ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 June: 28(2): 191~206 http://dx.doi.org/10.13088/jiis.2022.28.2.191

# KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용

**김동규**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(donggyukimc@kbfg.com)*

**이동욱**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(dylan@kbfg.com)*

**박장원**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(jangwon.park@kbfg.com)*

**오성우**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(swoh@kbfg.com)*

**권성준**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(sjkwon@kbfg.com)*

**이인용**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(iylee@kbfg.com)*

**최동원**
KB국민은행 테크그룹
금융AI센터 AI테크팀
*(dwchoi@kbfg.com)*

대량의 말뭉치를 비지도 방식으로 학습하여 자연어 지식을 획득할 수 있는 사전학습 언어모델(Pre-trained Language Model)은 최근 자연어 처리 모델 개발에 있어 매우 일반적인 요소이다. 하지만, 여타 기계학습 방식의 성격과 동일하게 사전학습 언어모델 또한 학습 단계에 사용된 자연어 말뭉치의 특성으로부터 영향을 받으며, 이후 사전학습 언어모델이 실제 활용되는 응용단계 태스크(Downstream task)가 적용되는 도메인에 따라 최종 모델 성능에서 큰 차이를 보인다. 이와 같은 이유로, 법률, 의료 등 다양한 분야에서 사전학습 언어모델을 최적화된 방식으로 활용하기 위해 각 도메인에 특화된 사전학습 언어모델을 학습시킬 수 있는 방법론에 관한 연구가 매우 중요한 방향으로 대두되고 있다. 본 연구에서는 금융(Finance) 도메인에서 다양한 자연어 처리 기반 서비스 개발에 활용될 수 있는 금융 특화 사전학습 언어모델의 학습 과정 및 그 응용 방식에 대해 논한다. 금융 도메인 지식을 보유한 언어모델의 사전학습을 위해 경제 뉴스, 금융 상품 설명서 등으로 구성된 금융 특화 말뭉치가 사용되었으며, 학습된 언어 모델의 금융 지식을 정량적으로 평가하기 위해 토픽 분류, 감성 분류, 질의 응답의 세 종류 자연어 처리 데이터셋에서의 모델 성능을 측정하였다. 금융 도메인 말뭉치를 기반으로 사전 학습된 KB-BERT는 KoELECTRA, KLUE-RoBERTa 등 State-of-the-art 한국어 사전학습 언어 모델과 비교하여 일반적인 언어 지식을 요구하는 범용 벤치마크 데이터셋에서 견줄 만한 성능을 보였으며, 문제 해결에 있어 금융 관련 지식을 요구하는 금융 특화 데이터셋에서는 비교대상 모델을 뛰어넘는 성능을 보였다.

**주제어** : 자연어 처리, 금융, 딥러닝, BERT, 사전학습언어모델

논문접수일 : 2022년 6월 16일 논문수정일 : 2022년 6월 21일 게재확정일 : 2022년 6월 21일
원고유형 : 학술대회용 Fast-Track 교신저자 : 최동원

## 1. 서론

구글 BERT(Bidirectional Encoder Representation Transformer)(Devlin, 2019) 사전학습 언어모델(Pre-trained Language Model)을 시작으로 텍스트 분류(Topic classification), 감성 분석(Sentiment analysis), 그리고 질의 응답(Question answering) 등의 다양한 자연어 처리 모델을 State-of-the-art 수준의 성능으로 학습시키기 위해서는 사전학습 언어모델의 역할이 필수적이게 되었다. 사전학습 언어모델은 비지도 방식을 활용하여 대량의 말뭉치를 기반으로 학습되며, 기존 영문 및 다국

어(Multilingual) 말뭉치 기반의 사전학습 언어 모델은 한국어 자연어 처리 역량에 있어 한계점을 가진다. 이를 극복하기 위해 최근 오픈소스 커뮤니티 및 학계에서는 한국어 말뭉치를 기반으로 학습한 다양한 사전학습 언어 모델을 공개하였으며, KoELECTRA[1](Park, 2020), KLUE-RoBERTa[2](Parketal, 2021) 등의 한국어 사전학습 언어모델이 대표적으로 널리 활용되고 있다.

이러한 State-of-the-art 한국어 사전학습 언어모델들은 위키, 뉴스(유소연, 임규건, 2021) 등의 범용적인 도메인에 대한 자연어 처리에서는 높은 성능을 보이지만 OOD(Out-of-Distribution)에 해당되는 의료, 법률, 금융 등의 특수 도메인에서는 취약한 면모를 보인다. 이는 해당 모델들이 위키 문서, 웹 페이지, 일반 뉴스 등으로 대부분 구성된 범용적인 말뭉치를 기반으로 학습되어 특정 도메인에서의 활용에 최적화되어 있지 않다는 한계가 있기 때문이다. 자연어 처리 모델에 있어 도메인 특화 능력의 부재로 발생할 수 있는 가장 일반적인 현상으로는 1) OOV(Out-of-vocabulary) 발생으로 인한 언어 이해 부족 2) 특정 도메인에서만 사용되는 단어 및 유의어 관계에 대한 이해 부족 등이 존재하는데, 이러한 문제들은 사전학습 언어모델을 기반으로 학습된 다양한 자연어 처리 태스크(Downstream task)의 활용 시 최종 성능에 매우 큰 영향을 준다.

