*The Korean Journal of Applied Statistics* (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.4.483
**33**(4), 483–497

# Prediction of stock prices using deep neural network models including an emotional predictor based on online news by industrial groups

Jun Hyeong Lim[a] · Young Sook Son[a,1]

[a]Department of Statistics, Chonnam National University

(Received April 28, 2020; Revised June 1, 2020; Accepted June 11, 2020)

**Abstract**

We used a deep neural network model for the prediction of the stock prices of Kia Motors and Shinsegae as listed in the KOSPI 100. We used an emotional variable derived from online news in addition to the various technical indicators most often used. The emotional variable used as a predictor variable was generated from the average of the emotional scores for companies in the industrial group after building an emotional dictionary specific to each industrial group classified in a social network analysis. The study was conducted with various combinations of predictors and confirmed that good predictive and profitable power could be expected when jointly using technical indicators and an emotional variable based on online news by industrial groups.

Keywords: prediction of stock prices, deep neural network, social network analysis, online news, emotional variable

## 1. 서론

한국예탁결제원(https://www.ksd.or.kr)의 국내 주식 투자자 현황 자료에 따르면, 2018년 12월 결산 상장법인 2,216사의 실질주주(중복주주 제외)는 약 561만명으로 전년 대비 약 55만명(10.9%) 증가하였다. 전체 실질주주가 보유한 주식 수는 약 868억 주로 주주 1인당 평균 약 15,463주로 전년 대비 4.9% 증가하였으며, 1인당 평균 보유종목은 4.27종목으로 전년 대비 8.4% 증가하였다. 또한 실질주주 연령별로 40대가 153만 명(27.6%)으로 가장 많았고 보유 주식 수는 50대가 135억주(33.0%)로 가장 많았다. Table 1.1의 2018년 기준 최근 5개년 12월 결산 상장법인 실질주주의 현황통계표에 의하면 주식 종목 수, 주주 수, 1인당 종목 수, 그리고 1인당 보유 주식 수는 모두 매년 증가하는 추세이다. 초유의 저금리 금융 환경이 지속되면서 많은 사람들이 주가 변동에 따른 위험을 부담하더라도 높은 수익률을 기대하며 주식투자에 참여하지만 수익을 얻는 것은 좀처럼 쉽지 않다.

주식투자에서 수익을 얻기 위한 주가예측 기법들은 오랫동안 투자자들의 지대한 관심을 받아왔다. 전통적인 주가예측 연구에서는 기본적 분석과 기술적 분석이 있다. 기본적 분석이란 주식의 가치를 분석하

[1]Corresponding author: Department of Statistics, Chonnam National University, 77, Yongbong-ro, Buk-Gu, Gwangju 61186, Korea. E-mail: ysson@jnu.ac.kr

**Table 1.1.** Current status of real shareholders of listed companies in the last five-year

| Year | Companies | Stockholder | Stocks item per person | Stocks held per person |
|---|---|---|---|---|
| 2014 | 1,836 | 4,415,830 | 3.36 | 12,476 |
| 2015 | 1,975 | 4,750,027 | 3.71 | 12,716 |
| 2016 | 2,070 | 4,939,465 | 3.75 | 13,670 |
| 2017 | 2,147 | 5,059,013 | 3.94 | 14,743 |
| 2018 | 2,216 | 5,611,764 | 4.27 | 15,463 |

여 미래의 주가 흐름을 예측하는 방법으로 해당 기업의 재무제표나 관련 업종의 통계 및 산업자료들을 활용한다. 기술적 분석이란 주식의 가격이나 거래량 등으로부터 파생된 다양한 기술적 지표들로부터 미래 주가의 등락을 예측하는 방법이다. 즉, 기본적 분석이 기업의 가치 위주로 분석을 한다면 기술적 분석은 주가 시계열 차트에 내포된 정보를 분석하는 기법이다.

주가는 투자자들의 수요와 공급에 의하여 가격이 형성된다. 투자자의 수요 및 공급에 영향을 미치는 요인은 다양하지만 초고속 인터넷 시대에 투자자들이 쉽게 접근할 수 있는 온라인 뉴스도 그 중 하나일 것이다. 이에 따라 최근 들어 온라인 뉴스를 기초로 주가를 예측하는 연구가 활발하게 이루어지고 있다.

Kim (2012)은 범용 감성사전이 아닌 주식 주제에 특화된 감성사전을 구축한 후 긍정 혹은 부정의 뉴스로 분류한 온라인 뉴스로부터 Korea Composite Stock Price Index (KOSPI)의 상승 혹은 하락을 예측하여 최대 54.8%의 정확도를 보였다.

Jeong 등 (2015)은 개별 기업의 온라인 뉴스에 기초하여 생성한 감성변수를 예측변수로 사용하여 KOSPI 200에 속하는 40개 기업의 주가예측에서 평균 약 56%의 예측 정확도를 보였다.

Choi (2016)는 신경망 알고리즘을 활용한 텍스트마이닝 기법인 Word2Vec을 사용하여 온라인 뉴스뿐만 아니라 12개의 기술적 지표(직전 주가 및 가격 변동량 등)를 예측변수로 사용한 support vector machine (SVM)에 의해 KOSPI 시가총액 상위 20개 종목의 주가를 예측하였다. 이 방법은 평균 53.4%의 정확도와 6.79%의 수익률을 보였다.

Lee와 Lee (2017)는 구조적 단어 추출 방법을 활용하여 온라인 뉴스를 예측변수로 사용한 SVM, boosting, 그리고 random forest와 같은 기계학습법에 의해 KOSPI 시가총액 상위 10개 종목의 주가 상승 및 하락을 예측하였는데 기존의 단어 추출 방법보다 구조적 단어 추출 방법을 활용한 경우가 평균적으로 60% 이상의 더 높은 정확도를 보였다.

Kim과 Kim (2017)은 명사뿐만 아니라 형용사, 동사, 유사어, 반의어 등을 포함하는 말뭉치 기초의 접근 방법에 의해서 감성사전을 구축하여 증권전문 웹사이트에 게시된 글에 대하여 구축된 감성사전을 적용하여 유의성을 검증하였다.

Kim과 Lee (2018)는 Word2Vec을 기반으로 뉴스 단어의 문맥을 고려한 감성사전을 구축하여 KOSPI의 방향성을 예측하였는데 평균 정확도는 55% 내외였다.

일반적으로 온라인 뉴스를 예측변수로 사용하는 주가예측 연구에서는 개별 종목의 온라인 뉴스를 기반으로 예측이 이루어졌다. 하지만 자동차 부품을 제조하는 회사의 긍정적인 뉴스가 출현하면 자동차 회사의 주가도 상승하듯이 예측하고자 하는 종목과 관련이 높은 종목들의 뉴스가 주가에 영향을 줄 수 있다. 실제로 Shynkevich 등 (2016)은 국제표준산업분류에서 같은 산업군에 속한 종목들은 서로 관련성이 있으므로 같은 산업군의 뉴스를 함께 통합하여 예측변수로 사용할 때 주가예측에서 더 높은 정확도를 보인다는 것을 밝혔다.

