

비정형, 정형 데이터의 이미지 학습을 활용한 시장예측

(MPIL: Market prediction through image learning of unstructured and structured data)

이운선*, 이주홍**, 최범기**, 송재원***

(Yoon Seon Lee, Ju Hong Lee, Bum Ghi Choi, Jae Won Song)

요약

금융 시계열 분석은 현대 사회의 경제적, 사회적으로 매우 중요한 역할을 하며 세계 발전에 영향을 미치는 중요한 과제지만 많은 잡음(noise)과 불확실성 등의 어려움으로 인해 금융 시계열 분석 예측은 어려운 연구 주제이다. 본 논문에서는 비정형 데이터와 정형 데이터를 함께 이미지로 변환하여 시장을 예측 하는 방법(MPIL)을 제안한다. 시장 예측을 위해 n일 기간의 비정형 데이터인 SNS, 뉴스 데이터를 감정분석하고 정형 데이터인 시장 데이터를 GADF 알고리즘으로 이미지 변환하고 이미지 학습을 통해 n+1일의 가격을 예측하는 초단기 시장을 예측한다. MPIL은 평균 정확도 56%로 기존 시장예측에 사용되던 감정분석을 활용하여 LSTM으로 시장을 예측하는 모델 평균 정확도 50%보다 높은 정확도를 보였다.

■ 중심어 : 감정분석 ; 시장예측 ; 딥러닝

Abstract

Financial time series analysis plays a very important role economically and socially in modern society and is an important task affecting global development, but due to difficulties such as a lot of noise and uncertainty, financial time series analysis prediction is a difficult research topic. In this paper, we propose a market prediction method (MPIL) by converting unstructured data and structured data into images. For market prediction, it analyzes SNS and news data, which is unstructured data for n days, and converts the market data, which is structured data, to an image with the GADF algorithm, and predicts an ultra-short market that predicts the price of n+1 days through image learning. MPIL has an average accuracy of 56%, which is higher than the 50% average accuracy of the model that predicts the market with LSTM by using sentiment analysis used for existing market forecasting.

■ keywords : Sentiment analysis ; Time series analysis ; Market prediction ; Deep learning

I. 서론

금융시장은 현대 사회에서 경제적, 사회적으로 매우 중요한 역할을 한다. 특히 경제성장의 중요한 역할을 하는 시장예측 연구는 다양한 분야에서 관심을 가지는 연구 주제이다. 시장을 예측하는 방법은 크게 2가지로 기업의 가치를 분석하여 종목을 선정하는 거시적 관점의 기본적 분석과 과거 주식의 가격이나 거래량 같은 자료를 이용하여 예측하는 기술적 분석으로 나뉜다[1].

기존 시장예측 방법으로 기술적 분석 중 수학적, 통계학적 접근 등의 다양한 시도가 존재하며 그 중 통계적 방법으로 시장 데이터와 같이 수치가 포함되어있는 정형 데이터를 활용한 시

장예측 연구가 활발히 진행되었지만, 시장 데이터는 불확실성과 잡음 및 변수가 많으며, 변동성도 높아 시장을 예측하기는 쉽지 않다[2].

최근에는 기계학습을 바탕으로 하는 유전자 알고리즘, 인공신경망 등의 시장예측 연구가 활발하게 이루어지고 있으며 예측에 사용되는 데이터는 다양해지고 있다[3]. Ribeiro 등[4]은 실시간으로 변하는 대중의 감정분석이 주가의 거래량 등과 연관성이 있다고 밝히고 있다. 이러한 이유로 SNS, 뉴스와 같은 비정형 데이터를 활용하여 시장을 예측하는 연구가 발표되었다[4-8]. 그러나 SNS 뉴스 등의 비정형 데이터는 정규화된 형태와 규칙이 없어서 의미 파악이 쉽지 않은 어려움이 있고 다른 형태의 정형 데이터들과 매우 이질적이므로 이러한 데이터들을 시장예측이라는 목적에 적합하게 통합하기가 매우 어렵기 때문

* 학생회원, 인하대학교 전기컴퓨터공학과 대학원생

** 정회원, 인하대학교 전기컴퓨터공학과 교수

*** 정회원, 밸류파인더스

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2019R1F1A1062094, 2017R1D1A1A02018319, NRF-2020R1F1A1069361).

