DBpia

코로나-19 데이터를 이용한 KOSPI 지수 예측 딥러닝 : LSTM을 이용하여

Forecasting KOSPI Index with LSTM Deep Learning Model using COVID-19 Data

저자 (Authors)

권도윤, 권소현, 변준영, 김미숙

Doyun Kwon, Sohyeon Kwon, Junyoung Byun, Misuk Kim

출처 (Source)

한국정보과학회 학술발표논문집 , 2020.12, 1367-1369 (3 pages)

발행처 (Publisher)

한국정보과학회

The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10529989

APA Style

권도윤, 권소현, 변준영, 김미숙 (2020). 코로나-19 데이터를 이용한 KOSPI 지수 예측 딥러닝 : LSTM을 이용하여. 한국정보과학회 학술발표논문집, 1367-1369.

이용정보 (Accessed)

고려대학교
163.***.133.25
2022/01/04 19:11 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s)and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s)for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

코로나-19 데이터를 이용한 KOSPI 지수 예측 딥러닝 : LSTM을 이용하여

권도윤 O, 권소현, 변준영, 김미숙*
세종대학교 데이터사이언스학과
{doyun317, sohyun42415, 2}@sju.ac.kr, misuk.kim@sejong.ac.kr

Forecasting KOSPI Index with LSTM Deep Learning Model using COVID-19 Data

Doyun Kwon, Sohyeon Kwon, Junyoung Byun, Misuk Kim
Department of Data Science, Sejong University

요 약

코로나-19로 인한 세계보건기구(WHO)의 팬데믹 선언은 세계 증시 및 국내 증시에 급락과 큰 변동성을 일으켰다. 본 연구는 주요국 가운데 가장 빠르게 회복한 코스피 지수와 코로나-19 데이터를 활용해 시계열 분석방법인 LSTM 모델로 주가예측을 진행한다. 2010년 1월부터 2020년 9월까지의 KOSPI200 종가, 시가, 거래량과 같은 기본지표와 2020년 1월부터 2020년 9월까지의 국내 코로나-19 데이터를 추가 지표로 사용하였다. 연구 결과 코로나-19 데이터를 사용했을 때 오차율이 최대 32% 감소되었고, 이는 통계적으로 유의했다. 신규확진자수는 주가에 더 많은 충격을 주는 것으로 확인되었다.

1. 서 론

주식시장의 주가지수는 경제 및 정치적 상황과 같이 다양한 주변 환경에 영향을 받기 때문에 주가지수의 정확한 예측은 매우 어려운 문제로 여겨진다. 최근 경제적 상황으로는 코로나-19가 글로벌 경제에 막대한 피해를 주었고 국내 경제 또한 많은 영향을 끼쳤다. 국내 경제에서 코로나-19 발생은 경기종합지수에 영구적·단계적 개입효과가 통계적으로 유의한 효과가 있었다[1].

코로나-19로 인한 세계보건기구(WHO)의 팬데믹 선언은 세계 증시 및 국내 증시에 지난 금융위기 이후 처음으로 전체적 급락과 큰 변동성을 일으키기도 했다. 코로나-19로 발생한 글로벌 경제위기 속에서 주요국 가운데 가장 빠르게 회복한 코스피 지수는 전 세계가 주목하는 시장이 되었다.

이에 본 연구는 국내 주식시장 데이터와 코로나-19 데이터를 활용해 시계열 분석 방법인 LSTM을 이용하여 코스피를 대표하는 KOSPI200 지수를 예측하고자 한다. LSTM 모델에서 코로나-19 데이터는 주가예측 성능 향상에 영향을 끼치는지 알기 위함이다. 이후 통계적 검증을 통해 성능 향상 여부를 판단하고 그 정도를 확인한다. 코로나-19 데이터와 같은 적절한 환경변수의 추가는 주식시장 주가 예측 모델링의 성능 변화가 있음을 알아내는 것이 본 연구의 목적이다.

2. 모 델 링

2.1 데이터 처리

본 연구에서는 2010년 1월부터 2020년 9월까지의 KOSPI200의 종가, 시가, 고가, 저가, 거래량, 전일대비 변화량을 사용하였다. 코로나-19 데이터는 2020년 1월부터 2020년 9월까지의 누적격리해제수, 누적사망자수, 누적확진자수, 신규격리해제수, 신규사망자수, 신규확진자수, 네이버 코로나-19 검색 트렌드를 사용하였다. 변수간의 선형관계를 확인하고, 적절한 변수를 사용하기 위해 다중공선성 검사를 진행하였다. 누적격리해제수, 누적사망자수, 누적확진자수는 VIF(Variance Inflation Factor)값이 10이상으로 나타나 변수에서 제거하였다[2].

