Journal of the Korean Institute of Industrial Engineers
Vol. 50, No. 4, pp. 240-250, August 2024.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2024.50.4.240

금융 특화 감정분석 모델과 딥러닝 시계열 예측 모델을 활용한 코스피 지수 예측

정가연[1] · 이혁제[2] · 이준영[2] · 이제혁[2†]

[1]국민대학교 경제학과, [2]국민대학교 AI빅데이터융합경영학과

Kospi Index Prediction using a Financial-specific Sentiment Analysis and Deep Learning-based Time Series Prediction Model

Gayeon Jung[1] · Hyeokje Lee[2] · Junyeong Lee[2] · Jehyuk Lee[2]

[1]Department of Economics, Kookmin University
[2]Department of AI, Big Data & Management, Kookmin University

This paper presents a methodology for predicting the KOSPI index using a news data-based sentiment analysis model and a deep learning-based time series prediction model. The closing price of the KOSPI index was used as a target variable, and macroeconomic indicators such as the gold price and market sentiment indicators such as sentiment scores were used as independent variables. We collected and preprocessed the KOSPI-related news data and used them in calculating the sentiment score by using the title or the summarized article. Subsequently, the KLUE-BERT model-based sentiment score by date and the KoFinBERT model-based sentiment score by date were extracted. LSTM, GRU, CNN-LSTM, and CNN-GRU were used as time series prediction models. As a result of conducting an experiment by combination of variables and models, the best performance was achieved when KLUE-BERT is applied on the summarized article and the CNN-GRU model were used.

Keywords: Deep Learning, BERT, Sentiment Analysis, LSTM, Kospi Index Prediction

1. 서 론

COVID-19 기간을 거치게 되면서 주식 시장의 변동성이 커지게 되었고 투자자들이 다수 유입되면서 주식 시장에 대한 관심이 커지고 있다. 변동 추세가 점점 복잡해지고 있는 주식 시장에서 안정적인 수익률을 얻기 위해서는 주식 시장의 거시적인 흐름을 파악하는 것이 중요하다. 즉, 코스피 지수의 전반적인 추세를 예측할 수 있다면 투자에 큰 도움을 줄 수 있을 것이다. 효율적 시장 가설(Efficient Market Hypothesis)이란 주식 시장의 가격이 형성될 때, 시장에 존재하는 모든 정보를 이미 내포하고 있다는 이론이다(Fama, 1965). 즉, 주가 외에도 환율, 금리와 같은 경제 지표, 재정 상태 및 경영 이슈 등 기업의 사업 현황도 주가에 영향을 주게 된다. 그러나 기존의 주가 예측 관련 연구들은 주로 기술적 지표 기반의 방법론을 활용해왔다(Agrawal *et al.*, 2019). 기술적 분석은 과거 주가 데이터를 활용해서 미래의 주가를 예측하는 방법론으로, 강한 추세를 보이는 종목에 투자하는 모멘텀 팩터 투자(Momentum Factor), 저평가 종목에 투자하는 밸류 팩터 투자(Value Factor) 등이 있다. 기존 예측 방법론은 복잡한 주식 시장의 변동성을 잘 설명하지 못하여 주가 예측에 한계가 있다(Kang, 2022).

최근에는 컴퓨팅 자원과 자연어를 처리하는 인공지능 기술이 발전하면서 SNS, 뉴스, 보고서 등 텍스트 분석과 관련한 연구가 활발하게 진행되고 있다(White, 1988; Schumacher and Chen, 2009; Han, 2021). 기업과 관련한 뉴스를 수집한 후 이를

† 연락저자 : 이제혁 교수, 서울특별시 성북구 정릉로 77 국민대학교 경영대학 AI빅데이터융합경영학과, Tel : 02-910-4537,
E-mail: jehyuk.lee@kookmin.ac.kr
2024년 1월 12일 접수; 2024년 2월 5일 게재 확정.

통해 기업에 대한 전반적인 긍정적 혹은 부정적 감정을 파악할 수 있다면 주가 예측에 도움이 되는 팩터로 활용할 수 있을 것이다. 뉴스 제목을 이용하여 감정 분석을 하고 시계열 데이터와 감정분석 결과를 함께 활용하였을 때 성능이 대폭 향상되었다(Jang, 2020). 지수 예측을 위해 단순 기술적 분석뿐만 아니라 텍스트 마이닝을 이용한 감정 분석을 활용한 연구가 진행되고 있다. 이 때, 뉴스 기사, 댓글, 소셜네트워크에 작성한 포스팅 등 다양한 데이터가 활용된다. 여기서, 뉴스 기사의 경우 일반적으로 본문의 양이 방대하여 주로 제목만을 사용하여 감정분석을 수행하는데, 이러한 접근 방식은 정확성이 제한적일 수 있다. 또한 감정분석 연구는 주로 단어별 감정 사전을 기반으로 진행되었으며, 이런 방식은 문맥을 제대로 고려하지 못하는 한계점이 있었다.

본 논문에서는 다양한 거시 경제 지표들과 주식 시장 관련 뉴스 데이터를 활용하여 주가 지수를 예측하는 방법론을 제안한다. 먼저, 주식 시장 관련 뉴스 데이터에 금융 특화 감정분석 모델을 활용하여 해당 뉴스 데이터의 감정 점수를 산출한다. 기존 연구의 한계점을 보완하고자 감정 분석을 수행할 때 뉴스 제목이 아닌 본문을 활용한다. 본문을 한 문장으로 요약한 후에 감정 분석을 진행하며, 단어별 딕셔너리 형태 대신 문맥을 고려한 한국어 BERT 기반 모델을 활용한다. 산출된 감정점수를 환율, 금값, 유가 등 거시경제 지표 데이터, 과거 주가 지수 데이터와 함께 결합하여 주가 지수 값을 예측하는 방법론을 제안한다. 그리고 이를 코스피 지수의 예측에 활용하여 제안 방법론을 검증한다. 논문의 구성은 다음과 같다. 제1장 서론에 이어서 제2장에서는 본 논문과 관련된 연구 들을 소개한다. 제3장에서는 본 논문에서 제안하는 연구 방법, 제4장에서는 정형 데이터와 비정형 데이터의 수집 및 처리와 함께 제안 방법들을 적용할 때 설정한 실험 환경을 소개한다. 제5장에서는 실험 결과, 마지막 제6장에서는 결론과 향후 연구 방향을 기술한다.

2. 관련 연구

2.1 통계적 모델을 활용한 주가 예측 연구

주가 데이터는 시계열 데이터라는 점에 착안하여 과거에는 시계열 예측 방법론을 활용하여 주가 예측을 하는 연구들이 수행되었다. 특히, 시계열 데이터에서 현재의 상태가 과거의 상태에 의존한다는 가정하에 시계열을 예측하는 자기회귀과정(Autoregressive Process)에서 파생된 방법론들을 많이 활용하였다. 대표적으로 자기회귀(Auto-Regressive, AR model), 이동평균(Moving Average, MA), 자기회귀이동평균(Auto-Regressive Moving Average, ARMA), 자기회귀누적이동평균(Auto-Regressive Integrated Moving Average, ARIMA) 모델이 존재한다. AR모델은 현재 시점의 값을 종속 변수, 과거 시점의 값들은 독립 변수로 설정하여 독립 변수의 선형조합으로 종속 변수의 값을 예측하는 모델이다(Yule, 1927). MA모델은 현재 시점의 값을 종속 변수, 과거 시점의 백색 잡음(white noise, $\epsilon_t \sim N(0,\sigma^2)$)들을 독립변수로 설정하여, 독립변수의 선형 조합으로 종속 변수의 값을 예측하는 모델이다. ARMA 모델은 AR모델과 MA모델에서 사용한 독립 변수들을 모두 활용하는 모델이다. 위 모델들은 전부 시계열 데이터가 시간과 관계없이 평균과 분산이 일정한 정상성(stationary)을 가정하므로 현실 데이터에는 적용하기 어렵다는 한계가 존재한다. 이러한 한계를 보완한 ARIMA모델은 비정상성(nonstationary) 시계열 데이터를 정상성을 나타내는 데이터로 변형한 뒤, ARMA 방법론을 적용한 모델이다(Box *et al.*, 2015).

