J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

데이터 세트별 Post-Training을 통한 언어 모델 최적화 연구 : 금융 감성 분석을 중심으로☆

Optimizing Language Models through Dataset-Specific Post-Training: A Focus on Financial Sentiment Analysis

정 희 도[1] 김 재 헌[1] 장 백 철[1*]
Hui Do Jung Jae Heon Kim Beakcheol Jang

요 약

본 연구는 금융 분야에서 중요한 증감 정보를 효과적으로 이해하고 감성을 정확하게 분류하기 위한 언어 모델의 학습 방법론을 탐구한다. 연구의 핵심 목표는 언어 모델이 금융과 관련된 증감 표현을 잘 이해할 수 있게 하기 위한 적절한 데이터 세트를 찾는 것이다. 이를 위해, Wall Street Journal에서 수집한 금융 뉴스 문장 중 증감 관련 단어를 포함하는 문장을 선별했고, 이와 함께 적절한 프롬프트를 사용해 GPT-3.5-turbo-1106으로 생성한 문장을 각각 post-training에 사용했다. Post-training에 사용한 데이터 세트가 언어 모델의 학습에 어떠한 영향을 미치는지 금융 감성 분석 벤치마크 데이터 세트인 Financial PhraseBank를 통해 성능을 비교하며 분석했으며, 그 결과 금융 분야에 특화된 언어 모델인 FinBERT를 추가 학습한 결과가 일반적인 도메인에서 사전 학습된 모델인 BERT를 추가 학습한 것보다 더 높은 성능을 보였다. 또 금융 뉴스로 post-training을 진행한 것이 생성한 문장을 post-training을 진행한 것에 비해 전반적으로 성능이 높음을 보였으나, 일반화가 더욱 요구되는 환경에서는 생성된 문장으로 추가 학습한 모델이 더 높은 성능을 보였다. 이러한 결과는 개선하고자 하는 부분의 도메인이 사용하고자 하는 언어 모델과의 도메인과 일치해야 한다는 것과 적절한 데이터 세트의 선택이 언어 모델의 이해도 및 예측 성능 향상에 중요함을 시사한다. 연구 결과는 특히 금융 분야에서 감성 분석과 관련된 과제를 수행할 때 언어 모델의 성능을 최적화하기 위한 방법론을 제시하며, 향후 금융 분야에서의 더욱 정교한 언어 이해 및 감성 분석을 위한 연구 방향을 제시한다. 이러한 연구는 금융 분야 뿐만 아니라 다른 도메인에서의 언어 모델 학습에도 의미 있는 통찰을 제공할 수 있다.

☞ 주제어 : BERT, FinBERT, 금융 감성 분석, post-training, 사전 학습 데이터 세트

ABSTRACT

This research investigates training methods for large language models to accurately identify sentiments and comprehend information about increasing and decreasing fluctuations in the financial domain. The main goal is to identify suitable datasets that enable these models to effectively understand expressions related to financial increases and decreases. For this purpose, we selected sentences from Wall Street Journal that included relevant financial terms and sentences generated by GPT-3.5-turbo-1106 for post-training. We assessed the impact of these datasets on language model performance using Financial PhraseBank, a benchmark dataset for financial sentiment analysis. Our findings demonstrate that post-training FinBERT, a model specialized in finance, outperformed the similarly post-trained BERT, a general domain model. Moreover, post-training with actual financial news proved to be more effective than using generated sentences, though in scenarios requiring higher generalization, models trained on generated sentences performed better. This suggests that aligning the model's domain with the domain of the area intended for improvement and choosing the right dataset are crucial for enhancing a language model's understanding and sentiment prediction accuracy. These results offer a methodology for optimizing language model performance in financial sentiment analysis tasks and suggest future research directions for more nuanced language understanding and sentiment analysis in finance. This research provides valuable insights not only for the financial sector but also for language model training across various domains.

☞ Keyword : BERT, FinBERT, Financial Sentiment Analysis, Post-training, Pre-training Dataset

1 Yonsei Graduate School of Information Seoul, South Korea
* Corresponding author: (bjang@yonsei.ac.kr)
[Received 26 December 2023, Reviewed 17 January 2024, Accepted 22 January 2024]

☆ This work was supported by the National Research Foundation of Korea Fund of RS-2023-00273751

1. 서 론

자연어처리 연구자들은 다양한 텍스트 데이터를 활용하여 금융 분야에 새로운 시각을 제공해왔다. 이러한 노력은 주가 예측[1, 2], 위험 예측[3], 금융 엔티티 추출[4]과 같은 응용 분야에서 널리 적용되었다. 이 분야의 연구는 규칙 기반 방법에서 딥러닝에 이르기까지 다양한 방법론을 통해 이루어졌으며, 금융 텍스트 분석은 꾸준히 주목받아 왔다.

금융 텍스트 분석은 일반적인 영역의 텍스트 분석과 구별되는 특징을 가진다. 첫 번째로, 실시간으로 업데이트되는 뉴스, 보고서, 소셜 미디어 게시글 같은 비정형 데이터가 금융 분야에서 중요한 정보를 제공한다. 두 번째로, 감성 점수와 같이 투자 정보와 밀접하게 관련 있는 지표들이 중요하게 다루어진다. 마지막으로, '상승' 또는 '하락' 과 같은 용어들이 금융 텍스트에서 더 자주 나타나며, 그 의미도 더욱 중요하다.

이러한 특성을 언어 모델이 잘 학습할 수 있도록, 연구자들은 주로 사전 학습된 언어 모델(PLM)을 사용한다. PLM들은 전이 학습을 통해 특정 도메인이나 과제에 맞춰 추가로 학습되며, 이 과정에서 모델은 해당 도메인이나 과제의 텍스트에서 중요한 의미 관계를 배우게 된다[5].