이러한 한계점을 극복하기 위해 의료, 법률 등 다양한 분야에서 해당 도메인에 특화된 사전학습 언어모델을 학습하고 활용하기 위한 연구를 시도하였으며 그 필요성 및 효율성을 입증하고 있다. 금융 또한 이러한 도메인 특화 모델의 필요성이 존재하는 분야의 하나로, 1) 일반적으로 사용되지 않는 금융 용어에 대한 이해 2) 상품명 등 고유명사의 이해 3) 문서내 여러 수치 및 값에 대한 이해(Numerical reasoning) 등 금융 전문 지식을 갖춘 언어 모델의 활용을 위해서는 이를 고려한 자연어 처리 모델의 학습이 필수적이다.

본 연구는 금융 특화 언어모델의 학습과 이를 응용한 금융 특화 자연어 처리 모델에서의 성능 향상을 목표로 한다. 금융 특화 언어모델의 학습을 위해 가장 일반적으로 사용되는 위키, 웹 문서, 뉴스에 더하여, 경제뉴스, 금융 상품/투자 설명서 등 금융 도메인 관련 문서들을 대량으로 추가한 말뭉치를 구성하여 학습에 사용하였다. 이렇게 학습된 금융 특화 사전학습 언어모델 KB-BERT의 언어이해 능력에 대한 정량적인 성능평가를 위해 위키, 뉴스 등으로 생성된 범용 데이터셋과 금융 특화 데이터셋 모두를 대상으로 모델의 학습 및 평가를 진행하였다. 성능평가 결과 금융 특화 KB-BERT는 범용 데이터셋에서 KoELECTRA, KLUE-RoBERTa 등의 State-of-the-art 한국어 사전학습 언어모델들과 유사한 성능을 보임과 동시에 금융 특화 데이터셋에서는 해당 모델들을 뛰어넘는 성능을 보이는 것을 확인할 수 있었다.

## 2. 관련 연구

### 2.1. BERT

BERT 사전학습 언어모델은 Transformer(Vaswani et al, 2017) 뉴럴 네트워크 구조를 활용한 첫 번째 언어모델로 기존 LSTM 기반 언어모델(Peters

1) https://github.com/monologg/KoELECTRA
2) https://huggingface.co/klue/roberta-base

et al, 2018)과 달리 Self-Attention 구조를 통해 자연어 텍스트로부터 더 강력한 contextual representation을 학습할 수 있다는 장점이 있다. BERT는 학습을 위해 Masked language modeling 방법을 사용하며, 이는 주어진 텍스트의 일부분을 임의로 마스킹(Masking)하고 이를 다시 원본 텍스트로 복원하는 디코딩(Decoding)을 수행하도록 모델이 학습된다. 이러한 학습 과정을 통해 사람으로부터 생성된 Labeled 데이터가 없어도 비지도 방식으로 언어 지식을 학습할 수 있다. 이렇게 학습된 BERT 모델은 다양한 자연어 처리 응용 태스크에 활용되며, 가장 대표적인 활용 방식은 사전학습 언어모델에 목표 태스크의 학습 데이터를 이용해 미세조정(Fine-tuning)으로 불리는 추가 학습을 거치는 것이다. 사전학습 언어모델을 이용해 미세조정 방식으로 학습된 자연어 처리 모델들은 그렇지 않은 모델들과 비교하여 주어진 태스크에 더 높은 성능 및 강건성(Robustness)을 보이는 등의 장점이 있다.

### 2.2. 도메인 특화 학습

BERT 등의 사전학습 언어모델은 학습에 사용된 말뭉치의 도메인 등 데이터 특성에 큰 영향을 받으며 이는 언어모델의 실제 사용에 있어 Downstream 태스크의 성능을 결정하는 중요한 요소이기에, 도메인 특화 언어모델 학습 연구가 활발히 이뤄지고 있다. 가장 일반적인 방식은 의료(Lee et al, 2020), 법률(Chalkidis et al, 2020), 과학(Beltagy et al, 2019) 등 목표 도메인 말뭉치를 수집 및 활용하여 From-scratch 방식으로 언어모델을 학습시키는 것이다. 이러한 도메인 특화 대상은 금융, 의료 등 분야의 구분에 한정되지 않으며 테이블 데이터(Yin et al, 2020), 지식 그래프(Wang et al, 2021) 데이터 처리를 위한 언어모델 등 목표 도메인 데이터가 갖는 특성에 따라 많은 응용 방법이 존재한다. 한편 이러한 도메인 특화 모델 학습의 효율성 향상을 위한 연구도 활발히 진행되고 있다. 가장 일반적인 접근법은 기존 학습된 범용 목적 언어모델을 기반으로 도메인 적응 기법을 활용하는 것이다. DAPT(Gururangan et al, 2020)는 소규모의 도메인 특화 말뭉치를 범용 모델에 추가적으로 학습하는 방법을 제안하였고 금융(Araci, 2019) 등 다양한 도메인에서 효과를 보였다. 이 밖에도 범용 모델의 OOV 방지를 위한 Adaptive Tokenizer(Sachidananda et al, 2021) 등 다양한 관련 연구가 존재한다. 이와 같은 범용 언어모델 기반의 Post-training 방법들은 앞서 설명된 From-scratch 방식과 비교해 성능상 뒤쳐지지만, 학습 시간 및 비용 등 효율성 측면에서 장점이 있다.

## 3. 금융 특화 사전학습 언어모델

본 장에서는 금융 특화 사전학습 언어모델인 KB-BERT의 모델 구조, 학습 방법 및 환경에 대해 서술한다.