2.2 LSTM 모델

전통적인 NN(Neural Network)는 이전의 일어난 사건을 바탕으로 나중에 일어나는 사건을 생각하지 못했다. 이런 문제를 해결하고자 나타난 모델이 RNN(Recurrent neural network)이다. RNN은 체인처럼 이어지는 성질로 시계열구조에서 탁월한 성능을 자랑한다.

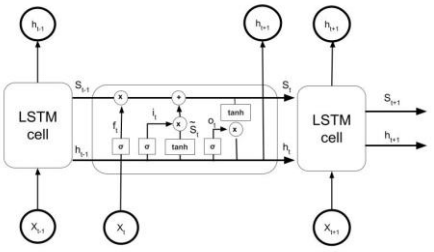

그림 1. LSTM 구조

고려대학교 | IP:163.***.133.25 | Accessed 2022/01/04 19:11(KST)

LSTM(Long Short-Term Memory Network)은 RNN의 특별한 종류 중 하나로, RNN의 단점인 긴 시계열 구조에서 과거의 학습 결과가 점점 사라지는 장기 의존성문제를 해결한 모델이다. LSTM의 핵심인 셀 스테이트(Cell state)는 컨베이어 벨트와 같이 과거의 정보가 전혀 바뀌지 않고 그대로 전달하는 구조로 장기 의존성 문제를 해결한다. LSTM은 단계적으로 진행되는데 첫번째 단계로 sigmoid 함수를 통해 삭제할 과거의 정보를 결정한다. 두번째 단계로 sigmoid 함수와 tanh 함수로 셀 스테이트에 업데이트하는 새로운 정보를 결정한다. 마지막으로 sigmoid 함수와 셀 스테이트의 출력을 tanh 함수로 계산하여 어떤 출력 값을 출력할지를 결정하게 된다. 그림 1에 LSTM의 구조를 나타냈다[3][4].

3. 주가예측 딥러닝 및 통계적 검증

3.1 주가예측 딥러닝

본 연구의 주가예측 딥러닝을 위해 KOSPI200의 종가, 시가, 고가, 저가, 거래량, 전일대비변화량 같은 6가지의 기본지표를 사용하였다. 또한 코로나-19의 신규격리해제자수, 신규사망자수, 신규확진자수, 네이버 코로나-19 검색트렌드와 같은 4가지의 추가 지표를 사용했다.

딥러닝을 위해 Python의 Tensorflow 라이브러리 중 하나인 keras를 사용하였다. 딥러닝은 입출력 데이터의 크기를 작게 해주는 것이 효과적이기 때문에, 모든 데이터를 [0,1]의 범위로 스케일링 하는 MinMaxScaler를 사용하였다.

모델의 학습을 위해 학습 데이터로 총 2654개의 데이터에서 70%인 1843개를 사용하였고 이는 2010년 1월부터 2017년 6월 12일까지이다. 테스트 데이터는 나머지 30%인 2017년 6월 13일부터 2020년 9월 29일까지 811개의 데이터를 사용하였다.

본 연구에서 사용하는 주가데이터와 코로나-19 데이터는 시계열 데이터이므로, 시계열 데이터에 적합한 딥러닝 알고리즘으로 LSTM을 선택하여 예측모델을 구축하였다. 최적화 알고리즘으로 AdamOptimizer를 선택하고 손실함수로 MAE(Mean Absolute Error)를 사용하였다.

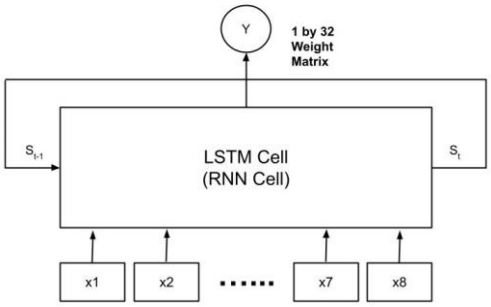

그림 2. LSTM 모델

최종 모델의 LSTM 입력 노드는 KOSPI200의 기본지표와 코로나-19 추가지표 8개로 이루어진다. 또한, 주식시장은 월요일부터 금요일까지 5일장이고, 주식시장의 한달은 20영업일이다. 본 연구에서는 타임 스탬프(time stamp)를 20개로, 20개의 셀로 한번에 160개의 데이터를 입력하여 20일씩 학습하였다. 중간층의 노드는 32개이고 출력 값은 다음날의 종가이다. 그림 2에 실험에 사용된 LSTM 모델에 대해 나타냈다.