과거의 연구에서 이미 일반 도메인에서 학습된 PLM들이 다양한 과제에서 효과적임을 입증했다[6]. 이에 착안하여, 경제 뉴스[7], 애널리스트 보고서[8, 9], 10-K 보고서[4, 9, 10], 웹 데이터[10], 레딧[10] 등을 활용하여 금융 분야에 특화된 PLM을 개발하는 연구가 진행되었다. 이러한 연구들은 금융 텍스트의 독특한 특징을 효과적으로 포착하고 이를 언어 모델이 이해할 수 있도록 하는 방법론을 제안하며, 금융 관련 과제에서의 성능 향상을 가져왔다.

그러나 완벽한 성능 달성은 여전히 도전적이었다. Araci는 FinBERT가 금융 뉴스의 감성 분석에서 BERT[11]보다 높은 성능을 달성하긴 했지만, 특정 분류에서 실패하는 예시를 보여주며 FinBERT가 직접적인 지시 단어가 없으면 수의 증감과 관련된 개념을 제대로 이해하지 못한다는 한계를 지적했다[7]. 본 연구는, 금융 분야에서 언어 모델이 잘 수행하지 못한 부분을 파악할 수 있다면, 그에 맞는 추가적인 사전 학습인 post-training을 수행함으로써 단점을 보완하고 성능도 향상시키는 것을 목표로 한다. 특히, FinBERT가 이해하기 어려웠던 증감 관련 용어들의 이해에 초점을 맞춘다.

본 연구에서는 post-training에 사용하는 데이터 세트로 금융 뉴스와 함께 문장 생성 능력이 뛰어난 GPT-3.5 turbo-1106[12] 를 사용해 생성한 문장을 각각 학습시켜 비교 분석해 성능 개선과 더불어 효과적인 post-training을 위한 데이터 세트를 찾고자 한다. 개선하고자 했던 언어 모델인 FinBERT[7]와 함께 BERT[11]도 성능 변화를 관찰하여 일반적인 도메인에서 학습된 모델도 본 연구에서 제안하는 방법으로 성능의 개선이 가능한 지 보이고자 한다.

2. 관련 연구

2.1 금융 감성 분석

감성 분석은 텍스트를 통해 사람들의 감정이나 의견을 추출하는 연구 분야다[13]. 특히, 금융 감성 분석은 일반적인 감성 분석과 달리, 시장 반응 예측에 중점을 두는 특화된 분야로 발전했다[14]. 금융 감성 분석은 주가 예측[15, 16], 금융 보고서 분석[17, 18], 투자 결정 지원[19, 20]과 같은 다양한 분야에서 중요한 역할을 한다. 이에 대한 연구가 시간에 지남에 따라 다양한 방법으로 이루어져 왔고, 시간이 지남에 따라 사전 기반 접근 방식, 기계 학습 기반 접근 방식, 딥러닝 기반 접근 방식으로로 발전해왔다[21].

사전 기반 접근 방식은 초기 금융 감성 분석 연구에 주로 사용된 방법론으로, 사전에 정의된 단어 사전을 이용해 감성을 분류하는 방법이다. 초기에는 충분한 정확도를 제공하지 못하거나 사전 내 감성을 나타내는 단어의 수가 불균형하게 분포해있다는 문제가 있기도 했지만[22], Loughran과 McDonald가 제안한 finance-specific word lists[23]가 이를 개선하는 데 큰 기여를 했다[24]. 이후, 도메인 적응을 위해 사전에 추가되는 단어들의 감성 점수를 업데이트하는 기술이 개발되었으나[25], 단어의 위치 정보를 반영하지 못하는 한계가 있었다. 이러한 한계는 금융 감성 분석에 기계 학습을 접목시키는 계기가 되었다.

기계 학습 기반 접근 방식에서는 SVM, Naive Bayes, Decision Trees와 같은 방법론을 이용하여 StockTwits 데이터를 'bullish'와 'bearish'로 분류하는 연구가 진행되었다[26]. 또한, 사전 기반 기법과 결합된 모델들도 개발되었다[27]. 이런 다양한 머신러닝 기법의 적용으로 성능이 개선되었지만, 위치 정보의 정확성과 데이터의 희소성 문제는 여전히 남아 있었다[28].

딥러닝 기반 접근 방식은 이러한 문제를 해결하고자 등장했다. 구글의 Transformer[29] 아키텍처의 등장 전에는 토큰을 벡터화해 정보를 저장해 정보를 저장하는 Word Embedding의 방법으로 GloVe[30], Word2Vec[31], ELMo[32]가 활발히 사용되었다. 이후 구글에서 Transformer 아키텍처의 인코더를 기반으로 개발한 BERT[11]의 등장은 더욱 정확한 단어 임베딩 표현을 추출할 수 있을 뿐더러 감성 분석 태스크의 성능을 대폭 향상시켰고, 연구자들은 이를 금융 분야에 접목시키는 방법에 대해 현재까지도 연구를 활발히 진행하고 있다.

특히 BERT[11]는NLU(Natural Language Understanding) 태스크에 적합함이 입증되었으며 이것을 금융 분야에 적용한 다양한 FinBERT(Financial BERT) 모델이 개발되었다. 이 모델들은 각각 독특한 방법론을 적용하여 수행하고자 하는 과제의 성능을 높이는데 집중했다.

