

# KF-DeBERTa: 금융 도메인 특화 사전학습 언어모델

전은광\*, 김정대\*, 송민상\*\*, 유주현\*\*

\*카카오뱅크, \*\*에프엔가이드 시팀

{weezy.j, jd.kim}@kakaobank.com, {mssong, onsemiro7988}@fnguide.com

## KF-DeBERTa: Financial Domain-specific Pre-trained Language Model

Eunkwang Jeon\*, Jungdae Kim\*, Minsang Song\*\*, Joohyun Ryu\*\*

\*Kakaobank, \*\*FnGuide AI Team

### 요약

본 논문에서는 금융 도메인 특화 사전학습 언어모델인 KF-DeBERTa(Korean Finance DeBERTa)를 제안한다. KF-DeBERTa는 대규모의 금융 말뭉치를 기반으로 학습하였으며, Transformer 아키텍처와 DeBERTa의 특징을 기반으로 구성되었다. 범용 및 금융 도메인에 대한 평가에서 KF-DeBERTa는 기존 언어모델들에 비해 상당히 높은 성능을 보였다. 특히, 금융 도메인에서의 성능은 매우 두드러졌으며, 범용 도메인에서도 다른 모델들을 상회하는 성능을 나타냈다. KF-DeBERTa는 모델 크기 대비 높은 성능 효율성을 보여주었고, 앞으로 금융 도메인에서의 활용도가 기대된다.

주제어: 자연어처리, 금융 언어모델, 사전학습 언어모델

### 1. 서론

2018년 이후로 GPT[1], BERT[2], RoBERTa[3], ELECTRA[4], T5[5], DeBERTa[6] 등의 대규모 사전학습 언어모델(PLM: Pre-trained Language Model)이 등장하였다. 이러한 사전학습 언어모델들은 작업별 레이블을 사용하여 미세 조정되었고, 많은 Downstream 자연어처리 작업에서 최고성능 기록을 세웠다. 자기 지도학습 방식의 사전학습 목표와 방대한 모델 파라미터 덕분에 대규모 사전학습 언어모델은 대량의 레이블이 없는 말뭉치 데이터로부터 효과적으로 지식을 추출할 수 있다. 이렇게 추출된 지식은 방대한 파라미터에 저장되고, 특정 작업에 대해 미세조정을 거치게 된다. 이렇게 방대한 파라미터에 암시적으로 인코딩된 풍부한 지식은 다양한 하위 작업에 이점을 제공할 수 있으며, 이는 실험적 검증과 경험적 분석을 통해 널리 입증되었다. AI 커뮤니티에서는 사전학습 언어모델을 하위 작업의 백본으로 채택하고, 처음부터 모델을 학습하는 것보다는 사전학습 언어모델을 활용하는 것이 일반적이다.

그러나 사전학습 언어모델 연구의 대부분은 영어를 기반으로 이루어지고 있다. 하지만 언어는 그 자체로 각 나라의 문화와 특성을 반영하고 있으며, 언어모델의 학습과정에서도 중요한 역할을 한다. 한글이 포함된 다국어 언어모델도 존재하지만 한국어는 그 구조와 문법이 다른 언어와 비교하여 많이 다르기 때문에, 다국어 언어모델을 그대로 적용하기에는 한계가 있다. 이에 따라 한국어 특성을 반영한 언어모델의 필요성이 대두되고 있다.

또한, 특정 도메인에 대한 언어모델은 그 도메인의 전문성을 반영할 수 있어, 보다 정확한 결과를 도출할 수 있다. 금융 도메인은 그 자체로 복잡한 전문 용어와 개념, 그리고 빠르게 변화하는 시장 상황을 반영해야 하므로, 일반적인 언어모델로는 이를 충분히 반영하기 어렵다. 따라서, 금융 도메인에 특화된 언어모델의 필요성이

점점 더 커지고 있다.

