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

감성 분석을 위한 FinBERT 미세 조정: 데이터 세트와 하이퍼파라미터의 효과성 탐구☆

FinBERT Fine-Tuning for Sentiment Analysis: Exploring the Effectiveness of Datasets and Hyperparameters

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

요 약

본 논문에서는 금융 뉴스 데이터로 추가적인 사전 학습이 진행된 BERT 기반 모델인 FinBERT 모델을 사용하여 금융 영역에서 감성 분석 시 학습시킬 데이터와 그에 맞는 하이퍼파라미터를 찾는 방법을 소개한다. 우리의 목표는 다양한 데이터 세트를 활용하고 하이퍼파라미터를 미세 조정하여 정확한 감성 분석을 위해 FinBERT 모델을 가장 잘 활용하는 방법에 대한 포괄적인 가이드를 제공하는 것이다. 이 연구에서는 제안된 FinBERT 모델 미세 조정 접근법의 아키텍처와 워크플로우를 개괄적으로 설명하고, 감성 분석 태스크를 위한 다양한 데이터 세트와 하이퍼파라미터의 성능을 강조한다. 또한, 감성 라벨링 작업에 GPT-3를 사용함으로써 GPT-3가 적절한 라벨러 역할을 하는지에 대한 신뢰성을 검증한다. 결과적으로 미세 조정된 FinBERT 모델이 다양한 데이터 세트에서 우수한 성능을 발휘한다는 것을 보여주었고, 각 데이터 세트에 대해 전반적으로 우수한 성능을 보이는 학습률 5e-5와 배치 크기 64의 최적의 조합을 찾았다. 또 일반 도메인의 뉴스보다 일반 도메인의 트위터 데이터 세트에서 성능이 크게 향상됨을 기반으로 금융 뉴스 데이터만으로만 추가적으로 학습시키는 FinBERT 모델에 대한 의구심을 제시한다. 이를 통해 FinBERT 모델에 대한 최적의 접근 방식을 결정하는 복잡한 프로세스를 간소화하고 금융 분야 감성 분석 모델을 위한 추가적인 학습 데이터 세트와 미세 조정 시 하이퍼파라미터 선정에 대한 가이드라인을 제시한다.

☞ 주제어: FinBERT, 금융 분야 감성 분석, 하이퍼파라미터 미세 조정

ABSTRACT

This research paper explores the application of FinBERT, a variational BERT-based model pre-trained on financial domain, for sentiment analysis in the financial domain while focusing on the process of identifying suitable training data and hyperparameters. Our goal is to offer a comprehensive guide on effectively utilizing the FinBERT model for accurate sentiment analysis by employing various datasets and fine-tuning hyperparameters. We outline the architecture and workflow of the proposed approach for fine-tuning the FinBERT model in this study, emphasizing the performance of various datasets and hyperparameters for sentiment analysis tasks. Additionally, we verify the reliability of GPT-3 as a suitable annotator by using it for sentiment labeling tasks. Our results show that the fine-tuned FinBERT model excels across a range of datasets and that the optimal combination is a learning rate of 5e-5 and a batch size of 64, which perform consistently well across all datasets. Furthermore, based on the significant performance improvement of the FinBERT model with our Twitter data in general domain compared to our news data in general domain, we also express uncertainty about the model being further pre-trained only on financial news data. We simplify the complex process of determining the optimal approach to the FinBERT model and provide guidelines for selecting additional training datasets and hyperparameters within the fine-tuning process of financial sentiment analysis models.

☞ Keywords: FinBERT, Financial Sentiment Analysis, Fine-Tuning hyperparameters

1. 서 론

OpenMarket은 금융시장에서 오고 가는 자산과 관련된 모든 속성을 반영한다. 따라서 OpenMarket에서 실시간으로 생성되는 다양한 속성들을 분석하여 인사이트를 도출하는 과정은 중요하고, 이러한 분석을 통해 시장의 상황을 예측하고 투자자에게 유용한 정보를 제공하는 분석에

1 Yonsei Graduate School of Information Seoul, South Korea
* Corresponding author: bjang@yonsei.ac.kr
[Received 03 July 2023, Reviewed 11 July 2023, Accepted 20 July 2023]
☆ This work was supported by the Yonsei University Research Fund under Grant 2023-22-0104.

대한 수요 또한 높다. 금융 도메인 감성 분석 역사에 의하면 뉴스, 트위터 등의 텍스트 데이터의 감성 지표가 투자자들의 심리에 영향을 미친다고 하며, 이들의 감성 분석을 통해 시장의 감성 변화에 대한 통찰력을 얻을 수 있음이 밝혀져 왔다[1].