### 3.1. KB-BERT

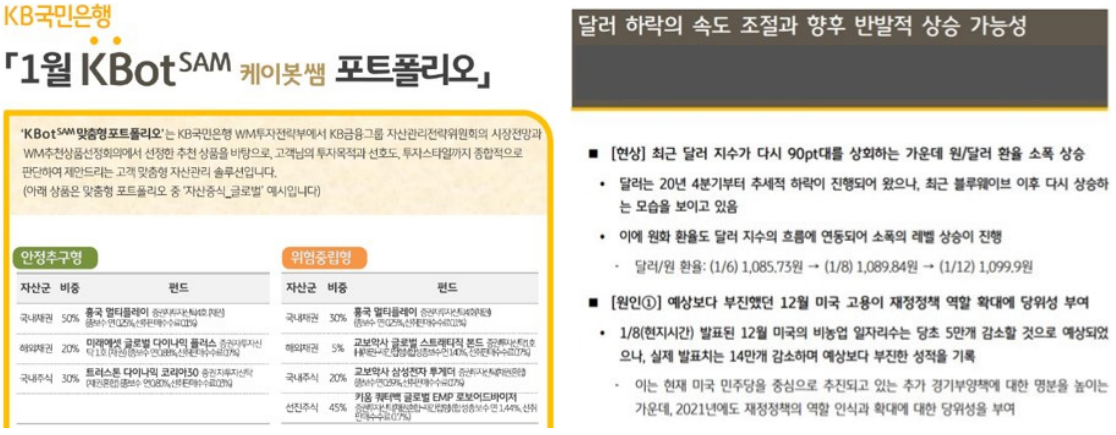
본 연구에서 제안하는 금융 특화 사전학습 언어모델인 KB-BERT는 BERT와 동일한 Transformer 뉴럴 네트워크로 구성되며, 모델 전체 구조는 <표 1>의 하이퍼파라미터 설정을 사용한다. KB-BERT는 약 110M개의 파라미터로 구성된 뉴럴 네트워크 모델로, KoELECTRA, KLUE-RoBERTa Base 사이즈 모델과 동일한 크기에 해당한다. 하지만

〈표 1〉 모델 하이퍼파라미터

| 이름 | Vocab | Word embedding | Layer | Hidden size | Self-attention heads |
|---|---|---|---|---|---|
| 크기 | 35,000 | 786 | 12 | 786 | 12 |

〈표 2〉 학습 말뭉치 크기

| 모델명 | 총 말뭉치 크기 (GB) | 금융 말뭉치 크기(GB) |
|---|---|---|
| KoELECTRA-v3 | 34 | - |
| KLUE-RoBERTa | 62 | - |
| KB-BERT | **90** | **40** |




〈그림 1〉 금융 문서 예시

기존 언어모델들과 달리 금융 특화 문서로 구성된 말뭉치를 구축하여 KB-BERT의 사전학습에 활용함으로써, 금융 도메인에 필요한 언어 지식의 학습을 목표로 한다. KB-BERT의 사전학습 과정은 NLU(Natural language understanding) 태스크에 강점을 보이는 BERT 모델의 Masked language modeling(Delvin et al, 2019) 방식을 통해 진행된다. GPT(Radford et al, 2019) 모델로 대표되는 NLG(Natural language generation)를 위한 Autoregressive 방식의 사전학습 언어모델은 본 연구에서 다루지 않는다.

## 3.2. 학습 말뭉치

### 3.2.1. 말뭉치 구성

KB-BERT 학습에 사용된 말뭉치는 기존 한국어 언어모델 학습에 사용되었던 기본적인 위키, 뉴스, 웹 문서를 포함하며 추가적으로 금융 관련

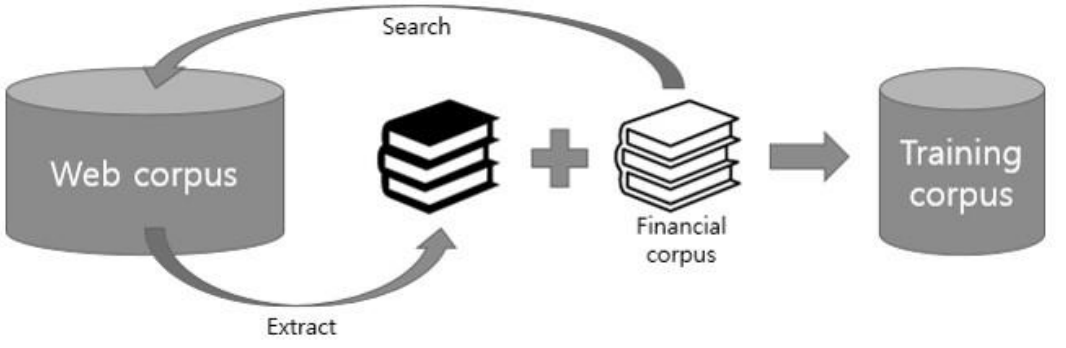
문서가 포함되었다. 대표적인 금융 관련 문서로는 금융 상품 설명서 및 투자 리포트 문서가 사용되었으며, 이에 대한 예시가 <그림 1>이다. KB-BERT의 학습에는 총 90GB 크기의 말뭉치가 사용되었으며 이는 기존 한국어 사전학습 언어모델과 비교했을 때 약 30GB 큰 크기이다. 총 용량의 약 40%에 해당하는 40GB는 경제 관련 뉴스, 금융 관련 문서로 구성되었으며 이는 기존 모델들의 총 용량과 비교해도 매우 큰 크기이다. 기존 모델 및 KB-BERT에 사용된 말뭉치 정보는 <표 2>에서 확인할 수 있다.