본 논문에서는 한국어 금융 도메인에 특화된 사전학습 언어모델을 제안한다. 이 모델은 한국어의 특성과 금융 도메인의 복잡성을 모두 반영하여, 보다 정확하고 효율적인 결과를 제공할 것으로 기대된다. 이를 통해 한국어 금융 도메인의 자연어 처리 성능을 향상시키고, 이를 바탕으로 다양한 금융 서비스와 응용 분야에 활용할 수 있을 것으로 전망된다.

### 2. 관련 연구

#### 2.1 사전학습 언어모델

BERT와 같은 사전학습 언어모델은 자연어처리 시스템의 표준이 되었다. BERT는 사전학습 단계에서 토큰을 [MASK]로 무작위로 교체하고 문맥 표현을 기반으로 손상된 단어를 복구하는 MLM(Masked Language Modeling) 작업을 통해 대규모 Transformer[7]를 사전학습 한다. 또한 NSP(Next Sentence Prediction)이라는 목표를 채택했는데, 이는 기계독해와 같은 문장 수준 작업에 도움이 된다고 저자들은 주장한다. 이후 사전학습 언어모델을 하위 작업에 대해 미세조정을 수행하여 처음부터 모델을 학습하는 것보다 더 높은 성능을 보여주었다. 이러한 BERT의 성공 이후 수 많은 확장이 등장했다.

RoBERTa는 BERT에서 말뭉치, 배치 크기, 학습 스텝을 확장하고 동적 마스킹 사용 및 NSP를 제거하여 BERT를 확장하였다. ELECTRA에서는 MLM 작업의 학습 효율성을 개선하기 위해 RTD(Replaced Token Detection) 작업을 제안하여 MLM을 사용한 BERT와 동일한 연산량으로 더 높은 성능을 보여주었고 DeBERTa에서는 한 토큰의 Attention이 토큰 뿐만 아니라 위치에도 의존한다는 견해를 바탕으로 기존 Attention 매커니즘에 상대적 위치정보도 추가하여 토큰-토큰, 토큰-위치, 위치-토큰 Attention을 수행하는

Disentangled Attention을 제안하여 더 높은 성능을 보여준 연구이다.

## 2.2 도메인 특화 사전학습 언어모델

언어 모델의 사전학습은 일반적인 도메인의 말뭉치, 예를 들어 뉴스 기사나 위키백과 등을 이용한다. 그러나 특정 도메인의 대규모 말뭉치를 이용하여 언어 모델을 사전학습 하는 것이 하위 작업의 성능을 더욱 향상시킬 수 있음이 입증되었다. 이에 따라, 여러 도메인 특화 BERT 모델들이 학습되고 공개되었다.

FinBERT[8]는 사전학습 된 BERT에 금융 말뭉치를 추가 학습하여 금융 도메인 하위 작업에서의 성능 평가를 수행하여 금융 특화 언어모델의 가능성을 보여주었다. BioBERT[9]는 대규모의 생물의학 말뭉치를 이용하여 생물의학 도메인 특화 언어모델을 사전학습 하였다. 비슷하게, ClinicalBERT[10]는 BERT 모델을 병원 재입원 예측 작업을 위한 임상 기록에 적용하였고, [11]에서는 BERT를 임상 기록과 퇴원 요약에 적용하였다. SciBERT[12]는 과학 도메인 자연어처리 작업의 성능 향상을 위해 과학 출판물의 대규모 도메인 말뭉치를 이용하여 과학 도메인 특화 BERT 모델을 학습하였다.

또한 최근 초거대 언어모델에서도 도메인 특화 데이터를 통해 학습한 금융 도메인의 BloombergGPT[13], 과학 도메인의 Galactica[14], 의료 도메인의 MedPaLM[15]이 존재한다.