한편, 대규모 데이터의 빠른 발전과 함께 딥러닝의 영역이 확대되면서 금융 부문에서 감성 분석을 수행하기 위한 다양한 모델들이 개발되었다[2,3]. 그중 금융 분야에서 NLP 전이 학습 방법을 구현하는 대표적인 모델로 FinBERT[1]가 있다. 이는 BERT[4] 모델에서 금융 도메인 데이터 세트를 사용하여 추가적인 학습 및 미세 조정을 거친 모델로, 효과적인 전이 학습을 구현하여 다양한 작업의 미세 조정 과정에서 발생하는 정보 손실 현상인 치명적 망각을 방지하기 위한 다양한 전략을 제공한다.

FinBERT 모델의 프로세스는 크게 세 단계로 설명할 수 있다:

1) 일반 도메인 말뭉치에 대한 초기 사전 학습
2) 금융 도메인 말뭉치에 대한 추가 사전 학습
3) 라벨링된 데이터를 사용한 미세 조정

이때 2단계 과정에서 2008년부터 2010년의 뉴스 말뭉치인 Reuters TRC2 Financial 데이터 세트가 사용되었는데, 이러한 데이터 세트로 추가적인 사전 학습을 진행한 것이 최신의 금융 시장 구조를 잘 반영할 수 있을지에 대한 의구심이 들었다. 또한, 사전 학습과 미세 조정을 위해 뉴스 데이터 세트만을 사용한 것에 의문을 제기한다. 뉴스 기사가 대중들의 심리를 움직인다는 점은 의심할 여지가 없지만, 뉴스만이 그렇게 작용하는 것은 아니기 때문이다. 그래서 우리는 뉴스 데이터와 함께 사람들의 심리 파악에 사용되는 SNS 데이터 세트 중 트위터 데이터 세트를 추가로 미세 조정해 어떠한 성능 변화를 보이는지 실험을 통해 보이고자 한다.

한편, 하이퍼파라미터에 따라 실험의 결과가 확연히 달라지므로 데이터 세트에 따른 적절한 하이퍼파라미터 선정도 중요하다[5]. 본 연구팀은 FinBERT 논문의 저자들이 제시한 하이퍼파라미터를 같은 도메인이지만 다른 유형의 데이터 세트에도 일반화시킬 수 있는지에 대한 문제도 함께 제기한다. 궁극적으로 본 연구 주제는 금융 도메인으로 추가적인 사전 학습된 FinBERT모델 위에서 미세 조정 했을 때 어떤 데이터 세트가 더 높은 성능을 보이는지, 또 다양한 데이터 세트별로 어떤 하이퍼파라미터를 설정해야 좋은 성능을 보이는지에 대한 연구이다. 우리는 데이터 세트의 선정뿐만 아니라 하이퍼파라미터 설정의 중요성을 함께 탐구하여 특정 도메인 감성 분석의 정확도를 향상시키는 방법론을 제안한다. 주요 기여는 다음과 같다:

- 금융 또는 다른 특정 도메인별 작업을 수행할 때 학습 데이터 세트를 선택하는 방법에 대한 이해를 넓힐 수 있다.
- 감성 분석 작업을 위한 데이터와 하이퍼파라미터 선택에 대한 안목을 기를 수 있다.
- 뉴스 데이터만을 학습 데이터 소스로 사용하는 모델들의 효율성에 대한 의문을 제기한다.
- GPT-3 API를 활용해 데이터 라벨링하는 것이 높은 정확도를 보여주며 라벨이 없는 데이터 세트로도 지도 학습이 가능하게 함을 보인다.

2. 관련 연구

2.1 금융 분야 감성 분석

감성 분석은 글로 표현된 언어에서 사람들의 감성이나 의견을 추출하는 작업이다[6]. 그중 금융 감성 분석은 영역뿐만 아니라 목적에서도 일반적인 감성 분석과 차이가 있다. 금융 감성 분석의 목적은 일반적으로 텍스트에 제시된 정보에 대해 시장이 어떻게 반응할지 추측하는 것이다[7].