위의 시계열 데이터 예측 모델을 활용하여 주가지수, 수익 등의 값을 예측하려는 다양한 연구가 수행되었다. ASEAN-5 시장의 수익을 예측한 연구에서는 AR모델과 비즈니스 주기에 맞게 제안된 모델인 Smooth Transition Auto-Regressive(STAR) 방법론을 활용하여 수익이 확률보행과정(Random Walk Process)적 특성이 나타나지 않음을 보였다(Liew *et al.*, 2003). 기술적 분석이 실제로 수익이 좋은 지 평가한 연구에서는 MA 모델을 활용하여 싱가포르 주가 지수의 값을 예측하는 과정을 수행하였다(Wong *et al.*, 2003). 주가 지수가 아닌 개별 종목의 주가를 예측한 연구도 수행되었는데, 이 연구에서는 ARIMA 모델을 활용하여 인도 주식시장에 존재하는 56개 종목의 주가를 예측하여 좋은 성능이 나타남을 확인하였다(Mondal *et al.*, 2014). 그러나, 이와 같은 통계적 모델 기반의 방법론은 데이터가 정상성, 주기성 등의 특징을 나타낸다는 가정이 필요하므로, 실제 데이터와는 맞지 않는 부분이 있다. 또한, 독립 변수들의 선형 조합으로 종속 변수들을 예측하기 때문에 모델의 비선형 모델을 구축하기 어렵다는 문제점이 존재한다.

2.2 딥러닝 모델을 활용한 주가 예측 연구

최근에는 시계열 데이터를 효율적으로 예측하는 신경망 구조 기반의 딥러닝 모델 주가 예측 연구들이 수행되고 있다. 순환신경망(Recurrent Neural Network, RNN)이란 일반 신경망 구조에 시계열 데이터 개념이 추가된 구조로 은닉층에 존재하는 과거 정보를 활용한다(Elman, 1990). LSTM(Long-Short Term Memory)은 RNN의 한 종류로 시퀀스 데이터를 처리하고 장기 의존성(Long-Term Dependencies)을 효과적으로 학습할 수 있으며, memory cell, input gate, forget gate, output gate로 구성되어 있다. LSTM은 memory cell을 통해서 각 시간 단계에서 내부 상태를 유지하고 필요한 정보를 유지함으로써 장기 의존성 문제를 해결했다(Hochreiter and Schmidhuber, 1997). GRU(Gated Recurrent Unit)는 RNN의 한 종류로 update gate와 reset gate로 구성된다. LSTM보다 더 간단한 구조를 가지고 있으며, 계산 비용이 적게 들지만 비슷한 성능을 보이며 효과적으로 작동한다(Chung *et al.*, 2014). LSTM 기반의 시계열 예측 모델을 학습할 때 합성곱 연산을 수행하는 층(convolution layer)을 결합한 연구가 있으며 단순

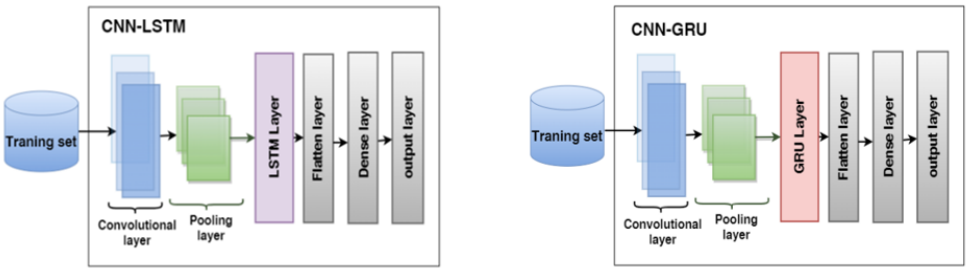

Figure 1. The Architecture of CNN+LSTM & CNN+GRU (Hwang and Shin, 2020)

LSTM을 활용했을 때보다 높은 예측 정확도를 가진다는 결과가 나왔다(Hwang and Shin, 2020). LSTM 및 GRU는 주기성이나 특정 시간 스케일의 특징을 감지하는 데 어려움이 존재한다. LSTM에 데이터를 입력하기 전 합성곱 연산을 수행하면 주요한 시간적 패턴을 학습할 수 있다. 이를 통해 보다 유용한 표현을 제공할 수 있으며 시계열 데이터에 포함된 잡음이나 이상치를 줄일 수 있다.

신경망 모델을 사용하여 금융 시계열 예측 모델링을 처음 시도한 연구에서는 다양한 딥러닝 모델을 적용해서 IBM의 주식 가격 변동성에서 패턴을 찾아내는 신경망 모델링을 진행하였고 효율적 시장 가설에 대한 증거를 확립하였다(White, 1988). 또한 주가 예측 모델로서 LSTM을 적용할 때 성능향상을 위해 고려해야 할 다양한 파라미터 설정과 함수들에 대한 적절한 조합 방법을 제안한 연구에서는 주가 예측을 위한 LSTM 적용 시 최적의 모델링 방법을 실증적인 형태로 제안하였다(Jung and Kim, 2020). CNN 기반 프레임워크를 활용해서 주가를 예측하기 위한 특징을 추출한 연구에서는 S&P 500, NASDAQ, Dow Jones Index(DJI) 시장 지수의 다음 날 이동 방향을 예측하였고 기존 알고리즘에 비해 예측 성능이 크게 증가하였다(Hoseinzade and Haratizadeh, 2019). 이러한 방법론들은 기존 통계적 모델을 활용한 모델보다 복잡한 모델을 사용할 수 있고, 모델링을 하기 위한 가정이 상대적으로 덜 필요하다는 장점이 있다. 그러나, 주가 예측시 시장 참여자들의 반응들을 전혀 고려하지 못하고 과거 주가 데이터만 활용한다는 한계가 존재한다.

2.3 감정 분석 모델을 활용한 주가 예측 연구

최근 기술의 발전에 따라 뉴스, SNS, 기업 보고서와 같은 텍스트 데이터 기반의 감정점수를 활용해서 주가 예측 연구에 활용하려는 시도가 늘어나고 있으며 기업의 주가와 뉴스 기사 간의 상관관계가 있음을 알 수 있다(Schumacher and Chen, 2009). 시간적 특성을 고려한 뉴스 데이터의 감성 분석을 이용한 딥러닝 기반 주가 예측 모델을 제안한 연구에서는 시간 크기와 감성 지표가 주가 예측에 미치는 영향에 대해 비교 및 분석을 진행하였다(Kang *et al.*, 2022). 텍스트에 대해서 감정점수를 매핑해주는 감정분석 모델에는 사전학습 모델인 BERT, 금융 corpus 특화 사전학습 모델인 FinBERT, FinBERT의 한국어 버전 모델인 KoFinBERT가 존재한다. KLUE-BERT는 대용량의 한국어 말뭉치를 수집 및 전처리한 후 Pre-Trained된 한국어 BERT 모델이다(Park *et al.*, 2021). 사용된 한국어 말뭉치는 473M 문장, 6.5B token으로 구성되어 있고, 목적에 맞게 Fine-Tuning을 진행한다. 이를 활용하여 뉴스 헤드라인 분류, 문장 유사도 비교, 자연어 추론, 개체명 인식, 관계 추출, 형태소 및 의존 구문 분석 등 다양한 작업을 수행할 수 있으며, 각 작업을 수행하기 위해서는 그에 맞는 학습 데이터로 Fine-Tuning을 수행한 후 활용한다. FinBERT는 Pre-Training된 BERT 언어모델에 더해 금융 특화 말뭉치로 Futrhter Pre-Training을 수행한 모델로 금융 도메인에서 NLP 관련 task를 수행한다(Araci, 2019). KoFinBERT는 FinBERT 모델의 개념을 한국어 task에 적용해 본 모델이다. KoFinBERT는 Pre-Training된 언어모델을 활용해서 금융 도메인 감정분류 task에 Fine-Tuning된 모델이다(Cho, 2023). KoFinBERT는 KoBERT 모델을 사용하여 재무 회계 관련 신문기사의 감정분석 정확도를 테스트한 연구를 기반으로 한다(Hyeon *et al.*, 2022). KoFinBERT는 사전 훈련된 ko-sober-ta-multitask(Gan, 2022) 모델로 임베딩을 진행한 후, 회계 재무 관련 기사에 대해서 금융 전문가들이 긍정/부정/중립을 직접 라벨링한 데이터셋을 기반으로 BERT기반 감정 분류를 수행하는 모델이다. “부채가 증가하고 있다”라는 문장이 입력되면 경영전략의 실패로 부채가 늘어난 것인지, 투자의 증가로 늘어난 것인지 문맥을 파악한 뒤 감정을 분류한다.