실험 결과는 예측 값과 실제 값 사이의 RMSE(Root Mean Square Error) 이며, 결과 값이 실행할 때마다 달라지기 때문에 10번 반복하여 평균으로 결과를 측정하였다. 해당 방법으로 추가지표의 최적 조합을 탐색하였고, 표 1에 다양한 추가지표의 조합을 나타냈다. 표 1에서 전체기간은 테스트 데이터의 전체기간을 의미하고, 코로나기간은 코로나가 발병한 2020년 1월부터 9월까지를 의미한다. (1)을 기준으로 생각하여 다른 결과들과 비교했을 때 (2)의 결과가 가장 우수했다. 주가예측에는 신규확진자수와 네이버 코로나-19 검색 트렌드가 두 개만 사용하는 것이 다른 코로나 지표들의 조합보다 가장 효율적으로 영향을 끼치는 것으로 나타났다. (1) 모델에 비해 (2) 모델은 전체기간에서 27.5%의 오차율 감소가 있었고 코로나 기간에서는 32%의 오차율 감소가 있었다. 그림 3과 그림 4는 각각 (1)과 (2)의 예측 값을 실제 값과 비교한 것이다.

표 1. 예측요소 실험결과

(1)
KOSPI 200 : 종가, 시가, 고가, 저가, 거래량, 전일대비변화량
COVID-19 : 없음

	Avg.	Std.	Max	Min
전체 기간	0.04	0.013	0.064	0.026
코로나기간	0.05	0.018	0.083	0.03

(2)
KOSPI 200 :종가, 시가, 고가, 저가, 거래량, 전일대비변화량
COVID-19 : 추가 확진자 수, 네이버 코로나-19 검색트렌드

	Avg.	Std.	Max	Min
전체 기간	0.029	0.002	0.034	0.027
코로나기간	0.034	0.003	0.041	0.032

(3)
KOSPI 200 :종가, 시가, 고가, 저가, 거래량, 전일대비변화량
COVID-19 : 추가 확진자 수, 네이버 코로나-19 검색트렌드, 추가 격리 해제자 수, 추가 사망자 수

	Avg.	Std.	Max	Min
전체 기간	0.032	0.003	0.039	0.276
코로나기간	0.038	0.006	0.051	0.032

(4)
KOSPI 200 :종가, 시가, 고가, 저가, 거래량, 전일대비변화량
COVID-19 : 추가 확진자 수, 네이버 코로나-19 검색트렌드, 추가 격리 해제자 수

	Avg.	Std.	Max	Min
전체 기간	0.032	0.003	0.04	0.027
코로나기간	0.038	0.006	0.052	0.031

(5)
KOSPI 200 :종가, 시가, 고가, 저가, 거래량, 전일대비변화량
COVID-19 : 추가 확진자 수, 네이버 코로나-19 검색트렌드, 추가 사망자 수

	Avg.	Std.	Max	Min
전체 기간	0.032	0.003	0.04	0.028
코로나기간	0.039	0.004	0.049	0.034

고려대학교 | IP:163.***.133.25 | Accessed 2022/01/04 19:11(KST)