접수일자 : 2021년 02월 23일

수정일자 : 1차 2021년 03월 25일, 2차 2021년 03월 26일

게재확정일 : 2021년 03월 30일

교신저자 : 이주홍 e-mail : juhong@inha.ac.kr

에 여러 이질적인 데이터의 통합이 시장예측에 장애 요소로 작용하고 있다. 이러한 시장예측에서의 어려움을 해소하기 위해서는 서로 다른 종류의 데이터를 통합하기 위한 전처리와 데이터 종류에 따른 가중치 변경, 통합, 연산, 학습 등의 장점이 있는 기법의 활용이 필요하다. 이러한 목적을 위해서 여러 이질적인 데이터들을 정교하게 통합하여 학습하는 기법[9]이 더 정확한 시장예측을 위해 활용될 수 있을 것이다.

따라서 본 논문에서는 비정형 데이터와 정형 데이터 등 여러 이질적인 데이터를 시장예측이라는 목적에 적합하도록 통합하기 위하여 여러 이질적인 데이터들을 이미지로 통합하여 변환하는 기법을 제안한다. II장에서 본 연구와 관련된 여러 기존 연구들을 설명하고 III장에서 제안 방법을 상세히 설명한다. IV장에서 실험방법과 실험 결과를 상세히 설명한다.

II. 관련 연구

감정분석은 특정 단어나 문맥을 분석하여 문장의 감정을 파악할 때 사용되는 기술이다[10]. BERT모델[11]은 기존 자연어 데이터 학습에 사용되던 단방향 학습이 아닌 양방향 학습을 통해 문장을 학습하는 방식을 제시하였으며 사전학습, 파인튜닝을 통해 다양한 언어학습 응용 분야에서 현재 최선의 기법 중 하나로 사용되고 있다.

[12]는 시장예측에 최적화된 방법은 존재하지 않으며 시장예측을 위해 다양한 시도를 할 필요가 있다고 주장하고 있다. 기존 연구는 대부분 LSTM과 RNN등의 딥러닝 기법을 이용하여 시장을 예측하였다[8,13]. 최근 시장예측을 위한 방법으로 기존 사용되던 기술들을 융합하여 사용하고 있으며 그 중 SNS와 뉴스 같은 비정형 데이터들이 활용되고 있다[4-6,14].

[4]는 대중의 감정을 분석하여 주식시장의 연관성을 확인하였다. 감정분석을 활용하기 위한 방법으로 감정 표현을 포함한 문

장을 식별하는 OpinionFinder와 텍스트 분석을 통해 6가지의 기분을 분류하는 GPOMS(Google-Profile of Mood States)를 사용하였으며 SNS와 시장과의 연관성을 확인하였다.

Xu Jiawei 등 [8]은 뉴스의 제목만을 추출해 단편적인 데이터의 감정을 분석하여 LSTM으로 시장예측에 사용하였고 [4]는 시장과 연관된 데이터가 아닌 SNS에 있는 모든 데이터를 활용하여 연구를 진행했다는 문제점이 존재한다. 이러한 연구들의 문제점은 시장을 예측하는데 알맞은 데이터가 아닌 잡음이 다수 포함된 데이터를 사용한다는 데에 있다. 또한 SNS 사용의 목적은 소통에 있으며 이를 위해 대부분의 SNS는 ‘좋아요’와 같은 게시글의 공감하는 공감 지수를 표현한다. 기존 감정분석을 활용해 시장을 예측하는 연구들에서는 SNS 데이터를 활용함에 이런 수치를 고려하지 않았다.

최근 시계열 데이터를 시각화하는 연구가 진행 중이며, 데이터의 정보를 손실하지 않고 이미지로 변환 가능한 경우 시각화된 이미지를 통해 다양한 문제에 응용할 수 있다[9, 14-18]. 이 중 [18]은 시장 데이터를 대량의 데이터 처리와 특성 파악이 쉽도록 데이터를 촛대 차트로 시각화하였다. 그러나 시각화한 데이터를 활용하여 시장을 예측하고자 한 연구는 진행되지 않았다. [22]는 시계열 데이터를 인코딩하여 스케일링 된 시계열의 단일 성분의 각 코사인을 계산하여 좌표 시스템으로 변환하는 방법인 GADF알고리즘을 제안했다.

