Journal of Korea Multimedia Society Vol. 24, No. 3, March 2021(pp. 448-459)
https://doi.org/10.9717/kmms.2020.24.3.448

온라인 뉴스와 거시경제 지표, 금융 지표, 기술적 지표, 관심도 지표를 이용한 코스닥 상장 기업의 기계학습 기반 주가 변동 예측

김화련[†], 홍승혜[††], 홍헬렌[†††]

Machine Learning Based Stock Price Fluctuation Prediction Models of KOSDAQ-listed Companies Using Online News, Macroeconomic Indicators, Financial Market Indicators, Technical Indicators, and Social Interest Indicators

Hwa Ryun Kim[†], Seung Hye Hong[††], Helen Hong[†††]

ABSTRACT

In this paper, we propose a method of predicting the next-day stock price fluctuations of 10 KOSDAQ-listed companies in 5G, autonomous driving, and electricity sectors by training SVM, XGBoost, and LightGBM models from macroeconomic·financial market indicators, technical indicators, social interest indicators, and daily positive indices extracted from online news. In the three experiments to find out the usefulness of social interest indicators and daily positive indices, the average accuracy improved when each indicator and index was added to the models. In addition, when feature selection was performed to analyze the superiority of the extracted features, the average importance ranking of the social interest indicator and daily positive index was 5.45 and 1.08, respectively, it showed higher importance than the macroeconomic financial market indicators and technical indicators. With the results of these experiments, we confirmed the effectiveness of the social interest indicators as alternative data and the daily positive index for predicting stock price fluctuation.

Key words: Stock Prediction, Google Trends, Opinion Mining, SVM, XGBoost, LightGBM

1. 서 론

2016년 세계 경제 포럼에서 4차 산업 혁명[1]이라는 용어가 주창된 이후 빅데이터 분석, 인공지능, 로봇공학, 사물인터넷, 무인 운송 수단, 3D 프린팅, 나노 기술 등 다양한 분야에서 기술 혁신이 진행 중이

※ Corresponding Author : Helen Hong, Address: (01797) Dept. of Software Convergence, Seoul Women's University, 621 Hwarang-ro, Nowon-gu, Seoul, Korea, TEL : +82-2-970-5756, FAX : +82-2-978-5981, E-mail : hlhong@swu.ac.kr
Receipt date : Dec. 9, 2020, Revision date : Mar. 5, 2021
Approval date : Mar. 9, 2021
[†] Dept. of Economics, Seoul Women's University (E-mail : ekfrl0124@swu.ac.kr)
[††] Dept. of Economics, Seoul Women's University (E-mail : mocha@swu.ac.kr)
[†††] Dept. of Software Convergence, Seoul Women's University
※ Haw-Ryun Kim and Seung-Hye Hong are co-first authors and contributed equally to this work.
※ This research was partly supported by the MISP (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW(2016-0-00022) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) and by a research grant from Seoul Women's University (2021-0057).

다. 또한 2017년 1월부터 4차 산업 관련 벤처 기업에 대한 코스닥(KOSDAQ) 상장 조건이 '테슬라 요건'[2]을 통해 간소화되었고, 이를 통해 중소· 벤처 기업들의 코스닥 상장과 투자가 더욱 주목을 받을 것으로 전망된다. 따라서, 4차 산업을 영위하는 중소기업의 주가 예측 관련 연구 필요성이 높아졌다.

주가 변동 예측 관련 연구는 정형 데이터를 사용한 연구와 비정형 데이터를 사용한 연구, 정형과 비정형 데이터를 함께 사용한 연구로 나눌 수 있다. 정형 데이터(structured data)는 미리 정해 놓은 형식과 구조에 따라 저장되도록 구성된 데이터이며, 정해진 형식과 저장 구조를 바탕으로 손쉽게 데이터에 대한 부분 검색 및 선택, 갱신, 삭제 등의 연산을 수행할 수 있다[3]. 비정형 데이터(unstructured data)는 정의된 구조 없이 정형화되지 않은 데이터이며 데이터 구조가 없어 비정형 데이터 자체만으로는 내용에 대한 질의 처리(query processing)를 할 수 없으므로 데이터 특징을 추출하여 반정형 데이터(semi-structured data) 또는 정형 데이터로 변환하는 전처리가 필요하다[3].

정형 데이터를 사용한 연구 중 [4]는 주식 가격 데이터 중 1분당 체결가를 이용하여 KOSPI 상장사 12개 기업의 주식 가격 동향을 HTM 모델을 통해 설정된 구간 간격에 따라 연속적으로 예측하였다. [5]는 주가 기본 데이터인 시가, 저가, 종가, 고가, 거래량을 이용하여 Google의 주가를 양방향 LSTM(Long Short-Term Memory) 모델을 사용하여 예측하였다. [6]은 기술적 지표 13가지(RSI, CCI, EMA, OBV, MACD, MINUS_DM, PLUS_DI, MINUS_DI, Wcl_p, Momentum, CMO, WILL_R, SSE)를 이용하여 KOSPI 지수의 등락을 인공 신경망을 통해 예측하였다. [7]은 기술적 지표 7가지(SMA, EMA, Stochastic K%, Stochastic D%, RSI, MACD, Disparity)를 이용하여 KOSPI200 주가지수 등락을 XGBoost 모델을 활용하여 예측하였다.

비정형 데이터를 이용한 연구 중 [8]은 온라인 뉴스와 소셜 네트워크 데이터로부터 기업별 감성 사전을 구축한 후 이를 바탕으로 감성 지표를 추출하여 KOSPI 상장 7개사의 주가 변동을 SVM(Support Vector Machine), 로지스틱 회귀분석(logistic regression), 베이지안 네트워크(bayesian network), 인공신경망(ANN, artificial neural network) 총 4가지 방법을 이용하여 예측하였다. [9]는 산업의 대표적 성격을 가지고 외부 요인에 의한 산업 주가 변화 특징을 대변하는 산업 대표 기업을 KOSPI 구성 종목 중 시가총액 기준 상위 10%를 기준으로 선정하고 이러한 기업과 예측하고자 하는 개별 기업의 온라인 뉴스를 이용하여 산업 대표 기업 및 개별 기업 감성 사전을 텍스트 마이닝을 통해 구축하였다. 이를 이용하여 제약, 소재, 자본재, 자동차, 내구소비재, 음식료, 하드웨어 업종 중 중소형 기업의 주가 변동을 SVM, MKL(Math Kernal Library) 모델을 이용하여 예측하였다. [10]은 증시 뉴스 데이터를 수집한 후 Word2Vec을 활용하여 감성 사전을 구축하고, 이를 바탕으로 뉴스에 감성 분석을 실시하여 KOSPI 종가 지수를 예측하였다. [11]은 무작위 행렬 이론으로 예측 대상인 기업과 관련된 온라인 뉴스를 사용하여 코스피 기업 중 소재 섹터에 해당하는 기업의 주가 등락을 MKL 모델을 통해 예측하였다.

