# KR-FinBERT 뉴스 감성분석을 활용한 KOSPI 주가지수 예측

장주현, 김재윤*
순천향대학교

kwack0202@sch.ac.kr, *kimym38@sch.ac.kr

# KOSPI Prediction Based on News Sentiment Analysis Using KR-FinBERT

Jang JooHyun, Kim Jaeyun*
Soonchunhyang Univ.

요 약

주식시장을 대표하는 종합주가지수는 경제적 상황을 대표하며 투자자의 매매 의사결정 지표로도 활용되므로 이를 예측하는데 투자자들은 오랜 관심을 가져왔다. 하지만 주가지수를 정형화된 지표만으로 예측하는 것은 지속적으로 양산되는 뉴스를 심리적 요인으로써 반영하지 못한다는 한계점이 생기게 된다. 따라서 본 논문은 KOSPI 상위 30 개 종목에 대해 뉴스들을 수집하고 KR-FinBERT 를 사용해 감성분석 기법을 적용하여 주가지수와 어떠한 관련이 있는지 분석하였다. 분석한 결과를 바탕으로 주가지수를 추정해본 결과 단순히 기술적 지표만 사용하여 주가지수를 추정할 때보다 뉴스 데이터를 고려하여 추정할 경우 예측 성과가 개선됨을 확인하였다. 더불어 시가총액의 규모가 큰 기업의 정보일수록 주가지수 추정에 더욱 큰 변동 요인으로 작용할 수 있다는 점을 확인하였다.

## Ⅰ. 서 론

주식시장에선 보통 기술적 지표와 거시적 지표 등 수치화된 지표들에 큰 변화가 발생하지 않았음에도 주가가 요동치는 현상을 볼 수 있는데, 이때 해당 종목에 관해 발생한 정보가 변동의 주 원인으로 작용하는 것을 알 수 있다. 여기서 뉴스는 투자자에게 정보를 전달해 주는 매개체 역할을 해주므로 주식시장과 뉴스는 밀접한 관계를 가지게 된다[1].

하지만 새로운 정보를 투자자가 일일이 분석하여 주가를 예측하는 것은 불가능한 일이다. 더군다나 정보의 내용이 긍정적인지, 부정적인지에 대한 판단은 보는 사람마다의 주관에 따라 달라지고 뉴스의 특성상 중립적 표현으로 전달되는 경우가 많아 그 저의를 정확하게 파악하는 것은 쉽지 않다. 이로 인해 기존의 주가 예측 연구들의 대다수는 주가지수에 영향력이 강한 수치형 지표를 선정하여 예측 모델에 적용하였다[2][3][4]. 이러한 연구들을 통해 기술적 지표와 재무정보 등이 결합된 모형을 주가 방향성 예측에 응용할 수 있음을 확인했다. 하지만 지속적으로 양산되는 뉴스 정보를 주가 예측에 반영하지 못하는 한계가 존재한다.

이러한 배경 속에서 본 논문은 주가지수를 추정하는데 뉴스에서 전해지는 감성의 정도를 반영하였을 때의 영향을 조사하기 위해 뉴스 데이터에 대한 감성분석과 XGBoost 모형을 사용하여 주가지수 예측을 진행하였다. 기존의 자연어 처리 분야에서 주로 사용한 Word2Vec 나 Bag-of-Words 보다 우수한 성능을 보여주는 BERT 를 감성분석 모델로 선정하였고, 그 중에서도 한국 금융 분야에 특화된 KR-FinBERT 를 최종 선정하였다.

## Ⅱ. 본론

본 논문에서는 기술적 지표만을 사용하여 주가지수를 예측할 경우와 기술적 지표에 감성분석의 결과를 고려하여 예측할 경우를 비교하여 감성분석의 결과가 주가지수를 예측하는 데에 유용함을 살펴보고자 한다.

### 2.1 Data

KOSPI 지수를 구성하는 종목 중 한국거래소 기준 시가총액 규모 상위 30 개의 종목을 선정하였다. 데이터 수집 기간 (5 년)에 맞춰 상장된 지 5 년 이하의 종목을 제외하고 30 개를 선정하였으며, 이들 종목은 코스피 전체 시가총액 규모의 약 60%를 구성하였다. 시가총액의 순위는 2022 년 마지막 거래일을 기준으로 하였다.