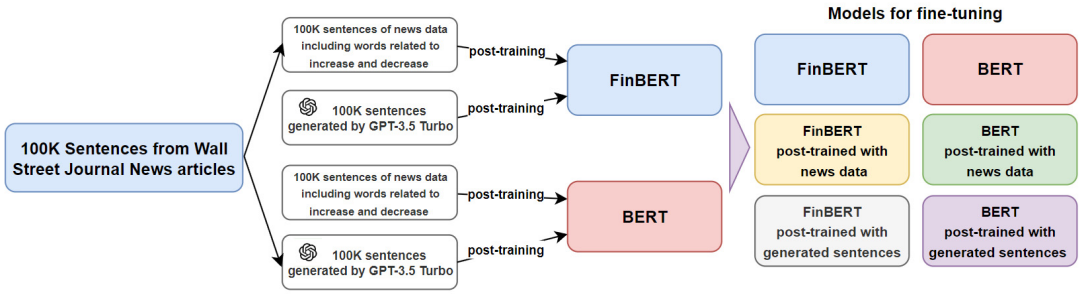
FinBERT[7]는 사전 학습된 BERT 모델을 금융 뉴스인 Reuters TRC-2 데이터셋으로 추가 학습시켜 금융 도메인에 대한 지식을 확장했다. 그 결과 금융 뉴스 감성분석 태스크에서 BERT에 비해 크게 우수한 성능을 보였다. FinBERT[9, 10]는 10-K, 10-Q 보고서, 실적 발표 녹취록, 애널리스트 보고서 텍스트 데이터를 활용해 BERT 모델을 처음부터 사전 학습하여 금융 전용 어휘 사전을 구축했다. 이 모델은 기존 BERT와 비교하여 다양한 금융 분야 태스크에서 높은 성능을 보여주었다.

2.2 Post-training

Post-training은 사전 학습된 모델의 가중치를 바탕으로 추가적인 학습을 수행하는 방법이다. 이는 모델이 처음부터 전체적으로 학습되는 것이 아니라 필요한 부분만을 집중적으로 업데이트함으로써 토큰 표현을 보다 효율적으로 개선하는 것을 기대할 수 있게 한다[33, 34]. 현재 여러 분야에서는 사전 훈련 과정에서 다루지 않은 새로운 도메인에 대한 언어적 이해를 강화하기 위해 post-training 방법을 활용하고 있다[33-37].

C. Du 등은 BERT 모델이 특정 도메인을 인식하는 능력이 부족하고, 도메인 간 지식 전이 시 원본과 대상 도메인의 특성을 구별하지 못한다는 문제점을 지적하며, 이를 해결하기 위해 도메인 구별을 위한 post-training을 제안한다[33]. 이를 통해 BERT 모델이 도메인을 인식하게 하고 자기 주도적으로 도메인의 특성을 파악하도록 돕는다고 주장한다. H. Xu 등은 고객 리뷰 데이터를 활용한 리뷰 이해 태스크에서 BERT를 post-training하여 성능을 향상시켰다[34]. T. Whang 등은 BERT 모델이 일반적인 도메인에 대해 사전 학습되어 있기에 미세 조정시 특정 단어와 구절에 대한 충분한 이해가 부족하다고 지적하며, 검색 기반 대화 시스템에서 multi-turn 응답 선택을 위한 post-training을 통해 성능을 개선한다[35]. 또한, J. Park 등은 기업의 사실적 지식을 언어 모델에 통합하기 위해 10-K 양식에서 기업명을 우선적으로 마스킹하는 방법을 제안하여 주식 관련 스팸 분류 태스크에서 우수한 결과를 보여준다[36].

본 연구에서는 금융 분야에서 중요하면서도 아직 충분히 다루어지지 않은 수치와 관련된 단어들에 대한 이해를 높이기 위해 세 가지의 데이터 세트로 post-training을 적용할 계획이다. 이를 통해 언어 모델이 금융 분야에서의 증감 표현에 대한 이해력을 향상시키는 것과 더불어 이 과정에서 어떤 데이터 세트가 효과적으로 작용하는지를 찾는데 기여할 것으로 기대한다.



(그림 1) 연구의 전체 워크플로우
(Figure 1) The overall workflow of our research

3. 연구 방법

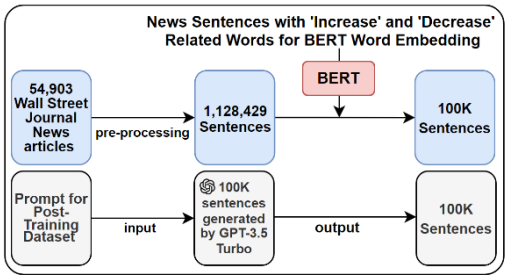
3.1 연구 개요

본 연구의 목적은 금융 뉴스 문장에서 증감과 관련된 정보를 보다 효과적으로 이해하고 예측하도록 언어 모델을 학습시키는 것이다. 여기서 post-training에 적합한 데이터 세트를 찾기 위해 증감 관련 단어를 포함하는 금융 뉴스와 GPT-3.5-turbo-1106으로 생성한 문장에 대해 실험한 후 성능을 비교한다. 본 연구의 프레임워크는 그림 1과 같고, 요약된 프로세스는 다음과 같다.

1. Wall Street Journal에서 증감과 관련된 단어를 포함하는 금융 뉴스 문장 10만개를 추출한다.
2. GPT-3.5-turbo-1106으로 금융 감성분석 성능 향상을 위한 post-training 데이터 세트 문장 10만 개를 생성한다.
3. FinBERT와 BERT모델에 대해 제작한 데이터 세트로 각각 subword masking 기법을 사용한 post-training을 진행한다.
4. Financial PhraseBank 데이터셋으로 미세 조정한 결과를 비교한다.