미국 다우존스지수 예측의 정확도 향상을 위해 뉴스 정보의 감정점수와 거시경제지표의 조합을 딥러닝 예측 모델을 통해 제시한 연구에서는 BERT 기반의 감정분석 모델링을 진행하였고 감정점수를 포함한 효과적인 지표 조합을 제시했다(Jang, 2020). 이와 비슷하게 NASDAQ 주가 예측을 위해 FinBERT로 New York Times의 일자별 감정점수를 도출하고 LSTM으로 시계열 예측을 수행한 연구가 있다. 해당 연구를 통해서 금융 감정 점수를 도출하였고 단일 예측 모델보다 좋은 예측 성능의 결과를 보였다(Shayan, 2022).

3. 연구 방법론

본 논문에서는 정형 데이터인 주가 및 거시경제 지표 데이터와 비정형 데이터인 뉴스 데이터를 활용하여 딥러닝 모델 기반 코스피 지수 예측 모델을 구현한다. 본 연구에서 제안하는 주가

지수 예측 모델에 대한 전체적인 구조도는 <Figure 2>와 같다. 먼저, 일자별 뉴스 데이터를 수집하여 전처리 및 요약 과정을 수행한 뒤, 이를 감정 분석 모델인 KLUE-BERT, KoFinBERT에 입력하여 뉴스의 감정 점수를 산출한다. 또한, 일자별 거시경제 지표와 코스피 지수 거래 내역 데이터를 수집하고 이를 전처리한 결과물을 산출한다. 그리고 위 두 개의 전처리 결과를 시계열 데이터 예측 모델의 일자별 입력값으로 활용하여 예측 모델링을 수행한다.

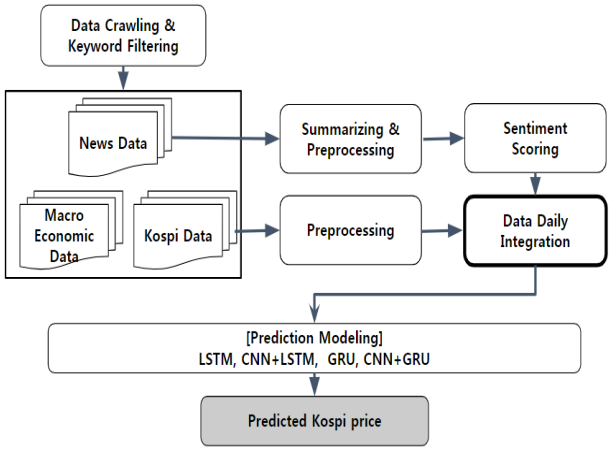

Figure 2. The Overall Architecture

3.1 뉴스 데이터 감정 지수 도출

(1) 뉴스 본문 요약 및 전처리(Summarization & Preprocessing)

PORORO(Platform Of neuRal mOdels for natuRal language prOcessing)를 활용하여 뉴스 본문 내용을 한 문장으로 축소하는 과정을 수행한다. 해당 라이브러리는 KaKaoBrain에서 개발한 한글 자연어처리 라이브러리로 Text Classification, Sequence Tagging, Seq2seq, Misc기능을 지원한다(Heo *et al.*, 2021). 이 중 Seq2seq의 Text Summarization을 활용했고, Abstractive summarization을 task로 설정하여 여러 문장을 한 문장으로 요약하는 기능을 사용한다.

(2) 감정 지수 도출(Sentiment Scoring)

뉴스 텍스트 데이터 감정 분석을 진행하기 위해 KLUE-BERT 모델과 KoFinBERT 모델을 활용한다. KLUE-BERT는 사전 학습된 한국어 BERT 모델로, 금융 도메인에 그대로 사용하기에는 무리가 있다. 따라서 금융 뉴스 기사와 함께 긍정/부정/중립으로 라벨링된 데이터셋을 활용한다(Malo *et al.*, 2014). 기존 데이터셋이 영문 버전으로 되어 있어, 번역 후 육안으로 검수된 한글 데이터셋을 이용하여 Fine-Tuning을 진행한다(Yoo, 2022). 이후 금융 감정 데이터셋으로 Fine-Tuning된 모델을 활용해서 뉴스 텍스트 데이터의 감정 점수를 도출한다. 각 긍정/부정/중립에 대한 확률값을

[Original article]
'President Moon Jae-in's Blue House▲Major Group Heads Emphasize Securing New Growth Engines in the New Year by "Turning Crisis into Opportunity" The heads of major conglomerate groups have emphasized securing new future growth engines in the new year 2021, saying they should turn the looming crisis into an opportunity amid growing uncertainty. They emphasized that change and innovation are urgently needed to survive in the global business environment, which has been disrupted by the novel coronavirus (COVID-19). Samsung Electronics Vice Chairman Lee Jae-yong began his first management action of the new year on the first working day of the new year by visiting a semiconductor business without delivering a separate New Year's speech. Samsung Electronics Vice Chairman Lee Jae-yong began his first management action of the new year on the first working day of the new year by visiting a semiconductor business without a separate New Year's speech. After attending a ceremony to inaugurate the foundry (semiconductor Wita Gongsan) production facility at Pyeongtaek Plant 2 in Gyeonggi Province, he reviewed the mid- and long-term strategy with the semiconductor division presidents. The Pyeongtaek 2 plant is an advanced complex production line that produces D-RAM, next-generation V-NAND, and ultra-fine foundry products. After producing memory semiconductors last year, the plant began to deliver facilities for foundry production this year. Vice Chairman Lee started his first management action in the new year. After attending the ceremony to inaugurate the foundry (semiconductor contract manufacturing) production facilities at the Pyeongtaek 2 plant in Gyeonggi-do, he reviewed the mid-term strategy with the semiconductor division presidents. The choice of Pyeongtaek Plant 2 as a high-tech complex that produces D-RAM, next-generation v-nand, and ultra-fine foundry products emphasizes the importance of the semiconductor business, which is its flagship business, and is an expression of its intention to further consolidate its position in the semiconductor business by improving the competitiveness of ultra-gap technology products. President Moon Remains Silent on Amnesty TheoryPresident Moon Jae-in has remained silent on former President Lee Myung-bak's amnesty theory and the COVID-19 outbreak at the Seoul East Detention Center. Instead, the Cheong Wa Dae has taken the lead in clarifying Moon's intentions. While opposition parties have been demanding that Moon clarify his stance on both issues, he is unlikely to do so for the time being. According to the Cheong Wa Dae on Thursday, Moon has no separate position on pardoning the two former presidents. The reasoning is that it is too early to discuss pardons as only Lee's sentence has been finalized. Park's sentencing by the Supreme Court is scheduled for Sept. 14.
Translated with www.DeepL.com/Translator (free version)

[Article Summary]
'Samsung Electronics Vice Chairman Lee Jae-yong kicked off the new year on Thursday with a visit to a semiconductor plant without a separate New Year's speech, and the KOSPI hit a record high in the afternoon as retail investors turned net buyers in the stock market.'

