Journal of The Korean Data Analysis Society (February 2025), 27(1), 31-43.
https://doi.org/10.37727/jkdas.2025.27.1.31

Fin-BERT 모델 기반 감성 점수 활용 주가 예측*

김영민[1], 강창완[2], 최용석[3]

요 약

경제 및 기술의 발전으로 인해 많은 사람이 주식에 대한 관심이 높아졌다. 이러한 관심을 발전된 머신러닝과 딥러닝 모형에 적용하여 많은 주가 예측 모델이 개발되었다. 최근의 연구는 주식에 대한 문자정보를 반영한 예측모형이 등장하였고 그 중 대부분의 연구는 뉴스 데이터를 이용하여 주가 예측 모델을 만들었다. 그러나 뉴스 데이터는 중립적인 경향이 너무 강해서 정확한 감성 분류가 어렵다고 판단되어 본 연구에서는 남녀노소 자유롭게 의견을 작성할 수 있는 커뮤니티 데이터인 네이버 종목토론실 데이터를 이용하여 주가 예측 모델을 만들고자 한다. 금융 도메인 데이터를 더 정확하게 감성 분류하기 위해 한국어 금융 데이터를 학습시킨 KR-FinBERT를 이용해서 감성 분류를 진행하였다. 주가는 최근 데이터에 영향을 가장 많이 받기 때문에 가중이동평균을 가중치로 이용하여 최종 감성 점수를 만들었다. 네이버 종목토론실 데이터로 만든 감성 점수가 주가 예측에 영향을 미치는지 확인하기 위해 Random Forest, XGBoost, LSTM 세 가지 분석 방법론을 이용하여 감성 점수가 있는 데이터로 만든 모델과 감성 점수가 없는 데이터로 만든 모델을 비교하였다. 모델 비교를 위해 사용된 평가 지표는 RMSE와 MAE가 사용되었다. 분석 결과 세 가지 분석 방법을 이용한 모델 비교에서 감성 점수가 있는 데이터의 평가 지표가 더 좋게 나와 감성 점수가 주가 예측에 영향을 미치는 것을 확인할 수 있었다

주요 용어 : 감성점수, Fin-BERT, Random Forest, XGBoost, LSTM

1. 서론

경제 및 기술의 발전으로 인해 많은 사람이 기업의 가치에 관심을 갖게 되고 이러한 회사에 투자하는 형식인 주식에 관한 관심이 높아졌다. 실제로 다양한 플랫폼들이 발달하면서 손쉽게 주식을 거래할 수 있는 환경이 만들어져 있다. 기업의 가치가 올라감에 따라 그에 따른 주식의 가격 또한 상승하게 되는데 이때 투자자들은 이익을 얻게 된다. 모두가 수익을 원하기 때문에 머신러닝 및 딥러닝을 이용하여 주가를 예측하는 연구가 활발하게 이루어져 왔다.

주가를 예측하는 것에는 전문적인 지식이 필요하여 증권분석가, 펀드매니저, 트레이더 등 다양한 전문 직종이 존재한다. 일반인의 경우에는 뉴스를 통해 쉽게 정보를 얻고 이를 이용한 연구들이 진행되었다.

---

*이 논문은 제1저자 김영민의 석사학위논문의 축약본입니다.

[1]46241 부산광역시 금정구 부산대학로 63 부산대학교 통계학과, 석사졸업. E-mail : rladudals51@naver.com
[2]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영빅데이터공학과, 교수.
E-mail: cwkang@deu.ac.kr
[3](교신저자)46241 부산광역시 금정구 부산대학로 63 부산대학교 통계학과, 교수. E-mail: yschoi@pusan.ac.kr
[접수 2025년 1월 8일; 수정 2025년 1월 21일; 게재 확정 2025년 1월 24일]

Chang, Yoo(2022)는 뉴스 제목을 이용하여 코로나 기간 동안의 제약회사 주가를 예측하는 연구를 진행했고, Kang et al.(2022)도 뉴스 데이터의 감성 분석을 이용한 딥러닝 기반 주가 예측 모델을 만드는 연구를 진행하였다. 연구 결과를 보면 뉴스 데이터를 통한 주가 예측 모델을 만들었을 때 성능이 좋은 것을 확인할 수 있다. 그러나 본 연구에서는 뉴스 데이터는 중립적인 성향이 짙어 긍정, 부정의 감성 분류에 적합하지 않다고 판단하여 긍정, 부정에 대한 구분이 확실한 커뮤니티 데이터를 이용하였다. 특히, 다양한 플랫폼 데이터 중 네이버 종목토론실을 채택하였는데 이는 잘 알려져 있고 많은 연구 사례로 사용되고 있다. Lee(2021)는 뉴스 데이터와 네이버 종목토론실 데이터를 Word2vec을 통해 주가 수익률을 설명하였고 종목토론실 데이터가 뉴스 데이터보다 안정적인 유의수준을 나타내었음을 보였다. 그러나 Word2vec에는 단점이 존재하는데 한 단어에 고정된 하나의 벡터를 사용하기 때문에 다양한 의미를 반영하지 못하고 모든 문맥에서 동일한 벡터를 사용하는 문제점이 존재한다. 그리고 한국어에서는 동음이의어와 같은 단어를 처리하는데 문제가 발생할 수 있고 또한 특정 도메인에 특화된 성능을 보일 수 있으며 시간이 오래 걸린다는 단점이 존재한다..

이러한 문제를 해결하기 위해 전체적인 문맥을 의미하는 컨텍스트를 반영한 트랜스포머 기반의 BERT 모델이 등장하게 된다. BERT는 Devlin et al.(2019)가 만든 모델로 대규모 텍스트에 사전 훈련되어 있어 다양한 자연어 처리 작업에 사용되고 미세 조정을 통해 특정 태스크의 성능을 향상시킬 수 있다. Choi, Kim, Jang(2023)은 뉴스 데이터 및 종목토론실 데이터를 KR-FinBERT를 통해 감성 분류한 뒤 공모주 가격 형성에 대한 영향력을 확인하였다.