Seong과 Nam (2018)은 한국표준산업분류에 의해 나누어진 산업군에 속한 종목들이 서로 이질적인 성

향을 띄는 종목들도 포함되어 있음을 확인하였다. 그래서 개별 기업의 뉴스뿐만 아니라 K-평균 군집 분석에 의해 주가의 흐름이 비슷한 동질적인 산업군에 속하는 종목들을 통합한 온라인 뉴스를 함께 예측변수로 사용할 때 보다 향상된 정확도를 보인다는 것을 밝혔다. 음식료품, 제약, 소재 산업군에 속하는 종목들의 주가예측에서 다중커널학습법을 적용하여 평균적으로 약 60%의 정확도를 보였다.

현재 우리는 빅 데이터를 활용하는 인공지능의 시대에 살아가고 있다. 2016년 관심을 모았던 인공지능 알파고와 이세돌 9단의 바둑 대결에서 알파고가 승리한 이후, 인공지능을 이용한 미래 예측 연구가 활발하게 진행되고 있다. 이에 따라 기업의 실적이나 각종 투자지표들을 참고하던 주식투자 분야에서도 빅 데이터를 활용한 인공지능을 이용하여 미래의 주가를 예측하는 연구가 진행되고 있다

본 연구에서는 인공지능 분야에서 널리 활용되는 심층 신경망모형(deep neural network model)을 사용하여 KOSPI 100에 속하는 개별 종목인 기아차와 신세계의 주가예측 및 수익률 계산을 수행하였다. 이때 사용되는 예측변수로는 기존에 많이 사용되는 51개의 기술적 변수들과 온라인 뉴스에 기초하여 만들어진 두 개의 감성변수이다. 감성변수를 만들기 위해서는 먼저 소셜 네트워크 분석(social network analysis; SNA)에 의하여 KOSPI 100 종목들을 동질적인 산업군으로 분류한 후, 산업군의 온라인 뉴스들에 기초한 감성사전을 구축한다. 두 개의 감성변수 중 하나는 산업군 감성사전에 기초한 감성분석(sentimental analysis)을 통하여 만든 개별 기업의 감성변수이고 다른 하나는 산업군에 속하는 개별 기업들의 감성변수를 평균한 산업군 감성변수이다. 정확도 및 수익률의 관점에서 기술적 변수와 SNA에 의한 산업군 감성변수를 예측변수로 사용했을 때가 다른 경우들보다 더 우수하였다.

논문의 구성은 다음과 같다. 2절에서는 주가예측에 사용한 데이터의 소개, 3절에서는 소셜 네트워크 분석에 의한 산업군 분류 절차, 4절에서는 온라인 뉴스를 감성변수로 만드는 과정, 그리고 5절에서는 심층 신경망모형을 학습시키는 절차를 설명한 후 기아차와 신세계의 주가예측에서 정확도 및 수익률의 결과를 제시한다. 마지막으로 6절에서는 연구 결과를 요약하고 연구의 한계점과 향후 연구방향에 대해 소개한다.

## 2. 데이터 소개

본 논문에서는 심층 신경망모형을 사용하여 KOSPI 100에 속하는 개별 종목 중 기아차(Kia Motors)와 신세계(Shinsegae)의 주가예측 및 수익률 계산을 하였다. 사용된 데이터는 2015년부터 2018년까지 총 4년간의 일별 주가 데이터와 온라인 뉴스 데이터이다.

심층 신경망모형에서 목표변수는 기아차(신세계)의 익일 시가 대비 종가가 상승인 경우는 1의 값을 하락인 경우는 0의 값을 갖는 이항변수이다. 예측변수로 사용된 51개의 기술적 변수(http://stat.jnu.ac.kr/dext5editordata/20200601_161839846_00748.pdf에서 51개 기술적 변수들에 대한 설명과 계산식을 참조한다)들과 온라인 뉴스에 기초하여 생성된 두 개의 감성변수가 Table 2.1에 정의되어 있다. 실제 자료분석에서는 Table 2.1에서 정의된 예측변수를 생성한 후 모두 표준화하여 사용하였다.

주식 시장은 공휴일을 제외하고 월요일부터 금요일까지 장이 열린다. 장 시작 동시호가는 8시 30분부터 시작해서 9시에 시가가 형성되며, 15시 20분부터 장 마감 동시호가가 시작되고 15시 30분에 종가가 형성된다. 투자자는 8시 30분부터 시작되는 장 시작 동시호가에 참여하게 되므로 온라인 뉴스 기사에 대한 분석이나 주가예측이 그 전에 완료되어야 현실적으로 예측이 의미가 있을 것이다. 장 시작 동시호가에 참여하기 전 실제로 자료 분석에 필요한 시간을 고려하여 8시까지의 뉴스 기사를 수집하여 분석에 이용하도록 하였다. 따라서 뉴스 데이터는 익일 주가예측을 위해 당일의 시가가 형성된 후인 9시부터 익일 시가가 형성되기 1시간 전인 8시까지의 시간 동안 인터넷에 업로드된 온라인 뉴스 기사를 수집하였다. 공휴일의 경우, 예를 들어 당일이 금요일인 경우에는 익일인 월요일의 주가를 예측하게 되는데 이

**Table 2.1.** Predictor variables

| | | |
|---|---|---|
| | Basic variables (10) | 1–5. Company stock price : Starting price, High price, Low price, Closing price, Trading volume<br>6–10. Company stock price compared to the previous day : Starting price, High price, Low price, Closing price, Trading volume |
| Technical indicators (51) | Volume indicators (9) | 1. Accumulation/Distribution (A/D)<br>2. Ease Of Movement (EOM)<br>3. Net Change Oscillator (NCO)<br>4. Negative Volume Index (NVI)<br>5. On Balance Volume (OBV)<br>6. Positive Volume Index (PVI)<br>7. Price and Volume Trend (PVT)<br>8. Volume Rate of Change (VRC)<br>9. Volume Oscillator (VO) |
| | Market characteristic indicators (17) | 1. Average True Range (ATR)<br>2. Chaikin's Oscillator (CO)<br>3. Chaikin's Volatility (CV)<br>4–6. Disparity (5-days, 10-days, 20-days)<br>7. Momentum (10-days)<br>8. Price Oscillator (PO) (9-days, 26-days)<br>9. Price Rate Of Change (PROC) (10-days)<br>10. Relative Strength Index (RSI) (14-days)<br>11. Vertical Horizontal Filter (VHF) (28-days)<br>12. William's %R (WR) (10-days)<br>13. Weighted Close (WC)<br>14. William's Accumulation/Distribution (WAD)<br>15. Fast %K (10-days)<br>16. Fast %D (5-days)<br>17. Slow %D (5-days) |
| | Trend indicators (15) | 1. Average Directional Movement Index (ADX)<br>2. Average Directional Movement Index Rating (ADXR)<br>3. Commodity Channel Index (CCI)<br>4–6. Directional Movement Index (PDI, MDI, DX)<br>7. Moving Average Convergence and Divergence (MACD)<br>8. Moving Average Convergence and Divergence Oscillator (MACDO)<br>9. Sonar Momentum Chart (SMC)<br>10–12. Simple Moving Average (5-days, 10-days, 20-days)<br>13–15. Exponential Moving Average (5-days, 10-days, 20-days) |
| Emotional variables(2) | | 1. Industrial group emotional variable<br>2. Company emotional variable |

경우에는 금요일 9시부터 월요일 8시까지의 뉴스 기사를 모두 수집해주었다. 따라서 Table 2.1의 모든 예측변수 값들은 익일 목표변수 값이 측정되기 전인 익일 8시까지 관측된 값들로서 과거의 예측변수 값들로부터 미래의 목표변수 값이 예측되는 원칙을 지키고 있다.

