KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용

김동규* 외

2022, 지능정보연구, VOL.28, NO.2

*KB국민은행

발제자: 조정효

개요

- 이 논문은 금융 분야에 특화된 한국어 사전학습 언어모델인 KB-BERT를 소개
- 대규모 금융 텍스트 데이터(금융 특화 말뭉치)를 기반으로 학습되었으며, 이를 통해 금융 관련 언어 처리 작업의 성능을 향상
- 또한, 다양한 언어 처리 작업에서의 KB-BERT의 성능을 평가하고, 그 결과를 기존 모델(KoELECTRA, KLUE-RoBERTa 등 state-of-the-art 한국어 사전학습 모델)과 비교 분석

관련연구-BERT

- BERT 사전학습 언어모델은 Transformer 뉴럴 네트워크 구조를 활용한 첫 번 째 언어모델
- Self-Attention 구조를 통해 자연어 텍스트로부터 더 강력한 contextual representation을 학습
- Masked language modeling → 비지도 방식으로 언어 지식 학습
- 사전 학습 언어모델에 목표 태스크의 학습 데이터를 이용해 미세조정 (Fine-tuning)으로 불리는 추가 학습을 거침

관련연구-도메인 특화 학습

- 목표 도메인 말뭉치를 수집 및 활용하여 From-scratch 방식으로 언어 모델을 학습
- **기존 학습된 범용 목적 언어모델을 기반으로 도메인 적응 기법을 활용**
  → DAPT(Gururangan et al, 2020)는 소규모의 도메인 특화 말뭉치를 범용 모델에 추가적으로 학습(post-training)하는 방법을 제안
- 범용 언어모델 기반의 Post-training 방법들은 앞서 설명된 From-scratch 방식과 비교해 성능상 뒤쳐지지만, 학습 시간 및 비용 등 효율성 측면 에서 장점이 있음

금융 특화 사전학습 언어모델 – KB-BERT

〈표 1〉 모델 하이퍼파라미터

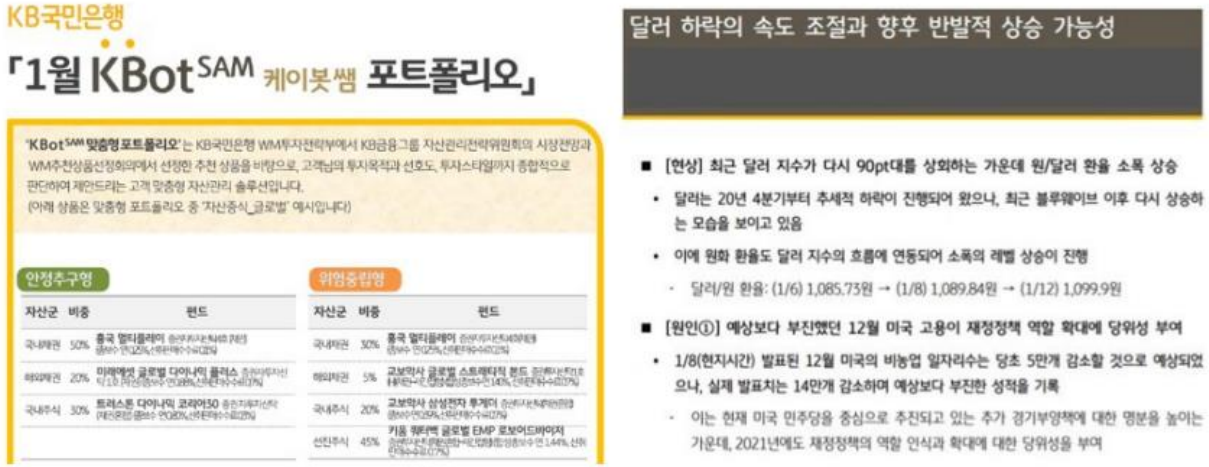
이름	Vocab	Word embedding	Layer	Hidden size	Self-attention heads
크기	35,000	786	12	786	12

