http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.4, November 2020, pp.61-75
https://doi.org/10.7838/jsebs.2020.25.4.061

SNS감성 분석을 이용한 주가 방향성 예측: 네이버 주식토론방 데이터를 이용하여

Stock Price Prediction Using Sentiment Analysis: from "Stock Discussion Room" in Naver

김명진(Myeongjin Kim)*, 류지혜(Jihye Ryu)**,
차동호(Dongho Cha)***, 심민규(Min Kyu Sim)****

초 록

주식의 가격을 이해하고 예측하기 위해서 활용되는 데이터의 범위는 기존의 정형화된 데이터에서 비정형화된 다양한 종류의 데이터로 확대되고 있다. 본 연구는 SNS에서 수집된 댓글 데이터가 주식의 미래 가격의 변동에 영향을 미치는지를 조사한다. 가장 많은 주식투자자가 참여하는 커뮤니티인 네이버 주식토론방에서 20개 종목에 대한 6개월 간의 댓글 데이터를 수집하여, 이들 데이터가 1시간 후의 가격 변동의 방향과 가격 변동의 폭에 대한 예측력을 가지는지 조사한다. 예측 관계는 LSTM과 CNN등의 딥뉴럴네트워크 기법을 활용하여 모델링 하였다. 20개 종목에 대해 조사하여 13개 종목에서 미래의 주가 이동 방향을 50% 이상의 정확도로 예측할 수 있다는 결과를 얻었고, 16개 종목에서 미래의 주가 변동폭을 50% 이상의 정확도로 예측할 수 있다는 결과를 얻었다. 본 연구는 네이버 주식토론방과 같은 SNS에서 형성된 여론이 주식 종목의 수급에 영향을 주어 가격의 변동 요인으로도 작용할 수 있다는 점을 확인한다.

ABSTRACT

The scope of data for understanding or predicting stock prices has been continuously widened from traditional structured format data to unstructured data. This study investigates whether commentary data collected from SNS may affect future stock prices. From "Stock Discussion Room" in Naver, we collect 20 stocks' commentary data for six months, and test whether this data have prediction power with respect to one-hour ahead price direction and price range. Deep neural network such as LSTM and CNN methods are employed to

이 연구는 서울과학기술대학교 교내연구비의 지원으로 수행되었습니다. This study was supported by the Research Program funded by the SeoulTech(Seoul National University of Science and Technology).

* First Author, Bachelor Student, Department of Industrial Engineering(ITM program), Seoul National University of Science and Technology(kevin_960819@seoultech.ac.kr)
** Co-Author, Bachelor Student, Department of Industrial Engineering (ITM program), Seoul National University of Science and Technology(jhryu48@seoultech.ac.kr)
*** Co-Author, Team Leader, Headquarter of Multi-Solution, KB Asset Management)(movieish@gmail.com)
**** Corresponding Author, Assistant Professor, Department of Industrial Engineering, Seoul National University of Science and Technology(mksim@seoultech.ac.kr)

Received: 2020-09-09, Review completed: 2020-10-14, Accepted: 2020-11-03

model the predictive relationship. Among the 20 stocks, we find that future price direction can be predicted with higher than the accuracy of 50% in 13 stocks. Also, the future price range can be predicted with higher than the accuracy of 50% in 16 stocks. This study validate that the investors' sentiment reflected in SNS community such as Naver's "Stock Discussion Room" may affect the demand and supply of stocks, thus driving the stock prices.

키워드 : 고빈도 금융 데이터, 텍스트 마이닝, 감성 분석, 딥뉴럴네트워크
High-Frequency Financial Data, Text Mining, Sentimental Analysis, Deep Neural Networks

1. 서 론

데이터를 이용하여 금융 자산의 움직임을 예측하는 것은 시장 참여자들의 오랜 관심이다. 주식 가격의 변동은 일반적으로 거시 변수에 해당하는 정치, 경제, 사회적 요인과 미시 변수에 해당하는 해당 기업의 주식 매매율, 기업의 이익과 성장 가치 등의 영향을 받는다는 것이 알려져 있다. 전통적으로 특정 기업의 주가를 예측하기 위해서 매출, 현금 흐름, 유무형 비용 등의 정형화(Structured)된 데이터들이 사용되어 왔다. 이들 정형 데이터는 대부분의 경우 전문가의 손을 거치며 정해진 규칙에 따라서 생성되었다는 특징이 있으며, 기업의 내재가치(Fundamental Value)를 추정하는 목적으로 활용된다. 정형 데이터가 기업의 주식 가격의 변동에 어떤 영향을 미치는지에 대해서 자산 가격 결정 이론에 대한 연구가 있다[4-5, 16].

전통적인 정형 데이터가 대부분 사업운영의 결과로서 전문가의 손을 거쳐 회계 장부 등으로 정형화되어 생성되었다면, 데이터 홍수의 시대라고 표현되는 4차 산업혁명 시대의 새로운 데이터(New Data)는 이들과 어떤 차이점이 있을까? 첫째, 새로운 데이터는 비단 사업 운영의 결과로 생성될 뿐만 아니라 그 과정에서 생성되기도 한다. 예를 들어, 사업의 운영 과정에서의 거래 내역, 구매 내역, 심지어 인사 기록들도 회사의 성장성이나 수익성에 영향을 줄 수 있다. 둘째, 새로운 데이터는 전문가나 회사 관계자가 아닌 투자자, 소비자, 그리고 심지어 일반 대중들에 의해서도 생성된다. 예를 들어, 온라인에서 영향력을 가진 사람(인플루언서)의 SNS 포스트, 회사 상품에 대한 온라인 리뷰, 검색 엔진에서의 회사에 관한 검색 횟수 등이 회사의 이미지에 영향을 주어 결과적으로 기업의 성장성이나 수익성에 영향을 줄 수 있다. 셋째, 새로운 데이터는 사람의 손을 거치지 않고 기계에 의해서 생성되기도 한다. 예를 들어 위성 사진의 이미지, 기업의 물류창고 근처의 교통량, 물류 운송 과정 중에 있는 선박의 위치 등도 기업의 수익성에 관한 함의를 가지고 있다. 요약하자면 새로운 데이터는 1) 다양한 과정에서 생성되며, 2) 비전문가에 의해서 생성되기도 하며, 또한, 3) 기계에 의해서 생성된다. 생성 주체의 범위가 넓고 기계가 생성하기도 하기에 생성되는 속도가 빠르고 양이 매우 크다.