## 3. 금융 도메인 특화 사전학습 언어모델

본 연구에서 제안하는 금융 도메인 특화 사전학습 언어모델 KF-DeBERTa의 말뭉치 수집 및 전처리, 모델구조, 학습 구성요소에 대해 설명한다.

### 3.1 말뭉치 수집 및 전처리

언어모델 사전학습 단계는 대규모 말뭉치를 기반으로 수행한다. 사전학습 말뭉치는 언어모델의 성능을 결정하는 요소 중 하나로써, 높은 성능의 언어모델을 위해선 말뭉치와 규모와 품질이 중요하다.

본 연구에서 사용한 사전학습 말뭉치는 표 1과 같이 다양한 도메인을 포괄하는 범용 말뭉치와 금융 도메인에 포함되는 금융 말뭉치로 구성된다. 범용 말뭉치는 일반적인 한국어 사전학습 언어모델에서 사용되는 모두의 말뭉치[16], 위키피디아, 나무위키, 뉴스댓글[17], AIHUB 말뭉치[18] 등을 포함한다. 금융 말뭉치는 금융

도메인에 해당되는 금융 리포트와 금융 뉴스 말뭉치가 포함된다. 범용 말뭉치와 금융 말뭉치를 함께 사용하여 한국어의 다양한 표현과 문맥을 포괄하면서도 금융 도메인의 전문성을 확보하고자 노력하였다.

표 1. 사전학습 말뭉치

Domain	Dataset	Disk size
범용 도메인	모두의 말뭉치	14.1 GB
	위키피디아	1.6 GB
	나무 위키	4.8 GB
	뉴스	4.6 GB
	뉴스댓글	12.5 GB
	AIHUB 말뭉치	11.3 GB
금융 도메인	Common Crawl	35 GB
	금융 리포트	5.9 GB
	금융 뉴스	26.5 GB

데이터 수집 이후 데이터 전처리를 다음과 같이 수행하였다.

- KSS(Korean Sentence Splitter)<sup>1</sup> 및 Kiwi(Korean Intelligent Word Identifier)<sup>2</sup>를 통해 말뭉치를 문장단위로 분할
- HTML tag 및 잘못된 문자(U+200B, U+FEFF, 등) 제거
- 뉴스와 리포트에 포함된 출처 정보 제거
- 한글이 아닌 언어의 비중이 높은 문장 제거
- 숫자비중이 높은 문장 제거
- 10 음절 이하로 구성된 문장 제거

토큰화는 SentencePiece[19] 라이브러리를 사용하여 형태소 인식 서브워드(Morpheme-aware Subword)[20] 방식으로 토큰화를 수행하였다. Vocab size는 128,000 크기로 설정하였다.

### 3.2 모델 구조

본 연구에서 모델 구조는 Transformer 아키텍처를 기반으로 한다. 이후 DeBERTa에서 제안된 Disentangled Attention 개선사항을 활용한다. DeBERTa-v2를 baseline으로 설정하며 MLM작업 대신 RTD 작업을 수행하는 DeBERTa-v3[21]도 함께 실험을 진행한다. 사전학습을 수행 후 하위 작업에 대해 미세조정 하여 실험결과 표 2와 같이 실험 단계에서 RTD 작업을 수행하는 DeBERTa-v3가 일부 검증 작업에서 매우 낮은 성능이 확인되어 최종 모델 구조는 DeBERTa-v2로 결정하였다. 이는 RTD 작업을

<sup>1</sup> <https://github.com/hyunwoongko/kss>

<sup>2</sup> <https://github.com/bab2min/Kiwi>

사전학습 작업으로 수행하는 언어모델에서 특정 작업에 대한 성능이 저하되는 것으로 확인되었다.