전통적인 금융 텍스트 분석으로는 Bag of Words와 어간 추출 등의 텍스트 처리와 같은 기법이 활용되었다[8]. 또한, '긍정적' 또는 '불확실한' 감성을 나타내는 값을 가진 금융 키워드의 빈도를 계산하여 금융 키워드 사전을 구성하기 위해 TF-IDF[9] 방법을 활용했다. 이와 함께 다양한 머신러닝 방법이 제안되었지만, 딥러닝이 현저히 우수한 것으로 입증되어 선호되는 방법론으로 확고히 자리 잡았다[10]. 예를 들어, 주식 동향 예측을 위해 LSTM[11] 모델을 활용하기도 하는데, 이 모델은 과거 주가와 당시 감성 스코어를 수집한 후, 여러 시간대 간의 시간적 의존성을 학습하고 이 지식을 활용해 미래 주가를 예측한다.

이러한 발전에도 불구하고, 라벨링된 금융 데이터 수집의 한계로 지도학습 기반의 딥러닝을 활용하는 것은 매우 어려운 과제였다. 이에 대한 해결책은 전체 모델의 파라미터의 대부분을 사전 학습된 값으로 초기화하고 분류 작업과 관련하여 해당 값을 미세 조정하는 전이학습을 사용하는 것이다[12]. 가장 최근에는 이러한 매커니즘이 잘 반영된 강건한 사전 학습 모델인 BERT를 주로 사용한다.

2.2 BERT

BERT는 Transformer[13]의 인코더 부분의 구조를 가져와 학습시켜 텍스트에 대한 이해를 돕는 양방향 언어 모델로, 2018년 구글에서 공개했다. 당시 감성 분석을 포함한 질의 응답, 개체명 인식, 문장 분류 등의 다양한 자연어 이해 벤치마크 데이터 세트에서 최고의 성능을 달성했다. BERT 모델은 라벨링이 되지 않은 대량의 BooksCorpus, Wikipedia의 텍스트 코퍼스를 MLM(Masked Language Modeling)과 NSP(Next Sentence Prediction) 기법을 사용해 사전 학습 시켜 텍스트에 대한 이해를 높인다. 이렇게 사전 학습 시킨 모델에 수행하고자 하는 태스크 처리를 위한 레이어를 추가해 라벨이 있는 데이터를 학습시킴으로써 미세 조정을 수행한다. 본 연구팀은 BERT 모델을 가져와 금융 도메인 데이터셋인 Reuters TRC2 Financial 로 추가적인 사전 학습을 수행한 FinBERT 모델을 실험에 사용한다. FinBERT는 영어 금융 뉴스 데이터셋인 Financial PhraseBank 데이터 세트를 대상으로 감성 분석 태스크를 진행했고, 당시 최고의 성능을 기록했다. 기본적으로 영어에 대해 학습된 모델이므로, 다른 언어에 대해 임베딩값이 최적화되어있지 않아 본 연구에서 역시 영어 데이터 세트만을 다룬다.