단기간에 다양한 형태로 생성되는 다수의 새로운 데이터는 기계학습과 딥러닝으로 대표되는 데이터 처리, 분석 기술에 힘입어 사회의 많은 모습을 바꾸고 있다. 마찬가지로 금융 자산 투자의 영역에서도 이런 새로운 데이터들을 얼마나 잘 활용하냐는 시장 참여자들의 경쟁 우위를

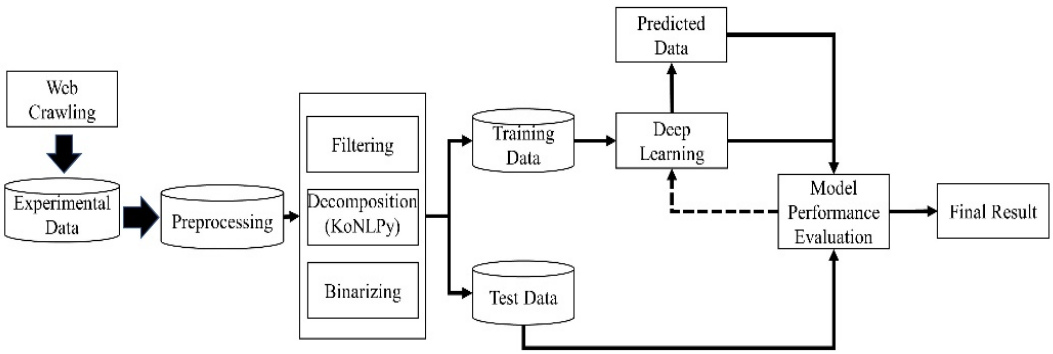
결정하는 요소가 될 것이다[19]. 여러 연구들에서 이에 따라 새로운 데이터들을 이용하여 투자에 활용하는 방안을 조사하였다[9, 10, 15, 18].

본 연구에서는 SNS에서 일반 주식투자자들이 생성한 텍스트 데이터를 수집하여 데이터에 드러난 투자자의 감성이 과연 미래의 자산 가격 움직임에 대한 예측력을 가지고 있는지를 조사한다. 대한민국에서 가장 많이 사용되는 검색 포탈이자 온라인 커뮤니티인 네이버의 "주식토론방"에서 KOSPI 200의 20개 종목에 대해 약50만개의 댓글 데이터를 9개월의 기간(2019년 11월~2020년 07월)에 대해 수집하였다. 미래 자산 가격의 움직임 중에 가격움직임의 방향성과 가격변동 폭의 크기에 대해 조사한다.

비정형 데이터에 해당하는 텍스트 데이터를 이용해 국내 주식시장의 움직임을 분석하려는 시도는 최근 활발히 이루어지고 있다. 데이터의 소스로 보면 크게 트위터, 페이스북과 같은 소셜 데이터를 이용하는 접근[8]과 뉴스 기사를 활용하는 접근[3]으로 나눌 수 있다. 그런데 트위터, 페이스북과 같은 소셜 데이터는 주식 투자를 하고 있거나 관심이 있는 사람뿐만 아니라 일반 불특정 대중의 데이터가 포함되기에 주식 가격에 실제로 영향을 미치는 시장 참여자를 대표한다고 보기 어려운 단점이 있다. 한편, 뉴스 기사의 경우에는 범용 감성 사전을 이용하게 되는데, 이러한 접근은 긍정과 부정의 극성 분류의 정확도가 다소 낮을 수 있다는 한계가 있다. 이에 반해 본 연구에서는 주식 투자를 하고 있거나 높은 관심도를 가진 사람들로부터 생성된 데이터를 활용하였기에 시장 참여자에 대한 대표성을 확보할 수 있고, 뉴스 기사에 비해서 훨씬 많은 양의 데이터를 활용하기에 극성 분류의 다소 낮을 수 있는 정확도의 이슈를 극복할 것이 기대된다.

비정형 데이터인 텍스트로부터 필요한 정보를 추출, 분석하기 위해서는 텍스트를 수치화하는 작업이 필수적으로 요구된다. 텍스트 데이터를 벡터, 행렬 등의 형태로 표현하면 이미 잘 구축되어 있는 데이터 마이닝 기법들을 활용하여 보다 쉽게 결과를 도출할 수 있는 장점이 있다. 데이터의 전처리 과정을 거친 후, 텍스트 데이터의 문맥을 파악하는 기능이 우수하다고 알려진 LSTM 딥러닝 알고리즘을 이용하여 분석을 하였다. LSTM 기법은 기존의 시계열 분석 기법인RNN(Recurrent Neural Network)의 문제점으로 알려진 Gradient Vanishing 문제를 극복한 기법이다[1]. 이때, Gradient Vanishing 문제란, RNN 학습의 역전파 과정에서 입력층으로 갈수록 기울기(Gradient)가 점차적으로 작아지는 현상으로, 입력층에 가까운 층 들에서 가중치의 업데이트가 제대로 되지 않는 문제이다.

본 연구는 개인 투자자가 주를 이루는 커뮤니티의 텍스트 데이터를 이용하여 주식 가격 예측력을 조사한다. 본 연구는 다음과 같은 의의를 가진다. 첫째, 대한민국의 개인 투자자의 감성을 가장 잘 대표할 수 있는 데이터를 활용하였기에, 이들 시장참여자들이 가진 시장에 대한 통찰력, 혹은 이들의 시장에의 영향력을 확인할 수 있다. 둘째, 많은 양의 데이터를 확보하였기에, 일간 가격 변동이 아닌 시간대별 가격 변동에 대한 예측력을 조사하였기에 본 연구의 결과는 감성 분석의 결과가 고빈도 매매에도 활용될 수 있다는 가능성을 제시한다. 셋째, 넓은 의미에서 전통적인 주식 관련 데이터뿐 아니라 비정형의 대체 데이터 역시 자산 가격의 움직임을 설명할 수 있음을 보인다.