이어서, 각 단계별 자세한 방법에 대해 설명한다.

3.2 데이터 세트 수집 및 전처리



(그림 2) post-training 데이터 구축 프로세스

(Figure 2) The process of constructing post-training dataset

그림 2는 금융과 관련된 증감 표현을 잘 이해할 수 있게 하기 위한 데이터셋을 제작하는 과정을 보여준다. 자세한 과정은 본 섹션에서 다룬다. 우리는 데이터 수집 플랫폼인 TDM Studio 를 통해 Wall Street Journal 54,903 개의 뉴스 기사를 수집했다. 감성 분석의 성능을 높이기 위한 post-training 데이터셋으로의 사용이 목적이었고, 미세조정을 위해 사용한 데이터 세트 역시 기간에 구애받지 않는 문장들이었기 때문에 수집 기간은 연구자들이 연구 환경을 고려해 2018년에서 2020년으로 선정했다. 이 성능과 관련 결측지 제거 후, spacy 라이브러리를 사용하여 문장 단위로 분리하고, 소문자로 변환한 뒤 특수문자와 불용어를 제거했다. 최소한 하나의 토큰을 마스킹할 수 있도록 5개 이상의 단어로 구성된 문장만을 선택하여, 총 1,128,429개의 문장을 선별했다. 이중 증감 관련 단어를 포함하는 문장으로 post-training 하는 것이 좋은 성능을 보이는지 확인하기 위해 증감과 관련된 단어를 추출한 실험 환경을 고려해 이를 포함하는 문장 10만 개를 추출해 사용한다. 증감 관련 단어 추출은 증감을 나타내는 대표적인 단어인 'increase'와 'decrease'를 기준으로 사전 학습된 BERT 모델의 단어 임베딩을 기반해 코사인 유사도가 가장 높은 100개의 단어를 각각 추출해 사용한다. 이때 BERT [11]의 사전 학습된 단어 임베딩을 사용하는 이유는 FinBERT[7]와 같은 vocabulary를 사용하면서 일반적인 범위의 증감과 관련된 단어들이 잘 학습되어 있기 때문이다. 또 FinBERT에서 같은 절차를 거쳐 increase, decrease와 유사도가 높은 단어들을 비교한 결과, 특수기호 등 일반적으로 증감과 관련되어 있지 않아 보이는 단어가 다수 포함되어 있어 사용하기 어려웠다. 실제로 BERT 모델에서는 "extend", "slowed", "improved", "weakening" 등 증가와 감소와 관련된 단어들이 잘 추출된 반면, FinBERT에서는 "generate", "adviser", "wept", "milo", 심지어는 일본어와 같은 관련 없는 토큰이 다수 포함되어 있었다. 이 과정에서 중복을 제외한 193개의 단어를 선정해 이를 포함하는 문장 10만개를 선별해 실험에 사용한다.

다른 세트로, 생성형 AI를 활용하여 하고자하는 태스크를 설명한 후 그에 맞는 역할을 부여해 문장을 생성하게 한 것을 post-training의 데이터 세트로 사용한다. 이때 사용하는 모델로 GPT-4-turbo가 아닌 GPT-4가 출시된 날 함께 업데이트된 GPT-3.5-turbo-1106 모델을 사용한다. 이는 GPT-3.5-turbo-1106이 간단한 작업을 수행하는데 적합하고 비용이 저렴하며 생성 속도도 빠르기 때문이다. 프롬프트는 시스템 메시지 2개, 콘텐트 메시지 2개를 조합해 사용한다. 시스템 메시지로는 상대적으로 넓은 범위에서 해야할 태스크를 잘 수행할 수 있을 만하다고 생각한 역할을 부여하는데 집중했고, 콘텐트 메시지로는 더 구체적인 태스크에 대한 설명을 자세히 하거나 FinBERT[7]가 올

바르게 감성 분류를 해내지 못한 문장을 예시로 직접 알려주어 생성된 문장의 다양성을 얻고자 했다. 이와 함께 같거나 유사한 문장 생성을 막기 위해 창의성을 조절하는 하이퍼파라미터인 temperature를 1로 설정했다. 또 콘텐트 메시지로는 주어진 태스크 외의 문장을 생성하지 않게 강한 명령형 문장을 사용했다. 문장의 수는 다른 비교 대상들과 마찬가지로 10만 개의 문장을 생성해 사용한다. 사용한 프롬프트는 표 1과 같다.

3.3 Post-Training

BERT 모델과 FinBERT 모델을 대상으로 앞서 구축한 10만 개의 증감 관련 단어를 포함한 금융 뉴스와 GPT-3.5 turbo로 생성한 10만 개의 문장을 각각 subword masking 기법을 사용해 post-training을 진행한다. 결과적으로 성능을 비교할 모델은 총 6개가 된다.

3.4 Fine-tuning

학습된 모델들을 바탕으로, Financial PhraseBank 데이터 세트을 사용하여 금융 감성 분석을 위한 미세 조정을 수행한다. 미세 조정을 위해 기존 히든 레이어 위에 새로운 임베딩 레이어를 추가하고, 문장에 대한 임베딩 벡터를 특징 벡터로 사용하여 분류 작업을 진행한다. 이 과정에서 각 학습된 모델의 성능을 기록하고 비교한다. Financial PhraseBank 데이터 세트에 대한 설명은 4장에서 자세히 다룬다.