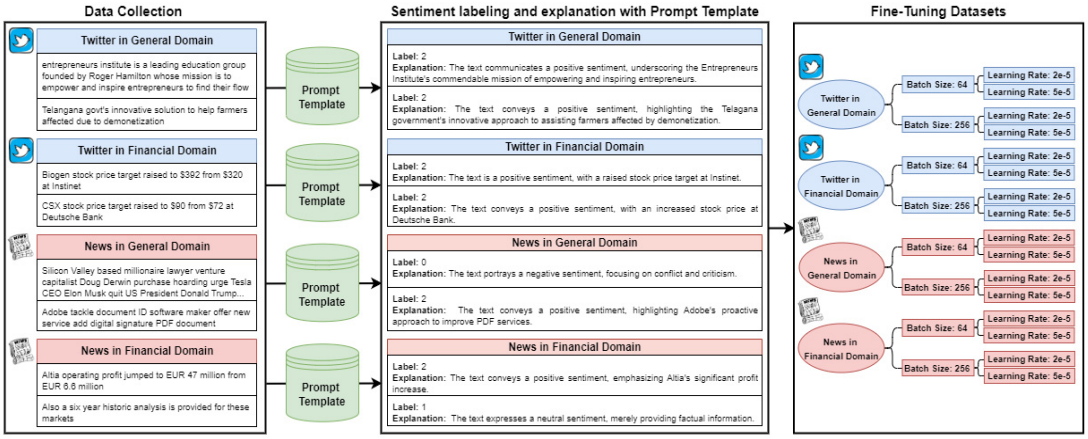
2.3 GPT-3 데이터 라벨링

전문가가 대규모 데이터 세트에 직접 라벨을 부여하는 것은 전문가 채용, 비용, 시간 측면에서 매우 비효율적인 방법이다. 따라서 최신 자연어 처리 연구에서는 데이터 라벨링을 위해 별개의 모델을 사용하는 일이 빈번하다. 본 연구에서는 라벨이 없는 데이터 세트에 라벨을 부여하기 위한 라벨러로 GPT-3[14] 모델을 선택했다. GPT-3는 OpenAI에서 개발한 초거대 언어 모델로, 1,750억 개의 파라미터를 통해 문맥 이해 및 생성에 능숙함을 보이며 다양한 자연어 처리 작업에서 인상적인 성능을 보였다. 또한, 사전 연구를 통해 GPT-3가 신뢰할 수 있는 라벨러 역할을 한다는 것이 입증되어[15] 본 논문에서는 데이터 라벨러로서 GPT-3를 사용하고 여기에 대한 성능도 추가로 평가해 보고자 한다.

3. 연구 방법

3.1 연구 개요

그림 1은 데이터 세트 수집부터 모델 미세조정에 이르기까지의 연구 개요를 요약한다. 일반 및 금융 도메인의 뉴스와 트위터 데이터 세트를 수집해 데이터 세트를 구성한다. 이후 GPT-3 모델과 프롬프트 템플릿을 활용해 감성 라벨링을 추출한다. 이 과정에서는 라벨에 대한 설명도 GPT-3 모델이 함께 생성하게 해 라벨의 신뢰도를 강화한다. 마지막으로 라벨링된 데이터 세트로 FinBERT를 미세 조정하고, 다양성이 성능에 미치는 영향을 분석한다.



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

3.2 데이터 세트

3.2.1 데이터 세트 수집

우리는 일반 도메인의 뉴스, 일반 도메인의 트위터, 금융 도메인의 뉴스, 금융 도메인의 트위터 네 가지 데이터 세트를 수집하고 구축하였다. 트위터 데이터는 금융 시장의 역동성을 자유롭게 반영하는 형태의 텍스트 데이터이기 때문에 뉴스 데이터와 함께 선택했다. 모든 데이터 세트는 캐글 (Kaggle)과 허깅페이스 (HuggingFace) 플랫폼에서 수집되었다. 일반 도메인의 트위터 데이터 세트는 2020년의 Tweet Sentiment Extraction 대회와 Tweet Sentiment Dataset 로부터 수집한 5만 개의 트윗을 포함하며, 이는 일반 사용자들이 트위터에 게시한 감성이 잘 드러나는 텍스트로 구성되어 있다. 일반 도메인의 뉴스 데이터 세트는 2017년부터 2020년 사이의 3만 건의 Reuters 뉴스 데이터를 바탕으로 허깅페이스의 Argilla 커뮤니티에 공개된 데이터를 활용하여 구축했다. 일반 도메인 데이터 세트들은 비즈니스, 정치, 문화, 스포츠 등의 다양한 분야를 다루는 뉴스, 그리고 감성이 잘 드러나는 트윗을 포함하고 있다. 금융 도메인 데이터는 금융 시장과 관련된 다양한 정보를 담고 있으며, 주식 시장 동향, 기업의 재무 상황, 경제 지표, 정책 변경 등을 포함하고 있다. 금융 도메인의 뉴스 데이터 세트는 허깅페이스의 Financial PhraseBank 데이터 세트, 2014년부터 2017년 사이의 Auditor sentiment 데이터 세트와 2019년부터 2021년 사이에 미국의 MarketWatch와 SeekingAlpha 등의 플랫폼에서 발행된 3만 건의 뉴스를 수집하여 구축하였다. 금융 영역의 트위터 데이터 세트는 허깅페이스의 Twitter financial sentiment 데이터 세트를 수집했고, 이는 금융 뉴스를 바탕으로 작성된 2만개의 트윗을 포함한다. 이 트윗들은 뉴스에 대한 사용자들의 반응과 견해를 포함한다. 모든 데이터 세트는 전체를 학습, 검증, 추론 데이터 세트로 각각 8:1:1의 비율로 나눠 실험에 사용한다.

3.2.2 데이터 세트 특징 및 전처리

표 1은 수집한 금융 도메인 뉴스 데이터 세트의 일부이다. 이들은 허깅페이스의 Datasets의 Auditor sentiment, Financial PhraseBank, Financial news sentiment 데이터 세트를 수집해 통합 후 중복을 제거해 사용했다. FinBERT의 추가적인 사전 학습에 사용된 데이터 세트와 가장 유사한 형식의 데이터 세트이다.

(표 1) 금융 도메인 뉴스 데이터 샘플.
(Table 1) A sample corpus of our News Data in Financial Domain.