#### 3.2.2. 말뭉치 전처리

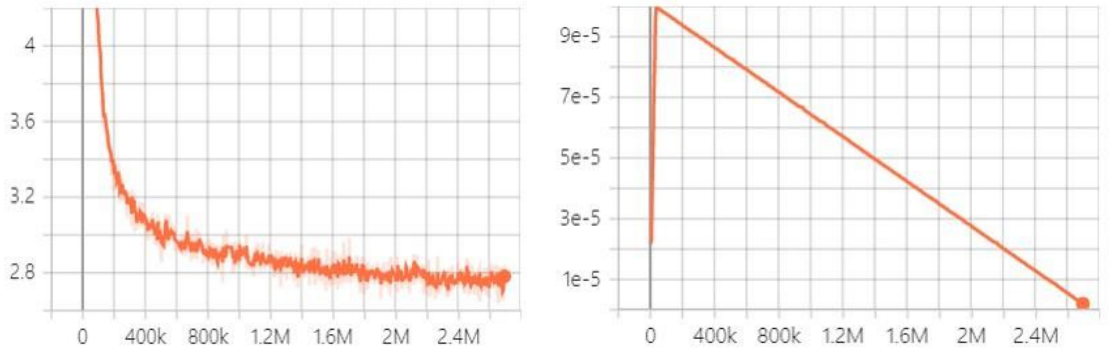
말뭉치는 언어모델의 성능에 매우 큰 영향을 주기 때문에 학습 말뭉치의 정제작업은 언어모델 사전학습을 위한 중요한 단계이다. KB-BERT를 포함한 많은 사전학습 언어모델에서 사용하는 가장 일반적인 말뭉치는 뉴스 및 웹 문서이며, 이 두 가지 유형의 문서들의 특징 중 하나는 광고성 텍스트가 빈번하다는 것이다. 이런 요소는 언어모델의 사실(Factual) 기반 예측 능력 저하 및 최근 언어모델의 윤리적(Ethical) 이슈 등 다양한 문제를 유발한다. 따라서 해당 문서들을 필터링하기 위해 스팸 텍스트 분류 모델이 활용되었다. 또한, 뉴스 및 웹 문서는 말뭉치 내 중복 문서가 존재하기 쉬우며 이로 인해 발생할 수 있는 언어모델의 성능 저하 및 불필요한 학습 시간 증가를 방지하기 위하여 해시(MinHashLSH) 기반 문서 중복 제거(Gao et al, 2021) 방식을 사용하여 문서 필터링이 수행되었다. 수집된 웹 문서 말뭉치 내에는 한국어 이외의 다양한 언어로 작성된 문서가 포함될 가능성이 있기 때문에 언어모델의 한국어 능력을 극대화하기 위해 언어 판별 모델을 적용하여 외국어 문서 필터링이 진행되었다.

#### 3.2.3. 금융 말뭉치 증강

KB-BERT를 위한 금융 도메인 학습 말뭉치 수량이 충분하지 않았으므로 추가적인 금융 문서 말뭉치를 획득하기 위해 문서 검색 기반의 말뭉치 증강 과정을 수행하였다. 이런 검색 기술을 기반으로 한 언어모델 학습은 도메인 적응(Domain adaptation)(Yao et al, 2021) 및 대규모 모델 학습(Borgeaud et al, 2021)을 위해 최근 연구에서 활발히 이용되고 있다. 초기 금융 말뭉치는 대량의 웹 말뭉치를 대상으로 검색 쿼리로 활




〈그림 2〉 금융 말뭉치 증강 과정




〈그림 3〉 학습 loss(좌) / Learning rate(우)

용되며, 이렇게 검색된 웹 문서들은 학습용 금융 특화 말뭉치에 추가되었다. 데이터 증강 과정은 <그림 2>에 표현되어 있다. 최종적으로 KB-BERT의 학습에는 금융 상품 설명서 및 투자 리포트 등을 포함해 총 40GB의 금융 도메인 말뭉치가 사용되었다.

### 3.3. 학습 환경

KB-BERT의 학습은 NVIDIA V100 32GB GPU 장치 8개를 사용하여 총 20일 동안 수행되었다. 512 배치사이즈의 학습 데이터를 기준으로 <그림 3>(좌)와 같이 약 2.5M iteration에서 모델의 loss가 수렴되었다. <그림 3>(우)의 learning rate 스케줄과 같이 학습 과정에서 learning rate는 점진적으로 증가 후 감소되는 warm up이 진행되며 이는 linear decay(Devlin et al, 2019) 방식을 통해 수행되었다.

## 4. 금융 특화 평가 데이터셋

본 장에서는 금융 특화 사전학습 언어 모델을 평가하기 위해 사용되는 토픽 분류, 감성 분류, 질의 응답 세 개의 자연어 처리 데이터셋에 대하여 서술한다.