수집된 데이터는 시간 순서에 따라 학습용(training), 평가용(validation), 그리고 검증용(test) 데이터

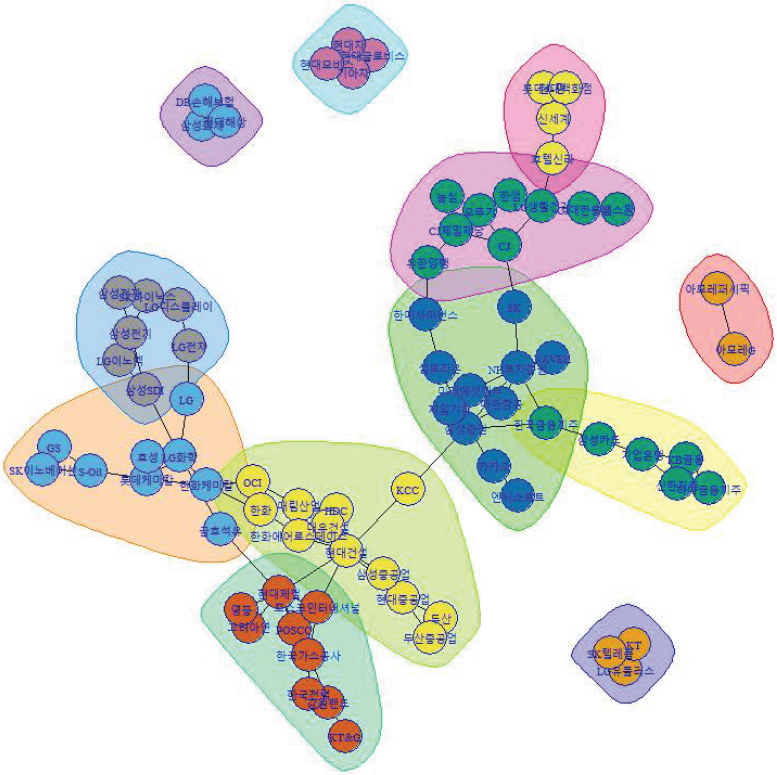


**Figure 3.1.** Network cluster classification using Greedy algorithm.

로 분할하여 분석에 사용하였다. 즉, 2015년 1월부터 2017년 12월까지 3년 동안의 데이터를 학습용, 2018년 1월부터 6월까지 6개월간 데이터를 평가용, 그리고 2018년 7월부터 12월까지 6개월간의 데이터를 검증용 데이터로 분할하였다.

학습용 데이터는 소셜 네트워크 분석을 통한 산업군 분류, 감성사전 구축, 그리고 심층 신경망모형의 학습 등에 사용되었고, 평가용 데이터는 심층 신경망모형의 시뮬레이션에서 최적의 옵션을 선택하는데 이용되었다. 학습용 및 평가용 데이터로 최적 옵션을 가지는 하나의 모형을 추정한 후, 이 추정모형으로 검증용 데이터의 매 시점에서 익일 시가 대비 종가의 상승 혹은 하락을 예측하여 정확도 평가와 수익률 계산에 사용하였다.

## 3. 소셜 네트워크 분석에 의한 산업군 분류

Table 2.1에서 정의된 두 개의 감성변수 중 하나는 개별 종목의 온라인 뉴스에 기초하여 생성된 감성변수(company emotional variable)이며 다른 하나는 개별 종목이 포함된 동질한 산업군의 온라인 뉴스에 기초하여 생성된 산업군 감성변수(industrial group emotional variable)이다

동질한 산업군으로 분류하기 위한 대상이 되는 주식 종목은 현재 KOSPI 100에 속한 100개 종목 중에서 분석 기간(2015년 1월부터 2018년까지 12월) 중에 상장된 종목 혹은 상장폐지된 종목을 제외한 총 82개 종목이다.

**Table 3.1.** Industrial group classification of Kia Motors and Shinsegae by KSIC and SNA

| Company | KSIC | SNA |
|---|---|---|
| 기아차 | 현대차, 현대모비스, 현대중공업, 삼성중공업 | 현대차, 현대모비스, 현대글로비스 |
| 신세계 | 롯데쇼핑, 호텔신라, 현대백화점, 포스코인터내셔널, 한샘 | 롯데쇼핑, 호텔신라, 현대백화점 |

본 논문에서는 82개 종목을 주가 흐름이 비슷한 산업군으로 분류하기 위하여 소셜 네트워크 분석을 이용하였다. 소셜 네트워크 분석이란 독립적인 개념들간의 대칭 혹은 비대칭적인 관계를 각 구성원을 나타내는 노드(node)와 구성원들 간의 관계를 나타내는 엣지(edge)라는 두 개의 요소로 이루어진 네트워크 그래프로 표현한 분석기법으로 정보통신기술의 발달과 더불어 최근 다양한 분야에서 많은 관심을 받고 있다.

82개 종목들의 네트워크 구성 방법으로는 각각의 종목들을 노드로 나타내고 종목들 간의 관계를 나타내는 엣지의 가중치로는 각 종목들 간의 로그수익률의 상관계수 $\rho$를 거리 $d = 10(1 - \rho)$로 변환하여 사용하였다.

네트워크 구성에서 모든 노드들이 서로 연결이 된다면 매우 복잡한 네트워크가 형성되므로 해석의 간결성을 위하여 각 종목에서 상위 2개의 상관계수를 갖는 종목들을 선정하여 네트워크를 구성한다. 네트워크를 구성한 후 탐욕 알고리즘(Greedy algorithm)을 이용하여 군집을 분류한다. 탐욕 알고리즘이란 최적의 해를 구하는데 사용되는 근사적 방법으로 여러 경우 중 하나의 선택을 해야 할 때마다 그 중 최적이라고 생각되는 것을 선택하며 진행하여 최종적인 결론에 도달하는 알고리즘이다. 즉, 각 노드에서 주가 흐름이 가장 가까운 노드를 하나씩 선택해 나가면서 군집을 분류하게 되며 분류된 군집 내 종목들의 주가의 흐름은 동질적인 성향을 띄게 된다.

탐욕 알고리즘 기법을 이용하여 분류한 산업군 분류 결과는 Figure 3.1과 같다. Table 3.1에서 SNA에 의한 기아차 및 신세계의 산업군 분류 결과는 기존의 한국표준산업분류(Korean Standard Industry Classification; KSIC)와 비교된다.