News in Financial Domain
Altia operating profit jumped to EUR 47 million from EUR 6.6 million
Also a six year historic analysis is provided for these markets
Kesko pursues a strategy of healthy focused growth concentrating on sales and services to consumer customers

표 2는 수집한 금융 영역 트위터 데이터 세트의 일부이다. 이들은 허깅페이스의 Datasets의 Twitter financial sentiment 데이터 세트를 수집해 사용했다. 문장 가장 앞에 관련 주식 티커가 있고 트위터 데이터 세트의 특성 상 짧고 간결한 문장이면서 감성을 나타내는 직설적인 단어가 많이 등장한다.

(표 2) 금융 도메인 트위터 데이터 샘플.
(Table 2) A Sample corpus of our Twitter Data in Financial Domain.

Twitter in Financial Domain
Biogen stock price target raised to $392 from $320 at Instinet
CSX stock price target raised to $90 from $72 at Deutsche Bank
$CX Cemex cut at Credit Suisse J.P. Morgan on weak building outlook
$ESS BTIG Research cuts to Neutral

표 3은 수집한 일반 뉴스 데이터 세트의 일부이다. 이들은 허깅페이스의 Datasets의 Argilla news dataset와 news summary dataset중 요약 전 본문 데이터 세트를 수집한 후 통합해 사용했다. 하나의 본문이 굉장히 길어 문장 길이에 대한 전처리가 필수적이다.

표 4는 수집한 일반 도메인 트위터 데이터 세트의 일부이다. 이들은 캐글의 Tweet Sentiment Extraction 대회와 Twitter Sentiment Dataset의 데이터 세트를 수집하여 통합해 사용했다. 길이가 대체적으로 짧고 감성을 나타내는 직설적인 단어를 다수 포함한다.

공통적으로 전처리 과정에서 영어가 아닌 언어, 하이퍼링크, 불용어와 같이 감성 분석에서 불필요한 요소를 제거했고, 너무 길거나 너무 짧은 데이터의 길이를 조정하는 작업을 수행했다.

(표 3) 일반 도메인 뉴스 데이터 샘플.

(Table 3) A sample corpus of our News Data in General Domain.

News in General Domain
Silicon Valley based millionaire lawyer venture capitalist Doug Derwin purchase hoarding urge Tesla CEO Elon Musk quit US President Donald Trump…
Adobe tackle document ID software maker offer new service add digital signature PDF document
Firefox Internet Explorer 39 Henhouse last time heard browser war Well 39 back reason first time seven year Microsoft lose Web browser market share

(표 4) 일반 도메인 트위터 데이터 샘플.

(Table 4) A sample corpus of our Twitter News Data in General Domain.

Twitter in General Domain
entrepreneurs institute is a leading education group founded by Roger Hamilton whose mission is to empower and inspire entrepreneurs to find their flow.
Telangana govt's innovative solution to help farmers affected due to demonetization
Business news and analysis from across the UK, with stories from Reach plc titles like the Birmingham Post, Western Mail, Bristol Post, The Journal, and more.

3.3 GPT를 활용한 감성 라벨링

수집된 데이터는 GPT-3 보다 언어 능력에서 뛰어난 성능을 가진 text-davincii-003 모델의 API을 활용해 라벨링한다. text-davincii-003은 OpenAI에서 제공하는 것으로, 모델의 프롬프트 생성 템플릿 기능을 사용해 zero-shot 감성 분류기를 구축했다. 이 프롬프트 템플릿은 모델이 라벨을 예측할 뿐만 아니라 그 이유를 구체적으로 설명하도록 하게 했다. 이렇게 함으로써 입력값에 대한 모델의 이해 능력을 향상시키고 감성 분석에 대한 추가 연구를 위해 설명이 타당한지 직접 검토할 수도 있다. 본 연구팀은 모델이 작업을 이해하고 정확하게 추론하기 위해 도메인 별로 강건한 프롬프트를 그림 2와 같이 구성한다.

금융 전문가들이 금융 뉴스를 대상으로 각 문장에 수작업으로 감성 라벨을 붙인 Financial PhraseBank 데이터 세트 중 모든 전문가들이 해당 라벨에 만장일치로 동의한 하위 집합을 선택해 샘플 테스트를 실시해 실제 라벨과 예측 라벨을 비교했다. 그 결과, Financial PhraseBank 데이터 세트에 대해 동일한 예측 결과를 보였으며, 이는 GPT-3 모델이 우리 연구에서 신뢰할 수 있는 라벨러로서의 역할을 할 수 있음을 의미한다. 그림 3은 감성 분류 결과의 샘플을 보여준다.