Figure 3. Examples of News Data Summary Using Pororo

Date	Content	positive	negative	neutral	Sentiment score
20210104	Samsung, Hynix Report Side-by-side…Semiconductor Strength	0.2	0.7	0.1	-0.5
20210104	Real estate market outlook negative	0.3	0.5	0.2	-0.2
20210104	KOSPI super strong on first trading day of the new year…above 2,900	0.5	0.4	0.1	0.1

Date	Daily Sentiment Score
20210104	0.5
20210105	0.3
20210106	-0.1

Figure 4. The Example of Sentiment Score by KLUE-BERT

Date	Content	Sentiment Score
20210104	Samsung, Hynix Report Side-by-side…Semiconductor Strength	1
20210104	Real estate market outlook negative	-1
20210104	KOSPI super strong on first trading day of the new year…above 2,900	1

Date	Daily Sentiment Score
20210104	0.33
20210105	0.31
20210106	-0.17

Figure 5. The Example of Sentiment Score by KoFinBERT

긍정-부정 점수로 변환하여 최종 점수로 설정한다. 뉴스 데이터는 총 4종류의 데이터를 활용하였으며, 각각 (1) 본문을 한 문장으로 요약한 데이터, (2) 제목 데이터, (3) 본문을 한 문장으로 요약한 후 전처리한 데이터, (4) 제목과 본문을 한 문장으로 요약한 데이터이다. 이후 각 뉴스 데이터별로 감정 점수를 도출하였고, 이를 통합하여 일일 감정점수를 도출하였다. 주식 시장 마감 시간인 15시 30분 이후 데이터들과 다음날 15시 30분까지의 데이터를 평균으로 통합하여 구성한다. 그 후 Z-score 정규화를 통해서 평균은 0, 표준편차는 1로 조정해준다. KoFinBERT 기반의 TextClassification Pipeline으로 뉴스 텍스트 데이터의 일자별 감정 점수를 도출한다. 모델의 output인 부정은 -1, 중립 0, 긍정은 +1로 매핑한 후 더해줌으로써 감정 점수를 계산한다. KLUE-BERT와 마찬가지로 총 4가지 버전의 감정 점수를 도출하였고, 뉴스 데이터별로 도출된 점수를 일별로 통합하는 과정을 거친다. 그 후 Z-score 정규화를 통해서 평균은 0, 표준편차는 1로 조정해준다.

3.2 주가 지수 예측 모델링

본 연구는 주가지수 및 일별, 월별 경제 지표 데이터 그리고 산출된 감정점수를 활용하여 코스피 지수 예측 모델을 구현한다. 전체적인 모델의 흐름도는 아래의 <Figure 6>과 같다. 사용되는 독립변수는 코스피 지수의 시가, 종가, 국제 유가, 금값, 환율, 감정점수에 해당하는 일별 변수와 물가지수와 같은 월별 변수가 있다. T시점의 코스피 예측값을 산출하기 위해서 독립 변수들을 5일치씩(T시점으로부터 T-5시점) 결합해서 구성하고 종속 변수는 5일 후(T시점)의 코스피 지수의 종가를 사용하여 구성하며 두 데이터를 합쳐 최종 데이터를 만든다. 이 때, 경제 지표 등의 값들은 모든 변수를 활용하지 않고 다중 공선성이 높은 변수와 중요도가 낮은 변수를 제거한다. 그 후 RNN 계열의 LSTM, GRU 모델로 학습 및 테스트 과정을 거쳐 과거 5일의 데이터를 바탕으로 다음 날의 주가를 예측하는 모델링을 수행한다.

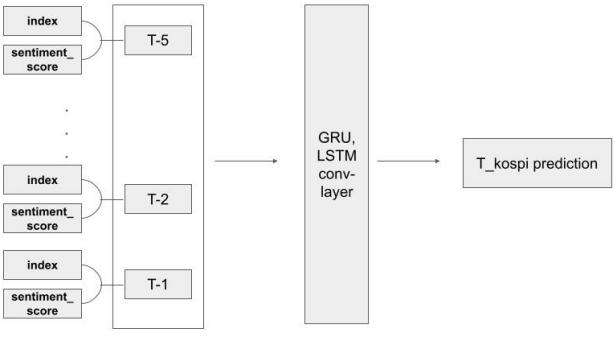

Figure 6. Prediction Model Flow

4. 실 험

4.1 거시경제 지표 및 주가 지수 데이터 수집 및 전처리

코스피 지수, 미국 주가 지수 및 환율, 유가 등 거시경제 지표들을 수집하기 위해 FinanceDataReader를 활용하였다. 주가 지수 데이터는 일자별 시가(open), 종가(close), 고가(high), 저가(low), 거래량(volume)을 수집하였다. 또한, 사람들의 관심도 추이를 반영하기 위해 Naver DATALAB을 이용하여 특정 검색어가 미리 지정된 기간 내에 검색된 비율 데이터를 수집하였다. 해당 데이터는 기간 내에 검색 수가 가장 많았던 날을 기준(100)으로 하여 일일 검색량을 상대적으로 표현한 데이터이다. 또한, 한국은행 api를 이용하여 월별 거시경제 지표 데이터(종합소비자물가지수, 이자율, 부동산지수)를 수집하였다. 단, 해당 데이터는 월 단위로 되어 있기 때문에, 일단위로 변환하는 과정을 수행하였다. <Table 1>은 이러한 과정으로 수집한 데이터 내의 변수 이름 및 설명을 나타낸다.

기업들의 주가 범위와 경제 지표의 범위가 다르므로 수치 데이터에 대한 전처리 작업으로 스케일링(Scaling)을 진행하였는데, 이는 데이터의 스케일을 맞춰 가중치의 스케일도 일정하게 맞춰주는 효과를 부여하기 위함이다. 사용한 방식은 데이터의 최소값을 0, 최대값을 1로 두는 MinMax Scaling으로 날짜 변수를 제외한 모든 변수에 적용하였다.

4.2 뉴스 데이터 수집 및 전처리

2021년 1월 1일부터 2023년 6월 30일까지의 네이버 뉴스 증권 섹터 부문의 기사들에 대하여 <Table 2>에 해당하는 정보를 수집하였다. 이 중, 910일 간 중에서 한국 주식 시장 개장 일자에 맞추어 공휴일과 토·일요일을 제외한 589일에 해당하는 데이터를 활용하였다. 수집된 뉴스의 원문 데이터에서 '코스피', '코스피200', 'kospi' 등을 키워드로 가지는 뉴스를 필터링해서 코스피와 관련된 뉴스 데이터만 활용할 수 있도록 하였고, 그 결과 총 135,910개의 데이터를 수집하였다. 수집된 데이터에 Cleaning 작업을 통해서 특수문자 제거, 이메일 주소 제거, URL 제거, 불용어 제거, 오타 제거 등의 전처리를 수행하였다. 이 때, 본 연구는 금융 텍스트에 대한 분석을 수행하는 것으로, 불용어를 처리할 때 불용어의 범위를 금융 특성에 맞게 설정할 필요가 있다. 특정 단어는 일반 도메인에서 유의하더라도, 금융 도메인에서는 유의하지 않은 단어일 수도 있다. 예를 들어, 코스피 관련 뉴스 기사에서 형식을 갖추기 위하거나 광고를 위해 흔히 출현하는 단어인 '입력', '수정', '기사', '기자', '신문사' 등이 있다. 이러한 특성을 고려하여 본 연구에서는 금융 도메인에서 유의하지 않

Table 1. Macroeconomic Indicatiors and Stock Index Data Schema

Date	Kospi	dji	us500	ex_AM	gold	oil	research	interest	consumer	real_estate

은 불용어를 추가로 선정하여 이를 처리하는 과정을 수행하였다.