정형 데이터와 비정형 데이터를 통합하여 사용한 연구 중 [12]는 트위터와 온라인 뉴스에 자연어 처리 기술 중 스팸 필터링 기술과 텍스트 마이닝 기술, 그리고 감성분석 기술 등을 이용하여 소셜 감성 데이터를 추출한 후 이를 각 기업의 재무비율(fundamental ratio)과 기술적 지표를 함께 SVM의 입력 변수로 하여 KOSPI200 주가지수 등락을 예측하였다. [13]은 GRU를 통해 계산된 온라인 뉴스의 감성 지표와 9개 기술적 지표(전일 종가 대비 당일 시가 변화율, DMI, Williams %R, RSI, Momentum, CCI, Bollenger Bands, MACD, Stochastic Slow K%)를 XGBoost 모델에 학습시켜 북한 테마에 속하는 KOSDAQ, KOSPI 상장 10개 기업의 익일 주가 등락을 예측하였다. [14]는 예측 대상 기업의 5개 기술적 지표(EMA, MACD, signal, Bollenger Bands, RSI), 업종 더미변수, 소셜 데이터에서 추출된 소셜 네트워크 지표를 이용하여 기업의 일주일 간 10% 이상 급등 여부를 랜덤포레스트(random forest), XGBoost, 그래디언트 부스팅(gradient boosting) 모델을 이용하여 예측하였다.

기존의 주가 또는 주가지수 예측 관련 연구는 주로 KOSPI200 주가지수 혹은 KOSPI에 상장된 기업을 연구 대상으로 함으로써 KOSDAQ에 상장된 중소형 기업의 주가 예측 관련 연구는 상대적으로 부족한 편이다. 또한, 온라인 뉴스를 이용하여 주가 예측

연구를 진행한 경우 기업의 주가가 기업 내외부 뉴스에 영향을 받음에도 불구하고 한 가지 종류의 뉴스를 고려한 연구가 대부분이다. 그리고 정형 데이터 중 주로 전통적인 주가 데이터, 기술적 지표를 이용한 연구를 진행하여 최근 증권, 자산 운용계에서 관심을 가지는 가격, 재무 데이터 및 경제지표 등의 전통적 데이터 범주에 속하지 않는 대체 데이터를 이용한 연구는 부족하다. 따라서, 본 연구에서는 KOSDAQ에 상장된 4차 산업 기업들 중 5G, 자율주행, 전기차 섹터에 해당하는 10개사의 익일 주가의 방향을 기술적 지표, 거시경제 지표, 개별 기업 뉴스에서 계산된 긍정 지수에 더해 관심도 지표로서 주식 커뮤니티 조회수와 검색지수, 섹터 뉴스에서 계산된 긍정 지수를 신규 고려하고 SVM, XGBoost, LightGBM 모델의 입력 변수로 하여 예측하고 투입 특징 별 성능을 비교한다. 또한 특징 선별(feature selection) 기법 중 RFE(Recursive Feature Elimination) 기법을 XGBoost, LGBM 모델에 적용하여 정형 데이터와 비정형 데이터가 통합된 데이터 세트에서 주가 예측 수행에 도움이 되는 특징을 선별한다. 마지막으로, 특징 선별 유무에 따른 모델의 성능 개선을 제시한다.

2. 제안 방법

본 논문에서는 2018년 1월 2일부터 2019년 12월 27일까지 약 490 거래일 간 KOSDAQ 150 상장 기업 중 4차 산업 관련 5G, 자율주행, 전기차 섹터에 해당하는 10개 기업을 대상으로 당일 종가대비 익일 종가의 등락을 정형 및 비정형 데이터를 이용하여 예측한다. Fig. 1은 제안 모델의 개요도로 정형 데이터에서 추출한 기술적 지표, 거시경제 지표, 금융 지표와 비정형 데이터에서 추출한 개별 기업 긍정 지수 및 본 연구에서 신규 투입하는 주식 커뮤니티 게시물 조회수, 검색지수, 섹터 긍정 지수를 가공하여 특징 추출 및 선별, 모델 훈련하는 과정을 나타낸다. 각 데이터는 데이터 수집 및 전처리, 특징 추출, 모델 훈련 과정을 거치며, 일부 실험에서는 특징 선별 과정 후 머신러닝 모델에 투입된다.

2.1 정형 데이터 획득 및 특징 추출

본 절에서는 주가 예측 및 투자 결정에 활용되어 온 기술적 지표에 더해 거시경제지표, 금융시장 지표를 개별 기업 및 산업의 경제활동에 영향을 미칠 수 있는 실물경제와 금융시장의 흐름을 반영하는 전통