4. 실험 설정

4.1 Financial PhraseBank

본 연구에서 사용한 평가용 데이터 세트는 금융 감성 분석 벤치마크 데이터 세트로 잘 알려진 Financial PhraseBank다. Financial PhraseBank는 금융 뉴스 문장을 대상으로 'positive', 'negative', 'neutral'로 감성 라벨링된 총 4845개의 문장이 포함되어 있다. 이 문장들은 금융 시장에 대한 배경 지식을 가진 Aalto University School of Business의 석사 학생 13명과 금융 시장에 대한 충분한 배경 지식을 갖춘 3명의 연구원들에 의해 주석이 달렸으며, 문장마다 5~8개의 주석이 부여되었고, 주석가 간의 문장 합의 강도에 따라 '50% agree', '66% agree', '75% agree', 'all agree'로 나뉜다. 본 연구에서는 모든 데이터 세트를 대상으로 진행하는 실험과 함께 정확한 라벨을 대상으로 하는 실험을 위해 'all_agree' 라벨이 부여된 2262개의 문장을 따로 사용해 성능을 비교한다. 데이터 세트의 개수와 라벨 비율은 표 2와 같다.

(표 2) Financial PhraseBank 데이터 세트의 합의 강도별 라벨 비율과 개수

(Table 2) The ratio and number of labels by consensus intensity in the Financial PhraseBank dataset

합의강도	긍정	부정	중립	개수
100%	25.2%	13.4%	61.4%	2262
75 - 99%	26.6%	9.8%	63.6%	1191
66 - 74%	36.7%	12.3%	50.9%	765
50 - 65%	31.1%	14.4%	54.5%	627
All	28.1%	12.4%	59.4%	4845

(표 1) post-training에 필요한 문장 생성을 위해 사용한 프롬프트

(Table 1) Prompts that were used for sentence generation required for post-training

System candidates	You are a developer working on financial language models. Specifically, you want to generate data to further pre-train a BERT-base financial model to improve the performance of financial sentiment analysis.
	You are helpful text generation assistant specialized...Use a variety of ways to start a sentence: sectors, directive nouns, conjunctions, prepositions, etc.
Content candidates	Please generate 100 sentences...ensure all information provided is generated by either utilizing different business sectors or using a variety of sentiment expressions such as increase, decrease, compare to, relative to, etc.
	Please generate 100 detailed, dynamic sentences. Here is an example: Pre-tax loss totaled euro 0.3 million, compared to a loss of euro 2.2 million in the first quarter of 2005. Do not generate the exact way as the samples.

4.2 실험 환경

실험에 사용한 하이퍼파라미터는 다음과 같다.

Post-training을 수행하기 위한 방법으로는 실험 환경을 고려해 배치 크기를 8로 설정하고 최대 에포크 수를 5개로 설정했다. 학습률은 5e-5인 아담 옵티마이저를 사용했다. NVIDIA GeForce RTX 3060 GPU 1대가 post-training에 사용되었고, 이는 문장 10만 개를 post-training하는데 모델 하나 당 약 10시간이 소요되었다. Fine-tuning 설정은 개선하고자 하는 메인 모델인 FinBERT(2019)[7]를 따른다. 학습 배치 크기를 64로 설정하고 학습률은 2e-5로, 0.2의 warm-up 비율을 사용했다. 총 6에폭의 학습을 마친 후, 가장 성능이 좋았던 에폭의 결과를 채택했다. 마찬가지로 NVIDIA GeForce RTX 3060 GPU 1대가 미세 조정에 사용되었다.

4.3 평가 척도

모든 실험에서, 데이터셋의 모든 부분에 대해 균등한 예측 성능을 입증하기 위해 데이터를 10개의 서브셋으로 나누고, 이중 일부를 훈련용 데이터로, 나머지를 테스트 데이터로 설정하여 훈련 데이터셋이 상대적으로 적은 환경과 충분한 데이터 환경 모두에서 실험을 진행한다. 훈련용 데이터 세트와 테스트 데이터 세트의 선택에 따라 결과가 달라질 수 있으므로, 모든 경우를 고려한 실험의 평균 성능을 사용한다. 예를 들어, 훈련용 데이터 세트와 테스트 데이터 세트의 비율이 2:8인 경우, 총 5회의 실험을 수행하고 이들의 평균값을 성능 지표로 사용한다. 불균형한 데이터셋을 대상으로 진행되는 실험이기 때문에 accuracy와 함께 가중 f1-score를 평가 지표로 사용한다. Accuracy의 자세한 식은 다음과 같다:

$$\text{Accuracy: } \frac{\sum_{i=n}^{n}(TP_i + TN_i)}{\sum_{i=n}^{n}(TP_i + TN_i + FP_i + FN_i)} \quad (1)$$

이때 TP_i는 i번째 클래스에 대해 올바르게 양성으로 식별된 사례들을 나타내고, TN_i는 올바르게 음성으로 식별된 사례들을 나타낸다. FP_i는 i번째 클래스에 대해 양성으로 잘못 식별된 사례들을 나타내고, FN_i는 음성으로 잘못 식별된 사례들을 나타낸다.

$$\text{Weighted f1-Score: } \sum_{i=n}^{n} w_i \cdot F1_i \quad (2)$$

Weighted f1-Score의 자세한 식은 위와 같다. 이때 w_i는 i번째 클래스의 가중치로, 해당 클래스의 실제값들의 비율이다. $F1_i$는 i번째 클래스에 대한 F1-score이다. 단일 클래스의 F1-score은 $F1_i = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i \cdot recall_i}$로 계산되며, 이때 정밀도(precision)와 재현율(recall)은 다음과 같이 정의된다:

$$precision_i : \frac{TP_1}{TP_1 + FP_1} \quad (3)$$

$$recall_i : \frac{TP_1}{TP_1 + FN_1} \quad (4)$$

이때 TP_i, FP_i, FN_i는 각각 i번째 클래스에 대한 TP, FP, FN의 수를 나타낸다.