## 4. 감성분석에 의한 감성변수 생성

감성분석이란 웹사이트, 소셜미디어, 혹은 문서 등에 나타난 텍스트(text) 자료를 대상으로 텍스트마이닝을 통하여 텍스트에 내포되어있는 사람의 감성을 추출하는 기법이다. 감성분석의 핵심은 텍스트를 이루고 있는 단어가 갖는 감성점수로부터 단어를 긍정, 부정, 혹은 중립의 감성으로 분류하는 것이다. 따라서 감성분석의 첫 단계는 단어별로 단어가 갖는 감성점수를 목록으로 정리해놓은 '감성사전'의 구축에 있다.

온라인 뉴스 속의 단어로부터 주가 상승(긍정), 하락(부정), 혹은 보합(중립)의 감성을 추출하는 감성분석이 주가예측에 활발하게 활용되면서 증시에 특화된 감성사전 구축 방법에 대한 연구도 다양하게 진행되어왔다 (Kim, 2012; Jeong 등, 2015; Choi, 2016; Lee와 Lee, 2017; Kim과 Kim, 2017; Kim과 Lee, 2018; Seong과 Nam, 2018).

KOSPI와 같이 전체 통합된 주가를 예측하고자 할 때는 전체 주식 시장에 대한 온라인 뉴스들을 대상으로 감성사전을 구축할 필요가 있다. 그러나 개별 종목의 주가를 예측하고자 할 때는 개별 종목의 뉴스뿐만 아니라 개별 종목이 속한 산업군에 특화된 감성사전이 필요하다. 예를 들어 '자율'이라는 단어는 운수장비 산업군에서는 자율주행 자동차에 관련한 뉴스에서 등장할 가능성이 높고 자율주행 관련 뉴스는 운수장비 산업군에 속한 종목들의 주가에 긍정적인 영향을 미칠 것으로 생각된다. 하지만 다른 산업군에서는 '자율'이라는 단어가 긍정적인 영향을 미칠 것으로 확실할 수는 없을 것이다. 본 논문에서는 개

**Table 4.1.** An example of emotional dictionary

| Word | Frequency | | | | Proportion | | | Emotion |
|---|---|---|---|---|---|---|---|---|
| | Overall | Up | Keep | Down | Up | Keep | Down | score |
| 가격경쟁 | 34 | 9 | 10 | 15 | 0.2647 | 0.2941 | 0.4412 | −0.1765 |
| 가격부담 | 34 | 10 | 9 | 15 | 0.2941 | 0.2647 | 0.4412 | −0.1471 |
| 가능성 | 8089 | 3074 | 1900 | 3115 | 0.3800 | 0.2349 | 0.3851 | −0.0051 |
| 활력 | 169 | 58 | 43 | 68 | 0.3432 | 0.2544 | 0.4024 | −0.0592 |
| 혁신적 | 286 | 111 | 80 | 95 | 0.3881 | 0.2797 | 0.3322 | 0.0559 |
| 상승한 | 4417 | 1814 | 991 | 1612 | 0.4107 | 0.2243 | 0.3650 | 0.0457 |
| 활성화 | 1074 | 438 | 268 | 368 | 0.4078 | 0.2495 | 0.3426 | 0.0652 |
| 힘입다 | 4407 | 1738 | 1077 | 1592 | 0.3944 | 0.2444 | 0.3612 | 0.0331 |

별 종목인 기아차 및 신세계의 주가예측을 목적으로 하므로 각 산업군에 특화된 감성사전을 구축하였다.

온라인 뉴스 데이터를 수집할 때 다양한 형식으로 되어 있는 기사들을 모두 추출하는 작업은 상당히 어려운 일이다. 네이버는 2017년 기준 뉴스/미디어 부분에서 가장 높은 점유율 60%를 차지한다. 따라서 네이버 형식으로 되어있는 기사들만 선정하여 뉴스 추출을 진행하였다. 또한 네이버에서 종목 이름으로 검색한 뉴스 기사들은 너무 방대하여 2017년 네이버 뉴스에서 가장 높은 점유율을 보이는 '연합뉴스', 그리고 경제신문사 중 가장 점유율이 높은 '머니투데이'와 '아시아경제'의 총 3개 언론사를 선정하여 뉴스 데이터를 수집하였다. 뉴스 데이터 수집에는 R 프로그램의 rvest 패키지를 사용하였다.

rvest 패키지를 이용하여 온라인 뉴스 데이터를 수집하면 특수문자나 html 관련 명령어들이 함께 추출된다. 이러한 불필요한 문자들이 문서에 섞이게 되면 뜻을 가지고 있는 가장 작은 말의 단위인 형태소를 분리할 때 잘못된 분리가 이루어질 가능성이 있을 뿐만 아니라 불필요한 문자들이 감성사전에 포함되게 되므로 특수문자, 숫자, 영단어 등을 제거해주는 전처리 과정을 거쳤다. 또한 대부분 뉴스 데이터의 마지막 3문장은 광고나 작성 기자의 이름, 이메일 주소 등이 포함되는 경우가 많으므로 마지막 3문장을 일괄적으로 제거하는 처리 과정을 거쳤다.

뉴스 데이터 전처리 과정을 마친 후에는 형태소 분리를 진행하였다. 형태소 분리는 R 프로그램의 KoNLP 패키지를 이용하였고, 패키지에 내장되어 있는 NIADic 사전을 사용하였다. 형태소 분리를 마친 단어들은 여러 품사를 가지게 되는데 문서의 감성을 잘 나타낼 것이라고 생각되는 5가지 품사로서 보통명사, 고유명사, 동사, 형용사, 그리고 보조용언을 감성사전 구축에 사용하였다.

5가지 품사의 단어를 추출한 후에는 단어들이 문서에 출현하는 빈도에 기초한 감성사전 구축 방법을 이용하였다. 즉, 당일 온라인 뉴스에 나온 특정 단어의 경우에 다음날 주가가 상승, 하락, 혹은 보합 일수를 각각 집계한 후 전체 출현 일수로 나눈 값을 각각 긍정, 부정, 혹은 중립점수로 계산하여 각 단어의 감성점수를 구한다. 이때 하나의 뉴스 기사에 특정 단어가 매우 높은 빈도로 출현하는 경우 그 뉴스 기사로 인해 단어의 감성점수가 왜곡될 수 있다. 따라서 기사에 특정 단어가 여러 번 등장하여도 한번으로 카운팅하여 계산하였다. 긍정점수와 부정점수는 다음과 같이 정의된다.

$$\text{긍정점수} = \frac{\text{익일 주가 상승시 뉴스 출현 빈도}}{\text{전체 뉴스 출현 빈도}}, \quad \text{부정점수} = \frac{\text{익일 주가 하락시 뉴스 출현 빈도}}{\text{전체 뉴스 출현 빈도}}. \tag{4.1}$$

최종적으로 단어의 감성점수는 긍정점수에서 부정점수를 뺀 값으로 계산되며 이 값은 −1에서 1까지의 값을 가진다. 이러한 과정을 거쳐 구축된 감성사전의 예시는 Table 4.1과 같다.

감성사전 구축 후 특정 뉴스 기사에 출현한 단어들의 감성사전 내의 감성점수를 가중평균하여 뉴스 기사의 감성점수를 구한다. 이때 '것', '년', '수', '때' 등과 같이 의미 없는 불용 단어의 감성점수가 뉴스의

감성 값을 왜곡할 수 있다. 따라서 뉴스 기사의 감성점수를 가중평균하여 계산할 때 불용 단어의 감성점수 계산에 끼치는 영향력을 줄이기 위해 Manning 등 (2009)의 TF-IDF 값을 가중치로 사용하였다.