III. 제안 방법

본 논문에서는 시장을 예측하기 위해 실시간으로 대중의 심리가 표현되는 SNS 데이터와 객관적 사실이 담긴 뉴스 데이터의 감정을 분석하고 감정 분석된 비정형 데이터와 시장 데이터를 이미지로 변환하고 학습하여 시장을 예측하는 방법(MPIL, Market prediction through image learning)을 제안한다. 제

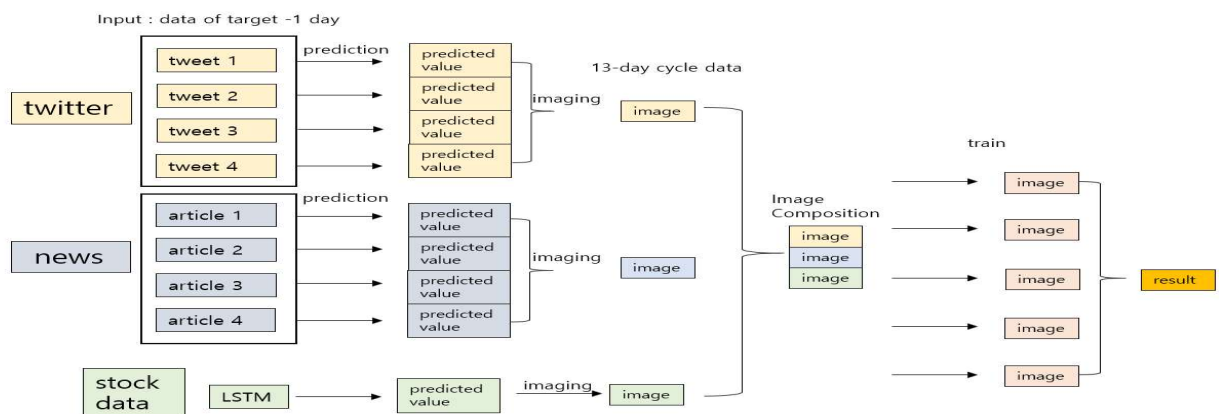


그림 1. 제안하는 방법의 전체 구조도

안한 구조는 그림 1과 같다. 수집한 SNS와 뉴스 데이터를 학습하여 시장을 예측하는 값을 출력하고 시장 데이터와 예측값을 합하여 이미지로 변환 후 이미지 학습을 통한 시장을 예측한다.

텍스트 데이터는 띄어쓰기와 맞춤법 등 많은 노이즈를 포함하고 있으며, 용도에 맞는 학습을 위해서는 전처리를 하여야 한다. 따라서 본 논문에서는 SNS와 뉴스의 감정을 분석하기 위해 기초적인 공백 처리, 광고, 특수 문자, 해시태그 제거 등의 자연어 데이터의 전처리를 진행하였다. 수집한 뉴스와 트위터 데이터 중 약 40,000개의 데이터에 대해 해당 종목에 대한 긍정(1), 부정(-1), 분류 불가(0)로 라벨링하였다. 그 중 SNS 데이터의 좋아요, 공유하기와 뉴스 데이터의 공유된 수를 파악하여 데이터에 추가한다. 40,000개의 뉴스와 SNS 데이터에 대해 BERT 모델[11]을 활용하여 문장 단위로 학습하고 예측값을 라벨링했다(1: 상승, 0: 분류 불가, -1: 하락). 동일 날짜 데이터들의 레이블 평균을 계산하여 11개의 수치가 포함된 주가 데이터와 결합한다. 그림 2는 동일 기간의 시가, 고가, 저가, 종가, 거래량 등 11개의 수치와 SNS 데이터를 결합하여 배치한 데이터이다. 13일 간격의 데이터를 데이터의 좌표를 인코딩한 후 Gram Matrix 연산을 통해 이미지로 변환하는 GADF 알고리즘[22]을 사용하여 그림 3과 같이 변환한 이미지를 생성한다.