5. 실험 결과

표 3과 표 4는 각각 합의 강도가 50% 이상, 합의 강도가 100% 인 Financial PhraseBank 데이터 세트를 대상으로 미세 조정한 결과이며, 프롬프트를 활용해 생성한 문장으로 post-training을 진행한 모델을 (P)로, 증감을 나타내 는 단어를 포함하는 뉴스 문장으로 post-training을 진행한 모델을 (N)으로 구분한다. 이들을 정확도 및 가중 f1-score로 비교한 것으로, 미세 조정된 데이터 세트에서 훈련-테스트 데이터 비율에 따라 구분된다. 각 표에서는 BERT류 모델과 FinBERT류 모델별로 훈련-테스트 비율별 최고 성능에 해당하는 수치를 굵은 글씨로 나타낸다.

표 3 을 보면, 모든 데이터셋으로 비교한 실험에서는 전반적으로 FinBERT 모델이 성능이 우수함을 알 수 있다. 이는 FinBERT가 대용량 금융 데이터 세트로 추가적인 사전 학습을 진행했기 때문에 미세 조정시 사용한 데이터 세트에 대해서도 모델이 잘 이해한 것으로 보인다. 하지만 학습하는 데이터 세트가 20%인 환경의 경우, FinBERT(P)가 성능이 가장 우수함을 알 수 있다. 이는 적절한 prompt를 사용해 생성한 문장으로 학습시켰을 때, 일반화 측면에서는 더욱 뛰어난 결과를 보일 수 있다는 것을 입증한다. 그러나 이와 같은 결과는 BERT 모델에서는 나타나지 않는 것으로 보아, 추가 학습의 효과는사전 학습된 모델의 도메인과의 일치 여부가 중요함을 시사한다.

(표 3) 합의 강도가 50% 이상인 Financial Phrasebank 데이터 세트 4,845개 대상 미세 조정 성능표
(Table 3) Performance chart for fine-tuning on 4,845 Financial PhraseBank dataset entries with consensus intensity over 50%

	Train : Test ratio	BERT	BERT(P)	BERT(N)	FinBERT	FinBERT(P)	FinBERT(N)
Accuracy	2:8	76.34	76.58	**79.33**	84.23	**85.06**	81.37
	5:5	**85.14**	82.58	83.80	**86.62**	83.72	86.00
	8:2	**85.76**	84.17	85.51	**87.87**	85.00	86.71
Average		82.41	81.11	**82.88**	**86.24**	84.59	84.69
Weighted f1-score	2:8	73.22	72.75	**76.65**	83.50	**84.39**	78.59
	5:5	**85.13**	82.66	83.90	**86.63**	83.78	86.08
	8:2	**85.81**	84.21	85.60	**87.88**	85.00	86.78
Average		81.39	79.87	82.05	**86.00**	84.39	83.82

(표 4) 합의 강도가 100%인 Financial PhraseBank 데이터 세트 2,262개 대상 미세 조정 성능표
(Table 4) Performance chart for fine-tuning on 2,262 Financial PhraseBank dataset entries with 100% consensus intensity

	Train : Test ratio	BERT	BERT(P)	BERT(N)	FinBERT	FinBERT(P)	FinBERT(N)
Accuracy	2:8	79.51	75.59	**80.05**	80.42	81.39	**83.82**
	5:5	94.30	91.65	**94.61**	94.21	95.54	**96.55**
	8:2	**96.51**	95.41	96.37	96.69	96.78	**97.17**
Average		90.11	87.55	**90.34**	90.44	91.24	**92.51**
Weighted f1-score	2:8	79.16	74.63	**79.92**	79.83	81.05	**83.75**
	5:5	94.29	91.54	**94.63**	94.21	95.53	**96.57**
	8:2	**96.52**	95.40	96.39	96.67	96.77	**97.18**
Average		89.99	87.19	**90.31**	90.24	91.12	**92.50**

더 정확한 감성 라벨을 대상으로 실험한 결과는 조금 달랐다. 표 4를 보면, FinBERT 모델을 추가로 학습시킨 모든 경우에서 FinBERT의 성능을 상회하는 결과를 보였다. GPT-3.5-turbo-1106으로 생성한 문장으로 추가 학습을 진행한 후의 성능도 기존 FinBERT보다 높은 성능을 보였지만, 증감 관련 단어를 포함하는 실제 뉴스 기사로 추가적인 학습을 진행한 경우가 가장 높은 성능을 보였다. 이러한 결과는 모델의 성능 개선을 위해 특정 목표에 부합하는 적절한 데이터 세트를 사용한 추가 학습이 매우 효과적임을 입증함과 동시에, 생성형 AI를 사용해 생성한 데이터로의 추가 학습의 의미가 분명함을 의미한다. 한편, BERT 모델에 대한 추가적인 학습은 성능 향상에서 효과가 없거나 미미했으며, 표 2의 결과와 마찬가지로 추가 학습 진행 시 도메인 일치를 고려해야 한다는 시사점을 남긴다. 종합적으로, 본 연구는 특정 도메인에 최적화된 모델이 해당 분야의 데이터에 대해 더 높은 이해를 보이며, 이에 대한 도메인 일치의 중요함을 확인시켜준다. 또 증감 관련 단어를 포함하는 금융 뉴스 문장으로 추가 학습을 진행한 것이 생성한 문장으로 추가 학습을 진행한 것보다 대체적으로 성능이 높았지만, 일반화를 더욱 요구하는 특정 환경에서는 적절한 프롬프트를 사용해 생성한 문장으로 학습한다면 이보다 더 높은 성능을 달성할 수 있다는 가능성도 시사한다. 이러한 결과는 향후 금융 분야의 감성 분석관련 연구에 있어 방향성을 제시하며, 모델 선택과 데이터 준비 과정에서의 실질적인 지침을 제공한다.