〈Figure 1〉 Process of the Study

본 연구의 수행 과정은 <Figure 1>과 같다. 우선 웹크롤링을 이용하여 텍스트 데이터를 확보한다. 이를 형태소 분리 등의 작업을 수행하는 전처리 과정을 거친다. 이 과정에서는 KoNLPy 알고리즘을 활용한다. 데이터 셋을 훈련 셋트와 검사 셋트로 분리하고 딥러닝 알고리즘을 적용해 실험을 수행하고 결과를 얻는다.

2. 선행 연구 논의

빅데이터는 양이 많고 수집되는 속도가 빠르며, 종류가 다양한 데이터이며 수집, 저장, 분석 과정을 통해 새로운 가치를 창출할 수 있는 프로세스와 기술이 매우 중요하다. 방대한 양의 데이터에서 가치를 창출하기 위해서 많은 빅데이터 분석 기법들이 사용되고 있으며, 대표적으로는 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 군집 분석, 감성 분석 등이 있다.

본 연구는 인간의 심리와 감성을 분석하는 감성분석(Sentiment Analysis)에 해당한다. 감성분석은 이미지 데이터 이외에도 텍스트 데이터, 영상 데이터 등을 활용하며, 주로 인간이 가질 수 있는 감성을 범주화 하여 모형화 하기에 기계학습의 분류 기법인 나이브베이즈(Naïve Bayes), 의사결정나무, KNN(K-Nearest Neighbors), SVM(Support Vector Machines)등의 기법들이 주로 활용된다. 텍스트 데이터를 이용하는 감성 분석에서는 최신 기계학습 기법인 인공신경망 알고리즘을 주로 사용하며, 문맥의 파악에 적합한 시계열 기반의 LSTM기법[2, 7, 17] 중첩 신경망을 활용하는 1D Convolution CNN 기법[6-7, 13-14] 등이 주로 활용된다. 인공신경망 기법을 분류문제에 활용하기 위해서 인공신경망의 마지막 활성함수를 분류를 위한 활성함수로 구성하면 된다. 본 논문에서는 1) LSTM, 2) 1D Convolution, 그리고 이들을 혼합한 3) LSTM+1D Convolution의 세 가지 방법으로 인공신경망을 구성하여 SNS에서 수집된 텍스트 데이터가 주가 변동에 대한 예측력을 가지는지 조사한다.

3. 데이터와 전처리

텍스트 마이닝 감성 분석을 통한 주가 예측에

관한 연구에는 크게 세 가지의 실험 설계요소가 있다. 첫 번째는 독립변수에 해당하는 텍스트 데이터의 소스이다. 많은 연구들이 트위터, 페이스북과 같은 소셜 미디어 데이터나 뉴스기사를 활용하였다. 가장 다수를 이루는 관련 연구는 정제된 정보인 뉴스기사가 주가의 흐름에 영향을 끼칠 것이라는 가설을 바탕으로 뉴스기사를 활용하였다[11, 12]. 본 연구에서는 직접적인 시장 참여자, 혹은 시장에 높은 관심을 가지고 있는 참여자들이 생산한 데이터를 활용하기에, 데이터의 투자자에 대한 대표성이 높다. 또한, 일반인들의 소셜미디어 활동의 기록이기 때문에 상대적으로 비속어나 은어들이 많이 포함된 구어체로 되어있다. 또한 뉴스와 같은 공식적인 소스에 비해서 데이터의 양이 매우 크다.

두 번째는 종속 변수의 선택이다. 대부분의 연구는 일간 예측력 (당일 대비 익일 종가의 등락을 예측하는 것을 목표로 함)을 조사하였으나, 본 연구에서는 거대한 양의 텍스트 데이터를 바탕으로 초단기 기간에 해당하는 1시간 단위에서의 예측력을 조사한다. 본 연구에서는 한 시간 뒤의 주가지수의 등락 뿐 아니라 주가의 변동성 지표에 해당하는 변동폭(Range)의 수준에 대한 예측도 포함하였기에, 본 연구의 결과는 비단 개인 주식 투자자를 포함한 변동성과 연관된 파생상품 거래자들도 활용할 수 있다.

세 번째 실험 설계 요소는 텍스트 데이터를 어느 수준까지 수치화 한 이후에 주식의 가격 변동과 연관을 시킬 지에 대한 결정이다. 텍스트 데이터를 잘 알려진 범용 감성분석 사전을 이용하는 방식[11]은 텍스트 데이터의 긍정/부정의 정도를 먼저 수치화 한 이후에 이들 수치와 시장의 움직임의 연관관계를 찾는다. 이는 텍스트 데이터가 주가에 관한 논의라는 점을 배재하고 범용적인 처리를 행하기에, 때로는 비속어를 섞어 정제되지 않은 형태의 댓글을 활용하는 본 연구의 데이터 셋에는 적합하지 않을 수 있다. 따라서 본 연구에서는 텍스트 데이터에서는 형태소 분리 작업만 한 이후에 이를 직접 미래의 주가 변동과 연관시키는 방식으로 실험을 설계하였다.

3.1 표본 기업의 선정

KOSPI 200 지수의 구성 종목 중에서 총 20개 종목을 선정하였으며, 이들 기업의 선정기준은 다음과 같다. 1) 데이터 수집 기간(7개월)에 관련 게시글과 댓글수가 6만개 이상이다. 댓글 수 6만개는 코스피 200 종목 가운데 상위 30%에 해당한다. 2) 일 평균 거래량은 14만 이상이다. 14만주의 거래량은 코스피 200 종목 가운데 상위 30%에 해당한다. 3) FnGuide가 제공하는 산업 분류 기준(FICS, FnGuide Industry Classification Standard)에 따라 편향되지 않도록 각 산업군당 3~4개의 종목을 무작위로 선정한다. FnGuide의 FICS는 주식을 10개의 섹터(Sector)와 25개의 산업군(Industry Group)으로 분류한다. 각 주식에 대해서 7개월(2019년 11월~2020년 5월 혹은 2020년 1월~2020년 7월)에 해당하는 데이터를 수집하였으며, 사용된 주식 종목은 다음 <Table 1>과 같다. 시가총액의 순위는 2020년 1월을 기준으로 하였으며, 데이터의 개수는 해당 기간의 게시글에 대한 댓글 수의 총 합이다.