Prompt Template
Classify the sentiment of the given financial text using the JSON format. Consider the financial implications and context when determining sentiment. Use positive, negative, and neutral in lowercase. {"prediction": sentiment label string, "explanation": sentence string describing why you think this is the sentiment considering the financial context}

(그림 2) 금융 영역의 감성 분류 작업에 대한 감성 라벨과 설명을 생성하기 위해 사용되는 프롬프트 템플릿.

(Figure 2) A sample sophisticated prompt template used to generate explanations for the sentiment classification task in financial domain.

	Sentence	Label	Prediction	Explanation
0	According to Gran, the company has no plans to move all productions to Russia, although that is where the company is growing.	1	neutral	The text does not express any sentiment, it is simply providing information.
1	For the last quarter of 2010, Componenta's net sales doubled to EUR 131m from EUR 76m for the same period a year earlier, While it moved to a zero pre-tax from a pre-tax loss of EUR 7m.	2	positive	The text describing a positive financial performance is, with net sales doubling and a move to a zero pre-tax profit from a pre-tax loss, indicating a strong financial performance.
2	In the third quarter of 2010, net sales increased by 5.2% to EUR 205.5 mn, and operating profit by 34.9% to EUR 23.5mn.	2	positive	The text is describing an increase in net sales and operating profit, which is a positive sign.
3	Operating profit rose to EUR 131mn from EUR 8.7mn in the corresponding period in 2007 representing 7.7% of net sales.	2	positive	The text is describing an increase in operating profit, which is a positive sign for the company.
4	Operating profit totalled EUR 21.1 mn, up from EUR 18.6mn in 2007, representing 9.7% of net sales.	2	positive	The operating profit has increased from previous year, indicating a positive financial performance.

(그림 3) GPT-3를 라벨러로 사용한 Financial PhraseBank에서의 분류 예측 및 생성된 설명의 결과.

(Figure 3) Result of classification prediction and generated explanation on Financial PhraseBank using GPT-3 API as annotator.

(표 5) 테스트 데이터 세트에 대한 실험 결과.
(Table 5) Table of our overall experimental results on the test dataset.

Dataset	Epochs	Batch Size	Learning Rate	F1 Score	Accuracy
Twitter in General Domain	6	64	2e-5	0.8588	0.8585
	6	256	2e-5	0.8448	0.8444
	6	64	5e-5	**0.8617**	**0.8620**
	6	256	5e-5	0.8557	0.8550
News in General Domain	6	64	2e-5	0.7861	0.8066
	6	256	2e-5	0.7953	0.8063
	6	64	5e-5	0.7794	0.8050
	6	256	5e-5	**0.7962**	**0.80963**
Twitter in Financial Domain	6	64	2e-5	0.9032	0.9034
	6	256	2e-5	0.8809	0.8816
	6	64	5e-5	**0.9191**	**0.9197**
	6	256	5e-5	0.9029	0.9034
News in Financial Domain	6	64	2e-5	**0.9188**	**0.9186**
	6	256	2e-5	0.889	0.8898
	6	64	5e-5	0.9130	0.9126
	6	256	5e-5	0.8843	0.8838

3.4 미세 조정

마지막으로 라벨이 부여된 데이터 세트로 미세 조정한다. 이때 본 연구팀은 모델 수준에서 하이퍼파라미터를 미세 조정하는 것이 아닌 서로 다른 데이터 세트와 하이퍼파라미터 간의 관계를 알고자 하이퍼파라미터를 변화시켜가며 사용하는 것에 초점을 둔다. 이를 통해 다양한 데이터 세트와 하이퍼파라미터가 금융 데이터 세트로 학습된 모델의 성능에 어떤 영향을 미칠 수 있는지에 대한 새로운 인사이트를 찾는 것을 목표로 한다.

4. 실험 설정

4.1 실험 환경

본 실험에서는 모델의 학습에 심층 학습을 위한 프레임워크인 Tensorflow와 Keras를 사용한다. CPU는 Intel® Core i7-13700K RAM 128GB를 사용했고, GPU는 NVIDIA RTX A6000 (VRAM 48GB) 모델 2개를 사용하였으며 GPU 병렬 연산 기능을 돕는 NVIDIA의 CUDA v11.7과 cuDNN v8.6.0을 활용하였다. 실험에서는 허깅페이스에 우리의 데이터 세트를 업로드해 사용했고, 모델 학습 시 고수준 API를 제공하는 자체 트레이너 클래스를 개발하여 하이퍼파라미터 튜닝을 용이하게 했다. 실험에서 일관된 결과를 보장하기 위해 모든 random seed를 고정하였다.

