https://doi.org/10.7840/kics.2023.48.6.748

뉴스의 감성 분석을 사용한 주가 예측 방법론

김 대 겸◆, 윤 민 혁*,
조 영 진°, 최 용 훈°°

Stock Price Prediction Method Using Sentiment Analysis of News

Daegyeom Kim◆, Min-Hyeok Yun*,
Young-Jin Cho°, Yong-Hoon Choi°°

요 약

딥러닝 모델을 사용하여 주식 가격을 예측하려는 연구들은 꾸준히 진행되고 있다. 주식 예측 딥러닝 모델들은 기본적으로 시계열 특성을 학습하고 이를 기반으로 미래값을 예측하는 구조를 가진다. 하지만 주식 가격은 외부 요인에 많은 영향을 받기 때문에 기술적인 데이터를 학습한 시계열 예측 모델로 주가를 예측하면 정확도가 떨어진다. 본 논문에서는 주가 데이터와 경제 뉴스 텍스트에서 추출한 감성 표현을 사용해 주식 가격을 예측하는 방법론을 제안한다. 주가 데이터와 감성 표현을 모델의 입력에 동시에 사용하는 것이 아닌 주가 데이터의 중간 표현과 감성 표현을 결합하여 사용하는 방식을 제안한다.

Key Words: stock price prediction, sentiment analysis, technical data, sequence-to-sequence model, transformer

ABSTRACT

Research using deep learning models to predict stock prices is constantly ongoing. Stock prediction deep learning models typically learn the time-series characteristics and use them to predict future values. However, because stock prices are heavily influenced by external factors, predicting stock prices using time-series prediction models that only learn technical data results in lower accuracy. In this paper, we propose a methodology for predicting stock prices using sentiment representations extracted from both stock price data and economic news text. Instead of using stock price data and sentiment representations as inputs to the model simultaneously, we propose a method of combining the intermediate representation of the stock price data and the sentiment representations.

Ⅰ. 서 론

주식 가격을 예측하는 것은 기존의 많은 연구에도 불구하고 좋은 성과를 내지 못했다. 주가 예측이 어려운 이유는 여러 가지가 있지만 가장 큰 요인 중의 하나는 외부 변수에 의해 시계열 특성이 무시될 만한 큰 영향을 받는다는 것이다.

전통적인 시계열 예측 기법은 데이터의 평균값과 샘플간의 상관관계를 분석하여 추세(trend)와 계절성(seasonality)과 같은 특성을 파악하고 예측 모형을 수립한 후, 다음 단계에 어떤 값이 가장 적절한지 예측한다. 최신의 기계학습 기반의 주가 예측 모델들도 기술적인 데이터들(technical data)의 잠재된 특성들을 파악하도록 학습하고 주어진 입력값에 대한 예측구간의 추론값을 출력하게 된다. 하지만 주식 가격은 추세나 계절성 같은 특성이 매우 불규칙하며 외부 이벤트에 의한 사람들의 관심에 많은 영향을 받기 때문에 기계학습 모델이 주어진 손실함수에 대한 오차를 최소화하도록 학습하더라도 테스트 입력에 대해서는 좋은 성능을 나타내지 못한다. 시계열 데이터 예측 분야에서 가장 우수한 성능을 나타내고 있는 Autoformer, Informer, N-HiTS와 같은 최신의 기계학습 모델들도 주가 예측 문제에서는 좋은 성능을 나타내지 못하였다.

본 논문에서는 수집한 뉴스 데이터에서 추출한 감성 표현을 주식 가격을 입력으로 하는 인코더의 출력과 결합하여 다음 시점의 주식 가격을 예측하는 방법론을 제안한다. 본 논문에서는 뉴스 감성 분석 모델을

※ 이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1064080).
◆ First Author : (0000-0003-1811-1529) Division of Robotics, Kwangwoon University, 0913ktg@kw.ac.kr, 학생회원
° Corresponding Author : (0009-0002-8810-6784) Division of Robotics, Kwangwoon University, yjaycho@gmail.com
°° Corresponding Author : (0000-0002-1460-0520) Division of Robotics, Kwangwoon University, yhchoi@kw.ac.kr, 교수, 종신회원
* (0009-0008-0137-238X) Division of Robotics, Kwangwoon University, gurals9368@gmail.com
논문번호 : 202304-086-C-LU, Received April 22, 2023; Revised May 2, 2023; Accepted May 2, 2023

다양한 기술데이터 예측 신경망 모델과 결합하여 실험을 수행하였다.