〈Table 1〉 List of Stocks

Mkt.Cap. Rank	Security Name(Code)	Sector	Industry Group	Period YY.MM-YY.MM	Num. of Obs.
2	SK Hynix(000660)	IT	Semiconductor	19.11-20.05	23,453
4	Naver(035420)	IT	Software	20.01-20.07	7,112
5	Samsung Biologics(207940)	Medical	Pharmaceutical and Bio	19.11-20.05	20,315
6	Hyundai Motors(005380)	Consumer Discretionary	Car and Parts	19.11-20.05	29,284
16	Korea Electric Power Corporation(015760)	Utility	Utility	19.11-20.05	25,363
23	Kakao(035720)	IT	Software	19.11-20.05	24,323
26	AmorePacific(090430)	Essential Consumer Goods	Household Goods	20.01-20.07	6,151
27	NC Soft(036570)	IT	Software	19.11-20.05	23,453
32	Samsung Electro-Mechanics (009150)	IT	Hardware	19.11-20.05	12,303
36	Netmarble(251270	IT	Software	19.11-20.05	11,135
44	LG Display(034220)	IT	Display	19.11-20.05	12,775
50	Hyundai Engineering & Construction(000720)	Industrial Goods	Capital Goods	20.01-20.07	4,807
51	Samsung Heavy Industries (010140)	Industrial Goods	Capital Goods	20.01-20.07	31,525
61	Hotel Shilla(008770)	Consumer Discretionary	Logistics	20.01-20.07	10,188
72	Korea Aerospace Industries (047810)	Industrial Goods	Commercial Service	19.11-20.05	10,395
80	Daewoo Shipbuilding & Marine Engineering(042660)	Industrial Goods	Capital Goods	20.01-20.07	3,818
87	Korean Air(003490)	Industrial Goods	Transportation	20.01-20.07	13,796
94	Hanjin Kal(180640)	Industrial Goods	Transportation	20.01-20.07	99,960
152	Doosan Infracore(042670)	Industrial Goods	Capital Goods	20.01-20.07	38,007
176	Bukwang Pharmaceutical (003000)	Medical	Pharmaceutical and Bio	20.01-20.07	72,850

3.2 텍스트 데이터 수집 및 전처리

텍스트는 파이썬 3.0의 Selenium라이브러리를 활용하여 크롤링 방식으로 수집하였다. 수집된 원본 데이터는 필터링 과정과 형태소 분석 과정의 두 단계를 거쳐 전처리 한다. 첫 번째 필터링 과정에서 광고글과 불용어와 같은 불필요한 글들을 제거하였다. 광고글들은 중복되는 성격을 띄기에 본 연구에 있어서 무의미하며, 불용어는 특별한 의미를 가지지 않는 조사(은", "는", "이", "가", "를" 등)나 숫자를 의미한다. 불용어 처리는 Ranks NL에서 제공하는 한국어 불용어 사전을 이용했다.

두 번째 과정인 형태소 분석 과정은 문장을 언어 단위인 어절 또는 형태소로 나누는 과정이다. 대표적인 형태소 분석기 Okt, Kkma,

오늘 털리면 안되요 오늘이 조정 막날입니다 ㅎㅎ 지금 갬들 많이들 털리시네요 ㅋㅋ
더럽게 못올리는듯요 ㅋㅋ 먼가 계속 누가누를려고 하는듯
■ 반드시 14000붕괴될듯 ㅎ 그때 매수 ㅎㅡLGD임을 잊지말것
몰빵 평단 15460원 만들었다
오늘 종가 14700원 예상 ㅎㅎ 갬들 많이 털리네요
4000주 장난질 폭등임박 개미들 뺏어먹으려고 어지간히 수고하네요
삭제된 게시물의 답글14550원에 매도후 장마감 매수대기중이니힘 좀 써 다오
12000만와라 그냥 내가 몰빵해버릴테니 ㅋㅋ

〈Figure 2〉 Example of Raw Data

['오늘', '털리다', '안되다', '오늘이', '조정', '막날', '이다', 'ㅎㅎ', '지금', '개다', '들다', '많이', '들다', '털리다', 'ㅋㅋ']
['더럽다', '못', '올리다', '듯', '요', 'ㅋㅋ', '멀다', '계속', '누가', '누르다', '하다']
['■', '14000', '붕괴', '되다', 'ㅎ', '매수', 'ㅎㅡ', 'LGD', '임', '잊다']
['몰다', '빵', '평단', '15460원', '만들다']
['오늘', '종가', '14700원', '예상', 'ㅎㅎ', '개다', '들다', '많이', '털리다']
['4000', '주', '장난', '질', '폭등', '임박', '개미', '뺏다', '먹다', '어지간하다', '수고', '하다']
['삭제', '되다', '게시', '물의', '답글', '14550원', '매도', '후', '장마', '감', '수', '대기', '중', '이니', '힘', '써다', '다오']
['12000만', '와라', '그냥', '내', '몰다', '빵해', '버리다', 'ㅋㅋ']

〈Figure 3〉 Separated Morphemes by Okt Analyzer

Komoran를 테스트하였으며, 처리해야할 데이터 양에 비해 속도가 느린 Kkma를 우선 제외하고, Komoran과 Okt 중에서 증권 관련 용어의 분석에 더 효과적으로 나타난 Okt를 사용하였다. 아래 <Figure 2>는 전처리 전의 raw 데이터이며 <Figure 3>은 Okt를 사용해 형태소를 나눈 결과이다.

분석 대상인 각 댓글을 형태소 단위로 나누면, 각 댓글에는 각각 다른 개수의 형태소가 포함되어 있다. 90% 이상의 댓글에서 70개 이하의 형태소를 가지고 있기에, 모든 댓글에 대해서 형태소의 길이를 70로 설정하였다. 즉, 70보다 길이가 긴 경우에는 70번째 이후의 형태소를 잘라내었고, 70보다 길이가 작은 경우에는 Zero Padding을 실시하였다. 각 댓글은 아래의 <Table 2>와 같은 데이터 구조를 가진다.