뉴스 데이터는 뉴스 정보를 제공하는 빅카인즈 (BIG KINDS)에서 경제부문의 표제 (headline)를 2018 년부터 2022 년에 거쳐 총 73,539 개를 수집하였다. 당일의 장 마감부터 다음날의 장 마감까지의 뉴스가 다음날의 주가지수 종가에 영향을 미치므로 시간대에 맞게 뉴스데이터의 일자를 바꾸어 주었다 (Table 1).

기술적 지표 데이터와 KOSPI 는 TA-Lib 와 파이썬 모듈 FinanceDataReader 를 사용해 수집하였다.

Table 1. Example of changing News Date

| Headline | Date | Prediction Date |
|---|---|---|
| News A | 2022-01-03 11:00 | 2022-01-03 |
| News B | 2022-01-03 20:00 | 2022-01-04 |

### 2.2 KR-FinBERT

뉴스 데이터의 표제에 대해 KR-FinBERT 모듈을 사용하여 감성분석을 진행하였다. KR-FinBERT 는 BERT 모듈 중 한국 금융 분야에 특화된 모듈이며 텍스트에 대하여 토큰화와 불용어 제거 과정을 수행할 수 있다. 전처리된 텍스트가 입력되면 긍정, 중립, 부정의 부문에서 각자 어느 정도의 비율을 내포하는지 환산해준다[5]. 뉴스 별로 가장 높은 비율의 감성을

주된 감성, 해당 비율을 감성 점수 (sentiment score)로 설정했고 수집된 73,539 개의 뉴스 중 긍정 38,590 개, 중립 22,106 개, 부정 12,843 개로 집계했다. 여기서 중립의 감성은 주가 변동에 기여하는 정도가 없다고 여겨 긍정과 부정의 뉴스만 주가지수 추정에 활용하였다.

### 2.3 감성도와 반응도의 이동평균선 생성

KR-FinBERT 를 통해 산출한 감성 점수 (sentiment score)에 대하여 긍정의 뉴스는 양수, 부정의 뉴스는 음수 처리를 해준 뒤 일별로 감성도 (sentiment degree)와 반응도 (sensitive degree)를 도출하였다.

Sentiment degree = Average (Sentiment score)*W
Sensitive degree = Average (|Sentiment score|)*W

여기서 가중치 (W)는 동일 가중치 (벤치마킹)와 시총 비율, 뉴스 빈도 비율로 설정했다. 시총 비율과 뉴스 빈도 비율로 산출한 지수들이 벤치마킹보다 주가지수 예측함에 있어 우수한지를 살펴보고자 한다.

생성한 지수들에 이동평균선을 생성하여 과거 감성 스코어를 활용하는 기간을 비교해 감성분석의 결과가 주가지수에 대해 갖는 영향력을 조사하였다.

### 2.4 실험 결과

감성도, 반응도와 주가지수 사이의 관계와 설명력을 파악하기 위해 상관분석을 진행하였다. Table2 와 같이 생성한 지수들이 주가지수와 양의 상관성을 보이며 시가총액을 가중치로 설정하여 환산한 지수가 가장 강한 상관성을 보이며 이동평균선의 기간이 늘어나면서 상관의 정도가 강해짐을 알 수 있다.

Table3 은 기술적 지표 (모멘텀)만 사용하여 예측한 결과와 가중치 별로 환산한 지수들을 사용하여 예측한 결과들의 평가지표 (RMSE)이다. 가중치를 시가 총액과 뉴스 빈도 비율로 하였을 때, 벤치마킹 (동일 가중치)보다 예측력이 더 우수한 것을 확인할 수 있다.