TF-IDF는 TF와 IDF 값의 곱으로 구해진다. Term frequency (TF) 값은 하나의 뉴스 기사 내에서 등장하는 특정 단어의 빈도를 나타낸 값으로 이 값이 클수록 뉴스 기사에서 특정 단어가 중요하다고 생각될 수 있다. document frequency (DF)는 특정 단어가 나타난 뉴스 기사의 수이며 inverse DF (IDF) 값은 DF 값의 역수에 로그를 취한 값으로 IDF 값이 클수록 DF 값은 작으므로 특정 단어가 나타난 뉴스 기사의 수는 적다. 따라서 TF-IDF 값이 클수록 현재 뉴스 기사에서는 자주 언급되면서 다른 뉴스 기사에서는 잘 언급되지 않음을 의미하고, 값이 작을수록 다른 뉴스 기사에서는 많이 언급되면서 현재 뉴스 기사와는 관련성이 작음을 의미한다 따라서 TF-IDF 값을 이용하면 특정 뉴스 기사에만 중요하게 생각되는 단어를 찾아낼 수 있을 뿐만 아니라 모든 뉴스 기사에 흔히 나타나는 불용 단어를 걸러내는 효과를 얻을 수 있다

이렇게 구해진 TF-IDF 값을 가중치로 하여 뉴스 기사 내 각 단어들의 감성점수를 가중평균 하면 하나의 뉴스 기사에 대한 감성점수를 구할 수 있다. 본 논문에서는 당일 주식 시장이 열린 9시부터 다음날 주식 시장이 열리기 전 8시까지의 각 온라인 뉴스 기사들의 감성점수들의 평균값을 '감성변수'라고 정의한다. Table 2.1에서 정의된 산업군 감성변수는 동질적인 산업군에 속하는 각 개별 종목의 감성변수 값을 평균하여 얻는다.

## 5. 심층 신경망에 의한 예측

### 5.1. 심층 신경망 개요

본 논문에서는 심층 신경망모형에 의한 주가예측을 위해서 R 프로그램의 h2o 패키지를 사용하여 분석하였다. 심층 신경망모형은 기존의 인공 신경망모형의 단점으로 지적되던 속도가 느리다는 점과 과적합 문제를 해결하면서 높은 예측 정확도를 보이는 기계학습법이다. 심층 신경망모형은 입력층과 출력층 사이에 다수의 은닉층이 있고 은닉층은 다수의 은닉노드를 포함하도록 구성되어서 목표변수와 예측변수들 간의 다양한 비선형적 관계를 학습할 수 있다. 심층 신경망모형에서 다층 퍼셉트론은 Backpropagation 알고리즘에 의해 학습되며 가중치들은 Stochastic gradient descent을 통해 갱신된다.

신경망 학습에 필요한 요소로 활성함수가 있다. 활성함수는 입력 신호의 총합을 출력 신호로 변환하는 함수를 말한다. 예측변수들을 $X_1, X_2, \ldots, X_n$이라 하고 예측변수들에 대한 가중치를 $w_1, w_2, \ldots, w_n$이라 할 때 예측변수들의 가중결합으로 표현되는 입력 신호의 총합 $\alpha$는 다음과 같다.

$$\alpha = \sum_{i=1}^{n} w_i x_i + b. \tag{5.1}$$

h2o 패키지 내에서 심층 신경망모형 분석에 사용할 수 있는 활성함수 종류는 Hyperbolic Tangent (tanh), Rectified Linear Unit (RELU), Maxout의 3가지가 있다.

tanh 함수는 시그모이드 함수의 대체재로 사용할 수 있는 활성함수이며 입력신호의 총합을 $-1$에서 1 사이의 값으로 변환한다. 원점 중심이기 때문에 편향 이동이 일어나지 않지만 입력신호의 절댓값이 클 경우 그래디언트를 소멸시켜버리는 문제를 갖는다. 입력신호의 총합 $\alpha$에 대한 tanh 함수의 식은 아래와 같다.

$$\tanh(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}. \tag{5.2}$$

**Table 5.1.** Definition of model number

| Hidden node | Epoch | L1, L2 normalization | | |
|---|---|---|---|---|
| | | 0.01 | 0.001 | 0.0001 |
| 20 | 10 | 1–10 | 11–20 | 21–30 |
| | 20 | 31–40 | 41–50 | 51–60 |
| | 30 | 61–70 | 71–80 | 81–90 |
| 50 | 10 | 91–100 | 101–110 | 111–120 |
| | 20 | 121–130 | 131–140 | 141–150 |
| | 30 | 151–160 | 161–170 | 171–180 |
| 100 | 10 | 181–190 | 191–200 | 201–210 |
| | 20 | 211–220 | 221–230 | 231–240 |
| | 30 | 241–250 | 251–260 | 261–270 |

RELU 함수는 시그모이드 계열과는 다른 활성함수이며, 입력신호의 총합이 0 이하이면 0을 출력하고, 0 이상이면 입력 신호의 총합을 그대로 출력하는 함수이다. 따라서 0 이상인 곳에서는 수렴하는 구간이 없다는 특징을 가지고 입력 값을 그대로 출력해주기 때문에 계산 속도가 빠르다는 특징이 있다. 입력신호의 총합 $\alpha$에 대한 RELU 함수의 식은 아래와 같다.

$$\mathrm{RELU}(\alpha) = \max(0, \alpha). \tag{5.3}$$

Maxout 함수는 위에서 설명한 RELU 함수와 Leaky RELU 함수를 일반화한 형태인데, Leaky RELU 함수란 RELU 함수에서 입력 신호의 총합이 0 이하일 경우에 미분 값이 항상 0으로 되어 해당 신경망이 활성화되지 않게 되는 문제를 보완한 것이다. Maxout 함수는 RELU와 Leaky RELU 함수에 의한 각 입력 신호의 총합 $\alpha_1$, $\alpha_2$을 계산하여 더 큰 값을 출력하는 형태이다. RELU 함수의 장점을 그대로 가지면서 단점을 보완하는 함수이지만, 계산 시간이 길다는 단점을 가진다. 입력 신호의 총합 $\alpha_1$, $\alpha_2$에 대한 Maxout 함수의 식은 다음과 같다.

$$\mathrm{Maxout}(\alpha) = \max(\alpha_1, \alpha_2). \tag{5.4}$$

### 5.2. 심층 신경망 학습

이 절에서는 심층 신경망모형을 학습시켜 좋은 성능을 보이는 모형의 옵션을 선택하는 과정을 설명한다. 심층 신경망모형 학습에는 h2o 패키지에 내장되어 있는 h2o.deeplearning 함수를 이용하였다.