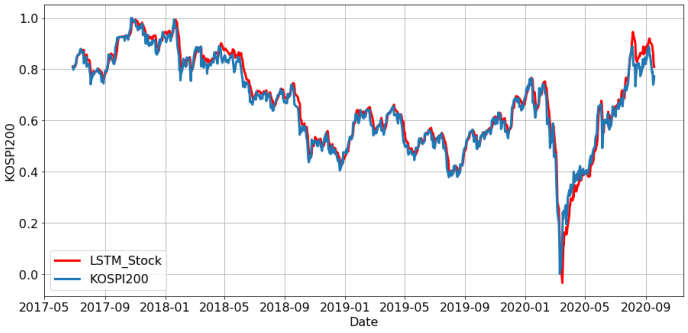

그림 3. KOSPI200과 (1) 모델

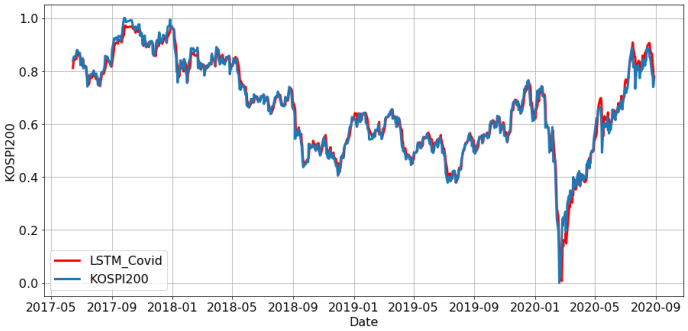

그림 4. KOSPI200과 (2) 모델

3.2 통계적 검증

표 2는 통계 검증을 실시한 결과이다. KOSPI200의 기본지표만 사용한 (1) 모델과 코로나-19 추가지표를 함께 사용한 최적모델인 (2) 모델의 통계적 검증을 위해 독립표본 T검정을 실시하였다.

독립표본 T검정을 실시하기 위한 Levene 등분산 검정에서 전체 기간과 코로나기간 모두 신뢰수준 99%하에서 유의수준보다 유의확률이 더 작기 때문에 두 표본의 분산이 동일하지 않은 것으로 확인되었다. 두 표본이 등분산이 아니기 때문에 이분산 독립 T 검정을 실시하였고, 신뢰수준 95% 하에서 유의수준보다 유의확률이 더 작기 때문에 모평균의 차이가 존재하지 않는다는 귀무가설을 기각하고 모평균의 차이가 존재한다는 대립가설이 채택되었다[5].

표 2. 통계 검증 결과

통계 검증	Levene등분산 검정		독립T검정(이분산)	
통계량	F 값	p-value	T 값	p-value
전체 기간	10.41	0.004	-2.629	0.026
코로나기간	8.559	0.009	-2.54	0.03

4. 결 론

본 연구는 Python의 Tensorflow를 이용해 LSTM 모델로 KOSPI200의 종가를 예측하는데 KOSPI200의 기본지표와 코로나-19의 추가지표를 사용해 어떤 코로나-19 지표가 모델의 예측력향상을 일으키는지 확인하였다. 연구 결과 주가 예측에 사용한 코로나-19의 추가지표인 신규격리해제자수, 신규사망자수, 신규확진자수, 네이버 코로나-19 검색 트렌드를 조합한 여러가지 모델 모두 KOSPI200의 기본지표만 사용한 모델보다 향상된 결과가 있었다. 그 중 신규확진자수와 네이버 코로나-19 검색 트렌드 두가지만 사용하여 조합한 모델이 제일 예측율이 우수했다. 이는 신규격리해제자수, 신규사망자수보다 신규확진자수가 주가에 더 많은 충격을 준다고 생각되며 그 영향이 모델에 적절하게 적용된 것으로 보인다.

코로나-19의 시계열 지표를 주가 예측에 사용하면 예측율이 상승하는 현상을 보여주었다. 이는 어떤 커다란 경제적 이벤트가 있을 때 그 관련 지표들을 적절히 조합해 이용하면 주가예측에 대한 오차율을 줄일 수 있는 것으로 보인다.

본 연구는 2020년 9월까지의 데이터를 사용하였으며 그 이후의 국내 코로나-19 진행사항은 고려하지 못하였는데, 꾸준한 데이터 업데이트로 연구를 이어갈 것이다.

또한 해외 주식시장의 데이터와 해외 코로나-19 데이터, 각 국의 구글 코로나-19 검색 트랜드 등을 이용해 해외의 주요 주가인 나스닥, 다우, 니케이, Euro Stoxx 등의 지수를 예측하는 연구를 진행할 예정이다.

5. 참고문헌

[1] 정우수, 한문승 (2020) 코로나19가 경제에 미치는 영향 분석, 한국정보통신설비학회 학술대회, pp.83-87

[2] 이승현. "다중공선성에 대한 연구." 국내석사학위논문 연세대학교 대학원, 2003. 서울

[3] Anita Yadav, C K Jha, Aditi Sharan, Optimizing LSTM for time series prediction in Indian stock market, Procedia Computer Science, Volume 167, Pages 2091-2100, 2020

[4] Thomas Fischer, Christopher Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research, Volume 270, Issue 2,2018,Pages 654-669,

[5] 신현준 (2009). t-검정을 이용한 웹 트래픽 임계치 검증 기법. 한국엔터테인먼트산업학회 학술대회 논문집, 3(2), 157-160

Acknowlegement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1G1A1101195) and Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2020040126).

* 교신저자

고려대학교 | IP:163.***.133.25 | Accessed 2022/01/04 19:11(KST)