http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.27, No.1, February 2022, pp.63-79
https://doi.org/10.7838/jsebs.2022.27.1.063

호가창과 뉴스 헤드라인을 이용한 딥러닝 기반 주가 변동 예측 기법

Deep Learning-based Stock Price Prediction Using Limit Order Books and News Headlines

류의림(Euirim Ryoo)*, 이기용(Ki Yong Lee)**, 정연돈(Yon Dohn Chung)***

초 록

최근 머신러닝 및 딥러닝 기법을 활용한 주식 가격 예측 연구가 다양하게 이루어지고 있다. 그 중에서도 최근에는 주식 매수 및 매도 주문 정보를 담고 있는 호가창을 이용하여 주가를 예측하려는 연구가 시도되고 있다. 하지만 호가창을 활용한 연구는 대부분 가장 최근 일정 기간 동안의 호가창 추이만을 고려하며, 호가창의 중기 추이와 단기 추이를 같이 고려하는 연구는 거의 진행되지 않았다. 이에 본 논문에서는 호가창의 중기와 단기 추이를 모두 고려하여 주가 등락을 보다 정확히 예측하는 딥러닝 기반 예측 모델을 제안한다. 더욱이 본 논문에서 제안하는 모델은 중단기 호가창 정보 외에도 해당 종목에 대한 동기간 뉴스 헤드라인까지 고려하여 기업의 정성적 상황까지 주가 예측에 반영한다. 본 논문에서 제안하는 딥러닝 기반 예측 모델은 호가창 변화의 특징을 합성곱 신경망으로 추출하고 뉴스 헤드라인의 특징을 Word2vec을 이용하여 추출한 뒤, 이들 정보를 결합하여 특정 기업 주식의 다음 날 등락 여부를 예측한다. 실제 NASDAQ 호가창 데이터와 뉴스 헤드라인 데이터를 사용하여 제안 모델로 5개 종목(Amazon, Apple, Facebook, Google, Tesla)의 일일 주가 등락을 예측한 결과, 제안 모델은 기존 모델에 비해 정확도를 최대 17.66%p, 평균 14.47%p 향상시켰다. 또한 해당 모델로 모의 투자를 수행한 결과, 21 영업일 동안 종목에 따라 최소 $492.46, 최대 $2,840.83의 수익을 얻었다.

ABSTRACT

Recently, various studies have been conducted on stock price prediction using machine learning and deep learning techniques. Among these studies, the latest studies have attempted to predict stock prices using limit order books, which contain buy and sell order information of stocks. However, most of the studies using limit order books consider only

이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1A2C1012543). 또한, 이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2019H1D8A2105513).

* First Author, Graduate Student, Department of Big Data Analysis and Convergence, Sookmyung Women's University (euirimryoo@sookmyung.ac.kr)

** Corresponding Author, Professor, Division of Computer Science, Sookmyung Women's University (kiyonglee@sookmyung.ac.kr)

*** Co-Author, Professor, Department of Computer Science & Engineering, Korea University (ydchung@korea.ac.kr)

Received: 2021-11-15, Review completed: 2021-12-22, Accepted: 2022-01-06

the trend of limit order books over the most recent period of a specified length, and few studies consider both the medium and short term trends of limit order books. Therefore, in this paper, we propose a deep learning-based prediction model that predicts stock price more accurately by considering both the medium and short term trends of limit order books. Moreover, the proposed model considers news headlines during the same period to reflect the qualitative status of the company in the stock price prediction. The proposed model extracts the features of changes in limit order books with CNNs and the features of news headlines using Word2vec, and combines these information to predict whether a particular company's stock will rise or fall the next day. We conducted experiments to predict the daily stock price fluctuations of five stocks (Amazon, Apple, Facebook, Google, Tesla) with the proposed model using the real NASDAQ limit order book data and news headline data, and the proposed model improved the accuracy by up to 17.66%p and the average by 14.47%p on average. In addition, we conducted a simulated investment with the proposed model and earned a minimum of $492.46 and a maximum of $2,840.93 depending on the stock for 21 business days.

키워드 : 주가예측, 호가창, 뉴스, 합성곱 신경망, Word2vec
Stock Price Prediction, Limit Order Book, News, CNN, Word2vec

1. 서 론

주식 가격의 예측은 경제, 수학 등의 다양한 영역에서 계속해서 도전되어온 문제이다. 그러나 복잡하고 유동적으로 움직이는 주가의 특성상 기존의 통계 기반 방법으로 주가의 패턴을 예측하기에는 많은 어려움이 있었다. 이에 최근 인공지능을 활용한 연구가 활발히 수행되면서 머신러닝 및 딥러닝 기법으로 주식 가격을 예측하려는 시도가 다양하게 이루어져 왔다[3, 12]. 기존의 딥러닝 기반 주가 예측 방법들은 주로 과거의 주식 가격 혹은 거래량을 사용하여 주가의 등락을 예측하였다. 하지만 최근 들어 주식의 매매 주문 정보를 담고 있는 호가창(limit order book)을 이용하여 주가를 예측하려는 연구가 시도되고 있다[10, 14]. 호가창은 주식 매수 및 매도 주문에 대한 호가 및 주문량을 나타내는 정보이다. 따라서 이러한 호가창 데이터를 사용하면 과거의 주식 가격과 거래량 등을 사용하는 것에 비해 보다 다차원적으로 주가를 예측할 수 있다는 장점이 있다. <Figure 1>은 실제 호가창 데이터를 나타내는 예시 화면이다.

한편 주가에 영향을 미치는 요소는 이러한 수치 정보 외에도 회사의 최근 상황을 포함하는 뉴스에서도 발견될 수 있다. 따라서 본 논문에서는 기업의 호가창 정보뿐만 아니라 해당 기업에 대한 뉴스의 헤드라인까지 사용하여 해당 기업의 주가 등락을 예측하는 딥러닝 기반 모델을 제안한다.

호가창을 이용하는 대부분의 기존 연구는 가장 최근 일정 기간 동안의 호가창의 변화만을 보고 다음 날 주가 등락을 예측한다. 따라서 본 논문의 선행 연구에서는 호가창의 중기 추이와 단기 추이를 모두 고려하여 주가 등락을 예측한 딥러닝 모델을 제시하였다[9]. 하지만 Ryoo

〈Figure 1〉 Example of Limit Order Book Data

et al.[9]는 뉴스 헤드라인에 대해서는 중기 추이와 단기 추이를 고려하지 못했다는 한계가 있다. 본 논문에서는 호가창과 뉴스 헤드라인 모두에 대해 중기 추이와 단기 추이를 모두 고려하여 주가 등락을 예측하는 딥러닝 기반 예측 모델을 제안한다. 중기 추이는 해당 데이터의 전반적인 추세를 나타내고 단기 추이는 바로 직전 추세를 나타낸다. 따라서 이들을 모두 활용하면 예측 정확도를 더욱 높일 수 있다. 본 논문에서 제안하는 모델은 호가창의 변화에 대한 특징(feature)을 합성곱 신경망으로 추출하고 뉴스 헤드라인의 특징을 Word2Vec을 이용하여 추출한 뒤, 이들 정보를 결합하여 특정 기업의 다음 날 주식 가격의 등락(상승, 하락, 유지)을 예측한다.