3.3 주가 데이터 수집

주가 데이터는 체크 단말기를 이용하여 수집되었으며, 거래일의 각 시간대(9, 10, 11, 12, 13, 14, 15시 정각)에 대해서 종가와 거래량을 수집하였다.

〈Table 2〉 Example of a Preprocessed Comment

Variables(unit)	Explanation
Datetime(YYYY-MM-DD HH:MM)	2020-03-12 13:17
Security name(security code)	Korean Air (003490)
Raw text(sentence)	“I ate the lunch”
Morpheme separated(list of words)	[I, eat, lunch]
Tokenized(a list)	[12, 153, 35]
Zero-padded(a list of length 70)	[12, 153, 35, 0, 0, …, 0]

4. 모델과 방법론

4.1 독립 변수와 종속 변수

독립 변수 X_t는 시각 t로부터 1시간 후인 시각 t+1사이에 수집된 댓글들의 집합이다. 주식의 t시점의 가격을 P_t라고 쓰며, 따라서 본 연구의 관심은 X_t와 P_s, $s \in (t+1, t+2]$의 연관관계이다. P_s, $s \in (t+1, t+2]$를 이용하여 아래와 같이 두 개의 종속 변수를 생성하였다.

첫 번째 종속 변수인 r_{t+1}은 시각 t+1부터 시각 t+2까지의 주가 변동의 방향을 의미한다. $P_{t+2} \geq P_{t+1}$인 경우에는 r_{t+1}가 1의 값을 가지고, 그렇지 않으면 0의 값을 가진다. 함수 $I_{[0]}$는 대괄호 안의 명제가 참이면1, 거짓이면 0의 값을 반환하는 binary indicator 함수이다.

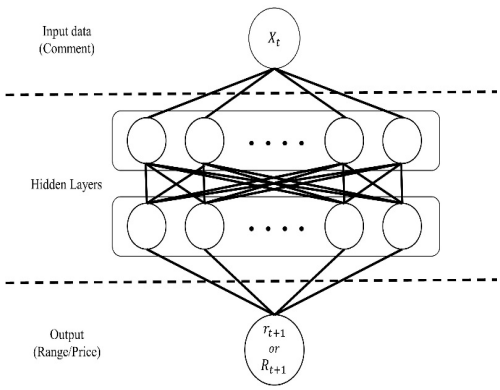
$$r_{t+1} = I_{[P_{t+2} \geq P_{t+1}]}$$

두 번째 종속 변수인 R_{t+1}은 시각 t+1부터 시각 t+2까지의 주가 변동의 최대폭(Range, 최고 주가와 최저 주가의 차이)이 상대적으로 컸는지 작았는지를 의미한다. 각 주식에 대해서 실험 기간에 대해서 매시간당 주가 변동의 최대폭(Range)을 수집하고 이중에 중간값을 찾아 δ라고 한다. 따라서 특정 1시간 동안의 주가 변동폭이 δ보다 클 확률이 50%, 작을 확률이 50%인 Status Quo가 성립된다. 즉, R_{t+1}의 값이 1이면 시각 t+ 1부터 시각 t+2까지의 주가 변동의 폭이 상대적으로 큰 것이고, 그렇지 않으면 0의 값을 가진다.

$$R_{t+1} = I_{\left[\max_{s \in (t+1,\ t+2]} P_s - \min_{s \in (t+1,\ t+2]} P_s \geq \delta\right]}$$

4.2 예측을 위한 딥뉴럴 네트워크

예측을 위한 딥뉴럴넷의 Input Layer에는 독립 변수인 1시간 분량의 댓글의 집합(X_t)이 들어가며, 최종의 Output Layer에는 종속 변수인 차후 1시간 동안의 주가 이동 방향(r_{t+1}), 혹은 주가 변동폭의 크기가 큰 지에 대한 Binary 변수(R_{t+1})가 들어간다. 이를 다이어그램으로 표현하면, 아래의 <Figure 4>와 같다.



〈Figure 4〉 Skeleton of Neural Network

Input Layer와 Output Layer의 사이에 위치하는 Hidden Layer를 1) LSTM, 2) 1D Convolution, 그리고 이들을 혼합한 3) LSTM+1D Convolution의 세 가지 방법으로 구성하여 SNS의 텍스트 데이터가 주가 변동에 대한 예측력을 가지는지 조사한다. 3가지 방식의 Hidden Layer 구성과 각 구성 내에서의 Neuron의 수 등의 하이퍼 파라미터를 조정하며, 예측을 위한 최적의 네트워크를 찾는다. 금융 시장 데이터를 이용한 정확한 예측은 실험적으로 매우 어렵다는 것이 잘 알려져 있다. 따라서 예측의 정확도가 통계적으로 유의미하게 50%를 넘을 수 있는지가 본 연구의 관심이다.

5. 결 과

5.1절에서는 본 연구의 가장 핵심적인 결과가 되는 예측 정확도에 대해 설명할 것이다. 5.2절에서는 최적화된 하이퍼 파라미터 조합을 설명할 것이다.

5.1 예측의 정확도

선정된 20개의 종목에 대하여 각각 학습을 진행하였다. 한 종목에 대한 시계열 데이터에 대해서 앞의 80%의 기간을 훈련 데이터 셋으로 하고, 뒤의 20%의 기간을 테스트 셋으로 하였다. 훈련 과정에서 훈련 데이터 셋의 후반 20%의 기간의 데이터는 검증 데이터셋으로 활용하였다. 테스트 셋에서의 성능을 요약하면 아래의 <Table 3>과 같다.

미래의 주가 이동 방향 (r_{t+1})에 대해서는 다수의 종목들에 대해서 통계적으로 유의미한 성능을 보였다. 20개의 종목 중에서 13개의 종목에 대해서 90% 이상의 신뢰수준에서 50% 이상의 예측 정확도를 보였으며, 12개의 종목에 대해서는 95%의 신뢰도, 9개의 종목에 대해서는 99% 이상의 신뢰도를 가지고 50% 이상의 정확도로 주가 이동 방향을 예측할 수 있다는 결론이 도출되었다.