Ⅱ. 관련 연구

관련 연구 [1]에서는 주가에 영향을 미치는 소셜 미디어 중 하나인 트위터 글의 감성을 분석하기 위해 이진으로 분류 가능한 OF (Opinion Finder)와 6가지로 감성 상태를 나타낼 수 있는 GPOMS (Google Profile of Mood States)를 사용해 소셜 미디어 정보를 고려하는 연구를 수행하였다. 또한 관련 연구 [2]와 [3]에서는 주가를 예측하기 위해 지표 데이터와 주가에 영향을 줄 수 있는 다른 데이터를 함께 사용했다. 그리고 관련 연구 [4]에서는 소셜 미디어 등의 웹 데이터에서 불안, 걱정, 두려움 등 부정적 감정 표현이 증가함에 따라 S&P500 지수가 하락한다는 발견하여 소셜 미디어의 반응이 주식의 가격과 연관 있다는 결과를 도출했다. 관련 연구 [5]는 GAN (Generative Adversarial Network) 모델을 이용하여 시장의 감성 상태를 주가 예측에 반영하였다.

인공지능 기술을 활용해 지표 데이터와 다른 데이터를 복합적으로 사용한 앞선 연구의 연구 결과는 단순 지표만을 사용해 예측하는 것이 아닌 뉴스와 소셜 미디어 등 다른 정보들을 복합적으로 사용해 데이터의 단편성을 낮춰 진행한 주가 예측이 더욱 의미 있는 결과를 도출하는데 유리하다는 결과를 보였다.

다양한 시행착오를 통해 우리는 감성 정보를 반영하기 위해 어텐션(attention) 모듈이 있는 인코더-디코더 구조를 사용한 시계열 예측 모델이 적합하다고 판단했다. 본 논문에서 적용한 인코더-디코더 구조의 기계학습 모델은 다음과 같은 3가지이다. 첫째는 전형적인 seq2seq 모델[6]이고, 둘째는 트랜스포머(transformer) 모델[7]이며, 셋째는 듀얼 스테이지 어텐션 모델[8]이다. 세 가지 모델을 모두 실험에 사용하였다.

Ⅲ. 뉴스 감성 표현을 사용한 주가 예측 방법론

본 논문에서는 뉴스 감성 표현을 사용한 주가 예측 방법론을 제안한다. 매일 경제 홈페이지에서 수집한 경제 뉴스와 네이버 종목토론방에서 수집한 종목 토론 데이터로 TF-IDF (Term Frequency Inverse Document Frequency) 사전을 생성한다. 만들어진 TF-IDF 사전을 사용해서 각 뉴스에 가장 유사한 주식 코드를 연결하고 경제 데이터로 사전 학습된 KR-FinBERT를 사용해 뉴스의 감성 표현을 추출한다. 감성 표현으로 분류층을 통과하기 직전 클래스 토큰의 마지막 히든 상태 (hidden state)를 사용한다.

표 1. 감성 표현을 사용하지 않은 것과 사용한 것에 따른 주가 예측 결과
Table 1. Stock Price Prediction Results Based on Not Using and Using Sentiment Analysis

적용모델	성능지표	감성미사용	감성사용
seq2seq [6]	Accuracy	0.48780	0.46770
	MSE	0.00005	0.00008
Transformer [7]	Accuracy	0.48630	0.50080
	MSE	0.02900	0.00321
DARNN [8]	Accuracy	0.45260	0.46240
	MSE	0.00013	0.00008

주가 데이터는 대신 증권의 CYBOS API를 사용해 2020년 1월부터 2022년 1월까지의 1분 봉 데이터를 수집했다. 뉴스가 사람들에게 읽히기까지 전파되는 시간이 필요하고 뉴스가 1분 단위로 사용될 만큼 충분치 않기 때문에 뉴스 데이터와 주가 데이터를 동일 시간 30분 단위로 묶어서 처리했다.

감성 표현을 모델의 중간 표현과 결합하여 학습하기 위해 인코더-디코더 구조를 가지는 seq2seq[6], Transformer[7], DARNN[8] 모델을 사용한다. 그림 1은 DARNN의 구조를 나타내고 주가 데이터와 뉴스 감성 표현이 제안한 방법론에서 학습에 사용되는 구조를 나타낸다. 모델 학습에 사용된 하이퍼파라미터들은 다음과 같다. 학습 데이터와 테스트 데이터는 9:1의 비율로 설정하였고, 데이터 정규화를 위한 스케일러는 MinMaxScaler(0~1), 손실 함수는 평균제곱오차(MSE: Mean Square Error), 입력 크기는 14, 출력 크기는 1, 학습율은 0.001, 학습회수(epoch)는 1000회, 배치크기는 128이다.