본 논문의 구성은 다음과 같다. 제2장에서 머신러닝 및 딥러닝을 활용한 기존의 주가 예측 모델들을 살펴보고, 제3장에서는 본 논문에서 주가 예측에 사용한 데이터를 설명한다. 제4장에서는 본 논문에서 제안하는 딥러닝 기반 주가 예측 모델을 상세히 설명하며, 제5장에서는 제안하는 모델의 성능평가 결과를 제시하고 모델의 수익성을 검증한다. 마지막으로 제6장에서는 결론을 맺는다.

2. 관련 연구

본 장에서는 본 논문의 관련 연구로서 기존의 주가 예측 모델들을 살펴본다. 미래 주식 가격의 움직임을 예측하는 것은 투자자들에게 매우 중요한 문제이기 때문에 다양한 데이터를 활용한 주가 예측 연구들이 수행되었다. 따라서 본 장에서는 주가 예측 기법에 사용된 데이터를 기준으로 기존의 주가 예측 연구를 살펴본다.

2.1 기술적 지표를 사용하는 주가 예측 연구

일반적으로 대부분의 주가 예측 연구는 과거의 주식 가격 및 기술적 지표들을 사용한다[2, 15]. 과거의 주식 가격을 사용하는 연구의 경우 대부분 시가(open price), 고가(high price), 저가(low price), 종가(close price)로 이루어진

OHLC 데이터와 거래량을 사용한다. 기술적 지표의 예로는 주가의 볼린저 밴드(Bollinger Bands), 이동평균(moving average), 이동평균 수렴확산(moving average convergence divergence, MACD) 등이 있다. Zhou et al.[15]은 이러한 기술적 지표들과 거래량 및 종가를 사용한 GAN(generative adversarial network) 기반의 주식 예측 모델을 제안했다. Bao et al.[2]는 OHLC 데이터, 13가지 기술적 지표와 거시변수들까지 고려하여 적층 오토인코더(stacked autoencoder)로 이들의 특징을 추출하고, LSTM(long short-term memory)을 사용하여 다음날의 종가를 예측하였다. 하지만 이들 모두 정량적 수치만 반영한 것으로 뉴스와 같은 정성적 데이터를 반영하지 못한다는 한계를 갖는다.

2.2 뉴스 데이터를 사용하는 주가 예측 연구

주가의 변화는 기술적 지표와 같은 수치들에 나타날 수도 있지만 회사의 최근 정보를 포함하는 뉴스에도 나타날 수 있다. 따라서 온라인 뉴스의 텍스트를 활용한 주가 예측 연구들이 진행된 바가 있다. Li et al.[6]는 뉴스가 주가에 미치는 영향을 감성분석을 통해 연구했고, Lee and Soo[5]는 뉴스 데이터를 사용하여 순환신경망(recurrent neural network, RNN)과 CNN을 합친 RCNN을 기반으로 한 주가 예측 모델을 제시하였다. Vargas et al.[11], Akita et al.[1], Peng and Jiang[8]은 과거의 주식 가격과 온라인 뉴스를 함께 고려하여 주가의 움직임을 예측한 연구이다. Peng and Jiang[8]는 이전 5일의 종가와 뉴스 헤드라인의 워드임베딩을 활용한 DNN(deep neural networks) 기반의 예측 모델을 제안했다. Akita et al.[1]는 Word2vec을 확장하여 한 단락을 단어와 동일하게 고려한 단락 벡터(paragraph vector)를 사용하여 뉴스의 정보를 나타냈고, 이에 주식의 고가와 저가를 함께 고려했다. 특히 Vargas et al.[11]은 기술적 지표와 단어 임베딩을 사용한 주가 예측 모델을 제안하였다. Vargas et al.[11]은 본 연구와 마찬가지로 Word2Vec 모델을 사용하여 뉴스 헤드라인의 단어들에 대한 임베딩 벡터를 생성하고, 그 벡터들의 평균 벡터를 뉴스 헤드라인을 나타내는 임베딩 벡터로 사용했다는 공통점이 있다. 위의 세 연구는 주가 관련 수치 데이터와 뉴스 텍스트 데이터를 결합하여 고려했다는 점에서 본 연구와 공통점이 있지만, 모두 고정된 한 기간에 대한 데이터만 고려하며 중기 추이와 단기 추이를 같이 고려하지 못한다는 한계를 갖는다.

2.3 호가창을 사용하는 주가 예측 기법

서론에서 언급한 바와 같이 최근에는 호가창을 사용하여 주가를 예측하려는 연구가 진행되고 있다. Tsantekidis et al.[10]은 LSTM으로 호가창 변화의 특징을 추출하여 주가를 예측하며, Zhang et al.[14]은 CNN 및 인셉션 모듈(Inception module)로 개별 호가창 데이터의 특징을 추출한 뒤, 이들의 변화를 LSTM으로 학습하여 주가를 예측한다. 하지만 Tsantekidis et al.[10]과 Zhang et al.[14]은 모두 뉴스 헤드라인과 같은 정성적 데이터는 반영하지 못했다는 점과 가장 최근 일정 기간 동안의 호가창의 변화만을 보고 주가 등락을 예측한다는 점에서 한계가 있다.

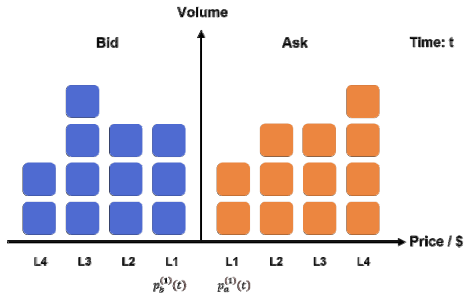
본 논문의 선행 연구인 Ryoo et al.[9]에서는 호가창의 중기 추이와 단기 추이를 동시에 고려한 모델을 제시하였다. 하지만 Ryoo et al.[9]에서는 뉴스에 대해서는 중기 추이와 단기 추이를 동시에 고려하지 못했다. 따라서 본 논문에서는 Ryoo et al.[9]의 내용을 발전시켜 호가창과 뉴스의 중기 추이와 단기 추이를 모두 고려하는 딥러닝 기반 예측 모델을 제시하고 그의 실험결과를 제시한다.