본 실험에서는 FinBERT 모델의 기본 하이퍼파라미터 설정인 학습률 2e-5, 배치 크기 64를 따른다. 여기에 우리는 좋은 컴퓨팅 환경이라면 더 큰 배치 크기를 사용하는 것이 옳은지 확인하기 위해 기본 설정보다 더 큰 배치 크기인 256과 미세 조정에서의 반영을 높이기 위해 더 큰 학습률인 5e-5를 추가로 도입해 데이터 세트에 가장 적합한 하이퍼파라미터 설정을 찾아보았다. 모든 실험은 FinBERT 모델의 미세 조정 실험과 같은 6 에포크 동안 진행되었으며, 이는 총 10시간의 학습시간을 필요로 했다. 나머지 모든 하이퍼파라미터는 FinBERT 모델의 설정을 따른다. 이렇게 함으로써 FinBERT 모델과 비교하여 우리의 접근법이 얼마나 효과적인지를 더욱 정확하게 평가할 수 있다.

4.2 평가 척도

본 연구는 정확도와 Macro F1 score를 주요 지표로 삼아 모델 성능을 평가한다. 정확도는 모델의 전반적인 분

류 성능을 나타내는 척도로, 전체 데이터 수 중 예측 결과와 실제 값이 동일한 수가 차지하는 비율이다. F1 score는 모델이 참이라고 분류한 것 중 실제 참인 것의 비율인 정밀도와 실제 참인 것 중 모델이 참이라고 예측 한 것의 비율인 재현율의 조화평균으로, 클래스 불균형이 있는 경우 사용하면 좋은 평가 척도다. 이중 다중 클래스 분류에 사용되는 Macro F1 score는 각 클래스의 F1 score를 계산하여 평균을 취한 값이다.

5. 실험 결과

표 5에서 금융 도메인 뉴스와 금융 도메인 트위터 데이터 세트가 각각 0.9186과 0.9197로 가장 높은 정확도를 달성했다. 이는 금융 관련 키워드가 밀집되어 있고 'bullish' 또는 'bearish' 등 강한 감성을 가지는 단어가 많아 이들이 정확한 감성 예측에 중요한 역할을 하기 때문이라고 추측한다.

일반 도메인 뉴스에 비해 일반 도메인 트위터로 미세 조정했을 때 더 나은 성능을 보였다는 점도 흥미로웠다. 이는 트위터 데이터의 감성 분석에 대한 잠재력을 보여주는 것으로, 'increased', 'decreased', 'upgraded', 'downgraded'와 같은 감성과 방향성의 의미를담고 있는 키워드가 상대적으로 다수 포함된 트위터의 비공식적이고 간결한 특성이 모델이 분류 작업을 더 쉽게 수행할 수 있게 단순화시키는 것으로 보인다.

본 연구팀은 4개의 데이터 세트별로 배치 사이즈를 64, 256으로, 학습률을 5e-5, 2e-5 각각 두어 총 16가지의 환경에서 실험했다. 실험 결과, 데이터 세트 별로 약간의 차이가 있지만 배치 크기를 64로, 학습률을 5e-5로 두고 미세 조정을 하는 것이 전체적으로 가장 견고한 성능을 보인다. 이는 배치 크기를 키워도 될 만큼의 학습 환경을 가졌더라고 성능 면에서는 배치 크기가 작을수록 더 효과적으로 학습할 수 있음과, 이와 함께 더 공격적으로 학습률을 설정하는 것이 미세 조정에 유리하다는 것을 시사한다.

또 금융 도메인 뉴스를 사용한 실험에서 배치 크기를 64, 학습률을 2e-5로 설정했을 때 가장 높은 성능을 달성했다. 이러한 결과가 나온 이유는 FinBERT 모델이 금융 뉴스 데이터로 추가적인 사전 학습을 진행할 때와 동일한 하이퍼파라미터를 사용했다는 것으로 추측된다. 이는 사전 학습과 추가적인 사전 학습, 또 미세 조정시 사용하는 하이퍼파라미터들을 동일하게 사용한다면 더 좋은 성능을 보일 수 있다는 인사이트를 제공한다. 전반적으로 실험을 통해 데이터 세트 선택과 하이퍼파라미터 튜닝의 중요성을 확인할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 금융 도메인 위에서 사전 학습된 FinBERT 모델을 대상으로 데이터 세트와 하이퍼파라미터가 미세 조정에 어떠한 영향을 미치는지 조사했다. 실험 결과, 금융 도메인 데이터 세트에서 트위터와 뉴스로 미세 조정했을 때 가장 높은 성능을 보여 일관적인 도메인 유지의 중요성을 보였다.