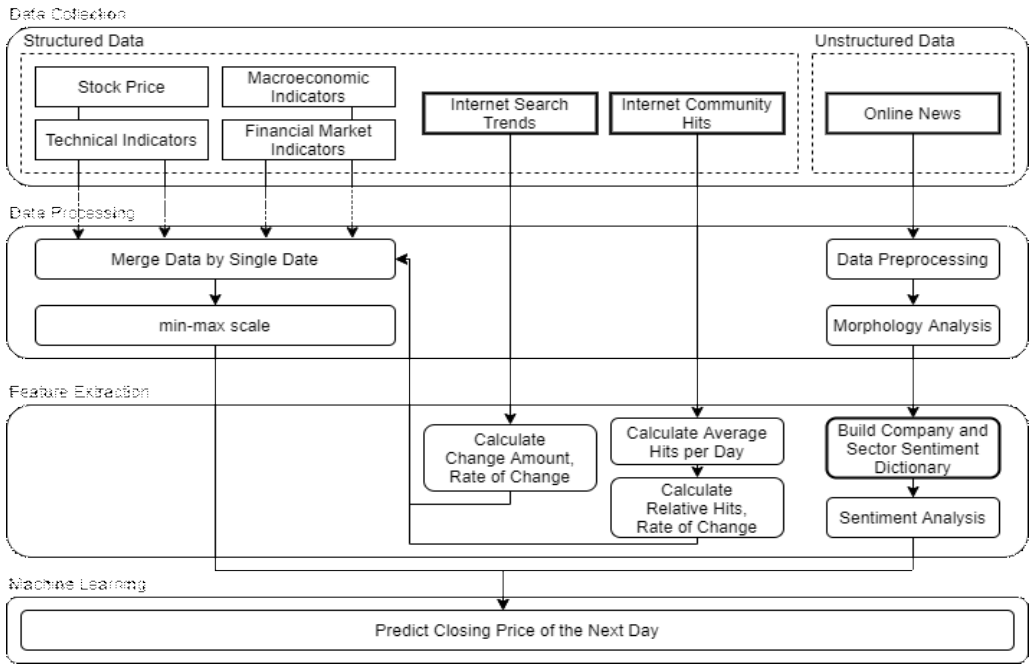


Fig. 1. Overview of Stock Price Fluctuation Prediction.

적 정형 데이터로 사용한다. 또한, 개별 종목에 대한 투자자의 관심도를 확인하기 위한 대체 데이터로 검색지수와 커뮤니티 조회수를 활용하며, 해당 대체 데이터 기반 지표를 관심도 지표라고 정의하여 사용한다.

기술적 지표는 Table 1과 같이 미래에셋대우 HTS 카이로스에서 제공하는 지표 중 추세, 모멘텀, 거래량, 변동성을 확인할 수 있는 25개 지표를 선정하였으며, 이를 통해 종목별 주가의 다양한 움직임을 파악한다. 이 때, 기술적 분석 기반의 투자에 활용되는 Signal 지표는 기술적 지표의 파생 지표라는 점을 고려하여 입력 변수에서 제외한다.

거시경제지표, 금융지표의 경우 한국은행 경제통계시스템(ECOS) 한눈에 보는 우리나라 100대 통계지표 중 분기, 연간 데이터를 제외한 일간, 월간 데이터 중 국민소득, 투자, 고용, 물가지수, 환율 등 실물경제와 통화량, 금리 등 금융시장 현황을 대표할 수 있는 지표를 선정하고, ECOS Open API를 통해 수집한다. 이를 통해 선정된 18개의 거시경제지표와 11개의 금융지표의 목록은 Table 2와 같다.

Table 3의 관심도 지표 중 검색량 지표는 대상에 대한 크고 작은 관심을 포괄하는 지표로 개별 기업의 일반적 관심도 및 이슈성과 종목에 대한 관심도를 반영하며 Google Trend에서 제공하는 기업명 검색지수 106주분과 Naver DataLab에서 제공하는 기업명 검색지수 731일분을 수집하고, 웹 크롤링을 통해 Naver 증권 해당 기업 종목 토론방에 해당 기간 업로드 된 84049개 게시물의 조회수 데이터를 수집한다. 커뮤니티 조회수 기반 관심도 지표는 개별 종목의 투자 관심도가 높은 적극적 시장 참여자의 관심도뿐 아니라 추세의 지속 가능성을 반영하는 지표로 주식 커뮤니티를 대상으로 한다. 주식 커뮤니티는 개인 투자자의 향후 주가에 대한 예측과 관심을 직접적

Table 1. Stock Price and Technical Indicators.

Category	Features
Stock Price	Open, Close, High, Low, Volume
Technical Indicators	SMA5, SMA20, SMA60, SMA120, CCI, MACD, EMA5, EMA20, Disparity5, Disparity10, Disparity20, MFI, Momentum, Stochastic Oscillator, Sto %K, Sto %D, William's %R, ROC, RSI, UpDI, DownDI, PVT, Bollinger Bands(Lower, Middle, Upper)

Table 2. Macroeconomic and Financial Market Indicators.

Category	Features
Macroeconomic Indicators	KRW/USD(Basic Exchange Rate), KRW/USD(Closing Rate), KRW/CNY(Basic Exchange Rate), KRW/CNY(Closing Rate), KRW/JPY(100 Yen), KRW/EURO, International Reserves, Manufacturing Operation Ratio Index, Unemployment Rate, Employment Rate, Consumer Price Index, Producer Price Index, Export Price Index, Import Price Index, Housing Sales Price Index, Current Account, Financial Account, Net Barter Terms of Trade Index
Financial Market Indicators	Bank of Korea Base Rate, Call Rate (Overnight), KORIBOR(3 month), CD(91 day), Monetary Stabilization Bonds(364 day), Treasury Bonds(3 year), Treasury Bonds(5 year), Corporate Bonds(3 year, AA-), M1(Narrow Money, Avg.), M2(Broad Money, Avg.), Lf(Avg.)

Table 3. Social Interest Indicators.

Category	Features
Internet Search Trends	Google Trend Score, Google Trend Change Amount, Google Trend Rate of Change, Naver DataLab Score, Naver DataLab Change Amount, Naver DataLab Rate of Change
Internet Community Hits Indicators	Average Hits per Day, Relative Hits, Average Hits Rate of Change

으로 표출할 수 있는 온라인 공간으로 게시물 수, 댓글 수, 조회수 등의 데이터를 확인할 수 있으며, 이를 바탕으로 개인 투자자의 관심도를 반영한 주가 등락 및 추세 예측이 가능하다.