심층 신경망모형을 학습시킬 때 사용자는 기본적으로 활성함수의 종류, 은닉층 수, 은닉노드 수, Epoch, L1 및 L2 정규화 등의 옵션을 설정해야 한다. 기본적인 6개 옵션 조합의 수는 매우 방대하므로 옵션값을 결정할 때 몇 단계의 시뮬레이션을 통하여 옵션 값을 정해 나갔다. 선행적인 시뮬레이션 결과로서 활성함수는 RELU, 은닉층의 개수는 3개로 고정하였다. 정규화의 경우 L1, L2 정규화를 혼합하여 사용하는 Elastic Net 정규화를 이용하였다. Elastic Net 정규화는 변수와 분산을 동시에 줄이고 싶을 때 사용하는 정규화 기법으로 h2o 패키지 내에서는 L1 및 L2 정규화 값을 동일하게 설정하여 적용할 수 있다.

Table 5.1은 옵션들의 조합을 달리하는 시뮬레이션 모형 번호를 정의한 표로서 Epoch 값은 10, 20, 30, 3개 은닉층의 각 은닉노드 수는 20, 50, 100, 그리고 L1 및 L2 정규화 값은 동일하게 0.01, 0.001, 0.0001로 값을 달리하여 정의한 모형 번호이다. 각각의 지정된 옵션값에서 10번씩 반복 실험하여 최적의 모형을 선정하는 과정을 거쳤다.

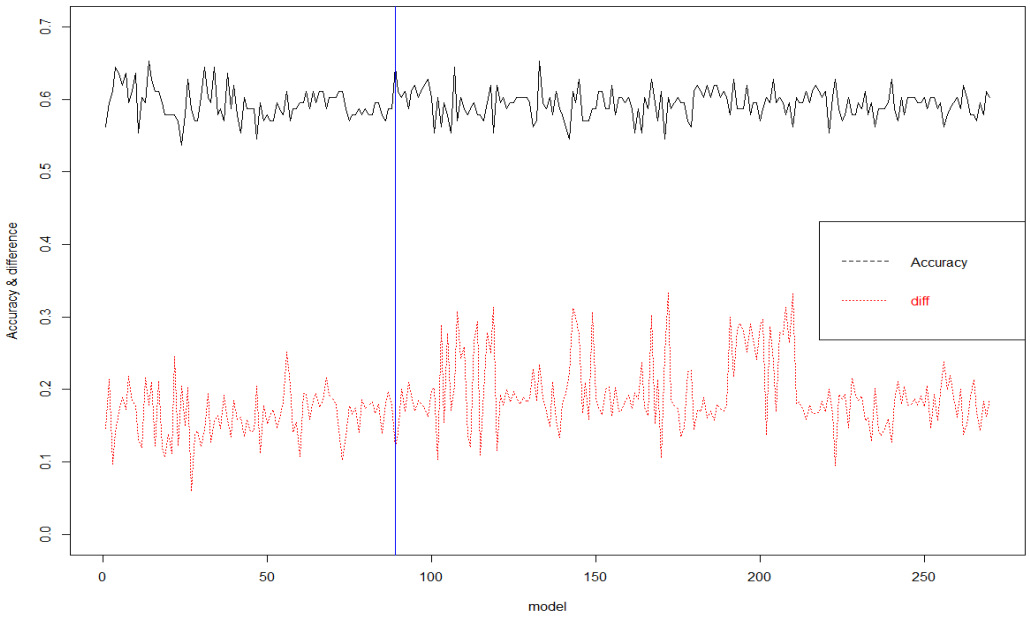


**Figure 5.1.** Accuracy for validation data and the difference between the accuracy of training and validation data.

Figure 5.1은 기아차의 익일 시가 대비 종가가 상승인 경우는 1의 값을 하락인 경우는 0의 값을 갖는 이항변수를 목표변수로 갖고 51개의 기술적 변수들과 SNA에 기초하여 분류한 산업군 감성변수를 예측변수로 갖는 심층 신경망모형에 대하여 Table 5.1에서 정의한 모형 번호에 따른 평가용 데이터의 정확도(accuracy) 그리고 학습용 데이터 정확도와 평가용 데이터 정확도의 차이(diff) 값을 나타내는 그래프이다. 최적의 모형을 선정하는 기준으로 과적합을 막기 위해서 하위 10% diff 값을 갖는 모형 중에서 평가용 데이터에서 가장 높은 정확도인 64.46%를 갖는 모형을 선정하였다. 이 모형은 Epoch = 30, hidden node = 20, L1 및 L2 정규화 옵션이 0.0001인 89번째 모형으로서 Figure 5.1에서 수직선으로 표시된 모형이다.

Table 5.1 및 Figure 5.1에서와 같은 최적 모형 선택방법은 batch size = 1로 해서 전체 270개 모형 중에서 최대 정확도를 갖는 모형을 선택한 방법이다. 이때 하나의 옵션 세트에 대하여 10회 반복을 하였기 때문에 batch size = 10으로 해서 10회에 대한 정확도의 평균 비교를 하여 최적 옵션을 택할 수도 있지만 분산이 큰 경우에 오히려 결과가 좋지 않았다.

### 5.3. 정확도 및 수익률 비교

이 절에서는 5.2절에서 설명한 절차에 의해 선정된 심층 신경망모형을 이용하여 기아차 및 신세계 주가의 예측 정확도 및 수익률을 비교해보기로 한다.

기아차의 종가 시계열을 나타내는 Figure 5.2에서 두개의 수직선은 전체 데이터를 시간 순서대로 학습용, 평가용, 그리고 검증용 데이터로 구분한다. 마지막 6개월 간의 검증용 데이터에 대하여 주가예측의 정확도와 수익률을 계산하였다.

수익률의 계산과정은 투자자들이 Table 5.2와 같이 매매한다고 가정했을 때의 결과이며 매매 수수료는 매수 및 매도 모두 동일하게 0.3%로 계산하였다. 모든 매매는 익일 8시까지의 정보를 가지고 익일 9시


**Figure 5.2.** Kia Motors' four-year closing time series.

**Table 5.2.** Stock trading strategy

| Stock holding status | Prediction | Behavior |
|---|---|---|
| Holding stock | Up(1) | Keep state |
| | Down(0) | Sell stock at starting price |
| Not holding stock | Up(1) | Buy stock at starting price |
| | Down(0) | Keep state |

시가에서 이루어진다. 익일 8시에 주식을 보유한 경우에는 목표변수 값인 익일 시가 대비 종가가 상승한다고 예측되면 현 상태를 유지하고, 하락한다고 예측하면 시가에 매도한다. 반대로 익일 8시에 주식을 보유하지 않은 경우에는 목표변수 값인 익일 시가 대비 종가가 상승한다고 예측되면 시가에 매수하고 하락한다고 예측하면 현 상태를 유지한다.

Table 5.3은 예측변수들의 조합에 따른 모형에 대하여 기아차 주가의 예측 정확도 및 수익률을 나타낸다. 각 모형은 앞서 설명한 심층 신경망모형의 최적 옵션 선정 방식에 의하여 선정된 모형들이다.