한국어 경제 관련 뉴스 데이터에 다양한 형태소 분석기를 적용한 이전 연구에서는 open-korean-text(Okt) 형태소 분석기가 가장 성능이 좋았다는 연구 결과가 존재한다(Hyeon *et al.*, 2022). 해당 연구에서는 형태소 분석기의 성능 평가 기준으로 형태소 분석기의 수행 시간과 분절 단위 측면을 고려하였다. 따라서, 본 연구에서도 형태소 분석기로 Okt(Twitter, 2014)를 적용하였다. 문장 토큰들의 품사를 태깅한 후, 무의미하다고 판단되는 부사, 접미사, 감탄사, 조사의 품사를 갖는 단어들을 제거하여 주어진 문장을 유의미한 토큰들의 집합으로 정규화 해주었다. 형태소 확인 및 품사 태깅, 정규화, 불용어 제거 등의 규칙 기반 전처리를 진행한 후 추가적인 처리가 필요한 부분은 수동으로 처리하였다. 또한, Pororo라이브러리의 Text Summarization을 활용해서 일반 본문과 전처리가 완료된 본문에 각각 추출식 요약을 진행하였다. 전처리 및 요약이 완료된 뉴스 데이터의 스키마는 <Table 2>와 같다.

Table 2. Preprocessed News Data Schema

pub_date_time	title	url	content	summarize_content	preprocess_content

4.3 데이터 통합 및 활용 변수 선택

최종적으로 사용하게 된 데이터는 주가 및 거시경제 데이터, 뉴스 데이터 감정 점수, 네이버 검색량 데이터이다. 일별 기준으로 데이터를 통합하였고, Date 변수를 포함하여 42개의 변수와 589개의 행으로 구성되었다. 이후 코스피 종가(kospi_close)를 제외한 코스피 시가(kospi_open), 코스피 고가(kospi_high), 코스피 저가(kospi_low), 코스피 거래량(kospi_volume) 변수를 하루씩 앞당겨서 데이터셋을 구성하였는데, 예측 시에 이전 날 값을 사용해야 하기 때문이다.

최종 데이터셋 구성을 위한 변수 선택 과정은 다음과 같다. 종가를 제외한 설명 변수들에서, 다중공선성이 높은 변수를 제거하였다. 그 후, 두 설명 변수 간 상관관계가 0.9 이상일 때, kospi_close와 상관관계가 더 작은 변수를 제거하여 변수의 개수를 11개로 축소했다. 마지막으로 Random Forest Regressor를 활용하여 변수별 중요도를 확인했는데, 0.00004의 매우 낮은 중요도를 나타내는 두 개의 변수를 추가로 제거하였다. 그 결과, 최종 데이터에 활용한 변수는 코스피 고가(kospi_high), 다우존스지수 거래량(dji_vol), S&P500지수 저가(US500_low), S&P500지수 거래량(US500_vol), 유가 시가(oil_open), 유가 거래량(oil_vol), 금값 저가(gold_low), 원/엔 환율 저가(ex_JP_low), 한국 기준금리(ko_interest), 코스피 검색 빈도수(research_kospi), 감정 점수이다.

4.4 모형 적용

(1) KLUE-BERT 모델 기반 감정분석

KLUE-BERT는 모두의 말뭉치, 나무위키, 뉴스 등을 사용해서 한국어를 사전학습 한 BERT기반 모델이다. 해당 사전학습 모델은 일반적인 텍스트 데이터를 활용하였기 때문에, 금융 분야에는 적합하지 않다. 따라서, 본 연구에서 활용할 수 있도록 감정 레이블이 존재하는 금융 관련 뉴스 데이터(Yoo, 2022)를 활용하여 Fine-Tuning 작업을 수행한다. 금융 라벨링 데이터셋으로 학습된 모델을 사용해서 본 연구의 뉴스 데이터를 기반으로 감정 다중 분류를 진행한다. 각 뉴스에 대해서 긍정, 부정, 중립에 대한 분류 확률을 계산한다. 뉴스에 대한 예측값을 일자별로 그룹화한 후 평균을 구해서 일자별 감정 확률을 계산한다. 일자 별 긍정 확률에서 부정 확률을 빼서 감정 점수를 계산한다. <Figure 7>은 KLUE-BERT를 활용하여 감정 점수를 산출하는 예시와 KLUE-BERT를 활용하여 일일 감정 점수를 도식화한 예시이다.

ㅋ

Date	Content	Positive	Negative	Neutral	Sentiment score
20210104	Samsung, Hynix Report Side-by-side…Semiconductor Strength	0.7	0.2	0.1	0.5
20210104	Real estate market outlook negative	0.3	0.5	0.2	-0.2
20210104	KOSPI super strong on first trading day of the new year…above 2,900	0.5	0.4	0.1	0.1

Date	Daily Sentiment Score
20210104	0.13

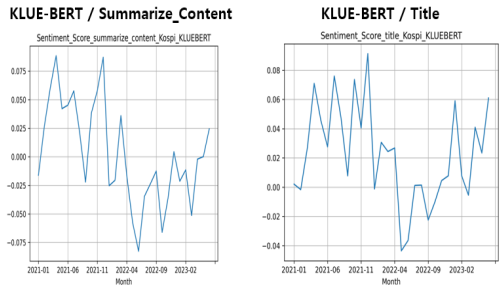

Figure 7. (top): Examples of Calculating Sentiment Score By KLUE-BERT, (bottom): Examples of Daily KLUE_BERT-based Sentiment Score

(2) KoFinBERT 모델 기반 감정분석

KoFinBERT는 금융 데이터셋으로 사전학습된 BERT 기반 모델을 가지고 감정분류 task를 Fine-Tuning한 모델이다. KoFinBERT를 불러와서 금융 뉴스 데이터를 기반으로 감정 다중 분류 task를 수행한다. 각 뉴스에 대해서 긍정일 경우 1, 부정일 경우 -1, 중립일 경우 0의 값을 매핑해서 예측값을 구성한다. 뉴스에 대한 감정 예측값을 일자별로 그룹화한 후 평균을 구해서 일자별 감정 점수를 계산한다.

Date	Content	Sentiment score
20210104	Samsung, Hynix Report Side-by-side…Semiconductor Strength	1
20210104	Real estate market outlook negative	-1
20210104	KOSPI super strong on first trading day of the new year…above 2,900	1

Date	Daily Sentiment Score
20210104	0.33

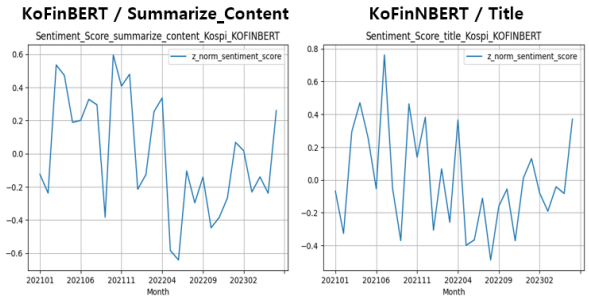

Figure 8. (top): Examples of Calculating Sentiment Score by KoFinBERT, (bottom): Examples of Daily KoFinBERT-based Sentiment Score

(3) 시계열 예측

시계열 예측 모델을 학습하기 위한 데이터 구성 과정은 다음과 같다. 실험 별로 주가 변수, 거시 경제 변수, 감정 점수 변수로 구성된 데이터를 구성한다. 이때 종속 변수인 코스피 종가를 독립 변수로 사용하지 않음으로써 외부 변수로 코스피 지수를 예측하도록 하였다. 이렇게 구성된 전체 데이터를 훈련 데이터(train)와 테스트 데이터(test)로 나눈다. 전체 데이터의 85%를 학습에 사용하고 나머지 15%를 테스트에 사용한다. 모델의 최적의 파라미터 값을 선정하기 위해서 훈련 데이터의 30%를 검증 데이터(validation)로 사용하였다. 과거 5일의 데이터를 바탕으로 다음 날의 주가를 예측하도록 데이터를 구성하였다. 시계열 예측 연구는 위의 4가지 모델을 기준으로 이루어졌으며 최종목적은 코스피 지수 예측의 정확도 증가가 아닌, 감성 점수의 포함 여부가 코스피 지수 예측에 어떠한 영향을 미치는지에 초점을 두고 있으므로 모델의 파라미터는 고정시킨 후 학습을 진행하였다. 모델의 파라미터 값의 경우 학습 횟수는 1000 epoch, 활성화 함수는 ReLU, Dropout 비율은 0.3으로 고정하였고 CNN 레이어의 커널 크기는 256, 커널 사이즈는 3으로 지정한 후에 max pooling(pooling size = 2)을 진행하였다. 구성한 학습 데이터 셋은 학습과 자체적인 평가를 위해서 train/validation/test로 나누게 된다. 위와 같이 데이터를 나눈 후 시계열 예측 모델링을 진행하였다.