〈표 2〉 학습 말뭉치 크기

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

금융 특화 사전학습 언어모델 – 학습 말뭉치

- 기본적인 위키, 뉴스, 웹 문서를 포함하며 추가적으로 금융 관련 문서 포함
- 금융 상품 설명서 및 투자 리포트 문서 (약 9GB)
- 총 용량의 약 40%에 해당하는 40GB는 경제 관련 뉴 스, 금융 관련 문서로 구성

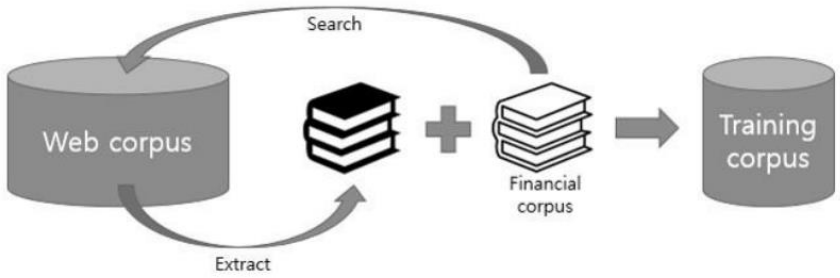


〈그림 1〉 금융 문서 예시

금융 특화 사전학습 언어모델 – 학습 말뭉치

- 전처리
  - 스팸 텍스트 분류 모델
  - 해시(MinHashLSH) 기반 문서 중복 제거

- 증강
  - 문서 검색 기반의 말뭉 치 증강 과정 수행
  - 초기 금융 말뭉치는 대량의 웹 말뭉치를 대상으로 검색 쿼리로 활용되며, 이렇게 검색된 웹 문서들은 학습용 금융 특화 말뭉치에 추가 → 총 40GB



〈그림 2〉 금융 말뭉치 증강 과정

금융 특화 평가 데이터셋

- 금융 특화 사전학습 언어 모델을 평가하기 위해 사용되는
  - 토픽 분류
  - 감성 분류
  - 질의 응답

금융 특화 평가 데이터셋 – 토픽 분류

- 본 연구에서 사용된 금융 특화 토픽 분류 데이터 셋은 정책, 상품 등 금융 관련 기사를 더 자세히 구분하기 위한 39개의 토픽 클래스로 구성

〈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

금융 특화 평가 데이터셋 – 토픽 분류

〈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 …

금융 특화 평가 데이터셋 – 감성 분석

- 가장 흔히 사용되는 감성 분석 데이터로는 영화 리뷰 데이터 기반의 NSMC (Naver sentiment movie corpus)가 존재

→금융 등의 특수 도메인 데이터를 대상으로 활용하기에 어려움이 있음

→ 본 논문에서 사용된 금융 특화 감성 분류 데이터셋은 **영문 멀티 레이블 감성 분석 데이터셋인 GoEmotions (Demszky et al, 2020)의 감성체계**를 기반으로 구축

→ 이에 더하여 금융 도메인 감성 분석 에 필요한 세분화 클래스가 추가적으로 포함

금융 특화 평가 데이터셋 – 감성 분석

<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 묶이자 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

금융 특화 평가 데이터셋 – 질의 응답

- 금융 특화 질의 응답 데이 터셋은 금융 상품 등에 대한 뉴스와 상품 설명서를 기반으로 생성된 샘플로 구성

〈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

성능 평가 – 범용 데이터셋

- NSMC (감성 분석), KLUE-YNAT(토픽 분류), KorQuAD (질의 응답) 세 개의 오픈 소스 데이터셋이 사용

〈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**

성능 평가 – 금융 데이터셋

- KB-BERT는 모든 금융 특화 데이터셋에서 KoELECTRA-v3, KLUE-RoBERTa와 비교하여 높은 성능을 보임!

〈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**