SNA 및 KSIC에 의해 분류된 기아차가 포함되는 운수장비 산업군은 Table 3.1을 참고한다. 예측 결과는 51개 기술적 변수(TI)와 함께 SNA 분류법에 의한 운수장비 산업군에 기초한 산업군 감성변수(SNA-g)를 예측변수로 사용하였을 때 59.35%의 가장 높은 정확도와 22.23%의 가장 높은 수익률을 보였다. 51개 기술적 변수(TI)와 KSIC에 따라 분류된 운수장비 산업군에 기초한 산업군 감성변수(KSIC-g)를 사용하였을 때는 두 번째로 높은 정확도 57.72%를 보였으나 수익률은 1.05%로 은행 수익률 1.5%보다도 더 낮았다. 이 모형의 결과는 51개 기술적 변수(TI)만을 예측변수로 사용하였을 때의 결과와 비슷하다. 51개 기술적 변수(TI)와 함께 SNA 분류법에 의한 운수장비 산업군에 기초한 기아차만의 감성변수(SNA-c)를 예측변수로 사용하였을 때는 수익률은 20.82%로 높지만 정확도는 52.03%로 다소 낮은 편이다.

**Table 5.3.** Predictive accuracy and rate of return for Kia Motors on models with selected predictors

| Predictors | TI SNA-g | TI SNA-c | TI KSIC-g | TI | Random | Bank deposit |
|---|---|---|---|---|---|---|
| Epoch | 30 | 30 | 20 | 10 | | |
| Hidden layers | 3 | 3 | 3 | 3 | | |
| Hidden nodes | 20 | 20 | 20 | 50 | | |
| L1 | 0.0001 | 0.001 | 0.01 | 0.001 | | |
| L2 | 0.0001 | 0.001 | 0.01 | 0.001 | | |
| Activation | RELU | RELU | RELU | RELU | | |
| Accuracy | 59.35% | 52.03% | 57.72% | 56.10% | 49.99% | - |
| Rate of return | 22.23% | 20.82% | 1.05% | 1.18% | −5.21% | 1.5% |

TI = 51 technical indicators; SNA-g = industrial group emotional variable using SNA; SNA-c = company emotional variable using SNA; KSIC-g = Industrial group emotional variable using KSIC; SNA = social network analysis; KSIC = Korean Standard Industry Classification.

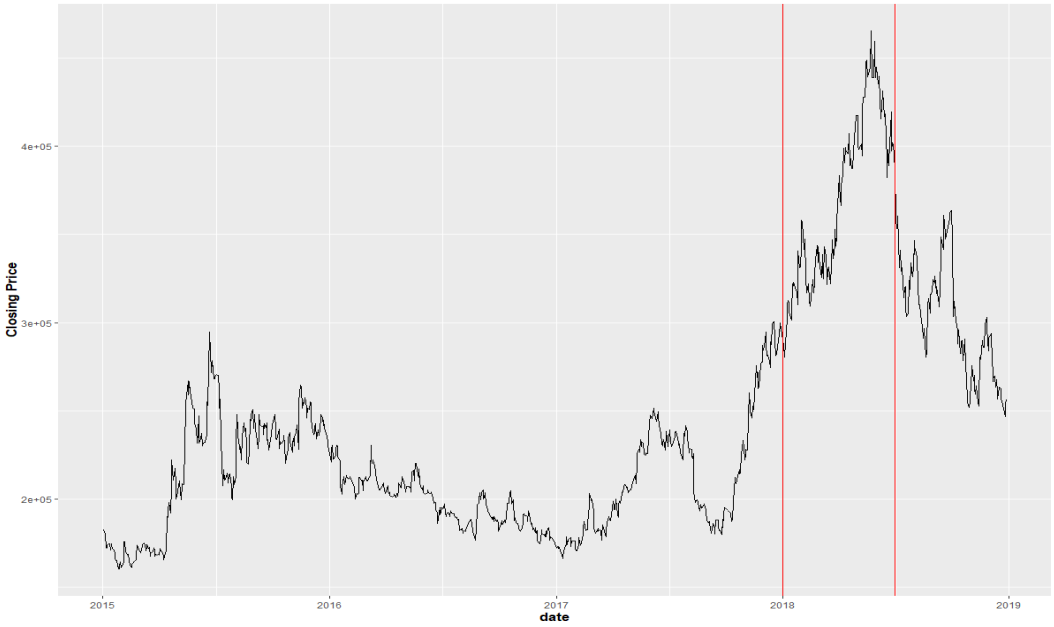


**Figure 5.3.** Shinsegae's four-year closing time series.

Table 5.3에서 Random은 랜덤하게 예측하였을 때를 의미하는데, 상승 및 하락을 각 50% 확률을 갖는 1,000회 반복하는 확률실험에서 평균 정확도는 49.99%이고 평균 수익률은 −5.21%이었다.

Figure 5.3은 신세계 종목의 종가 시계열을 나타낸다. 이 그림에서 두개의 수직선은 전체 데이터를 시간 순서대로 학습용, 평가용, 그리고 검증용 데이터를 구분한다.

Table 5.4는 예측변수 조합에 따른 모형에 대하여 신세계 종목의 예측 정확도 및 수익률을 나타낸다. SNA 및 KSIC에 의해 분류된 신세계가 포함되는 유통업 산업군은 Table 3.1을 참고한다. 예측 결과는 51개 기술적 변수(TI)와 함께 SNA 분류법에 의한 유통업 산업군에 기초한 산업군 감성변수(SNA-g)를 예측변수로 사용하였을 때 60.16%의 가장 높은 정확도와 2.36%의 가장 높은 수익률을 보였다. 산업군

**Table 5.4.** Predictive accuracy and rate of return for Shinsegae on models with selected predictors

| Predictors | TI SNA-g | TI SNA-c | TI KSIC-g | TI |
|---|---|---|---|---|
| Epoch | 30 | 30 | 20 | 10 |
| Hidden layers | 3 | 3 | 3 | 3 |
| Hidden nodes | 50 | 20 | 20 | 50 |
| L1 | 0.0001 | 0.001 | 0.0001 | 0.001 |
| L2 | 0.0001 | 0.001 | 0.0001 | 0.001 |
| Activation | RELU | RELU | RELU | RELU |
| Accuracy | 60.16% | 56.91% | 58.54% | 55.28% |
| Rate of return | 2.36% | 2.18% | −18.49% | −15.68% |

TI = 51 technical indicators; SNA-g = industrial group emotional variable using SNA; SNA-c = company emotional variable using SNA; KSIC-g = Industrial group emotional variable using KSIC; SNA = social network analysis; KSIC = Korean Standard Industry Classification.

감성변수 대신 신세계 개별 기업의 감성변수(SNA-c)를 사용하였을 때는 수익률은 2.18%로 비슷하나 정확도는 56.91%로 더 낮은 편이다. 51개 기술적 변수(TI)와 KSIC에 따라 분류된 유통업 산업군에 기초한 산업군 감성변수(KSIC-g)를 사용하였을 때는 두 번째로 높은 정확도 58.54%를 보였으나 수익률은 −18.49%로 원금을 손해보는 결과를 준다. 51개 기술적 변수(TI)만을 예측변수로 사용하였을 때도 결과는 비슷하다.