검색량 지표의 전처리 과정은 수집한 검색지수 데이터로부터 전일 대비 검색지수 변화량(수식 1)과 전일 대비 검색지수 변화율(수식 2)를 산출한다. 또한 커뮤니티 조회수 기반 관심도 지표의 전처리 과정은 커뮤니티 조회수 데이터를 바탕으로 일별 업로드된 게시물의 평균 조회수를 산출하고(수식 3), 평균 조회수의 전일 대비 변화율(수식 2)과 [15]의 방식을 활용한 1개월 평균 조회수 대비 상대 조회수(수식 4)를 산출한다. 수식 2의 변화율 산출 과정에서 전일 데이터가 0일 경우 1로 대체하여 계산한다. 주별, 월별 정형 데이터를 범위에 해당하는 일별 데이터에 병합한 후 결측치를 제거하고, 최종적으로 모든 데이터를 0에서 1 사이의 값으로 정규화한다.

$$Data\,of\,the\,Date - Data\,of\,\mathrm{Previous}\,Date \tag{1}$$

$$\frac{Data\,of\,the\,Date - Data\,of\,\mathrm{Previos}\,Date}{Data\,of\,\mathrm{Previos}\,Date} \tag{2}$$

$$\frac{\sum Hits\,of\,the\,Post}{Uploaded\,Posts\,of\,the\,Date} \tag{3}$$

$$\frac{\dfrac{Data\,of\,the\,Date}{\sum_{i=1}^{30} Data\,of\,(Date-i)}}{n=30} \times 100 \tag{4}$$

2.2 비정형 데이터 획득 및 특징 추출

본 절에서는 비정형 데이터에 해당하는 온라인 뉴스에서 감성 분석을 통해 특징을 추출한다. 감성 분석은 자연어 처리와 텍스트 분석, 전산 언어학 등을 이용해 텍스트 내에서 주관적인 정보를 확인하고 추출하는 기법으로 텍스트의 극성을 긍정, 부정, 중립 등으로 분류하고, 긍정과 부정의 대상이 되는 단어 혹은 개체를 추출하는 작업이 포함된다. 본 논문에서는 사전 기반 감성 분석을 사용하고, 증권시장 도메인 특화 감성 사전을 구축하기 위해 [8]의 방식을 이용하며, 개별 기업 감성 사전 뿐 아니라 개별 기업이 속한 섹터의 감성 사전도 섹터 뉴스를 통해 구축하여 개별 기업 및 섹터 소식에 영향을 받는 기업의 주가 예측 정확도를 개선하고자 한다.

데이터 수집 과정은 개별 기업 및 섹터별 감성 사전 구축을 위해 네이버 뉴스에서 웹 크롤링을 통해 온라인 뉴스를 수집하고, 선정한 키워드가 뉴스 제목에 포함된 경우 자동으로 해당 뉴스의 제목, 발행 일자, 요약본, 언론사를 수집한다. 개별 기업의 뉴스는 수집 대상 기업명이 뉴스 제목에 포함된 경우 수집하고, 섹터 뉴스는 수집 대상 섹터명이 뉴스 제목에 포함된 경우 수집한다.

데이터 전처리 과정은 섹터 뉴스는 섹터 고유의 뉴스 이외에 예측 대상이 아닌 기업의 뉴스가 추가적으로 수집된다. 이 때, KOSDAQ 150 상장 기업 중 동일 섹터에 속하는 기업의 전반적인 주가 흐름과 반대되는 흐름을 보이는 기업의 뉴스는 감성 사전 구축에 사용되면 예측 정확도를 낮출 수 있으므로 뉴스 필터링을 위해 수식 5를 통해 KOSDAQ 150에 상장된 기업 중 동일 섹터에 속하는 기업의 시가총액을 기업별로 구한 후 수식 6에서 지수 구성종목의 시가총액을 비교일자와 기준일자 기준으로 각각 더하고 이를 나누어 섹터별 주가지수 SSI(Sector Stock Index)를 도출한다. Table 4는 실제 각 섹터 지수에 포함된 기업을 나타낸다. 이후 동일 섹터에 속하는 KOSPI, KOSDAQ 상장 기업 중 특정 기업의 뉴스가 수집된 동일 섹터 뉴스 건수의 0.05% 이상 포함되면

Table 4. Firms included in each sector index.

5G	Electric Car	Self-Driving
RFHIC	L&F	MCNEX
SEOJIN SYSTEM	POWER LOGICS	CAMMSYS
ACE TECH	CIS	CHEMTRONICS
KMW	SINHEUNG SEC	I3 SYSTEM
WISOL	POSCO ICT	HYVISION SYSTEM
OE SOLUTIONS		
PI MATERIALS		

해당 기업 주가와 섹터 지수의 상관관계를 분석하고 음의 상관관계가 나오면 해당 기업명이 제목에 포함된 뉴스는 삭제하여 섹터 뉴스 필터링을 수행하고, 기업별 뉴스와 필터링 이후의 섹터별 뉴스는 정규표현식을 이용하여 기자명, 회사명, 특수문자, 숫자 등 주가 예측에 불필요한 정보를 제거하는 전처리 과정을 거친다.

$$Market\ Capitalization = Stock\ \ Price \times Shares\ \ \ Outstanding \tag{5}$$

$$Sector\ Stock\ Index = \frac{Market\ Capitalization\ at\ the\ time\ of\ comparison}{Market\ Capitalization\ at\ the\ base\ point} \times 100 \tag{6}$$

형태소 분석 과정은 특정 단어의 긍정, 부정을 판단하기 위한 사전 과정으로 Mecab[16] 형태소 분석기를 사용하여 여러 체언 중 실질적 의미를 가지는 명사를 추출한다. 추출된 명사는 감성 사전 후보 리스트에 포함된다. Mecab 형태소 분석기는 형태소 분석기인 OKT, Komoran, Khaii 보다 품사 태깅 시간이 빠르고, 데이터 증가 시에도 처리 시간이 크게 늘지 않는 장점[17]이 있으나 Mecab의 형태소 단어 사전에 유가증권시장 전문 단어는 많이 수록되어 있지 않아 사용자 단어 사전에 미래에셋대우에서 제공하는 증권용어사전[18] 중 한국어 명사 약 800여개를 추가하여 형태소 분석의 정확도를 개선한다.

감성사전 구축 과정은 기존 [8]과 [19]의 감성사전 구축 방식을 기반으로 기업의 주가가 기업 고유의 뉴스와 해당 기업이 속한 섹터의 뉴스에 모두 영향을 받는다는 가정에 따라 개별 기업 뉴스와 섹터별 뉴스를 이용하여 개별 기업 및 섹터별 감성사전을 구축한다. 감성사전에 수록되는 것은 형태소 분석 과정에서 추출된 명사 중 낮은 빈도수를 제외한 명사와 해당 명사가 익일 주가 등락에 미친 영향을 나타내는 단어 긍정 점수이다. 이 때, 전체 뉴스 데이터에서 출현한 빈도가 낮은 명사는 주가 상승 혹은 하락에 얼마나 영향을 미쳤는지 일반화하기 어려우므로 추출된 명사 리스트의 총 출현 빈도수를 평균한 후 해당 평균보다 낮은 빈도로 등장한 명사는 감성 사전 후보 명사 리스트에서 삭제한다.