표 2. KLUE benchmark 검증 데이터에 대한 성능결과

Model	KLUE-RE		KLUE-DP		KLUE-WoS	
	F1 <sup>mic</sup>	AUC	UAS	LAS	JGA	F1 <sup>S</sup>
DeBERTa-v2 (MLM)	69.70	75.07	94.05	87.97	50.21	92.59
DeBERTa-v3 (RTD)	62.77	61.97	92.01	87.52	40.68	90.91

### 3.3 학습 구성

최종 모델 구조는 표 3과 같으며 사전학습 과정에서 모델은 최대 512개 토큰의 시퀀스를 입력 받는다. MLM 작업에서 동적 마스크 전략(Dynamic Masking Strategy) 및 n-gram Masking을 사용한다.

표 3. KF - DeBERTa<sub>base</sub>의 Hyper-parameter

Hyper-parameter	KF - DeBERTa <sub>base</sub>
Number of Layers	12
Hidden size	768
Attention Heads	12
Attention Head size	64
Dropout	0.1
Warmup Steps	10k
Learning Rate	2e-4
Batch Size	1k
Weight Decay	0.01
Max Steps	1M
Adam $\epsilon$	1e-6
Adam $\beta_1$	0.9
Adam $\beta_2$	0.999
Gradient Clipping	1.0

## 4. 실험 및 결과

본 장에서는 사전학습 언어모델에 평가를 위한 데이터 및 금융 및 범용 도메인에 대한 평가 결과에 대해 서술한다.

### 4.1 데이터

KF-DeBERTa의 범용 및 금융 도메인 성능 평가를 위해 범용 및 금융 도메인에 해당하는 데이터셋을 구성한다. 범용 도메인 데이터셋은 공개되어 있는 다양한 작업

데이터셋을 사용하였으며 금융 도메인 데이터셋은 공개되어 있는 데이터가 없어 레이블링 작업 및 기계번역을 통해 데이터셋을 제작하였다.

**금융 도메인 데이터.** 금융 도메인 성능 평가를 위해 금융뉴스를 원문으로 하여 감성분류, 개체명 인식, 광고성 뉴스 분류 작업 데이터셋을 레이블링 하였다. 또한 기계번역을 사용하여 영문 금융 도메인 데이터셋인 FinancialPhraseBank[22], FiQA-SA[23], Headline[24] 한글로 번역하여 금융 도메인 평가 데이터로 활용하였다.

금융 도메인 데이터 리스트는 다음과 같다.

- Fin-Sentiment: 금융 도메인 감성분류
- Fin-NER: 금융 도메인 개체명인식
- Fin-Adnews: 금융 도메인 광고성 뉴스 분류
- KorFPB: FinancialPhraseBank 기계번역 데이터
- KorFiQA-SA: FiQA-SA 기계번역 데이터
- KorHeadline: Headline 기계번역 데이터

**범용 도메인 데이터.** 범용 도메인 성능 평가를 위해 KLUE Benchmark[25] 와 NSMC<sup>3</sup>, Paws-X[26], KorNLI[27], KorSTS[27], KorQuAD 1.0[28]을 사용하였다.

### 4.2 실험 환경

본 장에서는 실험을 위해 각 작업 학습 데이터셋으로 미세조정을 수행하고 테스트 데이터셋으로 평가를 진행하였다. KLUE의 경우 테스트셋 데이터가 공개되어있지 않으므로 검증 데이터셋을 통해 비교평가를 수행하였다.

공정한 평가를 위해 다음과 같이 제한된 하이퍼 파라미터 최적화만 수행하였다.

$$\begin{aligned} \text{batch size} &\in \{16,32\} \\ \text{learning rate} &\in \{1e^{-5}, 3e^{-5}, 5e^{-5}\} \\ \text{weight decay} &\in \{0, 0.01\} \\ \text{warmup proportion} &\in \{0, 0.1\} \end{aligned}$$

### 4.3 금융 도메인 평가 결과

표 4는 금융 도메인 데이터에 대한 사전학습 언어모델들의 결과이다. 표 1과 같이 KF-DeBERTa는 모든 금융 도메인 작업에서 가장 높은 성능을 보이는 것을 확인할 수 있으며 FiQA-SA를 제외한 평균 점수도 가장 높은 것을 확인할 수 있다.