6. 결론 및 향후 연구

본 연구는 금융 뉴스에서 중요한 의미를 지닌 증감과 관련된 정보를 효과적으로 이해하고 감성을 올바르게 예측하기 위한 언어 모델의 학습 방법론 및 학습 데이터 세트 선정에 관하여 탐구한다. 실험 결과, 증감과 관련된 단어를 포함하는 금융 뉴스 문장과 GPT-3.5-turbo-1106로

생성한 문장을 활용한 추가 학습은 모델의 이해도와 예측 성능을 상당히 향상시켰다. 이는 적절한 데이터 세트로 추가 학습을 진행한다면 언어 모델의 약점을 개선할 수 있고 해당 분야의 복잡한 언어적 뉘앙스를 포착할 수 있음을 시사한다.

이와 함께, post-training 과정에서 적절한 데이터 세트와 모델의 선택이 언어 모델의 성능에 결정적인 영향을 미침을 보여줌으로써, 실무자와 연구자들에게 모델 선택과 데이터 준비 과정에서 중요한 지침을 제공한다. 또한, GPT-3.5-turbo-1106과 같은 생성형 AI를 사용해 생성한 데이터 세트로 추가적인 학습을 진행하는 것이 높은 성능을 달성할 수 있음을 보인다. 따라서 본 연구는 향후 금융 분야의 더욱 정교한 언어 이해 및 감성 분석을 위한 연구에 방향성을 제시하며, 또 금융 도메인 안에서 감성 분석뿐만 아니라 다른 분류 태스크에서도 우수한 성능을 보일 것이라는 가능성을 입증한다. 특히, 본 연구는 개선된 모델의 감성 스코어를 활용하는 것뿐만 아니라, 투자 의견 분류나 시황 상승 및 하락 예측과 같은 감성 분석이 아닌 다른 분류 태스크에서도 모델이 학습해야 할 적절한 데이터 세트를 생성형 AI를 활용하여 생성함으로써 성능이 개선될 것이라는 기대를 가능하게 한다.

마지막으로, 본 연구는 BERT 기반 언어 모델의 텍스트 이해 능력 개선을 위해 post-training 시 사용하는 학습 데이터 세트에서의 인사이트를 찾고자 했다. 본 연구를 바탕으로 GPT와 같은 언어 생성 모델을 대상으로 한 투자 의견 생성, 기업 리스크 관리 등의 텍스트 생성과 관련된 태스크에서 유사한 실험을 진행해 결과를 관찰하는 것은 향후 연구 과제로 남겨둔다.

참고문헌 (References)

[1] E. J. De Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," Information Processing and Management, vol. 50, no. 2, pp. 426–441, Mar. 2014. https://doi.org/10.1016/j.ipm.2013.12.002

[2] R. M. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stockprediction using numerical and textual information," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Jun. 2016. https://doi.org/10.1109/icis.2016.7550882

[3] C. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang, "Financial sentiment analysis for risk prediction," Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 802–808, Oct. 2013. [Online] Available: https://www.researchgate.net/profile/Chuan-Ju_Wang/publication/258821644_Financial_Sentiment_Analysis_for_Risk_Prediction/links/551b6a920cf2fdce84389cf4.pdf

[4] L. Loukas et al., "FINER: Financial Numeric Entity Recognition for XBRL Tagging," Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jan. 2022. https://doi.org/10.18653/v1/2022.acllong.303

[5] K. R. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big Data, vol. 3, no. 1, May. 2016. https://doi.org/10.1186/s40537-016-0043-6

[6] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," Science China Technological Sciences, vol. 63, no. 10, pp. 1872–1897, Sep. 2020. https://doi.org/10.1007/s11431-020-1647-3

[7] Araci, Doug, "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models," arXiv preprint arXiv:1908.10063, 2019. https://doi.org/10.48550/arXiv.1908.10063

[8] A. Huang, H. Wang, and Y. Yang, "FinBERT: A Large Language Model for Extracting Information from Financial Text," Contemporary Accounting Research, vol. 40, no. 2, pp. 806–841, Jan. 2023. https://doi.org/10.1111/1911-3846.12832

[9] Y. Yang, M. C. S. Uy, and A. Huang, "FinBERT: a pretrained language model for financial communications," arXiv preprint arXiv:2006.08097, Jun. 2020. https://doi.org/10.48550/arxiv.2006.08097

[10] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining," Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Special

Track on AI in FinTech, Jul. 2020.
https://doi.org/10.24963/ijcai.2020/622

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.
https://doi.org/10.48550/arXiv.1810.04805

[12] OpenAI: J. Achiam et al, "GPT-4 Technical Report,"arXiv preprint arXiv: 2303.08774, Mar. 2023.
https://doi.org/10.48550/arXiv.2303.08774

[13] G. Vinodhini and RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," vol. 2, no. 6, 2012, [Online].
Available: https://www.researchgate.net/profile/Vinodhini-G-2/publication/265163299_Sentiment_Analysis_and_Opinion_Mining_A_Survey/links/54018f330cf2bba34c1af133/Sentiment-Analysis-and-Opinion-Mining-A-Survey.pdf

[14] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, vol. 69, pp. 14–23, Oct. 2014.
https://doi.org/10.1016/j.knosys.2014.04.022