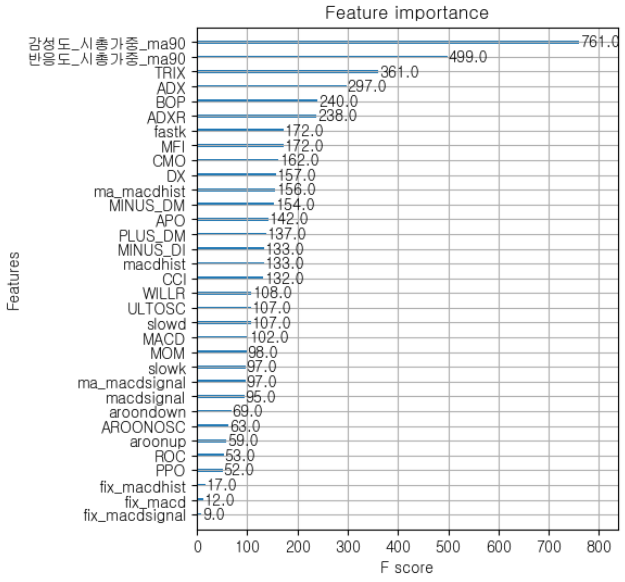


Fig. 1 Feature importance

Table 2. Summary of correlation results

| Period | 벤치마킹 (동일가중) | 뉴스빈도 가중평균 | 시가총액 가중평균 |
|---|---|---|---|
| 1 | 0.059 | 0.075 | **0.102** |
| 7 | 0.207 | 0.212 | 0.239 |
| 15 | 0.255 | 0.243 | 0.291 |
| 30 | 0.354 | 0.316 | 0.377 |
| 90 | 0.478 | 0.390 | **0.502** |

Table 3. Summary of prediction results

| 이동평균 기간 | 모멘텀 | 벤치마킹 (동일가중) | 뉴스빈도 가중평균 | 시가총액 가중평균 |
|---|---|---|---|---|
| 1 일 | **562.48** | 557.67 | 561.16 | 563.81 |
| 7 일 | - | 545.84 | 557.70 | 548.14 |
| 15 일 | - | 544.05 | 548.31 | 567.85 |
| 30 일 | - | 520.11 | 469.23 | 524.00 |
| 90 일 | - | 561.82 | **426.99** | **309.06** |

## Ⅲ. 결론

본 논문에서는 뉴스 데이터에 드러난 감성이 주가지수 예측의 활용 가능성에 대해 비교 분석하였다. 분석 결과 단순히 기술적 지표 (모멘텀 지표)만을 사용해 주가지수를 예측할 때 보다 감성분석의 결과를 고려한 경우의 예측 성과가 개선되는 것을 알 수 있었다. 시총 비율과 뉴스 빈도 비율을 가중치로 선정하여 환산한 지수들이 주가지수 예측에서 더욱 우수한 성능이 나타나는 것이 확인돼 주가지수는 시가총액의 규모가 크거나 뉴스의 발생이 빈번한 기업의 정보에 영향을 많이 받는 것으로 해석하였다. 하지만 본 연구에는 한계점이 존재하는데 첫 번째로 시가총액의 비율을 가장 최근 시점을 기준으로 설정하여 5 년간의 변화되는 시가총액의 비율을 고려하지 않았다는 점과 현재 국내 시장에서 가중치가 삼성전자에 크게 쏠려 나머지 기업들의 시총 규모가 작지 않음에도 기여하는 정도가 작아지게 되었다. 두 번째는 뉴스 기사의 표제 만을 사용하여 본문을 사용할 때 보다 감성분석의 정확도가 떨어질 수 있다는 점이다. 따라서, 향후 연구가 이루어진다면 앞서 말한 한계점을 보완하며, 기술적 지표 중 변수 선택법을 통해 감성분석의 결과와 최적의 조합식을 제시할 예정이다.

## ACKNOWLEDGMENT

본 연구는 2023 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2021-0-01399).

## 참 고 문 헌

[1] 송치영. (2002). 뉴스가 금융시장에 미치는 영향에 관한 연구. *국제경제연구, 8*(3), 1-34.

[2] 이우식. (2017). 딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측. *한국데이터정보과학회지, 28*(2), 287-295.

[3] 허준영, & 양진용. (2015). SVM 기반의 재무 정보를 이용한 주가 예측. *정보과학회 컴퓨팅의 실제 논문지, 21*(3), 167-172.

[4] 어균선, & 이건창. (2020). 합성곱 신경망을 이용한 주가방향 예측: 상관관계 속성선택 방법을 중심으로. *Information Systems Review, 22*(4), 21-39.

[5] https://huggingface.co/snunlp/KR-FinBert-SC