한편, 일반 도메인 데이터 세트에서 뉴스보다 트위터로 미세 조정하는 것이 훨씬 뛰어난 성능을 보였다. 따라서 대용량 말뭉치에 대해 사전 학습이나 추가적인 사전 학습을 진행할 때 FinBERT 모델과 같이 특정한 도메인의 뉴스에 국한된 데이터 세트가 아닌 트위터 등의 SNS 데이터 세트를 함께 활용하는 것이 더 좋은 모델 성능을 기대할 수 있을 것이라 본 연구팀은 제안한다.

이번 연구에서 가장 효과적인 하이퍼파라미터는 5e-5의 학습률과 64의 배치 크기다. 이와 함께 데이터 세트와 하이퍼파라미터에 따라 성능에 현저한 차이가 있음을 관찰한 이번 연구는 감성 분석을 위한 최적의 하이퍼파라미터를 선택하는 것이 얼마나 중요한지 보여준다. 마지막으로, GPT-3를 라벨러로 삼아 분석 작업에 필요한 감성을 효과적으로 추출하는 방법을 시연해 성능을 입증했다.

향후 작업에서는 미세 조정 성능만 비교했던 본 연구와는 다르게 BERT 모델의 구조만을 가져와 다양한 데이터 세트를 사전 학습을 하여 모델의 성능 변화를 관찰하고자 한다. 이와 함께, 뉴스나 트위터 외 사용할 만한 다른 종류의 데이터 세트를 찾아 함께 비교해 본 연구의 인사이트를 확장하는 데 집중할 것이다. 또 금융 분야 감성 분석은 주로 주가 예측과 연결된다. 비록 각 데이터 세트마다의 일별 데이터 수집에 한계가 있어 감성 분석만의 성능 비교에 그쳤지만, 주가 예측을 위한 도구로 감성 분석을 사용할 때 본 연구의 결과를 활용할 수 있게 하는 것이 향후 연구를 위한 탐구로써의 잠재력을 가지고 있다고 생각한다.

참고문헌(References)

[1] 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

[2] HOANG, Mickel; BIHORAC, Oskar Alija; ROUCES, Jacobo. “Aspect-based sentiment analysis using bert,” In: Proceedings of the 22nd nordic conference on computational linguistics. p. 187-196, 2019. https://aclanthology.org/W19-6120

[3] Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D., “Evaluation of sentiment analysis in finance: from lexicons to transformers” IEEE access, 8: 131662-131682, 2020. https://doi.org/10.1109/ACCESS.2020.3009626

[4] Devlin, J., Chang, M., Lee, K., & Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[5] Sun, C., Qiu, X., Xu, Y., & Huang, X., “How to fine-tune bert for text classification?,” In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18 (pp. 194-206). Springer International Publishing, 2019. https://doi.org/10.48550/arXiv.1905.05583

[6] Vinodhini, G., and R. M. Chandrasekaran, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012. https://www.researchgate.net/publication/265163299_Sentiment_Analysis_and_Opinion_Mining_A_Survey

[7] 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, 2014. https://doi.org/10.1016/j.knosys.2014.04.022

[8] Loughran, Tim and Bill Mcdonald., “Textual Analysis in Accounting and Finance: A Survey,” Journal of Accounting Research, 2016, 54.4: 1187-1230, 2016. https://doi.org/10.1111/1475-679X.12123

[9] Sparck Jones, Karen. “A statistical interpretation of term specificity and its application in retrieval,” Journal of documentation 28.1: 11-21, 1972. https://doi.org/10.1108/eb026526

[10] Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T.M., “Big Data: Deep Learning for financial sentiment analysis,” Journal of Big Data, 5.1-25, 2018. https://doi.org/10.1186/s40537-017-0111-6

[11] Hochreiter, Sepp, and Jürgen Schmidhuber. ″Long short-term memory″, Neural computation 9.8: 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[12] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I., “Improving language understanding by generative pre-training,” 2018. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., “Attention is all you need,” Advances in neural information processing systems, 30., 2017 https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[14] Brown, T.B., Mann, B., Ryder, N., Sub-biah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D., “Language Models are Few-Shot Learners,” Advances in neural information processing systems, 33, 1877-1901, 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

◐ 저 자 소 개 ◑

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

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

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