단어 긍정 점수 계산 과정은 수식 7과 같이 개별 기업 감성사전 구축 시에는 개별 단어가 발생한 날 익일에 주가가 상승한 경우 해당 뉴스 데이터에 1, 그 외의 경우 0의 값을 부여하고, 섹터별 감성 사전 구축 시에는 개별 단어가 발생한 날 익일에 섹터 주가지수가 상승한 경우 1, 그 외의 경우 0의 값을 부여한다. 수식 8의 단어 긍정값(word positive value)은 감성사전 수록 단어 리스트에서 개별 단어별로 해당 단어가 포함된 뉴스 데이터의 NSP 값을 모두 더해 개별 단어들의 단어 긍정값을 계산하고, 수식 9는 개별 단어의 총 출현 빈도수를 나눠 해당 단어의 긍정 점수를 계산한다. 출현 빈도 수 계산 시에는 특정 명사가 한 뉴스 데이터에 여러 번 등장해도 해당 뉴스에서의 등장 횟수를 1로 하고, 해당 명사가 출현한 모든 뉴스 데이터의 수를 합산한다. 이를 감성 사전 명사 리스트에 저장된 모든 단어를 대상으로 진행한 후 개별 기업 감성 사전과 섹터별 감성 사전을 구축한다.

$$NSP(j) = \begin{cases} 1\ (\mathrm{if}\ news\ j\ occured\ \mathrm{and}\ next\ stock\ price\ increased) \\ 0\ (else) \end{cases} \tag{7}$$

$$word.positive(i) = \sum NSP(\mathrm{if}\ word\ i\ is\ in\ news\ j) \tag{8}$$

$$word.score(i) = \frac{word.positive(i)}{frequency(i)} \tag{9}$$

감성 분석을 위해 구축된 개별 기업 감성 사전, 섹터별 감성 사전을 바탕으로 수식 10, 11을 통해 사전에 수록된 단어가 개별 뉴스 데이터에 포함된 경우 해당 단어의 긍정 점수를 match.positive에 모두 더하고, 해당 단어가 특정 뉴스 데이터에 포함되었는지 여부를 의미하는 match를 1로 변경한다. 그리고 수식 12에서 match.positive를 해당 뉴스에 등장한 사전 수록 단어의 개수로 나누어 개별 뉴스 데이터의 긍정 점수를 의미하는 news.positive를 도출한다. 수식 13에서 같은 날에 발행된 뉴스 데이터의 긍정 점수를 평균하여 일별 긍정 지수를 도출하는 방식으로 감성 분석을 수행한다.

$$match.positive(j) = \sum word.positive(i) \tag{10}$$

$$match(i,j) = \begin{cases} 1\ (\mathrm{if}\ word\ i\ is\ in\ news\ j) \\ 0\ (else) \end{cases} \tag{11}$$

$$news.positive(j) = \frac{match.positive(j)}{\sum match} \tag{12}$$

$$Daily\ Positive\ Index(k) = \frac{news.positive(k,j)}{n} \tag{13}$$

이 때, 수식 13에서 k는 특정일을 의미하고, n은 특정일에 발행된 뉴스의 개수이다.

2.3 특징 선별 및 분류

본 절에서는 Table 1, 2, 3에서 언급한 바와 같이 주가 데이터 및 기술적 지표 30개, 거시경제 및 금융 지표 29개, 관심도 지표 9개, 비정형 데이터 기반 긍정 지수 2개, 총 70개의 특징을 대상으로 특징 선별 방법인 RFE를 적용하여 중복 특징 및 중요도가 낮은 특징을 제외하여 분류 성능을 개선하며, 분류 과정에서 SVM, XGBoost, LightGBM의 세 가지 머신러닝 모델에 거시경제 지표, 기술적 지표, 관심도 지표, 긍정 지수를 입력하여 익일 주가 등락을 예측한다. 출력 값은 익일 주가가 상승하는 경우 1, 주가 유지 또는 하락일 때 -1로 설정한다.

SVM은 선형 분류를 수행하는 것 외에도 커널 트릭을 사용하여 비선형 분류를 효율적으로 수행하여 입력 데이터를 고차원 특정 공간에 매핑할 수 있는 분류 모델이며 XGBoost는 분산 경사 부스팅(distributed gradient boosting) 알고리즘으로 많은 문제를 빠르고 정확하게 해결하는 병렬 트리 부스팅(parallel tree boosting)을 제공한다. LightGBM은 XGBoost와 마찬가지로 분산 경사 부스팅 알고리즘이며 의사결정 트리 알고리즘 기반이지만 XGBoost보다 학습에 걸리는 시간이 적으며 메모리 사용량도 XGBoost에 비해 상대적으로 적은 것이 장점이다[20]. 본 논문에서 사용한 모델별 하이퍼파라미터 셋팅 범위는 Table 5와 같고, 지정된 셋팅 범위에서 Scikit-Learn에서 제공하는 GridSearchCV을 통해 최적의 하이퍼파라미터를 기업별로 지정하였다.

3. 실험 및 결과 분석

실험을 위해 프로그램 언어는 Python 3를 사용하였으며, 전처리와 모델 학습을 위해 라이브러리 TA-Lib, Scikit-Learn, XGBoost, LightGBM을 사용하였다. 검증을 위해 각 종목에 Stratified 5-겹 교차 검증을 적용하였으며, 한 겹(fold)은 약 98일로 구성되었다. 성능 평가 지표로는 각 겹의 등락 예측 평균 정확도를 사용하였다. 대체 데이터인 관심도 지표와 섹터 긍정 지수의 성능 검증을 위해 네 가지 비교 실험을 수행한다. 관심도 지표의 유용성을 알아보기 위해 정형 데이터에서 관심도 지표 사용 여부에 따른 결과를 비교하고, 섹터 긍정 지수를 평가하기 위해 비정형 데이터에서 섹터 긍정 지수 사용 여부에 따른 결과를 비교한다. 비정형 데이터의 유용성을 평가하기 위해 정형 데이터에 비정형 데이터 통합 여부에 따른 결과를 확인하고, 지표별 성능을 알아보기 위해 RFE 특징 선별 기법을 사용하여 성능 평가를 진행한다.

