientists and Engineer*, 41(11), 25-33.

Lee, C. H., Lee, Y. J. and Lee, D. H. (2020). A Study of Fine Tuning Pre-trained Korean BERT for Question Answering Performance Development, *Journal of Information Technology Services*, 19(5), 83-91, 2020.

Lee, H. (2023). Innovations and Risk Factors of Generative AI in the Financial Industry, *Global Financial Review*, 4(1), 91-121.

Nah, F. H., Zheng, F., Cai, R., Siau, J. K. and Chen, L. (2022). Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration, *Journal of Information Technology Case and Application Research*, 25(3), 277-304.

Park, S. and Kang, J. (2023). Analysis of Prompt Engineering Methodologies and Research Status to Improve Inference Capability of ChatGPT and Other Large Language Models, *Journal of Intelligence and Information Systems*, 29(4), 287-308.

Park, N. and Lee, M. (2023). Empowering Emotion Classification Performance through Reasoning Dataset from Large-scale Language Model, *Korean Computer and Information Society Academic Conference Papers*, 31(2), 59-61.

Seo, B., Lee, Y. H. and Cho, H. (2022).

Creation and Use of the News Sentiment Index (NSI) using Machine Learning, *National Accounts Review*, 1, 1-15.

Noh, S. (2024). Development of Large-Scale Language Models and Ways to Utilize Them in Financial Information Analysis, *Capital Market Research Institute Issue Report*, 19, 1-24.

Yu, Y. and Kim, H. (2023). Development of a Regulatory Q&A System for KAERI Utilizing Document Search Algorithms and Large Language Model, *Journal of Korea Society of Industrial Information Systems*, 28(5), 31-39.

**배 재 권 (Jae Kwon Bae)**

- 정회원
- 한남대학교 경영정보학과 경영학사
- 서강대학교 경영학과 경영학석사
- 서강대학교 경영학과 경영정보학박사
- (현재) 계명대학교 경영정보학과 정교수
- 관심분야: 데이터마이닝, 금융빅데이터분석, 인공지능, 생성형 AI 및 거대언어모델 등