### 4.1. 토픽 분류

토픽 분류(Topic classification)는 주어진 텍스트를 사전 정의된 토픽 클래스 중 하나로 분류하는 자연어 태스크이며 이러한 작업은 다양한 자

〈표 3〉 범용 및 금융 특화 데이터 토픽 클래스 비교

| 데이터 | 토픽수 | 토픽 클래스 |
|---|---|---|
| KLUE YNAT | 7 | 정치, 경제, 사회, 문화, 세계, IT/과학, 스포츠 |
| 금융 뉴스 토픽 | 39 | 경제정책, 수출/입, 투자, 금융상품 등 |

연어 응용 시스템에서 가장 기본적인 기능으로 활용된다. 자연어 텍스트 기반의 오피니언 마이닝(Han & Kando, 2019), 관심 키워드 추출(Karamanolakis et al, 2019) 등의 하위 작업을 수행하기 전 토픽 별 문서 분류를 통해 분석 결과를 세분화하여 분석하는 것이 대표적인 사용 예이다. <표 3>에서와 같이 범용적인 용도의 토픽 분류 데이터셋으로 KLUE(Park et al, 2021) 벤치마크에 포함된 YNAT은 정치, 경제, 사회 등 7가지의 가장 기본적인 뉴스데이터 토픽만을 클래스로 포함하는 반면, 금융 지식을 평가하기 위해 본 연구에서 사용된 금융 특화 토픽 분류 데이터셋은 정책, 상품 등 금융 관련 기사를 더 자세히 구분하기 위한 39개의 토픽 클래스로 구성되어 있다. <표 4>는 이러한 금융 특화 토픽 분류 샘플 데이터를 보여준다.

### 4.2. 감성 분석

감성 분석(Sentiment analysis)(김유영, 송민, 2016; 송민채, 신경식, 2018)은 자연어 텍스트에 내포된 사람의 감성 상태를 분석 및 예측하는 자연어 태스크로, 고객 피드백 데이터를 활용한 상품 분석(Tchalakova et al, 2011), 챗봇 답변 생성을 위한 페르소나(Persona) 기반 응답 생성(Firdaus et al, 2021) 등 다양한 형태로 실서비스에 적용되는 자연어 태스크이다. 한국어 자연어 처리 분야에서 가장 흔히 사용되는 감성 분석 데이터로는 영화 리뷰 데이터 기반의 NSMC[3] (Naver sentiment movie corpus)가 존재하는데, 영화 리뷰 댓글 기반의 말뭉치가 데이터로 사용되었으며 분류 체계 또한 긍정/부정의 이진 분류 형태이기에 금융 등의 특수 도메인 데이터를 대

〈표 4〉 금융 특화 토픽 분류 데이터 예시

| 토픽 | 뉴스 텍스트 |
|---|---|
| 경제정책 | 은행 개인사업자 대출에 대한 예대율 규제 완화가 연말까지 연장된다. 금융위원회는 15 일 은행 개인사업자 대출 신규취급분에 적용하는 예대율 가중치를 기존 100%에서 85%로 인하하는 조치를 12 월 말까지 연장하는 내용의 은행업 감독규정 개정안을 규정변경예고했다. |
| 수출/입 | 세계 수출 시장에서 점유율 1 위를 차지한 우리나라 제품이 70 개에 가까운 것으로 나타났습니다. 무역협회 국제무역통상연구원이 오늘(7 일) 내놓은 '세계 수출 시장 1 위 품목으로 본 우리 수출 경쟁력 현황' 보고서에 따르면 우리나라 세계 1 위 품목 수는 지난 2019 년 기준 69 개로 전년보다 7 개 … |
| 투자 | 암호화폐 시장이 달아오르는 가운데 바이낸스코인(BNB)이 큰 주목을 받고 있다. 올해 들어 1600% 가까이 상승하면서 BNB 는 사람들을 열광시키고 있다. BNB 는 24 시간 기준으로 약 25% 상승한 후 12 일(이하 현지시간) 시가총액 950 억 달러를 돌파했다. |
| 금융상품 | 삼성카드는 개인사업자에게 다양한 혜택을 제공하는 '삼성카드 BIZ LEADERS'를 출시했다고 15 일 밝혔다. 삼성카드 BIZ LEADERS 는 개인사업자들이 많이 사용하는 업종을 분석해 특화된 혜택을 제공하는 상품이다. 보험, 전기요금, 통신 업종에서 자동결제를 이용하면 결제금액의 10% 할인 ... |

3) https://github.com/e9t/nsmc

<표 5> 금융 특화 감성분류 데이터 예시

| 감정 | 텍스트 |
|---|---|
| 낙관 | 과거수익률 종목명 5 년 기준 연평균 수익률 당월 5 년 기준 연평균 수익률 전월 DHS, PEY, SPHD 모두 장기투자 했을때는 연 환산 수익률이 7 이상을 기록하고 있어서 매달 배당을 받는다는 점을 가만했을 때 캐시카우용 종목이라고 생각합니다. |
| | 외인이 던지는 건 미국 헷지펀드 등에서 고객 환매 요청을 대비해서 어쩔 수 없이 매도 하는거죠. 시장이 안정화 될 무렵 외인은 무조건 다시 삼전을 살겁니다. 그때 저렴하게 매수하기 위해 훈련 안된 개미들 공포에 손절매하게 할거고 가격 내려서 줍줍. 쓸 돈으로 투자한 개미들은 쫄아서 팔거고 … |
| 비난/반대 | 시장경제에 그냥 맡기면 될걸 억지로 규제하니 풍선효과로 이난리지. 불과 4 년전 미분양 나서 난리났던 수도권이 서울 몸이자 지금은 투기과열지구까지 됐자나. 다 규제 풀어버리면 더 내려간다니까 |
| | 리딩증권사로서 주식시장 전체에 대하여 그리고 주식투자자전체에 대하여 심각한 심리적 물질적 영향을 끼쳤다 따라서 계속 영업하려면 주식투자자 전체에 보상하던가 아니면 자진상폐해라 |

상으로 활용하기에 어려움이 있다. 본 논문에서 사용된 금융 특화 감성 분류 데이터셋은 영문 멀티 레이블 감성 분석 데이터셋인 GoEmotions (Demszky et al, 2020)의 감성체계를 기반으로 구축되었으며, 이에 더하여 금융 도메인 감성 분석에 필요한 세분화 클래스가 추가적으로 포함되었다. <표 5>는 금융 특화 감성 분석을 위한 대표적인 예시 중 하나로 금융 시장의 낙관 및 부정적 의견에 대한 감성 분석에 해당하는 샘플 데이터를 보여준다. 이러한 금융 특화 데이터들의 가장 큰 특징은 범용 감성 분석 데이터셋에서 사용되는 가장 일반적인 감성에 대한 긍정/부정 표현(좋다/싫다 등)이 존재하지 않더라도 텍스트내 금융 용어(매수/상장폐지 등)를 기반으로 감성을 판단할 수 있는지 평가할 수 있도록 설계되었다.