[15] A. Mittal, A. Goel, "Stock Prediction Using Twitter Sentiment Analysis,"Stanford University, 2011, [Online].
Available: cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[16] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Apr. 2019.
https://doi.org/10.1109/bigdataservice.2019.00035

[17] Á. Bernal and C. Pedraz, "Sentiment Analysis of the Spanish Financial Stability Report," International Review of Economics & Finance, vol.89, Part B, pp. 913–939, Jan. 2024.
https://doi.org/10.1016/j.iref.2023.10.037

[18] A. Levenberg, S. Pulman, K. Moilanen, E. Simpson, and S. Roberts, "Predicting Economic Indicators from Web Text Using Sentiment Composition," International Journal of Computer and Communication Engineering, vol. 3, no. 2, pp. 109–115, Jan. 2014.
https://doi.org/10.7763/ijcce.2014.v3.302

[19] D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis,"IEEE Transactions on Systems, Man, and Cybernetics, vol. 44, no. 8, pp. 1077–1087, Aug. 2014. https://doi.org/10.1109/tsmc.2013.2295353

[20] B. Hasselgren, C. Chrysoulas, N. Pitropakis, and W. J. Buchanan, "Using Social Media & Sentiment Analysis to Make Investment Decisions," Future Internet, vol. 15, no. 1, p. 5, Dec. 2022.
https://doi.org/10.3390/fi15010005

[21] X. Man, T. Luo, and J. Lin, "Financial Sentiment Analysis (FSA): A Survey," 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), May. 2019.
https://doi.org/10.1109/icphys.2019.8780312

[22] B. F. Green, P. J. Stone, D. Dunphy, M. S. Smith, and D. M. Ogilvie, "The General Inquirer: A Computer Approach to content analysis," American Educational Research Journal, vol. 4, no. 4, p. 397, Nov. 1967.
https://psycnet.apa.org/record/1967-04539-000

[23] T. Loughran and B. McDonald, "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," Journal of Finance, 66(1), pp. 35–65, 2011.
https://doi.org/10.1111/j.1540-6261.2010.01625.x

[24] H. Jangid, S. Singhal, R. R. Shah, and R. Zimmermann, "Aspect-Based Financial Sentiment Analysis using Deep Learning," WWW '18: Companion Proceedings of the the Web Conference 2018, Jan. 2018.
https://doi.org/10.1145/3184558.3191827

[25] F. Z. Xing, F. Pallucchini, and Z. Wang, "Cognitive-inspired domain adaptation of sentiment lexicons," Information Processing and Management, vol. 56, no. 3, pp. 554–564, May. 2019.
https://doi.org/10.1016/j.ipm.2018.11.002

[26] G.Wang et al., "Crowds on Wall Street: Extracting Value from Collaborative Investing Platforms,"CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing 2015, pp. 17–30, Feb. 2015.
https://doi.org/10.1145/2675133.2675144

[27] K. Cortis et al., "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News," Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 17–30, Aug. 2017.
https://doi.org/10.18653/v1/S17-2089

[28] S. Sohangir, D. Wang, A. Pomeranets et al., "Big Data: Deep Learning for Financial Sentiment Analysis," Journal of Big Data, vol. 5, article number 3, Jan. 2018.
https://doi.org/10.1186/s40537-017-0111-6

[29] A. Vaswani et al., "Attention is all you need," arXiv preprint arXiv: 1706.03762, Jun. 2017.
https://doi.org/10.48550/arXiv.1706.03762

[30] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, Oct. 2014.
https://doi.org/10.3115/v1/d14-1162

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv: 1301.3781, Jan. 2013.
https://doi.org/10.48550/arXiv.1301.3781

[32] M. E. Peters et al.,"Deep contextualized word representations," arXiv preprint arXiv: 1802.05365, Feb. 2018.
https://doi.org/10.48550/arXiv.1802.05365

[33] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4019-4028,Jan. 2020.
https://doi.org/10.18653/v1/2020.acl-main.370

[34] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT Post-Training for review reading comprehension and aspect-based sentiment analysis," arXiv preprint arXiv: 1904.02232, 2019.
https://doi.org/10.48550/arXiv.1904.02232

[35] T. Whang, D. Lee, C. Lee, K. Yang, D. Oh, and H. Lim, "An Effective Domain Adaptive Post-Training Method for BERT in Response Selection," arXiv preprint arXiv:1908.04812, 2019.
https://doi.org/10.48550/arXiv.1908.04812

[36] J. Park and S. Cho, "Incorporation of company-related factual knowledge into pre-trained language models for stock-related spam tweet filtering," Expert Systems With Applications, vol. 234, p. 121021, Dec. 2023.
https://doi.org/10.1016/j.eswa.2023.121021

[37] R. Luo, G. Huang, and X. Quan, "Bi-Granularity Contrastive Learning for Post-Training in Few-Shot Scene," arXiv preprint arXiv:2106.02327, 2021.
https://doi.org/10.48550/arXiv.2106.02327

◐ 저 자 소 개 ◑

정 희 도(Hui Do Jung)
2019년~2023년 서울시립대학교 수학과 학사
2023년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : jhd11j@yonsei.ac.kr

김 재 헌(Jae Heon Kim)
2015년~2019년 뉴욕대학교 (NYU) 컴퓨터공학과 학사
2023년~현재 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : jhk774@yonsei.ac.kr

장 백 철(Beakcheol Jang)
2009년 North Carolina State University 컴퓨터공학과(공학박사)
2021년~현재 연세대학교 정보대학원 교수
관심분야 : Natural Language Processing, Artificial Intelligence, Bigdata Analytics
E-mail : bjang@yonsei.ac.kr