본 연구에서는 조건에 맞는 기업을 선정한 뒤 해당 기업의 종목토론실 데이터를 수집 후 금융 데이터의 감성 분류에 특화된 KR-FinBERT를 이용하여 감성 점수를 만들어 낸다. 그 후 감성 점수가 주가 예측에 영향을 미치는지 머신러닝 및 딥러닝 모델을 이용하여 주가 예측 모델을 통해 알아보려고 한다. 2절에서 종목토론실에 대한 설명과 데이터 수집과정 및 전처리 그리고 Fin-BERT에 대한 설명을 진행한다. 그 후 3절에서 주가 예측 모델 설계 및 모델의 성능을 비교하며, 4절에서 결론으로 마무리하려 한다.

2. 데이터 설명 및 Fin-BERT 모델

2.1 데이터 설명

기존의 뉴스 데이터를 활용한 주가 예측과 달리 본 연구는 뉴스 데이터의 단점인 중립적 특성을 고려하여 누구나 접근할 수 있고 작성할 수 있는 커뮤니티 데이터를 주가 예측 모형에 반영하였다.

이러한 커뮤니티 데이터 중 네이버에서 제공하는 종목토론실을 선정하였다. 종목토론실은 해당 주식에 대해 개인적인 의견이나 정보를 실시간으로 공유할 수 있는 공간이다. 작성자가 익명으로 표시되다 보니 자유롭게 의견을 표현할 수 있는 장점이 있으나 무의미하거나 왜곡된 정보가 포함된 게시글이 발견되기도 하였다. 글마다 공감 및 비공감 버튼이 있는데 분석 결과, 의미 있는 정보들은 공감의 수가 높았고 쓸모없는 정보들은 비공감 수가 높은 경향이 있어 종목토론실의 글을 참고로 하고 싶다면 공감순이 높은 글들을 참고할 것을 추천한다.

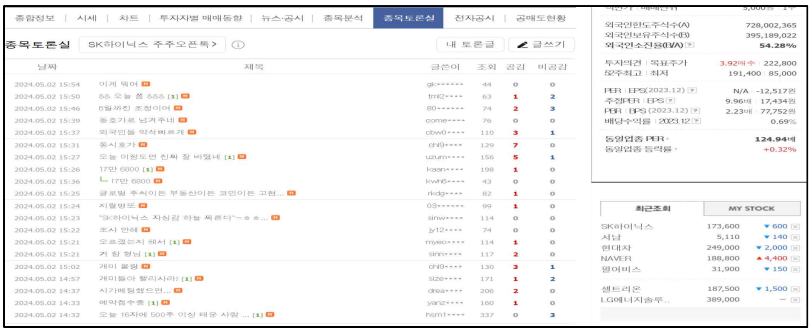

Figure 1. Naver Stock Discussion Board

Figure 1은 네이버 종목토론실 화면으로 각 글의 날짜, 제목, 글쓴이 등의 정보와 우측에는 해당 종목의 관련 자료들이 나타나 있다.

본 연구에서는 SK하이닉스, 셀트리온, 현대자동차, 네이버 총 4개의 기업을 선정하였으며 그 기준은 다음과 같다.

1) 모기업에 호재가 생길 때 관련 계열사 모두 주가에 영향을 받기 때문에 같은 계열사가 아닌 기업
2) 하나의 테마가 유행할 때 같은 테마라면 주가에 영향을 받기 때문에 서로 다른 테마를 가진 기업: SK하이닉스(반도체), 셀트리온(의약품), 현대자동차(자동차), 네이버(인터넷 서비스업)
3) 종목토론실 데이터의 경우 앞서 언급한 대로 의미 없는 데이터가 있어서 정확한 분석을 위해서 하루 업로드되는 게시물이 15개 이상인 기업

2.2 데이터 수집

종목토론실의 글을 얻기 위해 웹 크롤링(Web Crawling)을 진행하였는데 이는 인터넷의 웹 페이지들의 데이터를 수집하는 자동화된 방법을 뜻한다. 본 연구에서는 Selenium과 Selenium의 WebDriver를 사용하여 데이터를 수집하였다.

Table 1은 2021년 1월 1일부터 2023년 8월 31일 기간 동안 추출된 종목토론실 데이터의 시간(Times), 제목(Title), 내용(Contents)에 대한 정보로 데이터는 날짜를 기준으로 나열하였다.

Table 1. crawling data of stock discussion board

Times	Title	Content
2021.01.01 00:59	새해복많이받으세요 (Happy new year)	현대차 파이팅(Hyndai motor, fighting)
2021.01.01 08:19	우리주식 시장 (Our syock market)	새로운시장이 열린다 박스권 상향 돌파 깨어난 동학 개미의 출현 (A new market emerges: breaking out of the box pattern, the awakening of the donghak ants)
		...
2023.08.31 21:48	현대차조립공 (Hyundai motor assembly plant)	현대차조립공노조는 새겨들어라. 욕심이 도를 넘으면 반드시 재앙이 닥친다.(Hyundai motor assembly plant union, listen carefully: when greed goes too far, disaster is bound to follow.)
2023.08.31 22:07	내가 경영자라면 (If I were a manager)	이렇게 떼쓰는 직원들과는 빨리 헤어지고 싶을 듯. 정년연장불가다. (I'd want to part ways quickly with employees who make such unreasonable demands. Extending the retirement age is not an option.)