Date	Open	High	Low	Sns_avg	close	diff	fluctuation rate	volume	Institutional net sales	Foreign net sales	Foreign Number of shares held	Foreign Retention rate
20161114	3465	3670	3455	-0.400000	3560	-20	-0.0056	10482599	-10	-27058	145851	0.0060
20161115	3510	3520	3335	-0.500000	3375	-185	-0.0520	7211181	0	-63182	67969	0.0028
20161116	3365	3510	3345	-1.000000	3355	-20	-0.0059	7088921	0	3373	97342	0.0040
20161117	3375	3450	3265	-0.500000	3410	55	0.0164	4774817	0	152579	234221	0.0096
20161118	3410	3640	3200	2.333333	3295	-115	-0.0337	13415914	0	-166391	33549	0.0014
20161121	3380	3525	3300	0.000000	3315	-20	0.0061	6577332	0	8992	37641	0.0015
20161122	3315	3495	3235	24.142857	3270	-45	-0.0136	2681181	0	28705	70846	0.0029
20161123	3285	3420	3230	-1.500000	3370	100	0.0306	3419496	0	35341	102187	0.0042
20161124	3330	3515	3305	36.666667	3315	-55	-0.0163	5470748	0	-56555	46632	0.0019
20161125	3295	3445	3290	-0.888889	3410	95	0.0287	3044147	1309	323043	369675	0.0151
20161128	3410	3485	3310	-0.500000	3320	-90	-0.0264	2993730	0	-284567	96068	0.0037

그림 2. SNS, 뉴스 데이터와 시장 데이터를 결합한 형태의 데이터

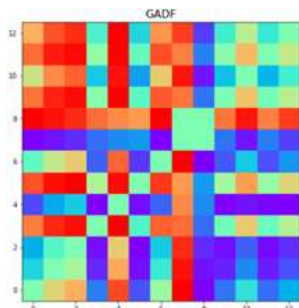


그림 3. 데이터의 이미지화

변환된 이미지를 Siamese Neural Network[19]를 통해 학습하고 시장을 예측한다.

IV. 실험

1. 데이터

본 논문에서는 이글벳, 유니슨, 빅텍, 아난티, 티플렉스, KOSDAQ지수 등 6개의 종목을 실험 대상으로 하여 성능을 평가하였다. 해당 종목들은 특정 종목군 내 주가 상승률 상위의 주식으로 테마로 분류된 주식들을 대표하는 대장주이다. 대장주는 특히 변동성이 크며 예측 시 특정 주식 테마의 전반적인 흐름 파악과 예측에 도움이 된다. 따라서 대중의 감정을 포함하여 시장을 예측하기 위해 약 5년의(2015-06-29 ~ 2020-07-27) 기간 동안 특정 산업군에서 대표하는 주식(대장주)의 관련 산업이나 종목을 언급한 키워드가 포함된 뉴스와 SNS(트위터) 텍스트 데이터를 수집하여 각 종목당 약 400,000개의 비정형 데이터를 수집하였다. 이 중 7개월의(2020-01-01 ~ 2020-07-27) 데이터를 성능 평가 테스트 기간으로 하였다. 주가 데이터는 동일 기간의 시가, 고가, 저가, 종가, 거래량 등 11개의 수치를 포함하는 데이터를 사용하였다.

2. 실험 개요

이미지로 변환된 주가 데이터를 활용하여 시장을 예측하기 위해 벡터 사이의 유사도를 계산해 학습하는 Siamese Neural Network를 이용하였다. 식 (1)은 각 네트워크를 통해 나오는 피처를 활용하여 L1 노름을 구하는 식으로, 식 (1)에서 σ 는 시그모이드 활성화함수이며 α_j 는 학습 과정에서의 파라미터이다. 식 (2)는 예측값과 타겟에 로스를 구해 식 (3)으로 가장 큰 값을 가지는 이미지를 같은 클래스로 분류하여 다음 날의 주가를 예측하였다[18]. 이미지로 변환하지 않은 감정분석한 SNS, 뉴스 데이터와 시장 데이터를 입력으로 하는 LSTM 모델[12]과 MPIL을 비교하였다.