3. 주가 예측에 사용된 데이터

3.1 호가창 데이터

호가창은 주식 거래를 위해 제출한 매도(ask) 주문과 매수(bid) 주문의 수량을 호가별로 기록한 정보이다. 어떤 종목에 대한 호가창은 각 시간 t 시점에서의 매도 주문호가, 매도 주문량과 매수 주문호가, 매수 주문량을 포함하고 있다. <Figure 2>는 어떤 시간 t에서의 호가창 데이터 일부를 나타낸다. 주문은 제출된 호가에 따라 여러 개의 레벨로 나뉘는데, <Figure 2>에서 L1은 가장 낮은 가격의 레벨(레벨 1)을 뜻하며, L4는 가장 높은 가격의 레벨(레벨 4)을 뜻한다. 주어진 어떤 종목에 대해, 본 논문에서는 시간 t에서 레벨 l의 매도 주문호가와 매도 주문량을 각각 $p_a^{(l)}(t)$과 $v_a^{(l)}(t)$로 표시한다. 이와 유사하게 시간 t에서 레벨 l의 매수 주문호가와 매수 주문량을 각각 $p_b^{(l)}(t)$과 $v_b^{(l)}(t)$로 표시한다.

일반적으로 나스닥(NASDAQ)과 같은 주식시장의 호가창 데이터는 나노초(nanosecond) 단위로 매우 빈번하게 변화가 기록된다. 하지만 본 논문에서는 그 정도까지의 상세한 정보는 필요하지 않다고 판단하여 1시간 단위로 호가창 데이터를 묶어 그들의 평균값을 주가 등락 예측에 사용하였다. 또한 레벨이 너무 높은 호가의 주문은 실제 거래로 연결되지 않기 때문에 레벨 1부터 10까지의 주문만 고려하였다.



〈Figure 2〉 Limit Order Book Example at Time t

3.2 뉴스 헤드라인

본 논문에서는 기업에 대한 뉴스를 검색하여 검색된 뉴스의 헤드라인을 모델 훈련에 사용한다. 이를 위해 본 논문에서는 뉴스 사이트에서 기업 뉴스를 검색하고, 검색된 뉴스의 헤드라인을 고정된 길이의 임베딩 벡터로 변환하였다. 검색된 뉴스의 헤드라인을 임베딩 벡터로 변환할 때는 뉴스 헤드라인에 포함된 단어들을 추출하고 불용어를 제거한 뒤, 남은 단어들 각각을 Word2vec을 사용하여 임베딩 벡터로 변환하였다. 본 논문에서 사용한 Word2vec 모델은 Google이 제공하는 사전 학습 모델로, Google News 데이터로 학습하였고 각 단어에 대해 300차원의 임베딩 벡터를 출력한다[13]. 이후 이 임베딩 벡터들의 평균 벡터를 해당 뉴스 헤

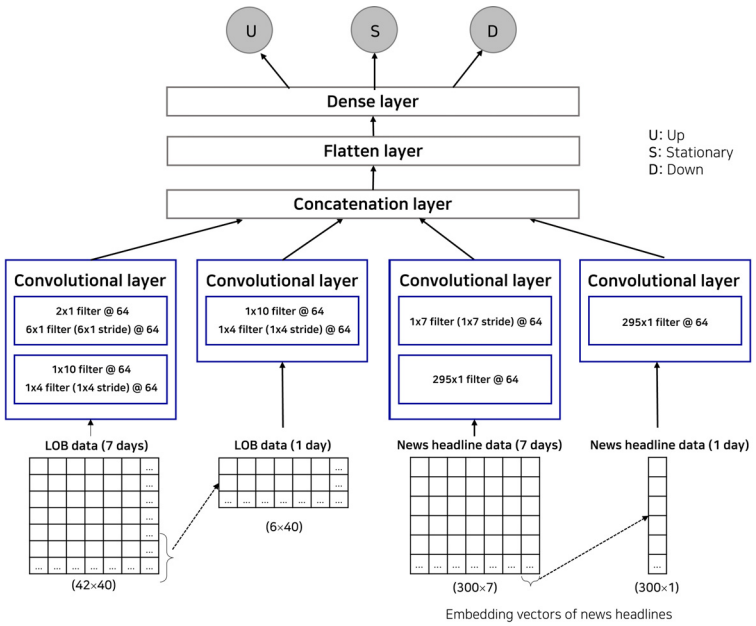
드라인을 나타내는 최종 임베딩 벡터로 사용하였다.

4. 제안하는 주가 등락 예측 모델

본 논문에서는 어떤 기업에 대한 주식 주문 정보를 담고 있는 호가창과 해당 기업과 관련된 뉴스 헤드라인을 사용하여 해당 기업의 주가 등락을 예측하는 딥러닝 기반 모델을 제안한다. 본 논문에서는 호가창의 중기 추이와 단기 추이를 모두 고려하는 한편, 주식에 영향을 주는 외부 요소까지 반영하기 위해 동기간 발생한 뉴스 헤드라인까지 고려하여 주가의 등락을 예측한다.

중기를 나타내는 기간은 유동적으로 정할 수 있으며 중기를 5일, 7일, 10일 등으로 설정하여 성능을 비교해 본 결과 중기를 7일으로 설정했을 때 가장 높은 예측 성능을 보였다. 따라서 본 논문에서는 호가창과 뉴스 헤드라인의 중기 추세를 고려하기 위해 최근 7일의 데이터를 중기 입력 데이터로 활용하고 해당 실험 결과는 5.4.3에서 제시한다.

<Figure 3>은 제안 모델의 전체적인 구조를 나타낸다. 제안 모델은 호가창의 중기 및 단기 추세를 모두 고려하기 위해 최근 7일 및 최근 1일의 호가창 데이터를 별도의 입력으로 받는다. 또한 해당 종목에 대한 중기 및 단기 뉴스 정보를 반영하기 위해 직전 7일과 1일의 뉴스 헤드라인 데이터를 역시 별도의 입력으로 받는다. 제안 모델은 호가창 변화와 뉴스의 중기 및 단기 특징을 각각 추출한 뒤, 이들을 결합하여 다음 날의 주가 변동(상승, 유지, 하락)을 예측한다.