### 4.3. 질의 응답

질의 응답(Question answering)은 주어진 문서(Context)와 사용자 질문을 바탕으로 문서 내 정답을 찾아 제공하는 자연어 태스크이며, 검색 엔진을 위한 정답 제공, 챗봇 등 대화모델에서 매

〈표 6〉 금융 특화 질의 응답 데이터 예시

| 본문 |
|---|
| KB 국민은행이 새롭게 단장한 스타뱅크 출시를 기념해 모바일뱅킹 전용 서비스를 시행한다고 28 일 밝혔다. 이번 환전 서비스는 미국 달러, 유로 등을 포함해 총 17 개 통화로 하루에 최대 3000 달러(미화 기준)까지 바꿀 수 있다. 특히 미국달러, 유로, 일본 엔화의 경우 3000 달러(미화 기준)까지 조건 없이 90% 우대 환율을 제공한다. 이는 전 금융권 모바일 환전 서비스 중 최대 우대 한도다. 스타뱅크 앱에서 환전을 신청하고 20 영업일 내에 기업은행 지점을 통해 외화를 찾아야 하며, 미국 달러, 유로, 일본 엔화, 중국 위안은 전국 모든 지점에서, 그 외 통화는 고객이 지정한 지점에서 수령 가능하다. 미화 1 만 달러까지는 여러 번 환전하고 한 번에 은행에서 찾을 수 있다. 외화 수령기간 내에는 스타뱅크 앱을 통해 외화예금에 입금하거나 원화로 재환전도 가능하다. |
| 질문: 스타뱅크 3000 달러까지 환금시 우대 환율은? 답변: 90% |
| 질문: 스타뱅크 4000 달러까지 환금시 우대 환율은? 답변: 답변불가 |

우 중요한 기능으로 활용된다. 한국어 질의 응답 모델 학습을 위해 KorQuAD(Lim et al, 2019) 등의 오픈소스가 공개되었지만, 대부분 위키, 뉴스를 기반으로 한 범용적인 질의 응답 데이터셋에 해당한다. 이와 달리 금융 특화 질의 응답 데이터셋은 금융 상품 등에 대한 뉴스와 상품 설명서를 기반으로 생성된 샘플로 구성된다. 대부분의 샘플들은 올바른 정답을 도출하기 위해서 금융 용어에 대한 이해와 금액 등의 수치적인 이해력을 필요로 하는 형태로 생성되어 언어모델의 금융 도메인 능력을 평가하는데 적합하다. <표 6>는 금융 특화 질의 응답 데이터셋에 포함된 일부 샘플을 보여준다.

## 5. 성능평가

본 장에서는 토픽 분류, 감성 분석, 질의 응답 세 개의 자연어 처리 태스크에 대하여, 범용 데이터셋과 금융 특화 데이터셋에서 KB-BERT의 성능을 분석한다.

### 5.1. 범용 데이터셋 평가

범용 데이터셋에서의 성능 평가를 위해 NSMC (감성 분석), KLUE-YNAT(토픽 분류), KorQuAD (질의 응답) 세 개의 오픈소스 데이터셋이 사용되었다. NSMC는 긍정/부정 이진 분류에 대한 정확도(ACC), KLUE-YNAT는 토픽 클래스 분류에 대한 F1, KorQuAD는 추출된 정답 텍스트 span과 실제 정답 텍스트 간의 F1 점수로 평가된다. <표 7>은 범용 데이터셋에서의 KoELECTRA-v3, KLUE-RoBERTa, 그리고 KB-BERT 사전학습 언어 모델들의 성능 평가 결과를 보여준다. 모든 데이터셋에서 KLUE-RoBERTa가 KoELECTRA-v3와 비교하였을 때 더 좋은 성능을 보인다. KB-BERT는 토픽 분류, 질의 응답 태스크에서 KLUE-RoBERTa와 비교했을 때 크지 않지만 더 좋은 성능을 보인다. 이와 같이, KB-BERT는 금융 특화 목적으로 학습되었지만 일반적인 언어 능력 측면에서 비교 대상 모델들에 뒤지지 않는 성능을 갖고 있음을 확인할 수 있다.