Table 6은 정형 데이터를 사용한 주가 변동 예측에서 관심도 지표의 효과를 확인하기 위해 관심도 지표를 제외한 기술적 지표, 거시경제·금융 지표를 우선적으로 투입하고 이후 관심도 지표와 통합하여 실험 후 비교한 결과이다. 실험 결과 XGBoost, LGBM 모델에서 관심도 지표를 투입하였을 때 평균 2.14%p, 2.73%p 향상되는 결과를 보였으며, SVM을 통한 실험에서는 평균 0.83%p 하락하였으나 3개 기업에서 평균 정확도가 향상되는 결과를 보였다. 특히 XGBoost, LGBM 실험 결과 평균 정확도가 상승한 9개 기업 중 3개 기업에서 평균 정확도가 4%p 이상, 최대 7%p 상승하였으며 하락한 1개 기업 역시 0.60%p, 1.07%p로 낮은 하락폭을 보였다.

Table 7은 비정형 데이터를 사용한 주가 변동 예측에서 섹터 긍정 지수의 유용성을 확인하기 위해

Table 5. Hyper Parameter Setting Range.

Model	Hyper Parameter Range
SVM	kernel: rbf C: [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 1000000] gamma: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10]
XGBoost	booster: gbtree max_depth: [3, 4, 5, 6, 7, 8], min_child_weight: [1, 3, 5] gamma: [0, 0.1, 0.2, 0.3, 0.4], n_estimators: [50, 100, 150, 200] learning_rate: [0.01, 0.05, 0.1, 0.15, 0.2], reg_alpha: [1e-5, 1e-2, 0.1, 1, 100]
LightGBM	boosting: dart learning_rate: [0.01, 0.05, 0.1], num_iterations: [1000] feature_fraction: [0.7, 0.8, 0.9, 1], bagging_fraction:[0.7, 0.8, 0.9, 1]

Table 6. Accuracy of Structured Data Based Stock Price Fluctuation.

Company	Without Social Interest Indicator			With Social Interest Indicator		
	SVM	XGBoost	LGBM	SVM	XGBoost	LGBM
A	84.09%	79.77%	77.05%	82.95%	**82.27%**	**81.59%**
B	84.90%	79.79%	77.60%	82.95%	**84.41%**	**81.00%**
C	85.78%	84.23%	82.17%	**85.79%**	**85.03%**	**84.00%**
D	79.73%	80.81%	78.92%	79.19%	**82.43%**	**79.19%**
E	83.91%	83.89%	81.08%	**84.47%**	**86.16%**	**83.90%**
F	88.50%	83.00%	81.44%	86.40%	**83.26%**	80.37%
G	82.84%	83.53%	81.05%	81.73%	**87.83%**	**85.80%**
H	86.19%	80.53%	77.38%	**86.81%**	**84.62%**	**84.92%**
I	84.78%	80.43%	79.94%	82.74%	**81.96%**	**80.69%**
J	82.73%	78.48%	73.64%	82.12%	77.88%	**76.06%**
Average Accuracy	**84.35%**	81.45%	79.03%	83.52%	**83.59%**	**81.75%**

Table 7. Accuracy of Unstructured Data Based Stock Price Fluctuation.

Company	Without Sector Positive Index			With Sector Positive Index		
	SVM	XGBoost	LGBM	SVM	XGBoost	LGBM
A	58.41%	61.36%	59.09%	66.82%	65.91%	64.09%
B	62.76%	64.24%	61.78%	66.42%	70.57%	67.39%
C	66.68%	68.99%	65.12%	69.77%	72.86%	72.09%
D	62.97%	61.35%	56.22%	65.95%	69.73%	67.03%
E	62.42%	67.80%	64.42%	68.95%	71.77%	69.23%
F	59.97%	60.76%	54.21%	64.40%	64.93%	59.42%
G	65.70%	65.92%	64.35%	73.60%	76.54%	73.84%
H	56.62%	63.23%	56.30%	64.79%	63.83%	61.63%
I	64.21%	65.98%	65.22%	67.75%	67.78%	67.00%
J	57.88%	58.18%	56.97%	58.18%	62.73%	58.18%
Average Accuracy	61.76%	63.78%	60.37%	**66.66%**	**68.67%**	**65.99%**

섹터 긍정 지수를 제외한 개별 기업 긍정 지수를 우선적으로 각 모델에 투입하고 이후 섹터 긍정 지수와 통합하여 실험 후 비교한 결과이다. 실험 결과 SVM, XGBoost, LGBM 3개의 모델에서 섹터 긍정 지수를 투입하였을 때 평균 정확도가 4.90%p, 4.89%p, 5.62%p 향상되는 결과를 보였다. 또한 실험 대상인 10개 기업의 정확도가 모든 모델에서 상승하고 SVM에서 최대 8.41%p, XGBoost에서 최대 10.62%p, LGBM에서 최대 10.81%p 상승하여 섹터 긍정 지수의 유용성을 입증하였다.

Table 8은 비정형 데이터의 효과를 측정하기 위해 정형 데이터를 이용하여 실험을 우선적으로 실험을 진행한 후 비정형 데이터에서 추출한 개별 기업 긍정 지수, 섹터 긍정 지수를 정형 지표에 통합하여 실험한 결과이다. 실험 결과 SVM, XGBoost, LGBM 3개의 모델에서 비정형 지표를 투입하였을 때 평균 정확도가 0.29%p, 0.03%p, 0.49%p 향상되는 결과를 보였다. 또한 SVM, XGBoost 모델에서 실험 대상 10개 기업 중 6개 기업의 평균 정확도가 상승하고 LGBM 모델에서는 7개 기업의 평균 정확도가 상승하여 과

Table 8. Accuracy of Structured and Unstructured Data Based Stock Price Fluctuation.