〈Figure 3〉 The Overall Architecture of the Proposed Model

4.1 호가창 변화 특징 추출을 위한 합성곱층

합성곱층(convolutional layer)은 입력 데이터를 스캔하면서 입력 데이터 내에 존재하는 특징을 추출하는 층이다. 본 논문에서는 호가창 데이터를 1시간 단위로 묶어 그들의 평균값으로 각 시간에 대한 호가창 데이터를 생성하였다. 따라서 주식 개장 시간이 매일 9:30부터 15:30까지라면 매일 6개(6시간)의 호가창 데이터가 생성된다. 본 논문에서는 호가창 데이터를 다음과 같이 표현한 뒤 합성곱 층으로 그의 특징을 추출하였다. 각 시간 t의 호가창 데이터는 식 (1)과 같이 레벨 1부터 10까지 각 레벨 i의 매도 주문호가($p_a^{(i)}(t)$), 매도 주문량($v_a^{(i)}(t)$), 매수 주문호가($p_b^{(i)}(t)$), 매수 주문량($v_b^{(i)}(t)$)을 나타내는 총 40개의 수치로 구성된다.

$$\left\{p_a^{(i)}(t), v_a^{(i)}(t), p_b^{(i)}(t), v_b^{(i)}(t)\right\}_{i=1}^{n=10} \quad (1)$$

따라서 1일치 호가창 데이터는 6(시간) × 40(개)의 크기를 가진다. 호가창의 매도 주문과 매수 주문 정보는 주식의 미래 가격을 결정하는데 매우 중요한 정보를 내포하고 있다. 본 논문에서는 6 × 40 크기를 가지는 1일치 호가창의 변화의 특징을 포착하기 위해 우선 첫 번째 합성곱층에서 크기가 1 × 4이고 스트라이드가 1 × 4인 필터를 64개 사용하여 각 시간 t에 대한 각 레벨별 특징을 추출하였다. 이후 이 결과에 다시 1 × 10 필터를 사용하여 각 시간 t에 대해 모든 레벨의 특징을 통합한 6 × 1 × 64 크기의 출력값을 얻는다.

반면 7일치 호가창의 변화의 특징을 포착하기 위해서는 먼저 7일치 호가창 데이터에 대해 1일치 호가창과 동일한 합성곱층을 적용하여 그 결과로 42 × 1 × 64 크기의 출력값을 얻는다. 이후 이 결과에 크기와 스트라이드가 6 × 1인 필터를 사용하여 날짜별로 통합된 특징을 추출하고, 마지막으로 크기가 2 × 1인 필터를 추가로 사용하여 최종적으로 1일치 호가창에 대한 합성곱층의 출력 결과와 크기가 동일한 6 × 1 × 64 크기의 출력값을 얻는다.

4.2 뉴스 헤드라인 특징 추출을 위한 합성곱층

본 논문에서는 뉴스 헤드라인을 구성하는 각 단어를 Word2vec을 사용하여 임베딩 벡터로 변환한 뒤, 이들의 평균 벡터를 뉴스 헤드라인을 나타내는 임베딩 벡터로 사용한다. 본 논문에서는 Google이 제공하는 사전 훈련된 Word2vec 모델[13]이 반환하는 300차원의 워드 임베딩 벡터를 사용하였으며, 따라서 1일의 뉴스 헤드라인에 대한 임베딩 벡터는 300 × 1 크기를 가진다.

제안하는 모델은 뉴스 헤드라인에 대한 임베딩 벡터에 크기가 295 × 1인 필터를 64개 사용하는 합성곱층을 적용하여 특징을 추출하였다. 따라서 1일치 뉴스 헤드라인 데이터는 이 합성곱층을 통해 호가창에 대한 합성곱층의 출력 결과와 크기가 동일한 6 × 1 × 64 크기의 출력값을 얻는다.

반면 중기 뉴스 헤드라인에 대한 임베딩 벡터는 300 × 7 크기를 가진다. 이에 1일치 뉴스 데이터와 마찬가지로 크기가 295 × 1인 필터를 64개 사용하는 합성곱층을 적용하여 특징을 추출하고, 이 후 크기가 1 × 7인 필터 64개를 통해 1일치 뉴스 헤드라인에 대한 합성곱층의 출력

결과와 크기가 동일한 6 × 1 × 64 크기의 출력값을 얻는다.

4.3 최종 예측

앞서 설명한 합성곱층을 통해 7일치(중기) 호가창 데이터, 1일치(단기) 호가창 데이터, 7일치(중기) 뉴스 헤드라인 데이터, 1일치(단기) 뉴스 헤드라인 데이터의 특징이 각각 6 × 1 × 64 크기의 결과로 추출되면, 이들을 결합층(concatenation layer)을 통해 결합하여 24 × 1 × 64 크기의 데이터로 만든다. 이후 이 데이터를 평탄화층(flatten layer)을 통해 1536 × 1 크기의 벡터로 만든다. 마지막으로 밀집층(dense layer)은 이를 입력으로 받아 상승(up), 유지(stationary), 하락(down) 각각에 대한 확률값을 출력한다. 이를 위해 밀집층에서는 출력 함수로 소프트맥스(softmax)를 사용한다.

제안 모델의 학습을 위해 손실함수로는 범주형 교차 엔트로피(categorical cross entropy)를 사용하였으며, 과적합(overfitting)을 피하기 위해 학습 과정에서 4개의 합성곱층 마지막에 각각 드롭아웃 확률을 0.5로 설정하였다.

5. 실험 결과

5.1 실험 환경 및 방법

본 장에서는 제4장에서 설명한 주가 등락 예측 모델을 실데이터로 훈련하고 그의 성능을 측정한 결과를 제시한다. 제안 모델의 훈련을 위해 최적화 알고리즘으로는 Adam을 사용하였으며, 배치(batch) 크기는 10으로 설정하였다. 또한 초기 학습률은 0.01로 설정하였으며, 학습 조기 종료를 사용하여 손실 값이 100번 에포크(epoch) 이상 연속으로 감소하지 않는 경우 모델 학습을 종료하도록 하였다. 실험에서 사용된 모델들은 모두 Python 3.8과 Tensorflow를 사용하여 구현되었으며, 모델 훈련 및 성능 측정은 Intel i7-6800K 3.40 GHz CPU, 32 GB RAM이 장착된 Windows 10 환경의 PC에서 수행하였다.