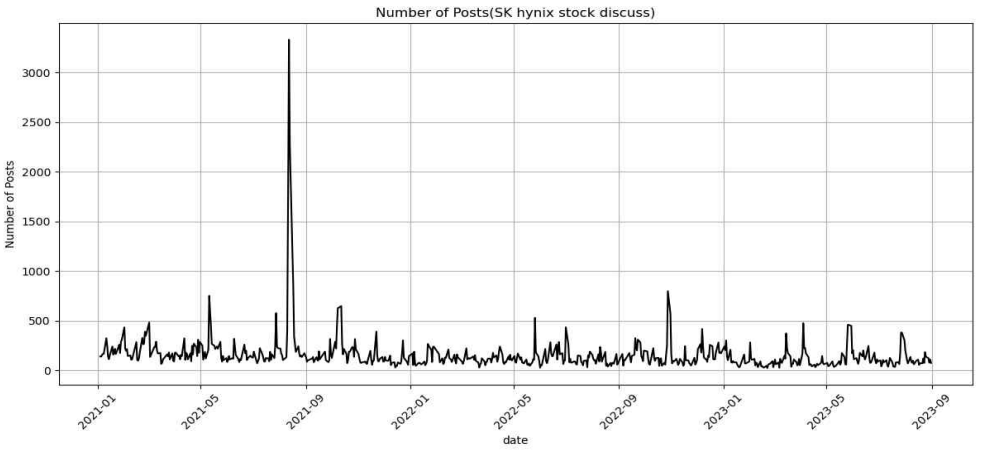

Figure 2. Daily number of posts about SK Hynix

Figure 2는 한 예로 SK하이닉스의 일별 게시물 수를 나타낸 그림이다. 그림을 보면 특정일에 그래프가 치솟는걸 볼 수 있는데 이는 특정일에 기업 관련 주요한 이슈가 발생하여 주가의 변동이 심해 종목토론실 글의 수가 폭발적으로 상승한 것을 확인할 수 있다. 이 경우는 SK하이닉스의 D램 주문량 하락으로 인한 주가 폭락 이슈가 존재했음을 볼 수 있다.

2.3 Fin-BERT

기존의 자연어 처리 분야에서 감정 분류하는 방법으로 다양한 방법이 사용되었다. 컴퓨터는 문자를 읽지 못하기 때문에 문자를 수치적 형태로 변환한 뒤 읽고 해석해야 한다. 문장을 구성하는 단어들을 수치적 형태인 벡터로 표현하는 방법을 문자 임베딩(Word Embedding)이라고 하며 이러한 표현을 통해 단어 간의 유사성을 포착할 수 있으며 각 단어의 고유한 특성 또한 포착할 수 있다. 단어 임베딩 방법에는 여러 방법이 존재한다. 특히 Word2Vec을 이용한 임베딩 연구들이 활발하게 진행되었지만, Word2Vec의 방법에는 몇 가지 문제점이 존재한다.

1) 단어가 많을수록 학습해야 하는 시간이 많아진다. 학습하는 데이터의 단어가 3억 개라고 한다면 3억 개를 학습하기 때문에 시간이 오래 걸릴 수밖에 없다.
2) Word2Vec는 각 단어에 대하여 하나의 벡터를 생성하기 때문에 같은 단어지만 의미가 다른 동음이의어 단어에 큰 문제가 발생한다. 예를 들어 “배”라는 단어는 먹는 음식과 선박을 의미하는데 먹는 음식의 배로 벡터가 학습된다면 선박의 의미를 가진 배를 사용할 수 없게 되어서 학습에 문제가 생기게 된다.

이러한 문제들을 해결하기 위해 기술들이 발전되었고 가장 최근에 구글에서 개발한 BERT라는 모델이 등장하는 계기가 된다. BERT는 Bidirectional Encoder Representation Transformer의 약자로 2018년 구글에 의해서 개발된 모델로 자연어 처리 분야에서 우수한 성능을 보인다. 이름에서 볼 수 있듯이 Transformer 모델의 Encoder 부분을 기반으로 설계되었으며 self attention 메커니즘을 사용하여 각 입력 토큰이 문장 내 다른 토큰과의 관계를 학습하도록 한다.

Table 2. Financial data used for training KR-FinBERT

Positive	Negative
현대바이오, '폴리탁셀' 코로나19 치료 가능성에 19% 급등 (Hyundai Bio surges 19% on the potential of 'Polytaxel' as a COVID-19 treatment.)	영화관 측, '코로나 빙하기' 언제 끝나나... "CJ CGV 올 4000억 손실 날 수도" (Movie Theater Industry: 'When Will the COVID Ice Age End?'... 'CJ CGV May Suffer a 400 Billion Won Loss This Year')
이수화학, 3분기 영업익 176억...전년 대비 80% 상승(Isu Chemical Reports Q3 Operating Profit of 17.6 Billion Won, an 80% Increase Year-on-Year)	C쇼크에 멈춘 흑자비행...대한항공 1분기 영업적자 566억(Profit Flight Halted by the 'C-Shock'... Korean Air Reports Q1 Operating Loss of 56.6 Billion Won)
"GKL, 7년 만에 두 자릿수 매출 성장 예상"(GKL Expected to See Double-Digit Revenue Growth for the First Time in 7 Years)	'1000억대 횡령,배임' 최신원 회장 구속... SK네트웍스 "경영 공백 방지 최선"(Chairman Choi Sin-Won Arrested for Embezzlement and Breach of Trust Worth 100 Billion Won... SK Networks Says 'Will Do Its Best to Prevent Management Vacuum)

BERT는 MLM(Masked Languages Model) 방법을 사용하여 모델의 양방향으로 문맥을 이해하는 능력을 강화한다. 특히, 금융 분야는 매도, 매수 등과 같은 전문 용어가 자주 등장하게 되는데 이러한 용어는 일상생활에서 쉽게 사용되지 않는다. 따라서 일반 BERT에 금융 데이터를 이용하여 분석할 때 성능이 좋지 않게 나온다. 이런 문제를 해결하기 위해 Yang, Uy, Hwang(2020)는 금융 코퍼스로 훈련하여 금융 분야에 특화된 BERT 모델인 Fin-BERT를 만들었다. 본 연구 진행 방향도 금융 분야의 텍스트 데이터를 이용한 주가 예측이기 때문에 한국어 기반 모델인 KR-FinBERT를 사용하였다. 일반적으로 Fin-BERT는 파이썬(Python) 환경에서 구현가능하다.