기아차 및 신세계 종목의 예측변수 조합에 따른 예측 정확도와 수익률을 비교해보았을 때, 두 종목 모두 기술적 변수와 SNA 분류법에 의한 산업군의 감성변수를 예측변수로 활용했을 때 가장 높은 정확도와 수익률을 나타냈다. 기아차와 신세계 종목의 수익률은 평균적으로 많은 차이를 보이는데 이는 신세계 종목의 경우 Figure 5.3에서 검증용 데이터의 기간인 2018년 7월부터 주가가 계속해서 하락하는 추세를 보이기 때문에 신세계 종목의 예측 정확도가 높더라도 높은 수익률을 나타내지 못하는 것으로 생각된다.

## 6. 결론

본 논문에서는 기아차 및 신세계 종목에 대하여 2015년 1월부터 2017년까지 12월까지의 일별 주가 시계열 데이터로 심층 신경망모형을 학습시킨 후, 2018년 7월부터 6개월 동안의 시가 대비 종가의 상승 혹은 하락을 예측하였다. 예측변수로는 기존 연구들에서 사용되던 51개 기술적 변수들과 온라인 뉴스로부터 도출한 감성변수를 사용하였다. 특히 KOSPI 100에 속한 종목들의 산업군을 소셜 네트워크 분석 방법을 이용하여 분류하고 각각의 산업군에 특화된 감성사전을 구축하여 종목이 속한 산업군의 온라인 뉴스 감성변수를 생성하였다.

본 논문에서 사용한 예측 절차의 성능을 확인하기 위하여 다양한 예측변수 조합을 갖는 심층 신경망모형에 대하여 시뮬레이션을 통해 찾아낸 최적의 옵션 값으로 학습시켰을 때 기술적 변수와 소셜 네트워크 분석방법을 이용하여 분류한 산업군에 기초한 감성변수를 함께 예측변수로 사용한 모형이 가장 높은 예측 정확도와 수익률을 주었다.

개별 기업만의 감성변수를 만들 때 문제점은 인지도가 높지 않은 기업의 경우는 관련된 뉴스량이 매우 적었다. 따라서 비교적 인지도가 높고 기사량이 많다고 생각되는 종목인 기아차와 신세계를 선택하여 분석하였다. 그러나 기사량이 많지 않은 종목을 주가 예측할 경우에는 산업군에 기초한 감성변수를 사용해도 좋은 결과를 줄 수 있다는 것을 자료분석 결과로부터 기대할 수 있다.

흔히 주가지수 예측을 다루는 연구에서 주로 다루는 주제는 익일 주가의 상승 혹은 하락을 예측하는 것이다. 그리고 대부분 60% 내외의 예측 정확도를 보여준다. 주식투자의 핵심은 수익률의 극대화이지만 높은 정확도가 수익을 보장해주지는 않는다. 정확도가 높더라도 수익률이 낮은 이유는 익일 주가의 상승 혹은 하락만을 예측하고 얼마만큼 상승하고 하락하는지를 나타내는 주가의 등락 폭에 대해서는 고려하지 못한다는 한계점을 갖기 때문이다. 흔히 사용되는 기술적 변수들에 더해서 예측하려는 종목의 산업군 감성변수를 포함한 모형에서는 주가의 상승 혹은 하락을 예측함에 있어서 등락의 폭이 큰 경우가 좀 더 반영되어 보다 더 높은 수익률을 주지 않았나 추측한다.

마지막으로 본 연구에서 한계점은 다음과 같다. 첫 번째, 단기의 주가를 예측하기 때문에 변동성이 큰 종목은 예측 정확도가 높더라도 수익률 면에서는 큰 손해를 볼 수 있다. 두 번째, 주가예측을 위해 전날 뉴스 데이터를 이용하기 때문에 주식 시장이 열린 이후 등장하는 뉴스를 즉각적으로 반영하지 못한다는 한계를 갖는다. 세 번째, 뉴스 데이터에 오탈자나 띄어쓰기가 잘못되어 있는 경우에 형태소 분리가 잘 이루어지지 않아 불용어가 감성사전에 포함될 수 있다는 한계를 갖는다.

후속 연구로 주가의 등락 폭을 예측하여 매매 기준을 정함으로서 수익률을 극대화 할수 있는 다양한 딥러닝 기법을 적용해보는 연구를 진행할 수 있을 것이며 텍스트 처리 기능을 보완하여 문서의 감성을 정밀하게 추출할 수 있는 텍스트마이닝 기법 연구도 진행할 수 있을 것이다.

## References


Choi, I. J. (2016). Prediction of stock fluctuation using web news text mining, *Pukyong National University Master's Thesis*.

Jeong, J. S., Kim, D. S., and Kim, J. W. (2015). Influence analysis of internet buzz to corporate performance: individual stock price prediction using sentiment analysis of online news, *Journal of Intelligence and Information Systems*, **21**, 37–51.

Kim, D. Y. and Lee, Y. I. (2018). News based Stock Market Sentiment Lexicon Acquisition Using Word2Vec, *The Korean Journal of Bigdata*, **3**, 13–20.

Kim, J. B. and Kim, H. J. (2017). A domain-specific sentiment lexicon construction method for stock index directionality, *Journal of Digital Contents Society*, **18**, 585–592.

Kim, Y. S. (2012). News big data opinion mining model for predicting KOSPI movement, *Kookmin University Master's Thesis*.

Lee, M. and Lee, H. J. (2017). Stock price prediction by utilizing category neutral terms: text mining approach, *Journal of Intelligence and Information Systems*, **23**, 123–138.

Manning, C. D., Raghavan, P., and Schutze, H. (2009). Introduction to Information Retrieval: Scoring, term weighting, and the vector space model, Cambridge University Press.

Seong, N. and Nam, K. (2018). Online news-based stock price forecasting considering homogeneity in the industrial sector, *Journal of Intelligence and Information Systems*, **24**, 1–19.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., and Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning, *Decision Support Systems*, **85**, 74–83.

# 산업군별 온라인 뉴스에 기초한 감성 예측변수를 포함하는 심층 신경망모형에 의한 주가 예측

임준형[a] · 손영숙[a,1]

[a]전남대학교 통계학과

(2020년 4월 28일 접수, 2020년 6월 1일 수정, 2020년 6월 11일 채택)

**요 약**

본 연구에서는 심층 신경망모형을 사용하여 KOSPI 100의 개별 종목인 기아차 및 신세계의 주가를 예측하였다. 예측변수로는 흔히 사용되었던 기술적 변수들과 함께 온라인 뉴스로부터 도출된 감성변수를 사용하였다. 특히 소셜 네트워크 분석을 활용하여 분류된 산업군에 특화된 감성사전을 구축한 후, 감성분석을 통하여 산업군에 속하는 각 기업들의 감성점수의 평균을 산업군 감성변수로 생성하였다. 여러 예측변수들의 조합으로 이루어진 모형들 중에서 기술적 변수와 산업군의 온라인 뉴스에 기초한 감성변수를 함께 사용하였을 때 우수한 예측력과 수익률을 보여주었다.

주요용어: 주가예측, 심층 신경망, 소셜 네트워크 분석, 온라인 뉴스, 감성변수

[1]교신저자: (61186) 광주광역시 북구 용봉로 77(용봉동), 전남대학교 통계학과. E-mail: ysson@jnu.ac.kr