〈Table 1〉 Examples of Embedding Vectors Representing News Headlines

Date	News Headline	Non-Stop Words	Final Embedding Vector
2019-10-07	Amazon launches bigger local online store in Singapore	[‘amazon’, ‘launches’, ‘bigger’, ‘local’, ‘online’, ‘store’, ‘singapore’]	[-0.16919097, 0.09529571, 0.00485535, 0.13308105, …]
2020-02-01	As coronavirus misinformation spreads on social media, Facebook removes posts	[‘coronavirus’, ‘misinformation’, ‘spreads’, ‘social’, ‘media’, ‘facebook’, ‘removes’, ‘posts’]	[-0.17208166, 0.10670873, 0.02502656, -0.14701977, …]
2020-06-04	Apple Must Face Shareholder Lawsuit Over CEO’s iPhone, China Comments: Judge	[‘apple’, ‘face’, ‘shareholder’, ‘lawsuit’, ‘ceo’, ‘iphone’, ‘china’, ‘comments’, ‘judge’]	[-0.16694918, 0.13148360,, -0.02254416, 0.14243941, …]
2020-06-19	Elon Musk tweets Tesla postpones annual shareholder meeting	[‘elon’, ‘musk’, ‘tweets’, ‘tesla’, ‘postpones’, ‘annual’, ‘shareholder’, ‘meeting’]	[-0.16451644, 0.13144810, -0.03523254, 0.13212077, …]

모델의 성능은 10-겹 교차검증(10-fold cross validation)으로 측정하였고, 본 논문에서 제안하는 주가 등락 예측 기법과 다른 비교 기법들의 성능을 정확도, 정밀도, 재현율, F1-Score를 측정하여 비교하였다. 이로부터 각 기법들이 얼마나 효과적으로 주가 등락을 예측하는지 비교 분석하였다.

5.2 실험 데이터

5.2.1 호가창 데이터

본 실험에서는 나스닥 증권 거래소에서 제공하는 실제 호가창 데이터를 제공받아 모델 훈련 및 성능 측정에 사용하였다[7]. 실험에서는 나스닥 증권 거래소의 대표적 종목인 Amazon, Apple, Facebook, Google, Tesla 5개 종목의 2019년 7월 22일부터 2020년 7월 21일까지 1년 동안의 실제 호가창 데이터를 활용하였으며, 3.1절에서 설명한 바와 같이 1시간 단위로 호가창 데이터를 묶고 레벨 1부터 10까지의 주문만 고려하였다.

본 논문에서는 미국 주식 개장 시간이 9:30부터 16:00까지인 점을 고려하여 매일 9:45부터 15:45까지 기록된 거래 데이터만 사용함으로써 장외거래의 영향을 최소화하였다.

5.2.2 뉴스 헤드라인 데이터

본 논문에서는 기업에 대한 뉴스를 검색하여 검색된 뉴스의 헤드라인을 모델 훈련 및 주가 등락 예측에 사용한다. 이를 위해 본 논문에서는 CNBC, 로이터통신(Reuters), 가디언지(The Guardian)에서 기업 뉴스를 검색하고, 검색된 뉴스의 헤드라인을 수집하였다. 수집한 뉴스의 헤드라인을 단어들로 나누고 불용어(stop words)를 제거한 뒤, 남은 단어들 각각을 사전 학습된 Word2Vec을 사용하여 300차원의 임베딩 벡터로 변환하였다. 이후 이 임베딩 벡터들의 평균 벡터를 해당 뉴스 헤드라인을 나타내는 최종 임베딩 벡터로 사용하였다.

실험에서는 앞서 선정한 5개 종목(Amazon, Apple, Facebook, Google, Tesla)에 대한 동기간 뉴스를 수집하여 이들의 뉴스 헤드라인을 사용하였다. 뉴스의 특성상 기업에 대한 뉴스가 존재하지 않는 날짜들이 흔히 존재한다. 본 논문에서는 주식 시장의 영업일 중 뉴스가 존재하지 않는 날짜에 대하여 해당 날짜의 뉴스 헤드라인 임베딩 벡터를 영벡터(zero vector)로 대체하였다. <Table 1>은 수집된 기사 헤드라인과 그에 대한 최종 임베딩 벡터의 예를 나타낸다.

5.2.3 정규화 및 예측변수 설정

본 논문에서는 모든 호가창 데이터에 대해 z-score로 표준화를 진행하였다. 또한 호가창 데이터와 뉴스 임베딩 벡터의 수치들은 그 값의 범위가 서로 크게 다르므로 둘 중의 하나가 예측 모델에 지나치게 큰 영향을 주지 않도록 최소-최대 정규화(min-max normalization)를 통해 모든 데이터 값을 [0, 1]의 범위로 변환하였다. 본 논문에서는 입력 데이터로 1시간 단위의 호가창 데이터를 사용하며, 각 시간 t의 주가 p_t는 아래 식 (2)와 같이 각 시간 t에 대해 레벨 1의 매도 주문 호가($p_a^{(1)}(t)$)와 레벨 1의 매수 주문 호가($p_b^{(1)}(t)$)의 평균 가격으로 나타낸다. 또한 본 논문에서는 주가의 예측 단위를 1일로 정하였기 때문에 어떤 날 d의 주가 P_d는 식 (3)과 같이 해당 일 d 내의 시간 $t(=1,2,...,6)$의

주가 $p_t(d)$들의 평균으로 나타낸다.

$$p_t = \frac{p_a^{(1)}(t) + p_b^{(1)}(t)}{2}, \quad t = 1, 2, \ldots, 6 \qquad (2)$$

$$P_d = \frac{p_1(d) + p_2(d) + \ldots + p_6(d)}{6} \qquad (3)$$

금융 데이터는 변동성이 매우 크기 때문에 본 논문에서는 단순히 바로 전날 주가와의 비교로 당일 주가의 상승, 하락을 판단하기보다 이전 k일 동안의 주가들을 모두 고려하여 당일 주가의 등락 여부를 판단한다. 본 논문에서는 어떤 날 d의 이전 7일 동안의 주가 평균을 아래 식 (4)와 같이 M_{d-7}라고 할 때 M_{d-7} 대비 당일 P_d의 변화율을 다음 식 (5)와 같이 r_d로 정의하였다.

$$M_{d-7} = \frac{P_{d-7} + P_{d-6} + \ldots + P_{d-1}}{7} \qquad (4)$$

$$r_d = \frac{P_d - M_{d-7}}{M_{d-7}} \qquad (5)$$

본 논문에서는 r_d가 0.02보다 큰 경우, 즉 이전 7일의 주가 평균에 비해 당일 주가가 2%보다 더 상승한 경우에 예측변수를 상승(up)을 나타내는 '2'로 설정한다. 마찬가지로 r_d가 -0.02보다 낮은 경우에는 예측변수를 하락(down)을 나타내는 '0'으로 설정한다. 그 외의 경우는 예측변수를 유지(stationary)를 뜻하는 '1'로 설정한다.

5.3 비교 기법

호가창과 뉴스를 결합하여 주가의 등락을 예측하는 기법은 아직 발견된 바가 없으므로, 본 논문에서는 제안 방법과 다음 세 가지 기법을 비교한다.