Kim et al.(2023)이 개발한 KR-FinBERT는 한국어 금융 분야에 특화된 BERT 모델이다. 이 모델은 한국어 위키피디아, 일반 뉴스 기사, 법률 텍스트, 한국어 댓글 데이터 등의 텍스트로 사전 훈련된 KR-BERT-MEDIUM 모델 기반에 금융 뉴스 기사와 증권 회사의 분석 보고서를 추가로 학습시켜 한국어 금융 분야에 대한 학습을 강화시켰다. KR-FinBERT의 성능을 확인하기 위해 5만 개의 데이터를 각 모델에 감성 분류를 진행하였더니 이는 한국어 금융 분야에 있어서 다른 BERT 모델들보다 정확하게 감성 분류작업을 진행하는 것을 확인할 수 있었다. Table 2는 KR-FinBERT 학습에 사용된 금융 데이터를 나타낸다. Positive는 긍정적인 리뷰, Negative는 부정적인 리뷰를 의미하며 이를 라벨링 하여 모델 학습에 활용하였다.

2.4 데이터 전처리

앞서 수집된 SK하이닉스, 셀트리온, 네이버, 현대자동차의 종목토론실 데이터를 이용해 긍정, 부정을 분류하여 감성 점수(Sentiment Score)를 만들어 연구를 진행하고자 한다.

한국의 주식 시장은 09:00부터 18:00까지 거래할 수 있다(시간 외 거래 포함). 이를 바탕으로 크롤링 데이터의 날짜를 재구분하였다. 예를 들어 1일 18:01 ~ 2일 18:00까지의 데이터는 2일에 반영하는 것으로 구분한다. 한국의 경우 공휴일, 주말에 주식 시장이 열리지 않으므로 공휴일, 주말의 글은 다음 주식 시장이 열리는 날에 반영한다고 가정하여 날짜를 조절하였다. 그 후 다음과 같은 기준으로 데이터를 전처리하였다.

1) 3글자 이하, 500글자 이상이거나 URL만 포함된 글은 의미가 없다고 판단하여 삭제
2) !,@,#,$,%,^,& 등과 같은 특수 문자 삭제

Table 3. Sentiment classification values obtained using KR-FinBERT

Text	Positive	Neutral	Negative	Sentiment Scores
상한가 한번 가자 제일파마홀딩스도 가자 PBR 세계최초 뇌졸중 치료제 임상결과 곧 발표(Let's Hit the Upper Limit! Let's Go, Jeil Pharma Holdings! Clinical Results for the World's First Stroke Treatment Drug to Be Announced Soon)	0.99989	0.00007	0.00004	0.99990
산사람 잡는AZ백신 복제연구 시작~ (AZ Vaccine Copycat Research Begins: Targeting the Living)	0.00004	0.99993	0.00003	0.00004
SK바이오 고점대비 분에 토막났다 현제 원대 셀트리온 원 겨우 반토막 조금 더 났... (SK Bio has dropped significantly from its peak. Currently, Celltrion has lost about half of its value, and it's still declining a bit further)	0.00008	0.00010	0.99983	-0.99983

전처리 된 본문 데이터를 KR-FinBERT로 만든 감성 분류기(Sentiment Classifier)의 입력값으로 사용하였고 최종적으로 소프트맥스 함수를 거쳐서 각 문장에 대한 긍정, 부정, 중립의 확률값을 얻었다. 여기서 얻어진 확률값 중 긍정과 부정만 선택하여 점수로 사용하였다.

긍정 점수를 s_i^p로, 부정 점수를 s_i^n로 정의하여 사용한다. 감성 점수를 만들 때 긍정, 부정에 대해서만 사용할 예정으로 긍정 점수가 부정 점수보다 높다면 긍정 점수를 사용하고, 부정 점수가 긍정 점수보다 높다면 부정 점수에 음수를 곱하여 값을 사용한다.

$$s_i = \begin{cases} s_i^p & \text{if } s_i^n < s_i^p \\ -s_i^n & \text{if } s_i^n > s_i^p \end{cases}$$

이러한 과정을 Table 3에 나타내었다. 어떤 본문 내용이 KR-FinBERT에 들어갔는지 나타내었고, 이러한 본문 내용이 KR-FinBERT를 거쳐 긍정(Positive), 중립(Netural), 부정(Negative)의 점수로 나타나게 된다. 그 후 앞서 설명한 대로 긍정 점수와 부정 점수를 비교하여 본문 내용에 대한 최종 감성 점수(Sentiment Scores)를 채택하게 된다.

각 게시물은 -1~1 값을 가지는 감성 점수로 채택되었다. 1의 값에 가까울수록 강한 긍정을 나타내는 게시물, -1의 값에 가까울수록 강한 부정을 나타내는 게시물이다.

본 연구에서 주가 예측할 때 단위를 일별로 진행하였다. 따라서 위의 감성 점수를 가지고 최종적으로 일별 감성 점수를 만들려고 한다. 먼저 날짜를 기준으로 감성 점수를 일별로 모두 합친다.

$$S_d = \sum_{i \in D_d} s_i$$

여기서 S_d는 특정 날짜 d에 대한 일별 감성 점수 합이며, D_d는 날짜 dd에 해당하는 감성 점수들의 집합, s_i는 날짜 d에 해당하는 각각 데이터의 감성 점수이다. 그리고 다음과 같은 특정 날짜 d에 해당하는 일별 게시물 수 c_d를 구한다.

$$c_d = count(D_d)$$

앞서 구한 일별 감성 점수 S_d에 일별 게시물 수 c_d를 나눠서 조정된 일별 감성 점수 A_d를 구한다.