Company	Structured Data			Structured Data + Unstructured Data		
	SVM	XGBoost	LGBM	SVM	XGBoost	LGBM
A	82.95%	82.27%	81.59%	83.64%	82.73%	81.36%
B	82.95%	84.41%	81.00%	83.44%	83.93%	81.74%
C	85.79%	85.03%	84.00%	86.55%	86.04%	85.80%
D	79.19%	82.43%	79.19%	79.19%	81.62%	78.65%
E	84.47%	86.16%	83.90%	83.33%	85.59%	84.18%
F	86.40%	83.26%	80.37%	86.13%	81.42%	81.17%
G	81.73%	87.83%	85.80%	83.53%	88.50%	87.36%
H	86.81%	84.62%	84.92%	86.18%	85.25%	83.97%
I	82.74%	81.96%	80.69%	84.27%	82.98%	81.45%
J	82.12%	77.88%	76.06%	81.82%	78.18%	76.67%
Average Accuracy	83.52%	83.59%	81.75%	**83.81%**	**83.62%**	**82.24%**

반 이상의 기업에서 비정형 데이터 투입 후 모델 성능이 향상되는 결과를 확인할 수 있었다. SVM, LGBM 모델에서 비정형 데이터 투입 시 최대 1.80%p의 상승률, XGBoost에서는 1.02%p의 상승률을 보였다. 각 모델에서 성능 지표가 하락한 기업의 하락률을 평균한 결과 SVM, XGBoost, LGBM에서 0.58%p, 0.92%p, 0.57%p로 낮은 하락률을 보였다.

Table 9는 RFE를 적용한 특징 선별 전후 모델, 기업별 평균 정확도 및 RFE 적용 결과 선택된 특징의 개수이다. RFE 적용 후 XGBoost 실험 결과 평균 정확도는 최소 0.00%p, 최대 1.08%p로 평균 0.28%p 상승하였으며, LGBM 실험 결과 평균 정확도는 최소 0.25%p, 최대 3.78%p로 평균 1.16%p 상승하였다. 또한 선택된 특징의 개수는 평균 41.6개, 25.4개로 전체 특징의 59.43%, 36.29%에 해당하였다.

Table 10은 RFE를 사용하여 측정한 특징 중요도 순위의 분류별 평균 및 특징 선별율이다. 특징 중요도 순위는 RFE 적용 이후 선별되는 특징의 순위를 1, 차순위로 판정된 특징의 순위를 2 이상의 자연수로 산정하였다. 중요도 순위 책정 결과 XGBoost 모델에서 기존 정형 데이터인 주가, 기술적 지표는 평균 15.67위, 거시경제·금융 지표는 평균 24.72위를 기

Table 9. Accuracy of Selected Feature Based Stock Price Fluctuation and Number of Features.

Company	XGBoost			LGBM		
	Before	After	n_features	Before	After	n_features
A	82.73%	82.73%	45	81.36%	82.27%	21
B	83.93%	83.93%	56	81.74%	83.69%	9
C	86.04%	86.04%	51	85.80%	86.83%	24
D	81.62%	82.70%	16	78.65%	82.43%	7
E	85.59%	85.59%	52	84.18%	84.75%	22
F	81.42%	82.20%	27	81.17%	81.43%	34
G	88.50%	88.50%	48	87.36%	87.81%	10
H	85.25%	85.25%	51	83.97%	85.23%	63
I	82.98%	83.24%	38	81.45%	81.71%	52
J	78.18%	78.79%	32	76.67%	77.88%	12
Average	83.62%	**83.90%**	41.6	82.24%	**83.40%**	25.4

Table 10. Average Feature Importance Ranking and Selected Ratio by Categories.

Company	XGBoost				LGBM			
	Stock Price, Technical	Macro-economic, Financial Market	Social Interest	Positive Index	Stock Price, Technical	Macro-economic, Financial Market	Social Interest	Positive Index
A	13.93	28.55	5.22	1.00	3.87	8.17	4.44	1.00
B	24.20	39.21	10.67	1.00	2.40	3.17	1.00	1.00
C	16.73	21.79	1.67	1.00	3.33	5.14	1.00	1.00
D	25.80	40.83	13.89	1.50	20.37	27.10	17.33	1.00
E	14.97	26.24	3.78	1.00	3.50	4.31	1.00	1.00
F	7.43	17.03	1.89	1.00	11.93	21.34	4.11	1.00
G	24.57	37.00	9.78	1.00	3.53	6.48	3.00	1.00
H	1.07	1.90	1.00	1.00	2.03	5.83	3.11	1.00
I	3.83	3.97	1.00	1.00	8.20	11.76	1.00	1.00
J	24.17	30.66	18.11	2.00	10.63	15.03	6.00	1.00
Average	15.67	24.72	6.70	1.15	6.98	10.83	4.20	1.00
Selected Ratio	38.39%	20.69%	63.33%	90.00%	63.67%	45.86%	80.00%	100.00%

록하며 각각 38.39%, 20.69%의 선별율을 보였으나 신규 정형 데이터인 관심도 지표는 평균 6.70위, 비정형 데이터 기반 지표인 긍정 지수는 평균 1.15위의 높은 중요성을 보이며 평균 63.33%, 90.00%의 선별율을 보였다. 또한 LGBM 모델에서도 주가, 기술적 지표는 평균 6.98위, 거시경제·금융 지표는 평균 10.83위로 각각 63.67%, 45.86%의 선별율을 보였으나 관심도 지표는 평균 4.20위, 80.00%, 긍정 지수는 평균 1.00위로 100% 선별되며 매우 높은 중요도를 보였다.