5.3.1 중기 호가창 데이터만을 사용하는 예측 모델

앞서 제2.3절에서 설명한 바와 같이 호가창을 활용하는 기존의 예측 모델들은 주로 최근 일정 기간 동안의 호가창 정보만을 사용한다. 본 논문에서는 이러한 기존 모델[10, 14]을 중기 호가창 데이터만을 사용하는 모델이라 부른다. 이 모델은 제4.1절에서 설명한 합성곱층과 동일한 합성곱층을 통해 7일간의 중기 호가창의 변화에 대한 특징을 추출한다. 이후 평탄화층을 통해 출력값을 벡터화하고, 소프트맥스를 출력 함수로 사용하는 밀집층을 통해 다음날 주식 가격의 움직임을 예측한다. 본 모델은 호가창의 최근 일정 기간 동안의 전반적인 추세는 고려하지만 호가창의 단기 추세 및 뉴스 헤드라인 정보까지는 반영하지 못하는 모델이다. 실험에서는 본 모델을 'LOB(mid)'로 표시한다.

5.3.2 중기와 단기 호가창 데이터를 모두 사용하는 예측 모델

본 모델은 본 논문에서 제안하는 아이디어인 중기와 단기 호가창 정보를 모두 사용하는 방법의 효과를 검증하기 위한 모델이다. 본 모델은 제4.1절에서 설명한 제안 모델의 중기 호가창과 단기 호가창의 합성곱층과 동일한 구조의 합성곱층을 사용하여 7일간의 중기 호가창과 1일간의 단기 호가창 변화에 대한 특징을 각각 추출한다. 이어서 결합층을 통해 두 출력값을 결합하고, 마찬가지로 평탄화층과 밀집층을 통해 다음날 주가의 등락을 예측한다. 본 모델은 중기와 단기의 호가창 정보는 고려하지만, 뉴

〈Table 2〉 Average Performance for Varying Periods of Mid-Term (Unit: %)

Period	Accuracy	Precision	Recall	F1 score
5 days	66.25	66.66	66.25	63.23
7 days	71.51	71.67	71.51	68.26
10 days	68.34	68.74	68.34	65.63
15 days	67.52	67.71	67.52	65.18

스에 담긴 주식 시장의 외부적인 요인을 고려하지 못하는 모델이다. 실험에서는 본 모델을 'LOB(mid)+LOB(short)'로 표시한다.

5.3.3 중기와 단기 호가창 데이터에 단기 뉴스 헤드라인까지 사용하는 예측 모델

본 모델은 (2)의 방법에 단기 뉴스 헤드라인을 추가적으로 활용하는 모델로서, 호가창의 중기 추세와 단기 추세는 물론 바로 전날의 뉴스 정보까지는 사용하지만, 중기적인 뉴스 정보는 사용하지 않는 모델이다. 실험에서는 본 모델을 'LOB(mid)+LOB(short)+News(short)'로 표시한다.

5.4 성능 평가 결과

5.4.1 중기 기간 설정

제안하는 방법은 호가창 데이터와 뉴스 헤드라인 데이터의 중기 추세를 반영한다. 중기를 나타내는 기간은 다양할 수 있으며, 본 실험에서는 다양한 기간의 데이터를 중기 입력 데이터로 설정하여 최적의 중기 기간을 찾는다. 본 실험에서는 중기 추세를 고려하기 위해 최근 5일, 7일, 10일, 그리고 15일의 데이터를 중기 입력 데이터로 활용하여 제안모델의 성능을 비교해보았다.

실험 결과, 5개 종목 중 Amazon, Facebook, Google 3개의 종목에서 중기를 7일로 설정했을 때 가장 높은 예측 성능을 보인 것을 확인할 수 있었으며, 5개 종목의 예측 성능의 평균을 구해본 결과, <Table 2>와 같이 중기가 7일일 때 가장 높은 평균 정확도를 보였다. 따라서 본 논문에서는 호가창과 뉴스 헤드라인의 중기 추세를 고려하기 위해 최근 7일의 데이터를 중기 입력 데이터로 활용하였다.

5.4.2 예측 성능 평가 결과

본 실험에서는 제5.3절에서 설명한 3가지 방법과 중단기 호가창 및 중단기 뉴스 헤드라인을 모두 사용하는 제안 방법의 성능을 비교하였다. 이를 위해 본 실험에서는 중기는 7일로 정의하고 단기는 1일로 정의한 뒤, 이에 맞추어 호가창 데이터와 뉴스 헤드라인 데이터를 4개의 모델에 각각 입력하였다. 물론 중기와 단기의 길이는 자유롭게 정의할 수 있다. <Table 3>는 Amazon, Apple, Facebook, Google, Tesla 5개 종목 각각에 대해 제안 모델과 3가지 모델로 주식의 등락을 예측한 정확도, 정밀도, 재현율, 그리고 F1-Score를 비교한 결과이다. <Table 3>에서 'LOB(mid) + LOB(short) + News(mid) + News(short)'는 제안 모델을 나타낸다.

중기 호가창 정보만을 사용하는 기존 방법('LOB(mid)')은 최소 53.90%, 최대 66.20%의 예측 정확도를 보였다. 이에 비해 중기 호가창 정보

와 단기 호가창 정보를 모두 사용하는 방법 (‘LOB(mid) + LOB(short)’)은 ‘LOB(mid)’ 대비 정확도를 평균 약 4.25%p 향상시킴을 볼 수 있다. 따라서 본 논문에서 제안하는 중기 호가창 정보와 단기 호가창 정보를 모두 사용하는 방법은 유의미한 전략임을 알 수 있다. 또한 호가창 정보 외에 뉴스 정보를 추가로 사용하는 방법 역시 ‘LOB(mid) + LOB(short)’와 ‘LOB(mid) + LOB(short) + News(short)’를 비교했을 때 후자가 정확도를 평균 약 6.54%p 향상시키므로 마찬가지로 효과적인 전략임을 확인할 수 있다. 마지막으로 제안 모델은 종목에 따라 최소 66.20%, 최대 77.48%의 예측 정확도를 보이며 다른 모든 모델들에 비해 가장 좋은 성능을 보임을 알 수 있다. 따라서 중기 호가창, 단기 호가창, 중기 뉴스, 단기 뉴스 정보를 모두 사용하는 제안 방법이 주가 등락 예측에 가장 효과적임을 알 수 있다.