Table 4. final Sentiment classification values (SK Hynix)

Date	S_d	c_d	A_d	WMA_d	$\lvert WMA_d \rvert$	f_d
2021-01-04	2.9750	130	0.0229	0.0150	0.0150	0.000343
2021-01-05	1.6809	133	0.0126	0.0150	0.0150	0.000189
…	…	…	…	…	…	…
2023-08-30	-6.6035	106	-0.0623	-0.0480	0.0480	-0.002988
2023-08-31	-0.7065	71	-0.0100	-0.0373	0.0373	-0.000371

$$A_d = \frac{S_d}{c_d}$$

주식의 가격을 결정할 때, 과거의 데이터보다 최근 데이터에 영향을 많이 받기 때문에 최근 데이터의 감성 점수에 높은 가중치를 두고 과거의 데이터에 낮은 가중치를 두고자 한다. 이때 가중치로 가중이동평균을 사용했다.

$$WMA_d = \frac{n \times A_d + (n-1) \times A_{d-1} + \cdots + 1 \times A_{d-(n-1)}}{n + (n-1) + \cdots + 1}$$

여기서 n은 이동 평균 기간으로 5로 설정하였다. 따라서 1일 ~ 4일 차의 가중이동평균은 없으므로 시점 d는 $5 \leq d \leq 660$의 범위를 가지게 된다. 이때 가중이동평균에 절댓값을 씌워 가중치로 만든 뒤, 일별 감성 점수에 곱하여 최종 감성 점수를 얻게 된다.

$$f_d = A_d \times \lvert WMA_d \rvert$$

단, 1일~4일 차 데이터는 가중이동평균이 없으므로 5일 차의 가중이동평균을 사용한다. Table 4는 최종 감성 점수가 만들어지는 과정 동안 만들어진 값들을 정리해둔 표이다.

3. 예측 모델 설계

이 절에서는 LSTM, XGBoost, Random Forest 등의 알고리즘을 활용하여 종목토론실에서의 감성 점수가 주가(종가) 예측에 미치는 영향을 분석하는 예측 모델을 설계하고자 한다. 본 연구에서의 목적이 감성점수 활용이 주가예측에 도움이 되는 지를 확인하는 데 있으므로 주가 예측 변수 선정은 주관적인 판단하에 원/달러 환율, 장단기 금리 차. 공포지수와 직접 만든 변수인 감성 점수를 산정하여 진행하였다. 다음은 주가 예측에 사용되는 변수들에 대한 설명이다.

Table 5에는 사용된 변수 Open(시가), Volume(거래량), USD/KRW(환율), T10Y2Y(장단기 금리 차), VIX(공포지수), Sentiment(감성 점수), Close(종가) 총 7개의 변수에 대한 설명이며, 종가를 제외한 나머지 변수로 종가 값을 예측하는 모델을 설계하려고 한다.

Table 6을 보면 변수 간 스케일 차이가 큰 것을 알 수 있다. 따라서 데이터에 대한 스케일 조정 작업이 필요하다. 스케일 조정 작업 중 대표적으로 최소-최대 정규화(min-max normalization)와 표준화(standardization)가 있다. 대부분 주가 예측 논문에서는 최소-최대 정규화 방법이 많이 이용되었다. 그러나 본 연구에서는 표준화 방법을 선택하였는데 그 이유는 다음과 같다. 주가의 경우 정해진 최댓값이 없어서 주가가 큰 폭으로 상승하는 날의 경우에는 이상치로 관측이 될 수 있다. 이 경우

스케일링하게 된다면 기존의 주가들은 모두 좁은 범위에서 형성되기 때문에 의미 있는 특성으로 고려할 수 없게 된다. 다른 데이터의 경우에는 이러한 이상치를 제거하면 되지만 주가의 경우 중요한 시점이기 때문에 제거를 할 수 없다.

Table 5. variables for forcasting stock price

Variables Name	Vaiables Explanation
Open(시가)	주식 거래가 시작될 때 첫 번째로 체결된 가격으로 당일 주가 움직임의 초기 방향성을 설정하는 중요한 요소이다.(The price at which the first trade is executed when stock trading begins is a crucial factor in setting the initial directional movement of the stock price for the day)
Volume(거래량)	주식 시장 시간 동안 거래된 주식의 총 수량을 의미하며 가격 변동의 강도와 방향을 예측하는 데 도움이 된다.(It refers to the total number of shares traded during market hours and helps in predicting the strength and direction of price fluctuations.)
USD/KRW(환율)	원/달러의 환율을 나타내는 것으로 환율이 상승하면 주가는 하락하는 경향을 보인다.(It represents the exchange rate of the Korean won to the US dollar, and when the exchange rate rises, stock prices tend to decline)
T10Y2Y(장단기 금리 차)	장기 금리에서 단기 금리를 뺀 것으로 장단기 금리 차가 역전되면 (마이너스 값을 가지면) 주가가 하락하는 경향을 보인다.(It is the difference between long-term and short-term interest rates, and when the yield curve inverts (i.e., when the difference becomes negative), stock prices tend to fall.)
VIX(공포지수)	시장의 변동성을 측정하는 지표로 공포지수라고 불린다. 공포지수가 높아질수록 주가는 하락하는 경향을 보인다.(It is an indicator that measures market volatility, commonly known as the Fear Index. As the Fear Index rises, stock prices tend to decline)
Sentiment(감성 점수)	종목토론실의 본문 내용을 KR-FinBERT를 이용하여 긍정, 부정으로 나눈 지표이다.(It is an indicator that divides the content of stock discussion boards into positive and negative using KR-FinBERT.)
Close(종가)	주식 거래 마감 전 마지막으로 체결된 가격(The price of the last trade executed before the stock market closes)