〈표 7〉 범용 데이터셋 성능평가

| 모델명 | NSMC (ACC) | KLUE-YNAT (F1) | KorQuAD v1 (F1) |
|---|---|---|---|
| KoELECTRA-v3 | 90.52 | 83.40 | 93.09 |
| KLUE-RoBERTa | **90.75** | 84.28 | 94.45 |
| KB-BERT | 90.72 | **84.52** | **94.66** |

〈표 8〉 금융 특화 데이터셋 성능평가

| 모델명 | F-sentiment (F1) | F-news (F1) | F-QA (F1) |
|---|---|---|---|
| KoELECTRA-v3 | 43.96 | 58.30 | 71.72 |
| KLUE-RoBERTa | 46.19 | 61.71 | 71.08 |
| KB-BERT | **47.86** | **64.10** | **72.94** |

〈표 9〉 범용 및 금융 특화 데이터 평균 성능

| 모델명 | 범용 데이터셋 | 금융 특화 데이터셋 |
|---|---|---|
| KoELECTRA-v3 | 89.00 | 57.99 |
| KLUE-RoBERTa | 89.82 | 59.66 |
| KB-BERT | **89.96** | **61.63** |

### 5.2. 금융 데이터셋 평가

<표 8>은 금융 특화 데이터셋에서의 사전학습 언어모델들의 성능 평가 결과를 보여준다. 표와 같이, KB-BERT는 모든 금융 특화 데이터셋에서 KoELECTRA-v3, KLUE-RoBERTa와 비교하여 높은 성능을 보이는 것을 확인할 수 있다. 특히, KB-BERT는 모든 태스크에서 KLUE-RoBERTa와 비교하여 2% 포인트에 가까운 성능 향상을 보인다. <표 9>는 범용 데이터셋과 금융 데이터셋에서의 각 언어모델 성능 평가 평균치로, KB-BERT의 평균 성능이 범용 도메인 대비 금융 도메인에서 1 % 포인트 이상 더 큰 폭으로 상승한 것을 볼 수 있다. 이를 바탕으로 금융 특화 사전학습 언어모델의 사용이 동일한 금융 도메인에서의 자연어 태스크 성능 향상에 큰 도움을 주는 것을 확인할 수 있다.

## 6. 응용

본 연구의 결과물인 금융 특화 사전학습 언어모델 KB-BERT는 다양한 자연어 처리 기반 서비스에 활용되고 있으며 다음과 같은 대표적인 응용 사례가 있다. 1) 지속적으로 수집 및 생산되는 금융 문서를 분석, 활용하기 위한 형태소 분석, 문서 분류, 감성 분석, 키워드 추출 등의 딥러닝 기반 자연어 분석을 수행하는 파이프라인을 구축하고 상품 및 고객 분석 등 다양한 분석 업무에 활용된다. 2) 경제 뉴스, SNS, 투자 및 경제 리포트로부터 중요한 이벤트를 추출하고 파악하기 위한 업무에 활용된다. 3) 직원 및 고객의 정보 탐색에 필요한 딥러닝 기반 문서 검색 및 질의응답 시스템 구축에 활용된다. 위 시스템들은 과거 수작업 규칙이나 ML 기반에서 운용되다가 사전학습 언어모델을 사용한 고도화를 통해 획기적인 성능 향상을 이루었다. 이 밖에도 KB-BERT는 다양한 기술 개발이 수반되는 인공지능 금융 비서 등 더욱 복합적인 서비스 개발에 필요한 기반기술로 활용될 예정되고 있다. 이와 같이, 금융 특화 언어모델은 다양한 금융 분야 응용 태스크(Downstream task)의 성능 향상에 크게 기여하고 있어 지속적인 발전을 위한 연구개발이 필요하다.

## 7. 결론

본 논문에서는 금융 특화 사전학습 언어모델을 구축하기 위한 과정 및 금융 특화 언어 모델을 활용하여 금융 도메인에 필요한 다양한 자연어 처리 모델의 상당한 성능향상을 이끌어 낼 수 있음을 보였다. 사전학습 언어모델을 만드는데 있어서 학습에 사용되는 말뭉치와 실제 응용 도메인을 매칭하는 비교적 단순한 작업이 의료, 법

률 등 다양한 도메인에서 언어 모델의 성능을 끌어올리는데 매우 중요하다는 기존 연구가 금융 도메인에서 또한 성립한다는 것을 확인하였으며, 향후 도메인 특화 말뭉치의 추가적인 수집 및 개선을 통한 언어모델의 지속적 성능 향상 가능성을 보였다. 본 논문의 성능평가에서 다뤄진 토픽 분류, 감성 분석, 질의 응답은 금융 도메인에서 활용 가능한 수많은 자연어 처리 태스크의 일부로 금융 특화 언어모델이 갖는 추가적인 태스크에서의 가능성 또한 향후 연구를 통해 확인해야 할 부분이다. 또한, 본 논문에서 다뤄진 Masked language modeling 기반의 NLU 사전학습 언어모델 외에 Autoregressive 기반 NLG 모델 또한 고객 상담을 위한 대화 모델 등 다양한 금융 서비스 구축을 위해 필요한 연구 방향의 하나로 향후 금융 특화 언어모델 학습의 대상으로 고려해볼 수 있다.

## 참고문헌(References)


**[국내 문헌]**

김유영, 송민. (2016). 영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축. 지능정보연구, 22(3), 71-89.

송민채, 신경식. (2018). 임베딩과 어텐션 매커니즘에 기반한 LSTM을 이용한 감성분석. *2018* 한국지능정보시스템학회 춘계학술대회 논문집, 107-108.

유소연, 임규건. (2021). 텍스트 마이닝과 의미네트워크 분석을 활용한 뉴스 의제 분석: 코로나 19 관련 감정을 중심으로. 지능정보연구, 27(1), 47-64.

**[국외 문헌]**

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17),* 6000–6010.

Tchalakova, M., Gerdemann, D., & Meurers, D. (2011). Automatic Sentiment Classification of Product Reviews Using Maximal Phrases Based Analysis. *In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011),* 111–117.

Devlin, J., Chang, M.-W. Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers), 4171–4186.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics,* Volume 36, Issue 4, 1234-1240.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long Papers), 2227–2237.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don’t Stop Pretraining: Adapt

Language Models to Domains and Tasks. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8342−8360.