KF-DeBERTa는 다음으로 성능이 좋은 KLUE-RoBERTa와 비교하여 평균 점수가 1% 이상 차이 나는 것을 확인할 수 있고 사전학습 단계에서 RTD 작업을 수행하는 KoELECTRA 와도 1% 이상 차이를 보여준다. 그리고 금융 개체명인식

<sup>3</sup> <https://github.com/e9t/nsmc>

표 4. 금융 도메인 성능 평가 결과

Model	Fin-Senti	Fin-NER	Fin-ADNews	KorFPB	KorFiQA-SA	KorHeadline	Avg (w/o FiQA-SA)
	Acc	F1 <sup>micro</sup>	Acc	Acc	MSE	F1 <sup>mean</sup>	
KoELECTRA <sub>BASE</sub>	90.56	89.81	96.98	92.36	0.0652	80.69	93.90
KLUE – RoBERTa <sub>BASE</sub>	91.21	90.31	96.34	90.97	0.0589	81.11	94.03
KF – DeBERTa <sub>BASE</sub>	<b>92.29</b>	<b>91.80</b>	<b>97.63</b>	<b>93.47</b>	<b>0.0553</b>	<b>82.12</b>	<b>95.27</b>

표 5. KLUE Benchmark Devset 성능 평가 결과. 굵은 글씨는 모든 모델에서 가장 높은 성능을 의미하고, 밑줄은 Base 크기의 모델 중에서 가장 높은 성능을 의미함.

Model	YNAT		STS		NLI		NER		RE		DP		MRC		WoS		Avg
	F1	RP	F1	ACC	F1 <sup>E</sup>	F1 <sup>C</sup>	F1 <sup>mic</sup>	AUC	UAS	LAS	EM	ROUGE	JGA	F1 <sup>S</sup>			
mBERT <sub>BASE</sub>	82.64	82.97	75.93	72.90	75.56	88.81	58.39	56.41	88.53	86.04	49.96	55.57	35.27	88.60	71.26		
XLM – R <sub>BASE</sub>	84.52	88.88	81.20	78.23	80.48	92.14	57.62	57.05	93.12	87.23	26.76	53.36	41.54	89.81	72.28		
XLM – R <sub>LARGE</sub>	87.30	93.08	87.17	86.40	82.18	93.20	58.75	63.53	92.87	87.82	35.23	66.55	42.44	89.88	76.17		
KR – BERT <sub>BASE</sub>	85.36	87.50	77.92	77.10	74.97	90.46	62.83	65.42	92.87	87.13	48.95	58.38	45.60	90.82	74.67		
KoELECTRA <sub>BASE</sub>	85.99	93.14	85.89	86.87	86.06	92.75	62.67	57.46	90.93	87.07	59.54	65.64	39.83	88.91	77.34		
KLUE – BERT <sub>BASE</sub>	86.95	91.01	83.44	79.87	83.71	91.17	65.58	68.11	93.07	87.25	62.42	68.15	46.72	91.59	78.50		
KLUE – RoBERTa <sub>SMALL</sub>	85.95	91.70	85.42	81.00	83.55	91.20	61.26	60.89	93.47	87.50	58.28	63.56	46.65	91.50	77.28		
KLUE – RoBERTa <sub>BASE</sub>	86.19	92.91	86.78	86.30	83.81	91.09	66.73	68.11	93.75	87.77	69.56	74.64	47.41	91.60	80.48		
KLUE – RoBERTa <sub>LARGE</sub>	85.88	93.20	86.13	<b>89.50</b>	84.54	91.45	<b>71.06</b>	73.33	93.84	87.93	<b>75.26</b>	<b>80.30</b>	49.39	92.19	82.43		
KF – DeBERTa <sub>BASE</sub>	<b>87.51</b>	<b>93.24</b>	<b>87.73</b>	<u>88.37</u>	<b>89.17</b>	<b>93.30</b>	<u>69.70</u>	<b>75.07</b>	<b>94.05</b>	<b>87.97</b>	<u>72.59</u>	<u>78.08</u>	<b>50.21</b>	<b>92.59</b>	<b>82.83</b>		