4.5 평가 지표

회귀 모델의 성능 평가 지표로 MSE(Mean Squared Error), RMSE(Root Mean Squared Error), MAPE(Mean Absolute Percentage Error)를 사용하였다. MSE는 실제값과 예측값의 차이를 제곱해 평균을 구한 값이다. RMSE는 MSE에 제곱근을 적용한 값으로 실제값과 유사한 단위로 변환하므로 이상치에 덜 민감하다는 이점이 있다. MAPE는 실제값과 예측값의 차이를 절대값으로 변환한 뒤 합산하여 평균을 구하고 이를 비율(%)로 표현한 값으로 오차를 퍼센트로 나타내기 때문에 직관적이다. 각각의 식은 식 (1)~식 (3)과 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (3)$$

위 3가지 평가 지표를 기반으로 예측 모델의 학습 오차(train loss), 검증 오차(validation loss)를 확인하였다. 검증 오차는 학습 데이터로 학습한 모델이 처음 본 데이터에서의 성능을 나타내므로 실제 학습이 얼마나 잘되었는지 확인하였다. 최종적으로 MSE가 가장 작은 모델을 선택하였고, 성능 평가를 할 때에는 MSE, RMSE, MAPE를 모두 활용하였다.

4.6 실험 설정

본 연구에서 적용한 모델을 구동한 서버의 하드웨어 환경으로는 CPU 8 Core, RAM 16GB, FGPU 1 UNIT(3.2 GiB)을 탑재한 서버이고, 사용된 딥러닝 라이브러리 및 소프트웨어는 PyTorch가 사용되었다. 본 연구에서는 총 6개의 연구 문제를 설정하여 실험을 구성했다.

Table 3. Examples of Labeled Sentence

Label	Example Sentence
+1 (Positive)	Stocks were positive today, with shares rising on the back of strong earnings.
0 (Neutral)	The financial markets have been a mixed bag lately.
-1 (Negative)	Stocks markets are volatile due o a slowing economy and uncertain political situation.

[실험 1] 감정분석 모델별 성능 확인

실험 1에서는 감정 분석 모델의 성능을 비교하기 위해서 총 4,846개의 금융 뉴스 기사 데이터에 감정분석 모델을 적용하여 성능을 확인한다. 이 때, 데이터는 긍정 28%, 부정 13%, 중립 59%로 구성되어 있다. <Table 4>는 위 데이터에 들어있는 예시를 나타낸다. 비교를 진행하기 위해 train:test = 8:2로 나누어서 검증을 하였고, KLUE-BERT의 fine-tuning시 100 epoch동안 학습을 진행하였고, mini-batch의 크기는 32, dropout은 0.5

로 설정하여 3-class classification 모델을 학습하였다.

Table 4. Compare to Previous Method

	Previous Method	Proposed Method
Method to calculate sentiment score	Derive sentiment scores for news titles	Summarize the text into one sentence and derive a sentiment score
Number of Macro economics index	3	5
Time-series Prediction Model	LSTM	CNN+GRU

[실험 2] 뉴스 감정 점수와 코스피 종가와의 상관관계

감정 점수는 KLUE-BERT 기반의 본문 요약 감정 점수, KLUE-BERT 기반의 제목 감정 점수, KoFinBERT 기반의 본문 요약 감정 점수, KoFinBERT 기반의 제목 감정 점수 버전으로 구성되어 있다. kospi_close와 가장 상관관계가 높은 감정 점수 버전을 확인해보는 실험을 진행했다.

[실험 3] 최종 데이터셋 변수 설정

4.3절에서 언급했던 바와 같이 본 연구에서는 주가 지수 예측시 모든 변수를 활용하는 대신, 다중공선성이 높은 변수, 중요도가 낮은 변수 등을 제거하는 과정 등 일부 변수를 선택하였다. 실험 3에서는 변수 선택 과정 수행 여부가 예측 성능에 어떠한 영향을 미치는지 확인하기 위해, 변수 선택을 했을 때와 하지 않았을 때 성능을 비교하였다.

[실험 4] 뉴스 감정 점수 포함/미포함 예측 성능 확인

실험 4에서는 감정 점수 변수를 사용했을 때와 사용하지 않았을 때의 예측 성능을 비교하였다. 이를 통해, 뉴스 기사의 감정 점수가 코스피 지수 예측에 주는 영향을 확인해보았다. 이때, 감정 점수는 [실험 2]에서 비교하였던 4가지 방법 중 가장 성능이 좋은 방법을 사용하였다.

[실험 5] 4가지 시계열 모델 & 4가지 감정 조합별 예측 성능 확인

시계열 모델로는 LSTM, GRU, CNN + LSTM, CNN + GRU를 사용했고, 감정 점수 변수로는 KLUE-BERT 기반의 본문 요약 감정 점수, KLUE-BERT 기반의 제목 감정 점수, KoFinBERT 기반의 본문 요약 감정 점수, KoFinBERT 기반의 제목 감정 점수를 사용했다. 총 16개의 조합을 생성한 후 각각 실험을 진행하여 가장 성능이 좋은 모델 & 감정 점수 조합을 확인했다.

[실험 6] 선행 논문과의 비교

기존 연구들 중, 뉴스의 감정 점수와 거시 경제 지표를 활용하여 주가지수를 예측한 연구가 존재한다(Han, 2020). 그러나, 해당 연구에서 제안하는 방법론과 본 논문에서 제안하는 방법론은 몇 가지 차이점이 존재하는데, 이는 <Table 4>에 명시하였다. 기존 방법론과 제안하는 방법론을 적용하여 성능을 비교하였다.

5. 실험 결과

5.1 감정분석 모델별 성능 확인

<Table 5>는 라벨링 데이터를 기반으로 KLUE-BERT, KoFinBERT 모델을 활용하였을 때, 감정 분류 성능을 평가한 결과이다. KoFinBERT를 활용할 때 보다 KLUE-BERT를 fine-tuning한 결과물을 활용하였을 때 좀 더 성능이 좋은 것을 확인할 수 있다.

Table 5. Sentiment Classifcation Result

	KLUE-BERT	KoFinBERT
Accuracy	0.854	0.776
Precision	0.838	0.802
Recall	0.837	0.777
F1	0.837	0.783

5.2 뉴스 감정 점수와 코스피 종가와의 상관관계

<Table 6>은 두 가지 버전의 감정 분석 모델과 두 가지 버전의 텍스트 데이터셋을 조합해서 생성한 4가지 감정 점수와 Kospi_close와의 상관관계를 나타낸다. 여기서 Text version열에 있는 값들 중, summarize_content는 뉴스 본문을 한 문장으로 요약한 결과에 감정 점수를 산출한 결과를 의미하며, title은 뉴스 제목에 대해 감정 점수를 산출한 결과를 의미한다. 감정 분석 모델은 'KLUE-BERT', 감정 분석 모델을 적용할 데이터는 '요약된 본문'을 사용해서 만든 감정 점수와의 상관관계가 가장 높은 것으로 확인된다.

Table 6. Correlation between Kospi Index and Sentiment Score

Model	Text version	Correlation
KoFinBERT	title	0.148
	summarize_content	0.196
KLUE-BERT	title	0.193
	summarize_content	0.243

5.3 최종 데이터셋 변수 설정

<Table 7>은 최종 데이터셋을 구성하기 위해서 변수 선택 방법론 4가지를 실험해 본 결과 다중공선성이 높은 변수 제거

와 함께 중요도가 낮은 변수를 제거했을 때 가장 좋은 성능을 보였다. 해당 방법론을 적용한 결과 42개의 변수에서 11개의 변수를 최종 선택하였다.