〈Table 3〉 Summary Statistics for Prediction Accuracy

Security Name	Number of observations		Future price(r_{t+1})		Future range(R_{t+1})	
	Train Set	Test Set	Acc. (%)	Pr (Acc. > 50%)	Acc. (%)	Pr (Acc. > 50%)
SK Hynix	8,385	2,092	49.90	0.536	51.39	0.102
Naver	4,715	1,168	57.28	< 1e-4***	77.31	< 1e-4***
Samsung Biologics	8,570	2,134	54.73	< 1e-4***	65.23	< 1e-4***
Hyundai Motors	12,675	3,145	52.50	0.003***	64.48	< 1e-4***
Korea Electric Power Corporation	10,470	2,606	50.61	0.264	76.02	< 1e-4***
Kakao	5,064	1,258	61.05	< 1e-4***	76.63	< 1e-4***
AmorePacific	2,905	719	55.35	0.002***	59.11	< 1e-4***
NC Soft	8,385	2,092	51.58	0.075*	63.00	< 1e-4***
Samsung Electro-Mechanics	5,158	1,283	60.41	< 1e-4***	52.46	0.039**
Netmarble	5,316	1,323	48.37	0.881	57.90	< 1e-4***
LG Display	5,177	1,292	48.92	0.782	49.69	0.587
Hyundai Engineering & Construction	2,340	579	51.81	0.192	62.18	< 1e-4***
Samsung Heavy Industries	16,544	4,119	51.69	0.015**	71.62	< 1e-4***
Hotel Shilla	4,209	1,048	47.42	0.952	44.94	0.999
Korea Aerospace Industries	4,092	1,017	52.61	0.049**	54.87	< 1e-4***
Daewoo Shipbuilding & Marine Engineering	1,876	462	42.64	0.999	49.13	0.644
Korean Air	5,787	1,441	54.06	0.001***	56.63	< 1e-4***
Hanjin Kal	49,941	12,447	52.53	< 1e-4***	60.83	< 1e-4***
Doosan Infracore	18,222	4,530	51.66	0.013**	65.14	< 1e-4***
Bukwang Pharmaceutical	38,159	9,490	53.47	< 1e-4***	75.82	< 1e-4***
Average	10,900	2,712	52.43	0.0057**	61.72	< 1e-4***

Asterisk marks imply statistical significance. (* implies >90%, ** implies >95%, *** implies >99%).

한편, 미래의 주가 변동폭의 크기(R_{t+1})에 대해서는 주가 이동 방향(r_{t+1})에 비해서 더욱 높은 성능을 관찰할 수 있었다. 20개의 종목 중 16개의 종목에서 95% 신뢰수준에서의 50% 이상의 정확도를 확인하였고, 15개의 종목에서는 99% 신뢰수준을 상회하는 정확도를 확인하였다.

일반적으로 알려진 바와 같이 주가의 방향 지표에 비해서 변동 지표의 예측력이 높았으며, 주가의 이동방향에 대해서 통계적으로 유의미한 예측력을 보였던 13개의 종목들은 변동 지표의 예측력 역시 통계적으로 유의미한 점을 확인할 수 있었다. 즉, 이들 종목의 경우에는 댓글이 주가의 이동방향과 변동성에 대한 예측력을 포함하고 있음을 의미한다.

5.2 최적화된 예측 네트워크

딥뉴럴네트워크를 구성하기 위해서 선택할 수 있는 많은 하이퍼 파라미터 중에서 아래 <Table 4>의 구성 요소에 대해서는 잘 알려진 하이퍼 파라미터의 값으로 고정하였다.

〈Table 4〉 Pre-Chosen Hyperparameters

Hyperparameter	Value
dropout	0.2
optimizer	Rmsprop
epochs	50 with early stopping
Batch size	128
activation function	sigmoid

네트워크의 3가지 구조와 각 구조의 Neuron의 수에 대해서는 하이퍼 파라미터 최적화 과정을 수행하였으며, 테스트한 조합과 최적의 파라미터는 아래의 <Table 5>와 같다. 즉, 종속변수를 미래의 가격의 이동방향으로 하였을 때에도 단층 LSTM 레이어를 사용하였을 때에 가장 정확도가 높았으며, 최적 Neuron의 개수는 64개이다. 종속변수를 미래 변동폭으로 하였을 때에는1D Convolution 레이어에서 첫 번째 레이어는 128개의 뉴런, 두 번째 레이어는 32개의 뉴런으로 구성하였을 때에 가장 정확도가 높았다.

〈Table 5〉 Optimal Hyperparameters

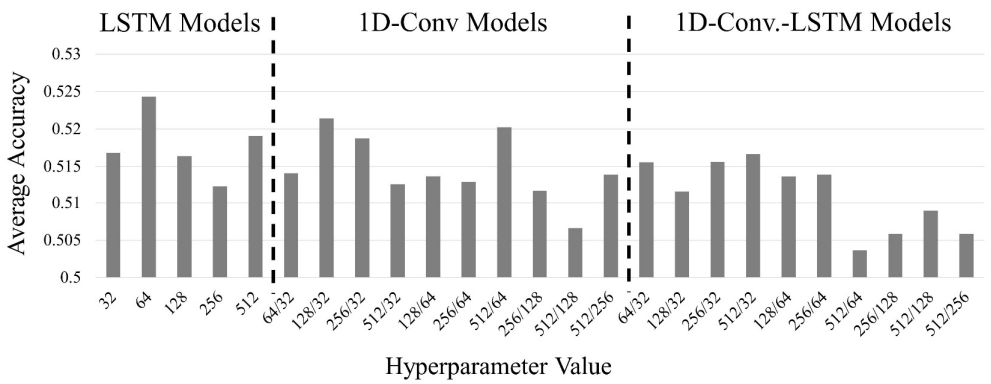
Network for Price(r_{t+1})					Network for Range(R_{t+1})				
[LSTM]					[LSTM]				
32	**64**	128	256	512	32	64	128	256	512
[1D-Conv.]					[1D-Conv.]				
1st layer:					1st layer:				
64	128	256	512		64	**128**	256	512	
2nd layer:					2nd layer:				
32	64	128	256		**32**	64	128	256	
[1D-Conv.-LSTM]					[1D-Conv.-LSTM]				
1st layer:					1st layer:				
64	128	256	512		64	128	256	512	
2nd layer:					2nd layer:				
32	64	128	256		32	64	128	256	

*Selected values are in underlined and bold-faced.