표 6. 범용 도메인 성능평가 결과

Model	NLMC	Paws	KorNLI	KorSTS	KorQuAD 1.0		Avg (w/o KorQuAD)
	Acc	ACC	Acc	spearman	EM	F1	
KoELECTRA <sub>BASE</sub>	90.63	84.45	82.24	85.53	84.83	93.45	85.71
KLUE – RoBERTa <sub>BASE</sub>	90.47	84.79	81.65	84.40	86.34	94.40	85.33
KF – DeBERTa <sub>BASE</sub>	<b>91.36</b>	<b>86.14</b>	<b>84.54</b>	<b>85.99</b>	<b>86.60</b>	<b>95.07</b>	<b>87.01</b>

작업에서는 KoELECTRA와 비교하여 약 2% 포인트 차이가 나는 것을 보여준다.

이와 같이 대량의 금융 말뭉치를 학습한 KF-DeBERTa는 금융 도메인 작업에서 기존의 범용 언어모델 보다 더 높은 성능을 보여주는 것을 확인할 수 있다.

#### 4.4 범용 도메인 평가 결과

표 5는 KLUE Benchmark에 대한 사전학습 언어모델들의 결과이다. 결과를 보면 KF-DeBERTa는 같은 크기의 언어 모델들과 비교하면 모든 작업에서 가장 높은 성능을 보여준다. 기존의 모델 중 KLUE-RoBERTa가 Base 크기의 언어 모델 중 가장 좋은 성능이 확인되었지만 KF-DeBERTa와 비교하여 평균 점수가 약 2% 포인트 이상 차이 나는 것을 확인할 수 있었다. Large 크기를 가진 언어모델과 비교하여도 NLI, RE, MRC 데이터셋을 제외하면 KF-DeBERTa

가 더 높은 성능을 보여주었으며 평균 점수도 KLUE-RoBERTa Large 모델 보다 0.4% 포인트 더 높은 것을 확인할 수 있었다.

표 6은 범용 도메인 데이터에 대한 결과로 모든 데이터셋에서 가장 높은 성능을 보여주었고 특히 Paws-X 데이터셋에서 기존 언어모델들 보다 높은 차이를 보여주었다. 평균 점수에서는 KoELECTRA와 비교하여 1.3% 포인트 더 높았고 KLUE-RoBERTa와 비교하여 약 1.7% 포인트 높았다.

#### 5. 결론

본 연구에서는 금융 도메인에 특화된 KF-DeBERTa를 제안하고 실험을 수행했다. 범용 말뭉치와 함께 금융 도메인의 성능을 확보하기 위해 금융 리포트, 금융 뉴스로부터 대량의 금융 말뭉치를 수집 및 정제하여 사전학습 데이터로 활용하였다. 이후 범용 도메인 및 금융 도메인에

대하여 철저한 성능 평가를 수행하였으며 그 결과 KF-DeBERTa는 기존의 언어모델보다 금융 도메인 및 범용 도메인 모두에서 가장 높은 성능을 확인할 수 있었다.

추가적으로 KF-DeBERTa는 각 도메인 평균 점수에서도 약 1% 포인트 이상의 점수차이를 보여주었고 같은 Base 사이즈 언어모델과 비교할때 성능 평가를 진행한 모든 데이터셋에서 기존 언어모델보다 높은 성능을 보여주었다. 뿐만 아니라 KLUE Benchmark에서는 Large 사이즈를 가진 XLM-R과 비교하여 모든 데이터셋에서 더 높은 성능을 보여주었고 KLUE-RoBERTa Large 사이즈와 비교할때 3개 데이터셋을 제외하고 5개 데이터셋 평가에서 더 높은 성능을 확인할 수 있었다.