$$\sigma((\sum_j \alpha_j |h_{1,L-1}^{(j)} - h_{2,L-1}^{(j)}|)) \quad (1)$$

$$L(x_1, x_2, t) = t \cdot \log(p(x_1 \circ x_2)) + (1-t) \cdot \log(1 - (p(x_1 \circ x_2))) + \lambda \cdot \|w\|_2 \quad (2)$$

$$C(\hat{x}, S) = \operatorname{argmax}_c P(\hat{x} \circ x_c), x_c \in S \quad (3)$$

3. 성능 평가

평가 지표로는 시장예측의 정확도(Accuracy)와 모델 성능 측정의 지표로 쓰이는 MAE (Mean absolute error), MAPE (Mean absolute percentage error)를 사용하였다. 정확도는 0과 1 사이의 확률값을 출력하며, 높을수록 좋은 성능을 나타내고 MAE는 실제 값과 예측값의 차이를 절대값으로 변환하여 평균화를 내며, MAPE는 MAE를 퍼센트로 변환한 값이다. MAE와 MAPE는 낮을수록 좋은 성능을 나타낸다. 그림 4, 5, 6은 각 종목별 [8]의 방법과 MPIL의 Accuracy, MAPE, MAE의 결과값을 나타낸 것이다.

$$Accuracy = \frac{1}{n} \sum_i Correct_i \quad (4)$$

$$correct = \begin{cases} 1, & y_1 \cdot \hat{y}_1 > 0 \\ 0, & else \end{cases}$$

$$MAE = \frac{1}{n} \sum_i |Actual - Prediction| \quad (5)$$

$$MAPE = \frac{1}{n} \sum_i \left| \frac{Prediction - Actual}{Actual} \right| \times 100 \quad (6)$$

그림4, 5, 6를 통해 본 논문에서 제안한 방법이 기존 연구[8]보다 평균적인 성능이 향상되었음을 알 수 있다. 그 중 그림 7에서와 같이 아난티는 기존 LSTM 모델[8]에서 51.42%의 정확도를 보였지만 SNS, 뉴스, 시장 데이터를 이미지화하여 이미지 학습을 통해 시장을 예측한 MPIL은 63.4%의 정확도를 보이며 좋은 성능을 보였다.