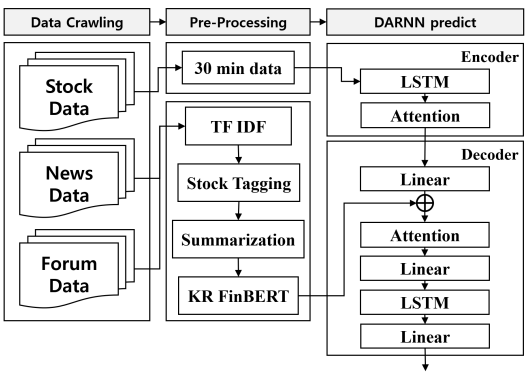

그림 1. 제안하는 모델 구조
Fig. 1. The proposed model architecture

Ⅳ.실험 결과

뉴스 데이터에서 가장 많이 언급된 3가지 종목(기업은행, KB금융, 하나금융지주)을 모델 평가에 사용했다. 감성 표현을 추가했을 때의 성능 향상을 확인하기 위해 주가 데이터만 사용해 예측한 결과와 감성 표현과 주가 데이터를 같이 사용해 예측한 결과를 표1에 비교하였다. 정확도(accuracy)는 모델이 예측한 주가의 등락을 테스트 데이터와 비교한 지표이다. 수식(1)은 정확도 a를 구하는 식이다.

$$a = \frac{1}{N-1}\sum_{t=0}^{N-1} b,\ b = \begin{cases} 1, & (L_{t+1} - L_t)(P_{t+1} - P_t) > 0 \\ 0, & otherwise \end{cases} \quad (1)$$

여기서 N은 테스트 데이터의 총길이이고, L은 테스트 데이터(주식 종가)의 라벨 리스트, P는 모델의 예측을 모아둔 예측 리스트이다. 정확도와 평균제곱오차(MSE) 모두 3가지 종목에서 가장 높은값을 표1에 나타내었다.

LSTM (Long-Short Term Memory) 기반 모델인 seq2seq[6]는 감성 표현을 추가했을 때 정확도가 2% 하락했고 LSTM과 어텐션(attention)을 결합한 DARNN[8] 모델은 정확도가 1% 상승했다. 자기 회귀 모델을 사용하지 않고 멀티 헤드 어텐션을 사용하는 트랜스포머[7] 모델은 감성 표현을 추가했을 때 정확도가 2% 향상되고 손실은 10배 감소하였다. 따라서 제안하는 방법론으로 감성 표현을 사용할 때 자기 회귀 모델보다 감성 표현에서 중요한 부분을 스스로 찾을 수 있는 어텐션 기반 모델에서 좋은 결과를 얻는 것을 확인하였다.

Ⅴ. 결 론

뉴스로부터 추출한 감정들을 주가 예측에 사용할 수 있는 기계학습 구조로 대표적인 3가지 인코더-디코더 기반의 모델들을 사용하였다. 주가의 기술 데이터(technical data)들을 학습하는 인코더의 출력에 감성 표현을 결합하여 디코더의 입력으로 사용하는 실험을 통해 자기 회귀 기반 모델보다 어텐션 기반 모델이 제안하는 방법론에 더 적합하다는 것을 확인하였다. 자기 회귀와 어텐션이 결합된 모델에서도 약간의 성능 향상을 관찰하였으므로 추가적인 연구를 통해 뉴스로부터 추출한 감정들을 주가 예측에 적용하는데 적합한 모델을 지속적으로 개발할 예정이다.

References

[1] R. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 12 pp. 11-19, Mar. 2009. (https://doi.org/10.1145/1462198.1462204)

[2] A. Kloptchenko, et al., "Combining data and text mining techniques for analyzing financial reports," *Int. Syst. in Accounting, Finance and Manag.*, vol. 12, no. 1, pp. 29-41, Mar. 2004. (https://doi.org/10.1002/isaf.239)

[3] K. S. Eo and K. C. Lee, "Predicting stock price direction by using data mining methods," *J. Korea Soc. Comput. and Inf.,* vol. 22, no. 11, pp. 111-116, Nov. 2017.

[4] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," *Int. AAAI Conf. Weblogs and Soc. Media*, vol. 4, no. 1, pp. 58-65, May 2010.

[5] R. Jadhav, S. Sinha, S. Wattamwar, and P. Kosamkar, "Leverag-ing market sentiment for stock price prediction using GAN," *2021 2nd GCAT*, pp. 1-6, 2021. (https://doi.org/10.1109/GCAT52182.2021.9587497)

[6] I. Sutskever, et al., "Sequence to sequence learning with neural networks," *Advances in NIPS*, vol. 27, 2014.

[7] A. Vaswani, et al., "Attention is all you need," *NIPS 2017*, vol. 30, 2017.

[8] Y. Qin, et al., "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. Twenty-Sixth IJCAI-17*, pp. 2627-2633, 2017.