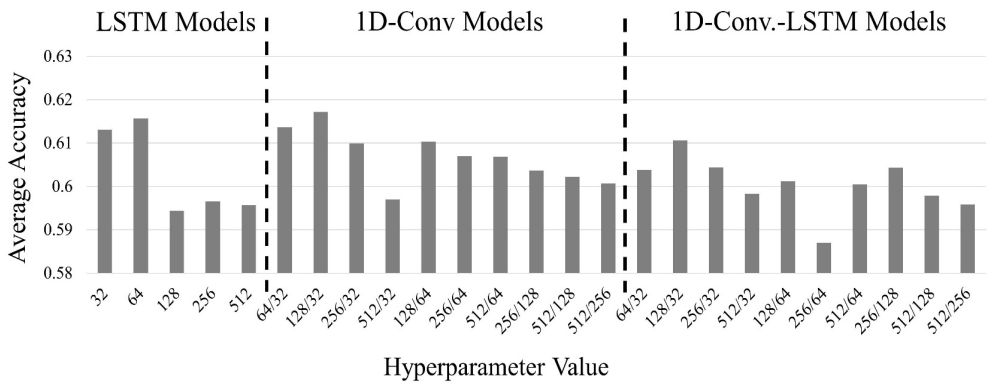
최적의 파라미터 조합을 사용하지 않은 경우에도 대체로 평균 정확도가 0.5를 넘는 것으로 나타났다. 아래 <Figure 5>는 주가 이동 방향에 대한 예측 네트워크에서 각각의 하이퍼 파라미터 별 평균 정확도를 보여준다. 1D-Conv. 모델과 1D-Conv.-LSTM모델에 비해 LSTM 모델의 예측력이 전반적으로 우수한 것으로 나타났다. <Figure 6>은 주가 변동폭에 대한 예측 네트워크에서 각 하이퍼 파라미터 별 평균 정확도를 보여준다. LSTM모델과 1D-Conv.-LSTM모델에 비해 1D-Conv.모델의 예측력이 전반적으로 우수한 것으로 나타났다. 특히 1D-Conv.모델은 하이퍼 파라미터의 변경에도 일관성 있는 예측력을 볼 수 있다.

6. 논의 및 결론

본 연구에서는 우리나라의 주식 투자자들이 가장 많이 참여하는 온라인 커뮤니티인 "네이버 주식토론방"에서 수집된 댓글을 활용하여 고빈도의 시간 축(High-Frequency Time Horizon)에서 주가의 이동 방향과 변동폭을 예측할 수 있는지를 조사하였다. 한글 텍스트 데이터를



〈Figure 5〉 Performance of Other Hyperparameter Values for Price Network(r_{t+1})



〈Figure 6〉 Performance of Other Hyperparameter Values for Range Network(R_{t+1})

시작으로 수치화 과정을 마친 후, 딥뉴럴네트워크 기법인 LSTM과 CNN을 활용하여 예측 네트워크를 구성하였다. 대부분의 주식에 대해서 이동 방향과 변동폭에 대한 예측이 50% 이상의 정확도로 가능함을 확인할 수 있었고, 변동폭에 대한 예측이 이동 방향에 대한 예측에 비해서 더 성능이 좋다는 점도 확인할 수 있었다.

수집된 데이터의 기간인 2019년 11월에서 2020년 7월까지의 기간이 국내 주식시장의 역사상 가장 많은 개인투자 자금이 유입된 시기인 강세장에 해당하는 점이 일종의 편향을 불러왔을 가능성이 있으나, 각 종목별로 6개월의 기간동안 1시간 단위로 추출하여 많은 샘플 수를 확보하였기 때문에, 해당 데이터 기간 중에서의 댓글의 정보와 가격 변동의 연관성은 통계적으로 입증이 된다. 향후 다양한 기간과 다양한 움직임을 가진 주식들을 선정해 연구가 이루어질 경우에 더욱 일반화된 결론을 내릴 수 있을 것이다. 국내 주식시장은 대체로 시가총액 상위의 대형주는 외국인과 기관 투자자의 영향을 많이 받고, 시가총액 하위의 중소형주는 개인 투자자들의 영향을 많이 받는 것으로 알려져 있다. 하지만, 본 연구에서 조사된 기업들에 대해서는 기업의 규모에 따라서 댓글의 정보력이 큰 차이가 나는 것이 발견되지는 않았다.

본 연구가 관찰한 사실을 활용하여 투자 수익을 얻기 위해서는 다음과 같은 조건이 선결되어야 한다. 첫째, 주가 이동 방향에 대한 예측력은 50%를 넘지만, 이는 매도-매수 가격 스프레드를 상회할 수준의 예측력은 아니다. 따라서 주가 이동 방향에 대한 예측력을 바탕으로 고빈도 매매에 활용하려면 시장가 주문이 아닌 지정가 주문의 형태로 구현된 정교한 매매 전략이 필요하다. 매도-매수 가격 스프레드 외에 거래 수수료나 세금으로 부터의 영향이 적은 기관 투자자의 경우에는 보다 높은 가능성이 있다. 둘째, 본 연구는 과거의 댓글 데이터와 가격 데이터를 수집하여 이루어졌지만, 이를 거래에 활용하기 위해서는 latency가 매우 낮은 크롤링 및 딥러닝 네트워크 적용이 거의 실시간으로 프로세싱 되어야 한다. 셋째, 본 연구는 6개월의 기간에 걸쳐 수집된 약 50여만 개의 댓글을 활용했다. 하지만, 하지만, 금융데이터의 경우에 시계열 적으로 특성이 바뀔 수 있다는 점을 고려한다면, 장기간의 데이터를 활용한 선행 연구와 여러 길이의 Subperiod 내에서의 효과성에 대한 검증도 선행되어야 한다.

본 연구의 후속 연구로서는 아래와 같은 방향을 제시한다. 첫 째, 바로 앞의 단락에서 제시한 바와 같이 투자 전략의 구현으로서의 현실적인 이슈(거래비용을 상회할 정교한 매매 전략, on-line learning, 강건성 확보를 위한 장기간의 데이터 수집)들에 대한 연구가 가능하다. 둘 째, 관찰된 조사 결과가 대한민국과 같이 인터넷 및 SNS의 사용이 활성화된 나라에 국한된 것인지에 대한 지역적인 연구가 가능하다.