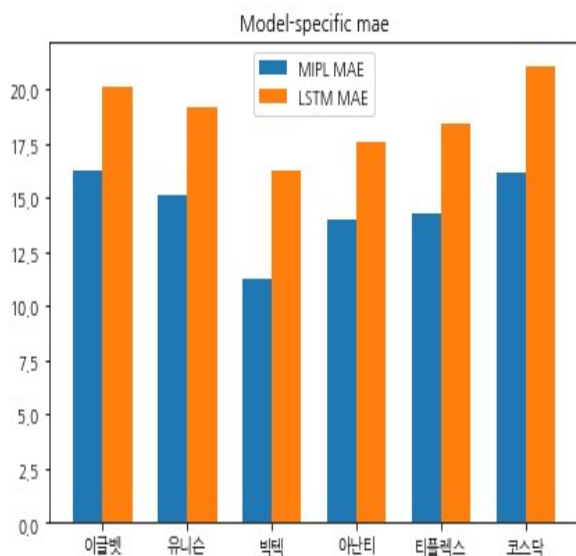


그림 4. 종목별 MAE 결과

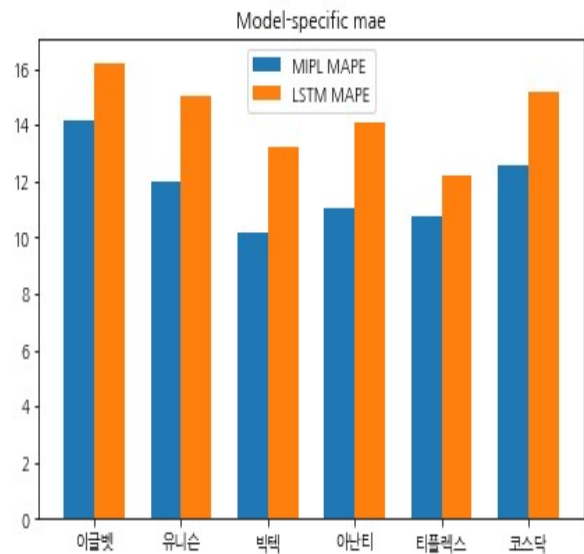


그림 5. 종목별 MAPE 결과

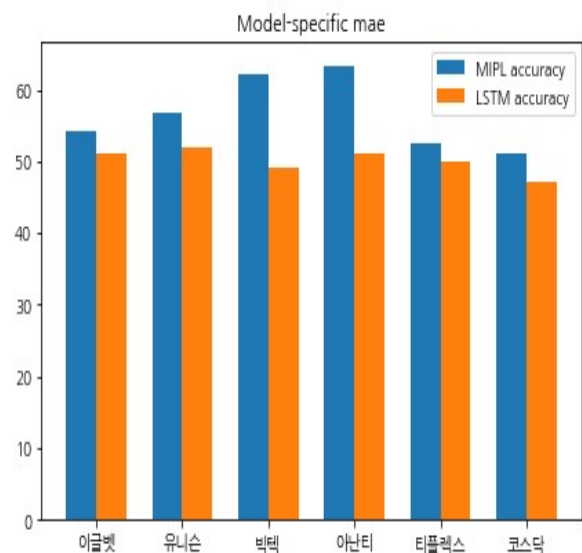


그림 6. 종목별 Accuracy 결과

V. 결 론

본 논문에서는 시장예측을 위해 n일 기간의 SNS, 뉴스 데이터인 비정형 데이터와 정형 데이터인 시장 데이터를 GADF 알고리즘으로 이미지 변환하고 Siamese neural network 모델을 활용하여 이미지 분류하여 n+1일의 시장을 예측하는 MPIL을 제안하였다.

기존 시장예측 모델과의 비교를 위해 기존 LSTM 모델[8]과 비교하였으며 제안한 방법의 우수함을 확인하였다. 또한 제안한 방법은 개별 종목과 종합 종목 모두에서 안정적이며 그 중 개별

종목의 예측에 좀 더 부합하는 방식이다. 특히 본 논문에서는 특정 테마를 대표하는 주식들을 예측하였으며 이를 토대로 특정 테마의 시장 흐름도 예측이 가능하다.

REFERENCES

- [1] Rodolfo C. Cavalcantea, Rodrigo C. Brasileiro, Victor L.F. Souza, Jarley P. Nobrega, Adriano L.I. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions," *Expert Systems with Applications*, Vol. 55, pp. 194-211, Aut. 2016.
- [2] Jonathan L. Ticknor, "A Bayesian regularized artificial neural network for stock market forecasting," *Expert Systems with Applications*, Vol. 40, No. 14, pp. 5501-5506, Oct. 2013.
- [3] 이형용, "한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형," *Enture Journal of Information Technology*, Vol. 7, No. 2, 33-43쪽, 2018년 7월
- [4] Johan Bollena, Huina Mao, Xiaojun Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, Vol. 2, No. 1, pp. 1-8, Mar. 2011.
- [5] 안성원, 조성배, "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측," *한국정보과학회 2010 한국컴퓨터종합학술대회 논문집*, 제 37권, 제1호, 364-369쪽, 강릉원주대학교, 한국, 2010년 6월
- [6] Tushar Rao, Saket Srivastava, "Analyzing Stock Market Movements Using Twitter Sentiment Analysis," *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, pp. 119-123, Istanbul, Turkey, Aug. 2012.
- [7] 김동영, 박재원, 최재현, "SNS와 뉴스기사의 감성 분석과 기계학습을 이용한 주가예측 모형 비교 연구," *Journal of Information Technology Services*, 제13권, 제3호, 221-233쪽, 2014년 9월
- [8] Xu Jiawei, Tomohiro Murata, "Stock Market Trend Prediction with Sentiment Analysis based on LSTM Neural Network," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019*, pp. 475-179, Kowloon, Hong Kong, Mar. 2019.
- [9] 이충현, "Data Digest: 데이터 이미지화와 응용," *정보과학회 컴퓨팅의 실제 논문지*, 제24권, 제11호, 614-622쪽, 2018년 11월
- [10] Theresa Wilson, et al., "Opinion Finder: A system for subjectivity analysis," *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pp. 34-35, Vancouver, Canada, Oct. 2005.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805*, Oct. 2018.
- [12] S.K. Aggarwal, L.M. Saini, A. Kumar, "Price forecasting using wavelet transform and LSE based mixed model in Australian electricity market," *Int. J. Energy Sector Manage*, vol. 2, No. 4, pp. 521 - 546, Nov. 2008.
- [13] Qun Zhuge, Lingyu Xu, Gaowei Zhang, "LSTM Neural Network with Emotional Analysis for Prediction of Stock Price," *Engineering Letters*, Vol. 25, No. 2, pp. 64-72, May 2017.
- [14] 박지연 외 2인, "시계열 데이터의 이미지화 알고리즘을 이용한 기계학습 성능 비교 및 분석," *2019년도 한국통신학회 추계종합학술발표회 논문집*, 135-136쪽, 국민대학교, 한국, 2019년 11월
- [15] Nima Hatami, Yann Gavet, Johan Debayle, "Classification of time-series images using deep convolutional neural networks," *arXiv:1710.00886*, Oct. 2017.
- [16] Zhiguang Wang, Tim Oates, "Imaging Time-Series to Improve Classification and Imputation," *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 3939-3945, Buenos Aires, Argentina, Jul. 2015.
- [17] Zhiwei Gao, Tiangang Zou, Lulu Zhang, Juan Chen, Jing Guo, Weixing Ji, "Rapidly Generate and Visualize the Digest of Massive Time Series Data," *2017 IEEE Third International Conference on Big Data Computing Service and Applications(BigDataService)*, Vol. 1, pp. 157-164, Redwood City, USA, Apr. 2017.
- [18] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," *International Conference on Machine Learning*, vol. 37, 2015.
- [19] Rashmi Sutratti, Dr. D.A. Torse, "Stock Market Forecasting Techniques: A Survey," *International Research Journal of Engineering and Technology, International Research Journal of Engineering and Technology (IRJET)*, vol. 6, No. 5, pp. 4842-4844, May 2019.
- [20] 이모세, 안현철, "효과적인 입력변수 패턴 학습을 위한 시계열 그래프 기반 합성곱 신경망 모형: 주식시장 예측에의 응용," *지능정보연구*, 제24권, 제1호, 167-181쪽, 2018년 3월
- [21] Feiyun Xiao, Yanyan Chen, Yanhe Zhu,