Table 6. data structure of SK Hynix

Date	Open	Volume	usd/krw	VIX	T10Y2Y	Sentiment	Close
2021-01-04	124500	7995016	1084.73	26.97	0.82	0.020	126000
2021-01-05	124500	7180224	1086.62	25.34	0.83	0.011	130500
...	...	...	...	...	...	...	...
2023-08-29	118300	2682352	1322.82	14.45	-0.75	-0.009	118600
2023-08-30	122400	2629711	1321.07	13.88	-0.78	-0.044	119400
2023-08-31	121000	4403695	1323.03	13.57	-0.76	-0.005	121800

Table 7. standardized data of SK Hynix

Date	Open	Volume	usd/krw	VIX	T10Y2Y	Sentiment	Close
2021-01-04	0.9012	2.2937	-1.7240	1.1311	0.6903	0.9814	1.0022
2021-01-05	0.9012	1.8623	-1.7026	0.8053	0.7023	0.9566	1.2625
...	...	...	...	...	...	...	...
2023-08-29	0.5450	-0.5189	0.9675	-1.3714	-1.1955	0.8016	0.5742
2023-08-30	0.7806	-0.5468	0.9477	-1.4854	-1.2316	0.4432	0.6204
2023-08-31	0.7002	0.3924	0.9698	-1.5473	-1.2075	0.8660	0.7593

Table 8. summary of train, validation and test data

Data	Period	Size
Train data	2021-01-04 ~ 2022-11-11	460
Validation data	2022-11-14 ~ 2023-04-06	100
Test data	2023-04-07 ~ 2023-08-31	100

표준화된 SK하이닉스, 셀트리온, 네이버, 현대자동차 데이터를 XGBoost, Random Forest, LSTM으로 모형화하여 주가 예측 모델을 만들려고 한다(Kim, Jeong, Song, Kim, Lee, 2023, Kim, Kim, Song, Jeong, Lee, 2023, Hwang, Yoon, 2023, Choi, Lee, Kim, 2022). 이때 XGBoost와 Random Forest는 머신러닝 모델이고 LSTM은 딥러닝 모델이기 때문에 데이터 형태를 변형하여 모델링을 해야 한다. 본 연구에서 고려할 모델은 다음과 같다.

- Model 1 : 감성 점수를 포함한 데이터를 통해 만든 주가 예측 모델(With Sentiment)
- Model 2 : 감성 점수가 없는 데이터를 통해 만든 주가 예측 모델(No Sentiment)

XGBoost와 Random Forest의 진행 과정은 사용되는 모수를 제외하면 똑같이 진행되며, LSTM은 입력 데이터 형태가 달라서 변형하여 모델링을 진행하였다. 먼저 XGBoost와 Random Forest를 통한 모델을 만들기 위해서 표준화된 데이터를 훈련 데이터, 검증 데이터, 평가 데이터로 나누는 작업을 진행한다. 이들 데이터의 비율은 각각 70%, 15%, 15%로 시간순으로 할당된다.

Table 8은 나눠진 훈련 데이터, 검증 데이터, 평가 데이터의 데이터 개수와 시간 순서로 나누어져 각 데이터의 기간을 나타낸 표이다.

모델링을 진행할 때, X에는 독립변수(Open, Volume, USD/KRW, T10Y2Y, VIX, Sentiment)가 있으며, Y에는 종속변수(Close)를 분리하여 진행한다.

Random Forest 모델에서 사용된 모수는 n_estimators, min_samples_split, min_samples_leaf이고 이들의 범위를 다음과 같이 고려하였다. n_estimator는 100, 200, 300, 500으로 설정하였고, min_samples_split은 10, 12, 16으로, min_samples_leaf는 2, 4, 6, 8, 10으로 설정하였다. 이때 최적의 모수 조합을 찾기 위해 GridSearch CV를 이용하였다.

다음으로 XGBoost 모델에서 사용된 모수는 n_estimaotrs, max_depth, subsample, colsample_bytree이다. 본 연구에서는 이선구 외(2024)의 모수 범위를 참고하여 n_estimators는 100, 200, 300, 500으로 설정하였고 max_depth는 5, 6, 7, 8로 subsample은 0.6, 0.7, 0.8, 0.9, 1로 설정하였다. 마지막으로 colsample_bytree는 0.6, 0.7, 0.8, 0.9, 1의 범위로 설정하여 GridSearch CV를 통한 최적의 모수 탐색을 진행하였다.

마지막으로 LSTM을 사용하여 예측 모델을 만들었다. XGBoost, Random Forest와 전처리 과정은 동일하게 진행한 후, 고정 사이즈의 윈도우가 이동하면서 윈도우 내의 Close를 제외한 변수들을 이용해 다음 날의 Close 값을 예상하는 슬라이딩 윈도우 방법을 이용하여 데이터를 재구축한다. 여기서 사용된 윈도우 크기는 20이다. 사용된 layer는 LSTM layer 2개와 Dense layer 2개로 총 4개의 layer를 사용하였다. 첫 번째 LSTM layer의 unit 수는 512개, 두 번째 LSTM layer의 unit 수도 512개로 설정하였으며 층마다 활성화 함수로 하이퍼볼릭탄젠트(hyperbolic tangent)를 사용하였다. 다음으로 쌓은 첫 번째 Dense layer의 unit은 128이고 마지막 Dense layer에는 Close를 출력하기 위해 unit

수를 1로 지정하였다. 또한 과적합 방지를 위해 LSTM 층마다 드롭아웃(Drop-out) 옵션을 사용하였으며 비율은 0.6이다. 드롭아웃은 딥러닝에서 과적합을 방지하기 위한 옵션으로 모델 훈련 과정에서 일부 뉴런을 임의로 제외하는 방법이다. 예를 들어 드롭아웃의 옵션을 0.3으로 지정할 때 30% 비율의 뉴런이 임의로 선택되고 선택된 뉴런의 출력을 0으로 설정한다.