또한 <Table 3>에서 ‘LOB(mid)’, ‘LOB(mid) + LOB(short)’, ‘LOB(mid) + LOB(short) + News(short)’, ‘LOB(mid) + LOB(short) +

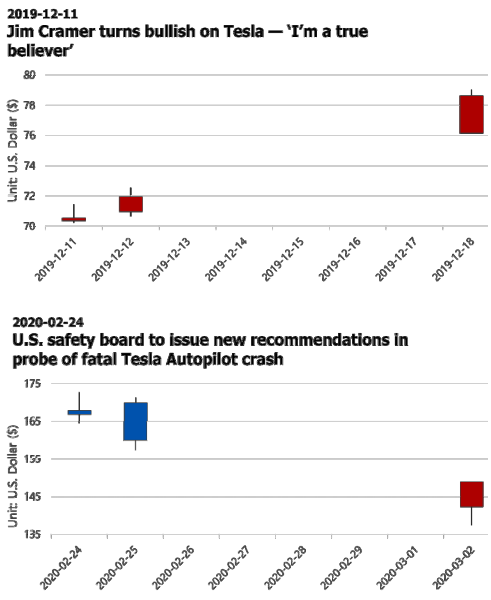
〈Table 3〉 Performance Evaluation Results (Unit: %)

Models	Accuracy	Precision	Recall	F1 score
Amazon				
LOB(mid) + LOB(short) + News(mid) + News(short)	77.48	76.35	77.48	74.14
LOB(mid) + LOB(short) + News(short)	70.22	65.99	70.22	63.60
LOB(mid) + LOB(short)	63.72	60.59	63.72	59.42
LOB(mid)	59.82	57.77	59.82	53.78
Apple				
LOB(mid) + LOB(short) + News(mid) + News(short)	68.52	71.62	68.52	66.44
LOB(mid) + LOB(short) + News(short)	64.90	63.61	64.90	62.02
LOB(mid) + LOB(short)	61.58	60.90	61.58	57.64
LOB(mid)	58.35	62.92	58.35	54.52
Facebook				
중LOB(mid) + LOB(short) + News(mid) + News(short)	66.20	65.79	66.20	61.30
LOB(mid) + LOB(short) + News(short)	64.43	60.27	64.43	58.62
LOB(mid) + LOB(short)	59.18	45.49	59.18	49.27
LOB(mid)	53.90	46.98	53.90	45.62
Google				
LOB(mid) + LOB(short) + News(mid) + News(short)	75.57	72.26	75.57	70.46
LOB(mid) + LOB(short) + News(short)	75.17	74.34	75.17	70.09
LOB(mid) + LOB(short)	64.45	60.27	64.45	57.55
LOB(mid)	58.03	59.69	58.03	54.05
Tesla				
LOB(mid) + LOB(short) + News(mid) + News(short)	69.78	72.32	69.78	68.97
LOB(mid) + LOB(short) + News(short)	64.40	65.75	64.40	62.05
LOB(mid) + LOB(short)	57.50	61.68	57.50	53.78
LOB(mid)	55.08	53.32	55.08	51.11

News(short) + News(mid)'로 갈수록 점점 성능이 증가하는 경향이 있음을 볼 수 있다. 이것은 중기 정보만 사용하는 것보다는 중단기 정보를 모두 사용하는 것이, 호가창 정보만 사용하는 것보다는 뉴스 정보까지 사용하는 것이 효과적임을 의미한다. 제안 모델은 Google 한 종목의 정밀도를 제외하고는 모든 지표에서 다른 모든 모델들과 비교하여 가장 좋은 성능을 보인다. 제안 모델은 유일하게 Google 주식 등락 예측 정밀도에서 'LOB(mid) + LOB(short) + News(short)'의 정밀도인 74.34%보다 소폭 감소한 72.26%을 보이지만, 이 경우에도 중기 호가창 정보만을 사용하는 기존 모델인 'LOB(mid)'의 정밀도인 59.69%보다 훨씬 높은 성능을 보임을 알 수 있다.

5.4.3 뉴스 헤드라인에 따른 주가 등락 예

제5.4.2절의 <Table 3>을 보면 뉴스 헤드라인이 주가 예측에 큰 영향을 미친 것을 파악할 수 있다. 이에 예시로 종목 Tesla에 관한 기사 두 개를 살펴보았다. <Figure 4>는 Tesla를 포함한 뉴스 헤드라인의 예시와 뉴스 발행일자로부터 7일 동안의 주가 변화를 나타낸 그림이다. 먼저 2019년 12월 11일의 뉴스 헤드라인은 CNBC의 앵커인 Jim Cramer가 Tesla에 대해 긍정적으로 코멘트 한 내용으로 이후 다음 날인 12월 12일과 7일 뒤인 12월 18일 모두 12월 11일과 비교하여 상승한 것을 확인할 수 있다. 또한 2020년 2월 24일의 헤드라인은 Tesla의 자율주행 기능 실행 중 일어난 사고를 포함하는 내용으로, 발행일 다음날인 2월 25일과 3월 2일에 모두 하락한 것을 확인할 수 있다.



〈Figure 4〉 Example of News Headlines and Stock Price Afterward

5.5 모의 투자 결과

본 실험에서는 제안 모델의 예측 성능 측정에서 한걸음 더 나아가 제안 모델이 예측한 결과에 따라 투자하는 모의 투자를 통해 제안 모델의 수익성을 측정하였다. 본 실험에서는 Lavrenko et al.[4]에서 사용한 다음 투자 전략을 따라 투자하였다. 만약 예측 모델이 다음 날 주가가 상승한다고 예측하면 해당 종목의 시가로 $10,000만큼을 투자한다. 투자 기간은 하루로, 하루 동안 투자금액의 α%만큼의 수익이 나면 즉시 매도하고, 그렇지 않으면 종가로 매도한다. 제안 모델이 다음 날 주가가 하락할 것으로 예측하면 해당 종목의 시가로 $10,000만큼을 매도한다. 하루 동안 그 금액보다 β% 낮은 금액으로 매수할 수 있다면 즉시 매수해서 차익을 얻는다. 만약 하루 동안 β%만큼 하락하지 않는다면 종가로 해당 주식을 매수한다.

〈Table 4〉 Profits for Each Stock Obtained by the Proposed Model

Stock	Average Profit	
	α=2, β=2	α=8, β=4
Amazon	$595.77	$586.50
Apple	$549.14	$492.46
Facebook	$619.29	$1,265.88
Google	$691.21	$650.25
Tesla	$720.03	$2,840.83
Average	$635.088	$1,167.184

〈Table 5〉 Profits for Each Stock Obtained by A Random Model

Stock	Average Profit
Amazon	-$97.797
Apple	$96.678
Facebook	-$418.806
Google	-$179.285
Tesla	-$82.486
Average	-$136.339

실험을 위해 기존 1년 치의 데이터에서 11개월의 데이터를 훈련 데이터로 사용하고, 나머지 1개월의 데이터를 테스트 데이터로 사용하여 제안 모델의 주식 등락 예측 결과에 따라 투자하였다. 투자에 사용된 테스트 데이터는 2020년 6월 22일부터 2020년 7월 21일까지의 데이터로 총 21영업일이다. 본 논문에서 제안하는 예측 모델은 주식 등락을 상승, 하락, 유지로 예측하기 때문에 만약 다음날 주가가 유지된다고 예측할 때는 투자하지 않았다. <Table 4>는 제안 모델로 투자하여 발생한 수익을 나타낸다. α, β 모두 2로 설정하여 실험했을 때는 종목에 따라 최대 $720.03, 최소 $549.14, 평균 $635.088의 수익이 발생했고 α, β를 각각 8, 4로 설정하여 더욱 공격적으로 투자한 경우에는 최대 $2,840.83, 최소 $492.46, 평균 $1,167.184의 수익을 달성했다. 해당 수익이 통계적으로 유의미함을 보이기 위해 랜덤 투자와의 비교를 수행하였다. 랜덤 투자는 1,000번 동안 무작위로 상승, 하락 또는 유지로 가정하고 위와 같은 전략으로 투자하는 것으로, 실험 결과 <Table 5>와 같은 수익을 기록했다. <Table 4>과 <Table 5>에서 볼 수 있듯이 제안 방법은 모든 종목에 대하여 랜덤 투자에 비해 현저히 높은 수익을 발생시킨다. 따라서 제안 방법은 주식의 등락을 효과적으로 예측함을 확인하였다.