"DGADF/GASF-HOG: feature extraction methods for hand movement classification from surface electromyography," *Journal of Neural Engineering*, Vol. 17, No. 4, Jul. 2020.

- [22] 천성길, 이주홍, 최범기, 송재원, "대규모 외생 변수와 Deep Neural Network를 사용한 금융 시장 예측의 성능 향상에 관한 연구," *스마트미디어저널*, 제9권, 제4호, 26-35쪽, 2020년 12월
- [23] 유경호, 노주현, 홍택은, 김형주, 김판구, "Self-Attention을 적용한 문장 임베딩으로부터 이미지 생성 연구," *스마트미디어저널*, 제10권, 제1호, 63-69쪽, 2021년 3월

저 자 소 개



이윤선(학생회원)

2018년 한남대학교 컴퓨터공학과 학사 졸업.
2018년~현재 인하대학교 전기컴퓨터공학과 석사과정

<주관심분야 : 자연어처리, 딥러닝, 시계열 분석>



이주홍(정회원)

1983년 서울대학교 전자계산기공학과 학사
1985년 서울대학교 전자계산기공학과 석사
2001년 KAIST 컴퓨터공학전공 박사
2001년 ~ 현재 인하대학교 컴퓨터공학과 교수

<주관심분야 : 머신러닝, 금융투자공학>



최범기(정회원)

1986 서울대학교 수학과 학사
1995 Florida State university computer science 석사
2017 ~ 현재 인하대학교 컴퓨터공학과 겸임교수

<주관심 분야: 머신 러닝, 금융투자공학>



송재원(정회원)

2005년 성공회대학교 컴퓨터공학 학사
2007년 인하대학교 컴퓨터공학 석사
2013년 인하대학교 컴퓨터공학 박사
2013~2014 대구경북과학기술원 개발팀장
2014~2017 인천국제공항공사 선임연구원

2017~2017 가천길대학병원 연구교수
2017~현재 (주)벨류파인더스 대표이사

<주관심분야: 머신러닝, 데이터마이닝, 금융IT>