그 후 최적화 함수로 Adam을 사용하며 learning rate는 0.001로 설정하여 진행했다. Epoch는 100, batch size는 32로 진행하였고 과적합 방지를 위해 early stop 옵션을 사용하여 loss가 20번 상승했을 때 학습을 중지하였다.

XGBoost, Random Forest, LSTM을 이용하여 감성 점수 특성이 있는 데이터를 이용한 예측 모델, 감성 점수 특성이 없는 데이터를 이용한 예측 모델 각각 만든 뒤 비교하여 감성 점수가 주가 예측에 영향을 미치는지 비교를 진행하였다.

3.1 모델 성능 비교

본 연구에서 감성 점수가 주가 예측에 영향을 미치는지 알기 위해 실제 종가와 예측값을 비교하여 다양한 평가 지표를 사용하였고 감성 점수가 있는 데이터의 평가 지표의 값이 감성 점수가 없는 데이터의 평가 지표보다 더 좋으면 감성 점수가 주가 예측에 영향을 미친다고 판단하였다.

평가 지표는 RMSE(Root Mean Squared Error), MAE(Mean Absolute Error)을 이용하여 Random Forest, XGBoost, LSTM을 비교하였고 그 결과는 Table 9, Table 10, Table 11에 요약해 두었다.

Table 9. model comparison using Random Forest

	Random Forest							
	SK Hynix		Hyundai motor		Naver		Celltrion	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
With Sentiment	0.1188	0.0889	0.1052	0.0800	0.0636	0.0497	0.0676	0.0505
No Sentiment	0.1322	0.1039	0.1213	0.0951	0.0753	0.0623	0.0714	0.0636

Table 10. model comparison using XGBoost

	XGBoost							
	SK Hynix		Hyundai motor		Naver		Celltrion	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
With Sentiment	0.1466	0.1128	0.1119	0.0858	0.1080	0.0862	0.1099	0.0853
No Sentiment	0.1601	0.1280	0.1311	0.0992	0.1205	0.0956	0.1343	0.1124

Table 11. model comparison using LSTM

	LSTM							
	SK Hynix		Hyundai motor		Naver		Celltrion	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
With Sentiment	0.1036	0.0801	0.1673	0.1355	0.1220	0.0899	0.1539	0.1255
No Sentiment	0.1219	0.0953	0.2223	0.1803	0.1312	0.1110	0.1682	0.1393

세 가지 분석 방법에서 With Sentiment는 감성 점수를 포함한 데이터를 사용한 예측 모델이며, No Sentiment는 감성 점수를 제거한 데이터를 사용한 예측 모델이다. Table 9의 Random Forest 결과를 보면 네 가지 종목에서 감성 점수가 있는 모델이 없는 모델보다 RMSE, MAE 값이 더 낮게 나타난 걸 알 수 있다. Table 10의 XGBoost의 결과에서도 동일하게 감성 점수가 있는 데이터의 평가 지표가 더 좋게 나왔으며, Table 11의 LSTM도 마찬가지로 감성 점수가 있는 데이터의 평가 지표가 좋게 나타났다.

XGBoost와 Random Forest는 지정한 범위 안에서 최적의 모수를 찾았기 때문에 추후의 연구에서 범위를 더 늘리고 변수를 추가한다면 성능이 좋게 나올 수 있을 것이다. 또한 LSTM도 layer 부분의 범위를 늘려서 최적의 모수를 찾는다면 더 좋은 성능이 나올 수 있다.

결론적으로 본 연구에서 종목토론실 글들이 정확하게 주가 예측을 함에 영향을 미치는 것을 확인할 수 있었다.

4. 결론

본 연구는 남녀노소 자유롭게 글을 올릴 수 있는 커뮤니티인 네이버 종목토론실의 글을 KR-FinBERT를 이용해서 감성 점수화하여 감성 점수가 주가 예측에 영향을 미치는지 머신러닝과 딥러닝 모델인 Random Forest, XGBoost, LSTM을 이용하여 예측 모델을 만들었다. 그 결과 세 가지 모델에서 전부 감성 점수가 포함된 데이터가 주가 예측에 영향을 미치는 것을 확인할 수 있었다. 감성 점수가 포함되지 않을 때보다 감성 점수가 포함되었을 때 주가 예측이 정확하였다. 기존에 활용되던 Word2Vec의 방식이 아닌 동음이의어 처리에 능숙한 BERT 모델을 사용했다는 점에서 큰 의미가 있다. 또한 금융 데이터들을 미세 조정하여 금융 전문 용어에 대한 성능이 향상되었고 더욱 정확한 감성 분류를 진행할 수 있었다. 이러한 결과는 주식 초보자들이 주가의 흐름을 파악하기 위해 가벼운 참고용으로 종목토론실의 글들을 확인한다면 도움을 받을 수 있을 것으로 보인다.

그러나 본 연구에 몇 가지 한계점이 있다. 첫 번째로 데이터 수집 기간이 2021년 1월부터 2023년 8월까지 총 32개월로 한정된 데이터를 수집하였으며 실제 주식 시장 개장일 수는 660일밖에 되지 않아 표본데이터의 수가 적은 편이다. 그러므로 수집 기간을 늘려 더 많은 표본데이터를 통한 향후 추가 연구가 필요하다. 두 번째로 변수의 수가 적다는 것이다. 본 연구에서 감성 점수를 제외하고 사용한 변수는 Open, Volume, USD/KRW, T10Y2Y, VIX 5개이다. 이러한 변수들 외에도 주가에 영향을 미치는 변수들이 더 존재하기 때문에 많은 변수를 가지고 변수 중요도를 산정하여 변수를 채택하는 방식으로 진행한다면 더 정확한 결과가 나올 것으로 예상된다. 마지막으로 모델 학습에서 사용된 모수 범위를 더 크게 늘려 최적의 모수를 찾는다면 더욱더 정확한 주가 예측이 가능할 것이라 기대한다.