Sachidananda, V., Kessler, J., & Lai, Y.-A. (2021). Efficient Domain Adaptation of Language Models via Adaptive Tokenization. *In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing,* 155−165.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions," *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4040−4054.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. L., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., & Sifre, L. (2021). Improving language models by retrieving from trillions of tokens. *arXiv*.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics,* 9:176−194.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Pack, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J., & Cho, K. (2021). KLUE: Korean Language Understanding Evaluation. *arXiv*.

Park, J. (2020). KoELECTRA: Pretrained ELECTRA Model for Korean. *GitHub repository*. https://github.com/monologg/KoELECTRA.

Lim, S., Kim, M., & Lee, J. (2019). KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension. *arXiv*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. *In Findings of the Association for Computational Linguistics: EMNLP,* 2898−2904.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., (2019). Language Models are Unsupervised Multitask Learners. *arXiv*.

Karamanolakis, G., Hsu, D., & Gravano, L. (2019). Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 4611-4621.

Yao, X., Zheng, Y., Yang, X., & Yang, Z. (2021). NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework. *arXiv*.

Han, W. -B., & Kando, N. (2019). Opinion Mining with Deep Contextualized Embeddings. *In Proceedings of the 2019 Conference of the North American Chapter of the Association*

*for Computational Linguistics: Student Research Workshop,* 35–42.

Firdaus, M., Jain, U., Ekbal, A., & Bhattacharyy, P. (2021). SEPRG: Sentiment aware Emotion controlled Personalized Response Generation. *In Proceedings of the 14th International Conference on Natural Language Generation,* 353–363.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3615–3620.

Yin, P., Neubig, G., Yih, W., & Riedel, S. (2020). TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8413–8426.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2021). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv*.

Abstract

# KB-BERT: Training and Application of Korean Pre-trained Language Model in Financial Domain

Donggyu Kim* · Dongwook Lee* · Jangwon Park* · Sungwoo Oh*
Sungjun Kwon* · Inyong Lee* · Dongwon Choi**

Recently, it is a de-facto approach to utilize a pre-trained language model(PLM) to achieve the state-of-the-art performance for various natural language tasks(called downstream tasks) such as sentiment analysis and question answering. However, similar to any other machine learning method, PLM tends to depend on the data distribution seen during the training phase and shows worse performance on the unseen (Out-of-Distribution) domain. Due to the aforementioned reason, there have been many efforts to develop domain-specified PLM for various fields such as medical and legal industries. In this paper, we discuss the training of a finance domain-specified PLM for the Korean language and its applications. Our finance domain-specified PLM, KB-BERT, is trained on a carefully curated financial corpus that includes domain-specific documents such as financial reports. We provide extensive performance evaluation results on three natural language tasks, topic classification, sentiment analysis, and question answering. Compared to the state-of-the-art Korean PLM models such as KoELECTRA and KLUE-RoBERTa, KB-BERT shows comparable performance on general datasets based on common corpora like Wikipedia and news articles. Moreover, KB-BERT outperforms compared models on finance domain datasets that require finance-specific knowledge to solve given problems.

**Key Words** : Natural language processing, Finance, Deep learning, BERT, PLM

Received : June 16, 2022 Revised : June 21, 2022 Accepted : June 21, 2022
Corresponding Author : Dongwon Choi

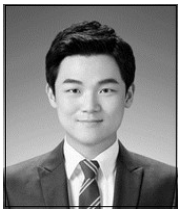
* Financial AI Center, Tech Group, KB Kookmin Bank
** Corresponding author: Dongwon Choi
Financial AI Center, Tech Group, KB Kookmin Bank
13, Uisadang-daero, Yeongdeungpo-gu, Seoul 07237, Korea
Tel: +82-2-2073-6529, Fax: +82-502-306-5403, E-mail: dwchoi@kbfg.com

# 저 자 소 개

**김 동 규**
현재 KB국민은행 금융AI센터에 재직 중이다. 세종대학교에서 컴퓨터공학 석사 학위를 취득하였다. 주요 관심분야는 자연어 처리, 정보검색 등이다.

**이 동 욱**
현재 KB국민은행 금융AI센터에 재직 중이다. 성균관대학교에서 컴퓨터공학을 전공했으며, 주요 관심분야는 자연어 처리, 대화형 인공지능 등이다.

**박 장 원**
현재 KB국민은행 금융AI센터에 재직 중이다. 연세대학교에서 경영학과 컴퓨터과학을 전공했으며, 주요 관심분야는 자연어 처리, 언어모델이다.



**오 성 우**
현재 KB국민은행 금융AI센터에 재직 중이다. 연세대학교 정보대학원에서 머신러닝, 딥러닝을 전공하여 정보시스템학 석사 학위를 취득하였다. 주요 관심분야는 자연어 처리, 딥러닝, 메타러닝, 동형암호 등이다.

**권 성 준**

현재 KB국민은행 금융AI센터에 재직 중이다. 한국공학대학교에서 경영학과 컴퓨터공학을 전공했으며, 주요 관심분야는 자연어처리, 문서분류, 감성분류 등이다.

**이 인 용**

현재 KB국민은행 금융AI센터에 재직 중이다. 연세대학교 정보산업공학 전공 후, 서울종합과학대학원에서 빅데이터MBA 학위를 취득하고 스위스 경영대 AI 빅데이터 박사 과정에 재학 중이다. 주요 관심분야는 자연어 처리, Learning Analytics, 추천 시스템 등이다.