Table 7. Performance by Feature Selection Method

Selected Features	MSE
All Features	10404.85
w/o multicollinearity features	4836.22
w/o low importance features	6478.19
w/o multicollinearity & low importance features	**4121.39**

5.4 뉴스 감정 점수 포함/미포함 예측 성능 확인

<Table 8>은 감정 점수 변수를 포함했을 때와 포함하지 않았을 때의 예측 성능을 비교한 결과이다. 감정 점수를 모델에 포함하였을 때 MSE, MAPE 값이 감소한 것으로 보아 뉴스 데이터로부터 산출된 감정 점수가 지수 예측 성능에 긍정적인 영향력을 미치며, 모델의 안정성을 더해준다는 것을 알 수 있다. 즉, 시계열 주가 예측 연구에 있어서 자연어 처리 기술이 중요하게 활용된다고 해석할 수 있다.

Table 8. Performance of Including/excluding Sentiment Score

sentiment score usage	MSE	MAPE
w/o sentiment score	32752.79	5.83%
w/ sentiment score	2176.14	1.52%

5.5 4가지 시계열 모델 & 4가지 감정 조합별 예측 결과 확인

<Table 9>는 시계열 모델과 감정 점수 변수의 조합 16가지의 성능을 나타냈다. 전반적으로 CNN 레이어를 쌓았을 때 모델의 성능이 좋아졌고 KLUE-BERT 감정 분석 모델을 사용했을 때 성능이 좋은 것을 확인할 수 있다. 또한 GRU가 LSTM보다 좋은 성능을 보여준다. 결론적으로 KLUE-BERT 기반의 본문 요약 감정 점수와 CNN-GRU 모델을 사용했을 때 가장 좋은 예측 성능을 보였다. 약 590일간의 시계열 예측에서는 더 간단한 구조를 가지고 있는 모델인 GRU의 전반적인 안정성을 확인할 수 있었다. 또한, CNN 레이어를 먼저 거치면서 변수의 특징을 추출하고 이후 시계열 모델로 진행될 때 학습이 더 원활하게 이루어짐을 알 수 있다. 따라서 기본 LSTM, GRU 같은 모델에 다양하게 층을 추가한다면 성능 향상에 기여할 수 있을 것으로 해석된다. <Figure 9>는 KLUE-BERT로 본문을 요약한 문장에 대한 감정 지수를 도출한 뒤, CNN-GRU모델을 활용하여 주가 지수를 예측한 결과에 대한 그래프이다.

Table 9. Performance of Sentiment Score & Prediction Model Combination

Sentiment Model	Text Version	Prediction Model	MSE	MAPE
KoFinBERT	title	LSTM	3386.22	1.89%
		GRU	3446.27	1.96%
		CNN-LSTM	2261.03	1.61%
		CNN-GRU	2015.43	1.16%
	summarize content	LSTM	4440.08	2.11%
		GRU	3781.84	2.06%
		CNN-LSTM	4796.32	2.28%
		CNN-GRU	1978.98	1.43%
KLUE-BERT	title	LSTM	7011.60	2.66%
		GRU	2362.33	1.59%
		CNN-LSTM	5127.31	2.49%
		CNN-GRU	2477.74	1.63%
	summarize content	LSTM	14193.22	4.02%
		GRU	4212.11	1.93%
		CNN-LSTM	3410.06	1.88%
		CNN-GRU	**1909.83**	**1.38%**

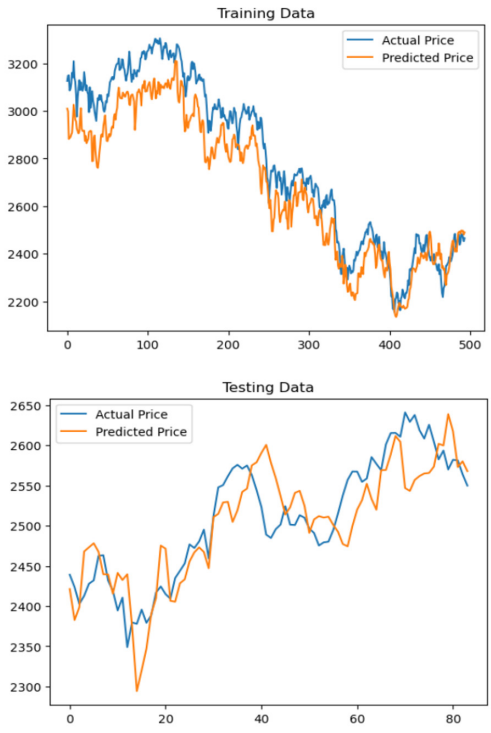

Figure 9. Actual & Predicted Price Prediction Result Using KLUE-BERT and CNN-GRU Model (top): Training Data, (bottom): Test Data

5.6 선행 논문과의 비교

제4.6절에 있는 <Table 4>는 선행 논문과 본 논문의 3가지 차이점을 서술한 표이다. 여기서는 <Table 4>에 있는 내용을 바탕으로 선행 논문과 본 논문을 비교하는 3가지 실험을 진행한 결과를 서술한다. <Table 10>은 감정 점수를 산출하는 모델을 적용하는 대상에 따라서 비교한 결과이다. 기존 논문에서는 신문 기사의 제목에 적용하였지만, 본 논문에서는 본문을 한 문장으로 요약한 결과에 적용하였다. 실험 조건을 동일하게 하기 위해서 시계열 모델은 CNN-GRU을 사용하였다. 그 결과, 본 연구에서 제안한 방법이 더 좋은 결과가 나오는 것을 알 수 있다. <Table 11>은 주가지수 예측 시 활용한 거시경제 지표의 종류에 따른 성능 평가 결과이다. 기존 논문에서는 3가지의 지표만 활용하였고, 본 논문에서는 5개의 지표를 활용하였다. 이 때, 실험 조건을 동일하게 하기 위해 예측 모델은 CNN-GRU를 사용하였고, 감성 점수는 뉴스 제목만을 활용하여 산출하도록 설정하였다. 그 결과 본 논문에서 제안한 방법이 더 좋은 성능을 나타냄을 확인하였다. <Table 12>는 주가지수 예측 시 사용한 모델에 따른 차이를 나타낸다. 기존 논문은 단순히 LSTM을 활용하였으나 본 논문에서는 CNN-GRU 모델을 적용하였다. 실험 조건을 동일하게 하기 위해, 감성 점수는 본문을 한 문장으로 요약한 KLUE-BERT 기반 결과에서 산출하도록 설정하였다. 그 결과, CNN-GRU를 적용하였을 때 단순 LSTM을 적용하였을 때보다 크게 성능이 향상됨을 알 수 있다.

Table 10. Performance Evalution Results Based on the Application of the Emotional Score Model

Model	MSE	MAPE
Previous Method (sentiment score from the title)	2477.74	1.63%
Proposed Method (sentiment score from the summarized text)	1909.83	1.38%

Table 11. Performance Evalution Results Based on Macroeconomic Indicators Used

Model	MSE	MAPE
Previous Method (Index: gold, oil, currency)	3283.81	2.48%
Proposed Method (Index: gold, oil, currency, interest rate, kospi index)	1655.14	2.51%

Table 12. Performace Evaluation Results Based on Model Structure

Model	MSE	MAPE
Previous Method (Model: LSTM)	14193.22	4.02%
Proposed Method (Model: CNN-GRU)	2477.74	1.63%

6. 결 론

본 논문은 뉴스 데이터 기반 감정 분석 모델과 딥러닝 기반의 시계열 예측 모델을 사용해서 코스피 지수를 예측하는 방법론을 제시하였다. 주식 시장의 전반적인 흐름을 나타내는 코스피 지수의 종가를 종속 변수로 사용했으며 금값, 유가, 환율 등과 같은 거시 경제 지표와 감정 점수와 같은 시장 심리 지표를 독립 변수로 사용해 주었다.