따라서 향후 연구에는 다양한 변수와 넓은 모수 범위를 포함하고 추가적인 커뮤니티 데이터(다른 플랫폼의 커뮤니티 데이터, 기업 리뷰 등)를 사용하여 주가 예측 모델을 개발함으로써 본 연구의 한계점을 줄여나갈 수 있을 것이다.

References

Chang, Y. I. and Yoo, S. J. (2022). Stock Prices Prediction Using Machine Learning with News Headlines for a Pharmaceutical Company during the COVID-19 Pandemic, *The e-Business Studies*, 23(4), 237-255.

Choi, N. H., Kim, K. S., Jang, H. S. (2023). Prediction of Initial Price Formation and Impact Analysis of Pubilc Offering Stock Listing Data Using News and Comment Text: Based on XGBoost, *The Korean Journal of Financial Management*, 40(1), 81-104. DOI : 10.22510/kjofm.2023.40.1.005

Choi, S. H., Lee, S. H., Kim, M. S. (2022). Analysis of the degree of damage according to typhoon information using the influential sphere, *Journal of The Korean Data Analysis Society*, 983-993. (in Korean). DOI: https://doi.org/10.37727/jkdas.2022.24.3.983

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

Hwang Y. E., Yoon, S. H. (2023). Influencing factors of pedestrian traffic accident for junior and senior in Daegu, *Journal of The Korean Data Analysis Society*, 923-936. (in Korean). DOI: https://doi.org/10.37727/jkdas.2023.25.3.923

Kang, D. W., Yoo, S. Y., Lee, H. Y., Jeong, O. R. (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. (in Korean). DOI : 10.9708/jksci.2022.27.08.031

Kim, J., Jung, H. and Jang, B. (2023). FinBERT Fine-Tuning for Sentiment Analysis: Exploring the Effectiveness of Datasets and Hyperparameters. *Journal of Internet Computing and Services*, 24(4), 127-135

Kim T. H., Jeong, H. J. Song, J. Y. Kim, N. N., Lee, E. M. (2023). Analysis of influencing factors of suicide ideation using random forest model : Focusing on the national health and nutrition examination survey, *Journal of The Korean Data Analysis Society*, 1121-1132. (in Korean). DOI: https://doi.org/10.37727/jkdas.2023.25.3.1121

Kim, T. H., Kim, N. N., Nani, Song, J. Y., Jeong, H. J., Lee, E. N. (2023). Predicting depression om adults using a random forest model : Focusing on the 8th Korean national health and nutrition examination survey, *Journal of The Korean Data Analysis Society*, 1449-1462. (in Korean). DOI: https://doi.org/10.37727/jkdas.2023.25.4.1449

Lee, J. H. (2021). Stock flunctuation prediction study using news and social media text data : Based on deep learning, master's thesis, Sungkyunkwan University.

Lee, S. K., Yoo, S. J. (2024). Predicting Real Estate Fractional Investment Prices with the XGBOOST Model : Centered on the Kasa TE Logistics Center, *GRI REVIEW*, 26(1),1-22. (in Korean). DOI : 10.23286/gri.2024.26.1.001

Nam, K. S., Cho, H. J., Bae, J. G., Jung, J. S., Kim, S. J. (2020). A Study on the correlation and Granger causality between stock market and major interest rates in Korea and USA, *Regional industry review,* 43(3), 381-397.(in Korean). DOI : 10.33932/rir.43.3.17

Yang, Y., Uy, M. C. S., Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications, arXiv preprint arXiv:2006.08097

Yun, W., Seo, D., Min, S., Nam, H. (2021). Breast cancer classification using RandomForest and XGBoost, *The Korean Institute of Communications and Information Sciences*, 113 - 114. (in Korean).

Stock Price Prediction utilizing Sentiment Scores Based on Fin-BERT Models*

Young Min Kim[1], *Changwan Kang*[2], *Yong-Seok Choi*[3]

Abstract

Economic and technological advancements have led to a growing interest in stocks for many people. Investors want to make profits, which has led to the creation of many stock price prediction models using machine learning and deep learning models. Most studies have used news data to create stock price prediction models, but news data tends to be too neutral to accurately classify sentiment. In this study, we aim to create a stock price prediction model using NAVER Stock Discussion Room data, a community data where people of all ages can freely write opinions. In order to more accurately classify sentiment in the financial domain, we used KR-FinBERT, which is trained on Korean financial data, to perform sentiment classification. Since stock prices are most affected by recent data, we used a weighted moving average as a weight to create the final sentiment score. To check whether the sentiment scores generated from the NAVER stock discussion board data have an impact on stock price prediction, we compared the models generated from the data with sentiment scores to the models generated from the data without sentiment scores using three different analysis methodologies. Random Forest, XGBoost, and LSTM. The evaluation metrics used to compare the models were RMSE and MAE. The results of the analysis showed that the evaluation metrics of the data with sentiment scores were better in the model comparison using the three analysis methods, confirming that sentiment scores have an impact on stock price prediction.

Keywords : Sentiment score, Fin-BERT, Random Forest, XGBoost, LSTM.

---

*This Work was based on the master's thesis of first author(Young Min Kim)

[1]Graduate student, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: rladudals51@naver.com

[2]Professor, Dept. of Industrial Management & Big Data Engineering, Dongeui University, Busan, 614-714, Korea. E-mail: cwkang@deu.ac.kr

[3](Corresponding author) Professor, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63 beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: yschoi@pusan.ac.kr

[Received 8 January 2025; Revised 21 January 2025; Accepted 24 January 2025]