4. 결 론

본 논문에서는 주가 데이터, 기술적 지표, 거시경제 지표, 금융 지표, 관심도 지표와 온라인 뉴스에서의 개별 기업 일별 긍정 지수, 섹터 일별 긍정 지수를 활용하여 주가 등락을 예측하는 모델을 제시하였다. 또한 XGBoost, SVM, LGBM 모델을 활용하여 실험을 진행하였으며, 특징 간 비교를 위해 RFE 기반의 특징 선별을 실시하였다. 정형 데이터 실험의 관심도 지표 통합 실험 이후, SVM 모델을 제외하고 XG Boost 모델에서 10개사의 정확도가 상승하며 평균 정확도가 2.14%p 상승하였고 LGBM 모델에서 9개사의 정확도가 상승하며 평균 정확도가 2.72%p 상승하였다. 이 실험을 통해 Boosting 모델에서 검색지수와 커뮤니티 조회수와 같은 관심도 지표가 주가의 등락 강화[15] 뿐만 아니라 실제 주가의 방향성 예측에도 유의미한 영향력을 가지는 것을 확인하였다. 비정형 데이터 실험 과정에서는 섹터 긍정 지수 투입 이후 투입 이전에 비해 모든 모델에서 실험 대상 10개사의 정확도가 상승하고 3개 모델 평균 정확도가 5.13% 상승하였다. 이 실험은 기업 특화 사전에 산업 우량기업 사전을 추가로 사용해 기업 특화 사전을 단독으로 사용한 것에 비해 정확도를 최대 1.4%p 상승시킨 기존 유사 연구[9] 보다 모델의 성능이 큰 폭으로 향상되는 결과를 보였다. 특징 간 비교 실험 결과로는 관심도 지표와 비정형 지표의 평균 중요도 순위는 5.45위, 1.08위로 주가 데이터 및 기술적 지표, 거시경제 및 금융 지표의 11.33위, 17.78위 대비 높은 중요도를 보였다. 특히 비정형 데이터 기반 긍정 지수는 전체 실험 중 95%의 특징 선별율을 보였으며, 관심도 지표 역시 71.84%의 특징 선별율을 보이며 해당 지표의 유용성을 입증하였다. 따라서 관심도 지표와 섹터 긍정 지수를 추가로 활용하여 KOSPI 상장 기업에 비해 뉴스 데이터가 부족해 연구하기 어려웠던 KOSDAQ 중소형 기업의 주가예측을 시도한 해당 실험을 통해 투자 업계에 기여하고자 한다. 다만 4차 산업혁명 관련 전 산업이 아닌 특정 산업군에

대한 예측을 진행한 것과 비정형 데이터로 온라인 뉴스만을 이용한 점은 본 연구의 한계로 남는다.

REFERENCE

[1] The Fourth Industrial Revolution: What It Means, How to Respond. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond (accessed November 16, 2020).

[2] Financial Service Commission, *Reorganization of Listing and Public Offering System to Build a Dynamic Capital Market,* October 5, 2016.

[3] ISO/IEC 20546:2019(en) Information technology —Big data—Overview and Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso-iec: 20546:ed-1:v1:en (accessed November 24, 2020).

[4] D.H. Seo, S.G. Bae, S.J. Kim, H.S. Kang, and J.M. Bae, "Development of a Continuous Prediction System of Stock Price Based on HTM Network," *Journal of Korea Multimedia Society*, Vol. 14, No. 9, pp. 1152-1164, 2011.

[5] I.T. Joo and S.H. Choi, "Stock Prediction Model based on Bidirectional LSTM Recurrent Neural Network," *Journal of Korea Institute of Information, Electronics, and Communication Technology,* Vol. 11, No. 2, pp. 204-208, 2018.

[6] H.H. Ha and K.M. Ha, "Forecast of the Stock Market Price Using Artificial Neural Network and Wavelet Transform," *Journal of KIISE,* Vol. 46, No. 12, pp. 1249-1261, 2019.

[7] D.W. Hah, Y.M. Kim, and J.J. Ahn, "A Study on KOSPI 200 Direction Forecasting Using XGBoost Model," *Journal of the Korean Data & Information Science Society,* Vol. 30, No. 3, pp. 655-669, 2019.

[8] D.Y. Kim, J.W. Park, and J.H. Choi, "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] N.K. Lee, *Forecasting of Stock Prices Based on Blue-Chip Keyword Dictionary of News Articies,* Master's Thesis of Korea Advanced Institute of Science and Technology, 2017.

[10] D.Y. Kim and Y.I. Lee, "News Based Stock Market Sentiment Lexicon Acquisition Using Word2 Vec," *Journal of Korea Bigdata Society,* Vol. 3, No. 1, pp. 13-20, 2018.

[11] N.Y. Seong and K.H. Nam, "Predicting Stock Movements Based on Financial News with Systemic Group Identification," *Journal of Intelligence and Information Systems,* Vol. 25, No. 3, pp. 1-17, 2019.

[12] S.H. Song, K.S. Han, S.J. Choi, and S.M. Park, "A Study on Improving Financial Asset Trading Model Using Social Big Data Analysis Based on Machine Learning," *Entrue Journal of Information Technology,* Vol. 16, No. 1, pp. 51-62, 2017.

[13] D.Y. Park and K.H. Lee, "Price Fluctuation Prediction for Theme Stocks Using Online News and Technical Indicators," *Journal of KIISE Database Society of Korea,* Vol. 35, No. 3, pp. 66-76, 2019.

[14] M. Choi, *Stock Investment Strategy Using Social Media Data and Machine Learning,* Master's Thesis of Seoul National University, 2020.

[15] H. Ahn, *Big Data Quant Strategy*, Korea Investment & Securities Co.,Ltd., 2016.

[16] Eunjeon Project(2013). http://eunjeon.blogspot.com/2013/02/blog-post.html (accessed November 17, 2020).

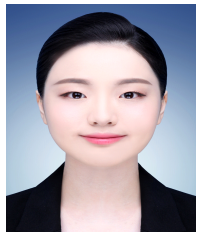
[17] Y.J. Lee, S.B. Kim, H.S. Hong, and J.W. Gim, "Comparison and Evaluation of Morphological Analyzers for Patent Documents," *Proceeding of Korean Institute of Information Technology,* pp. 264-265, 2019.

[18] Mirae Asset Daewoo Securities Term Dictionary. https://www.miraeassetdaewoo.com/hki/hki3028/r01.do (accessed November 17, 2020).

[19] E.J. Yu, Y.S. Kim, N.G. Kim, and S.R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems,* Vol. 19, No. 1, pp. 95-100, 2013.
[20] E.A. Daoud, "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset," *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering,* Vol. 13, No. 1, pp. 6-10, 2019.

김 화 련

2021년 서울여자대학교 경제학과 4학년
2021년 서울여자대학교 소프트웨어융합학과 4학년
관심분야 : 금융공학, 인공지능, 딥러닝



홍 승 혜

2021년 서울여자대학교 경제학과 학사
2021년 서울여자대학교 소프트웨어융합학과 학사
관심분야 : 인공지능, 데이터베이스

홍 헬 렌

1994년 이화여자대학교 전자계산학과 학사
1996년 이화여자대학교 전자계산학과 석사
2001년 이화여자대학교 컴퓨터학과 박사
2001년~2003년 서울대학교 컴퓨터공학부 BK 박사후연구원
2003년~2006년 서울대학교 컴퓨터공학부 BK 계약조교수
2006년~현재 서울여자대학교 소프트웨어융합학과 교수
관심분야 : 의료 인공지능, 영상처리 및 분석