감성 점수는 코스피 관련 키워드로 네이버 금융 뉴스 본문을 크롤링한 뒤 전처리 및 요약 작업을 수행함으로써 제목 데이터셋과 본문 요약 데이터셋을 구성하였다. 이어서 KLUE-BERT 모델 기반의 일자별 감정 점수와 KoFinBERT 모델 기반의 일자별 감정 점수를 추출하였다. 시계열 예측 모델로는 LSTM, GRU, CNN-LSTM, CNN-GRU를 사용했다. 변수들과 모델들의 조합 별 실험을 진행한 결과 KLUE-BERT 기반의 본문 요약 감정 점수와 CNN-GRU 모델을 사용했을 때 가장 좋은 성능을 보였다.

본 논문은 기여는 첫 번째, 금융 특화 감정 분석을 통해 금융 시장 심리를 나타내는 감정 점수를 도출했다는 점이다. 두 번째, 코스피 지수에 영향을 주는 다양한 거시 경제 지표를 활용함으로써 주식 시장의 특성을 반영한 것이다. 세 번째, 기존 논문에서는 뉴스의 제목 기반으로 감정 점수를 도출했다면 본 연구에서는 뉴스의 본문 요약 추출 본 기반으로 감성 점수를 도출함으로써 성능을 개선했다. 네 번째, 기존 연구는 주로 과거의 코스피 지수를 독립 변수로 함께 사용했지만 본 연구에서는 독립 변수로 코스피 지수를 사용하지 않고 다른 지표들로 코스피 지수를 예측했다는데 의의가 있다. 본 논문의 한계점은 첫 번째, 텍스트 데이터 수집의 비용으로 인해 분석 기간이 2년 6개월로 상대적으로 짧다는 것이다. 10년 치 데이터를 확보한다면 코스피 지수의 전반적인 추세를 포함해서 예측할 수 있을 것으로 기대된다. 두 번째, 고정된 파라미터와 학습 방법을 사용했다는 것이다. 추후에는 최적화 함수, 활성화 함수, 초기화 방법 등 다양한 학습 방법과 하이퍼 파라미터를 사용해서 모델의 예측 성능을 높이도록 노력할 것이다. 따라서, 향후 연구는 더 긴 시간의 데이터를 확보하고 다양한 실험을 하여 본 논문의 한계점을 보완해 나갈 예정이다.

참고문헌

Agrawal, M., Khan, A. U., and Shukla, P. K. (2019), Stock Price Prediction using Technical Indicators: A Predictive Model using Optimal Deep Learning, *International Journal of Recent Technology and Engineering(IJRTE)*, **8**(2).

Araci, D. (2019), FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, arXiv preprint arXiv:1908.10063.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015), *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.

Cho, K. (2023), KoFinBERT, GitHub Repository, https://huggingface.co/k

woncho/KoFinBERT.
Cho, K., Van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014), Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734.
Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805.
Elman, L. J. (1990), Finding Structure in Time, *Cognitive Science*, **14**(2), 179-211.
Fama, E. F. (1965), Random Walks in Stock Market Prices, *Journal of Financial Analysts*, **21**(5), 55-59.
Gan, J. (2022), ko-sroberta-multitask, Hugging Face, https://huggingface.co/jhgan/ko-sroberta-multitask.
Heo, H., Ko, H., Kim, S., Han, G., Park, J., and Park, K. (2021), PORORO: Platform Of neuRal mOdels for natuRal language prOcessing, Github repository, https://github.com/kakaobrain/pororo.
Hochreiter, S. and Schmidhuber, J. (1997), Long Short-Term Memory, *Neural Computation*, **9**(8), 1735-1780.
Hoseinzade, E. and Haratizadeh, S. (2019), CNNpred: CNN-based stock market prediction using a diverse set of variables, *Expert Systems with Applications*, **129**, 273-285.
Hwang, C. and Shin, K. (2020), CNN-LSTM Combination Method for Improving Particular Matter Contamination (PM2.5) Prediction Accuracy, *Journal of the Korea Institute of Information and Communication Engineering*, **24**(1), 57-64.
Hyeon, J., Lee, J., and Cho, H. (2022), Sentiment Analysis of News on Corporation Using KoBERT, *Korean Accounting Review*, **47**(4), 33-54.
Jang, E. (2020), LSTM Combination of BERT Sentiment Analysis and Time Series Macro economy Index for Predicting Stock Price, (Doctoral Dissertation, Master Thesis, Korea University).
Jung, J. and Kim, J. (2020), A Performance Analysis by Adjusting Learning Methods in Stock Price Prediction Model Using LSTM, *Journal of Digital Convergence*, **18**(11), 259-266.
Kang, D., Yoo, S., Lee, H., and Jeong, O. (2022), A study on Deep Learning-based Stock Price Prediction using News Sentiment Analysis, *Journal of The Korea Society of Computer and Information*, **27**(8), 31-39.
Liew, K. S., Lim, K. P., and Choong, C. K. (2003), On the Forecast Ability of Asian-5 Stock Markets Returns Using Time Series Models, *ICFAI Journal of Applied Finance*, **10**, 17-29.
Malo, P., Shinha, A., Korhonen, J., and Takala, P. (2014), Finance Phrase Bank, Hugging Face, https://huggingface.co/datasets/financial_phrasebank.
Mondal, P., Shit, L., and Goswami, S. (2014), Study of Effectiveness of Time Series Modeling (ARIMA) in Forecasting Stock Prices, *International Journal of Computer Science, Engineering and Applications*, **4**(2), 13.
Park, S., Moon, J., Kim, S., Cho, W-I., Han, J., Park, 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., and Cho, K. (2021), Klue: Korean Language Understanding Evaluation, arXiv preprint arXiv:2105.09680.
Schumacher, R. P. and Chen, H. (2009), Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZ Fintext System, *ACM Transactions of Information Systems(TOIS)*, **27**(2), 1-19.
Shayan, H. (2022), FinBERT-LSTM: Deep Learning Based Stock Price Prediction Using News Sentiment Analysis, arXiv preprint arXiv:2211.07392.
Song, S., Kim, J., Kim, H., Park, J., and Kang, P. (2019), Development of Early Warning Model for Financial Firms Using Financial and Text Data: A Case Study on Insolvent Bank Prediction, *Journal of the Korean Institute of Industrial Engineers*, **45**(3), 248-259.
Twitter (2014), Okt: open-korean-text, GitHub repository, https://github.com/open-korean-text/open-korean-text.
White, H. (1988), Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, *Proc. 1988 Int. Conf. on Neural Networks*, **2**, 451-458.
Wong, W. K., Manzur, M., and Chew, B. K. (2003), How Rewarding is Technical Analysis? Evidence from Singapore Stock Market, *Applied Financial Economics*, **13**(7), 543-551.
Yoo, W. (2022), Finance Sentiment Corpus, GitHub Repository, https://github.com/ukairia777/finance_sentiment_corpus.
Yule, G. U. (1927), VII. On a Method of Investigating Periodicities Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers, *Philosophical Transactions of the Royal Society of London Series A*, **226**, 267-298.

저자소개

정가연: 국민대학교 경제학과/AI빅데이터융합경영학과에서 학사과정에 재학 중이다. 연구 분야는 머신러닝, 딥러닝, 자연어처리, 시계열 예측이다.

이혁제: 국민대학교 AI빅데이터융합경영학과에서 학사과정에 재학 중이다. 연구 분야는 시계열 예측, 딥러닝이다.

이준영: 국민대학교 AI빅데이터융합경영학과에서 학사과정에 재학 중이다. 연구 분야는 머신러닝, 딥러닝, 컴퓨터비전, 자연어처리이다.

이제혁: KAIST 산업및시스템공학과에서 2012년 학사, 2014년 석사학위를 취득하고 서울대학교 산업공학과에서 2020년 박사학위를 취득하였다. 이후 삼성전자에서 근무하다가 2022년부터 국민대학교 AI빅데이터융합경영학과의 조교수로 재직 중이다. 연구 분야는 머신러닝, Industrial AI이다.