References

[1] Hochreiter, S. and Schmidhuber, J., “Long short-term memory,” Neural computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

[2] Hong, S. H., “A study on stock price prediction system based on text mining method using LSTM and stock market news,”

Journal of Digital Convergence, Vol. 18, No. 7, pp. 223-228, 2020.

[3] Jeong, J. S., Kim, D. S., and Kim, J. W., "Influence analysis of Internet buzz to corporate performance: Individual stock price prediction using sentiment analysis of online news," Korea intelligent information Systems Society, Vol. 21, No. 4, pp. 37-51, 2015.

[4] Kang, Y. J. and Jang, W. W., "The Five-Factor Asset Pricing Model: Applications to the Korean Stock Market," Eurasian Studies, Vol. 13, No. 2, pp. 155-180, 2016.

[5] Kim, D. H., "Asset Pricing Model in Korean Stock Market," Association of financial engineering, Vol. 13, No. 2, pp. 87-119, 2014.

[6] Kim, D. S., Kim, K. T., and Kim, J. W., "Character-based multi-category sentiment analysis on social media using deep learning algorithms," Korean Institute Of Industrial Engineers, Vol. 2017, No. 4, pp. 5082-5084, 2017.

[7] Kim, D. Y. and Lee, Y. I., "News based Stock Market Sentiment Lexicon Acquisition Using Word2Vec," The Korea Journal of BigData, Vol. 3, No. 1, pp. 13-20, 2018.

[8] Kim, D. Y., Park, J. W., and Choi, J. H., "A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles," Journal of Information Technology Services, Vol. 13, No. 3, pp. 221-233, 2014

[9] Kim, H, G., Kim, S. D., and Kim, H. W., "A Case Study on the Establishment of an Equity Investment Optimization Model based on FinTech: For Institutional Investors," Korea Knowledge Management Society, Vo. 19, No.1, pp. 97-118, 2018.

[10] Kim, J. Y. and Kim, C. S., "An Analysis on Mediating Effect of Participant Activity in Investment Crowdfunding," The Journal of Society for e-Business Studies, Vol. 25, No. 1, pp. 65-82, 2020.

[11] Kim, Y. S., Kim, N. G., and Jeong, S. R., "Stock-Index Invest Model Using News Big Data Opinion Mining," Journal of Intelligence and Information Systems, Vol. 18, No. 2, pp. 143-156, 2012.

[12] Lee, H. J., "Analysis of News Big Data for Deriving Social Issues in Korea," The Journal of Society for e-Business Studies, Vol. 24, No. 3, pp. 163-182, 2019.

[13] Lee, M. S. and Ahn, H. C., "A Time Series Graph based Convolutional Neural Network Model for Effective Input Variable Pattern Learning: Application to the Prediction of Stock Market," Korea intelligent information Systems Society, Vol. 24, No. 1, pp. 167-181, 2018.

[14] Park, H. J., Song, M. C., and Sim, K. S., "Sentiment Analysis of Korean Reviews Using CNN-Focusing on Morpheme Embedding," Korea intelligent information Systems Society, Vol. 24, No. 2, pp. 59-83, 2018.

[15] Seo, I. S., Yeo, S. S., and Kang, H. J., “A Study on the Suggestion of Domestic Stock Market Analysis Scheme using Big Data,” Korean Institute of information technology, Vol. 2014, No. 5, pp. 550-554, 2014.

[16] Son, S. H., Kim, T. H., and Yoon, B. H., “Testing the Linear Asset Pricing Models in the Korean Stock Market,” Korean Journal of Financial Studies, Vol. 38, No. 4, pp. 547-568, 2009.

[17] Song, S. H., Kim, J. H., Kim, H. S., Park, J. S., and Kang, P. S., “Development of Early Warning Model for Financial Firms Using Financial and Text Data: A Case Study on Insolvent Bank Prediction,” Journal of the Korean Institute of Industrial Engineers, Vol. 45, No. 3, pp. 248-259, 2019.

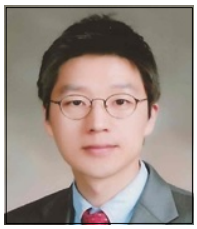
[18] Suh, M. S. and Kim, D. H., “A Study on the Changing Direction of FinTech Service Model based on Big Data,” The e-business studies, Vol. 20, No. 2, pp. 195-213, 2019.

[19] Yoo, H. S., “What are the core competitiveness and alternative data in the digital age?,” Available at: https://2e.co.kr/news/articleView.html?idxno=209967, 2019.

저 자 소 개

김명진 (E-mail: kevin_960819@seoultech.ac.kr)
2015년～현재 서울과학기술대학교 산업공학과 ITM전공 (학사과정)
관심분야 빅데이터, 금융공학, 텍스트 마이닝, AI

류지혜 (E-mail: jhryu48@seoultech.ac.kr)
2016년～현재 서울과학기술대학교 산업공학과 ITM전공 (학사과정)
관심분야 빅데이터, 경영, 텍스트 마이닝



차동호 (E-mail: movieish@gmail.com)
2003년 한양대학교 경영학부 (학사)
2010년 KAIST 경영대학원 금융MBA (석사)
2011년～2015년 유리자산운용, 퀀트 펀드운용
2016년～현재 KB자산운용, ETF/EMP펀드 운용
관심분야 금융공학, 자산배분, ETF, 고빈도 매매

심민규 (E-mail: mksim@seoultech.ac.kr)
2007년 서울대학교 산업공학과 (학사)
2008년 University of Chicago 수학과 금융수학전공 (석사)
2014년 Georgia Institute of Technology 산업공학과 (박사)
2015년～2017년 삼성자산운용, KB자산운용 계량연구 및 펀드운용
2018년～2019년 경희대학교 산업경영공학부 연구교수
2019년～현재 서울과학기술대학교 산업공학과 조교수
관심분야 금융공학, 강화학습, 확률모형, 스마트 그리드, 스마트 시티