6. 결 론

본 논문에서는 호가창의 중기 추이 외에도 호가창의 단기 추이와 동기간 뉴스 헤드라인을 추가로 활용하여 예측 성능을 높이는 딥러닝 기반 주가 등락 예측 모델을 제안하였다. 본 논문의 제안 방법은 호가창의 최근 일정 기간 동안의 전체적인 추이만 고려했던 기존 연구와는 달리 단기 호가창 데이터를 추가로 사용하여 예측일 바로 직전의 추이를 고려하고, 동기간의 뉴스 헤드라인을 활용하여 외부적인 영향까지 함께 고려하여 다음 날 주가의 등락을 보다 정확하게 예측한다. 본 논문에서 제안하는 딥러닝 기반 예측 모델은 중단기 호가창의 정보와 중단기 뉴스 헤드라인의 특징을 각각 합성곱층으로 추출하고 이들의 결과를 종합하여 최종적으로 주가 등락을 예측한다.

본 논문에서는 나스닥의 대표적인 종목인 Amazon, Apple, Facebook, Google, Tesla에 대한 실제 호가창 데이터와 뉴스 헤드라인 데

이터를 사용한 실험을 통해 본 논문에서 제안하는 방법이 호가창의 중기 추이만을 고려하는 기존 방법에 비해 정확도를 평균 약 14.5%p 이상 향상시킴을 확인하였다. 또한 제안 방법으로 모의 투자를 수행한 결과 평균 -$136.34의 수익을 내는 랜덤 투자와 비교하여 종목에 따라 최소 $492.46, 최대 $2,840.83의 수익을 달성하였다. 따라서 본 논문의 제안 방법은 중단기의 호가창 및 뉴스 정보를 사용하여 더욱 정확하게 주가의 등락을 예측함을 확인하였다.

References

[1] Akita, R., Yoshihara, A., Matsubara, T., and Uehara, K., "Deep learning for stock prediction using numerical and textual information," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1-6, 2016.

[2] Bao, W., Yue, J., and Rao, Y., "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," PLOS ONE, Vol. 12, No. 7, 2017.

[3] Kim, M., Ryu, J., Cha, D., and Sim, M. K., "Stock price prediction using sentiment analysis: From 'stock discussion room' in Naver," The Journal of Society for e-Business Studies, Vol. 25, No. 4, pp. 61-75, 2020.

[4] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J., "Mining of concurrent text and time series," In KDD-2000 Workshop on Text Mining, pp. 37-44, 2000.

[5] Lee, C. and Soo, V., "Predict Stock Price with Financial News Based on Recurrent Convolutional Neural Networks," 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 160-165, 2017.

[6] Li, X., Xie, H., Chen, L., Wang, J., and Deng, X., "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, Vol. 69, pp. 14-23, 2014.

[7] Nasdaq TotalView-ITCH, https://www.nasdaqtrader. com/Trader.aspx?id=Totalview2.

[8] Peng, Y. and Jiang, H., "Leverage financial news to predict stock price movements using word embeddings and deep neural networks," in Proc. NAACL-HLT. San Diego, CA, USA: Association for Computational Linguistics, pp. 374 - 379, 2016.

[9] Ryoo, E., Kim, C., and Lee, K. Y., "Deep learning-based stock price prediction using limit order books and News Headlines," Annual Conference of KIPS (ACK) 2021, pp. 541-544, 2021.

[10] Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A., "Using deep learning to detect price change indications in financial markets," 2017 25th European Signal Processing Conference (EUSIPCO), pp. 2511-2515,

2017.

[11] Vargas, M. R., de Lima, B. S. L. P., and Evsukoff, A. G., “Deep learning for stock market prediction from financial news articles,” 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 60-65, 2017.

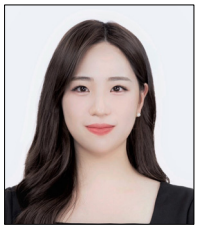
[12] W. Jiang, “Applications of deep learning in stock market prediction: Recent progress,” Expert Systems with Applications, Vol. 184, 2021.

[13] Word2vec, https://code.google.com/archive/p/word2vec.

[14] Zhang, Z., Zohren, S., and Roberts, S., “DeepLOB: Deep convolutional neural networks for limit order books,” IEEE Transactions on Signal Processing, Vol. 67, No. 11, pp. 3001-3012, 2019.

[15] Zhou, X., Pan, Z, Hu, G., Tang, S., and Zhao, C., “Stock market prediction on high-frequency data using generative adversarial nets,” Mathematical Problems in Engineering, Vol. 2018, Article ID 4907423, 11 pages, 2018.

저 자 소 개



류의림	(E-mail: euirimryoo@sookmyung.ac.kr)
2020년	숙명여자대학교 수학과 (학사)
2020년~현재	숙명여자대학교 빅데이터분석융합학과 (석사과정)
관심분야	데이터마이닝, 빅데이터

이기용	(E-mail: kiyonglee@sookmyung.ac.kr)
1998년	KAIST 전산학과 (학사)
2000년	KAIST 전산학과 (석사)
2006년	KAIST 전산학과 (박사)
2006년~2008년	삼성전자 소프트웨어연구소 책임연구원
2008년~2010년	KAIST 전산학과 연구조교수
2014년~현재	숙명여자대학교 소프트웨어학부 교수
관심분야	데이터베이스, 데이터마이닝, 빅데이터, 데이터스트림

정연돈	(E-mail: ydchung@korea.ac.kr)
1994년	고려대학교 전산과학과 (학사)
1996년	한국과학기술원 전산학과 (석사)
2000년	한국과학기술원 전산학과 (박사)
2006년~현재	고려대학교 정보대학 컴퓨터학과 교수
관심분야	데이터 프라이버시, 데이터베이스 시스템