모델의 파라미터 개수를 감안할때 KF-DeBERTa는 KLUE-RoBERTa Large와 비교하여 더 적은 파라미터 수를 가지고 있다. KF-DeBERTa는 더 적은 파라미터 개수로 더 높은 성능을 보여주어 KLUE-RoBERTa와 비교할 때 파라미터 대비 더욱 효율적인 모델이다.

향후 과제로는 금융 도메인 특화 말뭉치를 추가확보하여 지속적인 성능 향상을 위해 시도할 것이며 금융 도메인에 자연어처리 관련 작업에 기초가 되는 언어모델로 활용할 계획이다.

### 참고문헌

- [1] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
- [2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [3] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
- [4] Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).
- [5] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.
- [6] He, Pengcheng, et al. "Deberta: Decoding-enhanced bert with disentangled attention." *arXiv preprint arXiv:2006.03654* (2020).
- [7] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- [8] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." *arXiv preprint arXiv:1908.10063* (2019).
- [9] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.
- [10] Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." *arXiv preprint arXiv:1904.05342* (2019).
- [11] Alsentzer, Emily, et al. "Publicly available clinical BERT embeddings." *arXiv preprint arXiv:1904.03323* (2019).
- [12] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." *arXiv preprint arXiv:1903.10676* (2019).
- [13] Wu, Shijie, et al. "Bloomberggpt: A large language model for finance." *arXiv preprint arXiv:2303.17564* (2023).
- [14] Taylor, Ross, et al. "Galactica: A large language model for science." *arXiv preprint arXiv:2211.09085* (2022).
- [15] Singhal, Karan, et al. "Large language models encode clinical knowledge." *arXiv preprint arXiv:2212.13138* (2022).
- [16] National Institute of Korean Languages. NIKL CORPORA 2020 (v.1.0), 2020. URL <https://corpus.korean.go.kr>.
- [17] Lee Junbum, KcBERT: Korean Comments BERT, Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology, pp. 437-440, 2020.
- [18] National Information Society Agency. NIA AI-HUB 대규모 웹데이터 기반 한국어 말뭉치 데이터, 2022.
- [19] Kudo, Taku, and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." *EMNLP 2018* (2018): 66.
- [20] Park, Kyubyong, et al. "An empirical study of tokenization strategies for various Korean NLP tasks." *arXiv preprint arXiv:2010.02534* (2020).
- [21] He, Pengcheng, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing." *arXiv preprint arXiv:2111.09543* (2021).
- [22] Malo, Pekka, et al. "Good debt or bad debt: Detecting semantic orientations in economic texts." *Journal of the Association for Information Science and Technology* 65.4 (2014): 782-796.
- [23] aia, Macedo & Handschuh, Siegfried & Freitas, Andre & Davis, Brian & McDermott, Ross & Zarrouk, Manel & Balahur, Alexandra. (2018). WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. WWW '18: Companion Proceedings of the The Web Conference 2018. 1941-1942. 10.1145/3184558.3192301.
- [24] Sinha, A., & Khandait, T. (2021, April). Impact of News on the Commodity Market: Dataset and Results. In Future of Information and

- Communication Conference (pp. 589-601). Springer, Cham.
- [25] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." *arXiv preprint arXiv:2105.09680* (2021).
- [26] Yang, Yinfei, et al. "PAWS-X: A cross-lingual adversarial dataset for paraphrase identification." *arXiv preprint arXiv:1908.11828* (2019).
- [27] Ham, Jiyeon, et al. "KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding." *arXiv preprint arXiv:2004.03289* (2020).
- [28] 임승영, 김명지, and 이주열. "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋." *한국정보과학회 학술발표논문집* (2018): 539-541.