JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 8, pp. 31-39, August 2022
https://doi.org/10.9708/jksci.2022.27.08.031

A study on Deep Learning-based Stock Price Prediction using News Sentiment Analysis

Doo-Won Kang*, So-Yeop Yoo*, Ha-Young Lee*, Ok-Ran Jeong*

*Student, Dept. of Software, Gachon University, Seongnam, Korea
*Visiting Professor, Dept. of Software, Gachon University, Seongnam, Korea
*Student, Dept. of Software, Gachon University, Seongnam, Korea
*Professor, Dept. of Software, Gachon University, Seongnam, Korea

[Abstract]

Stock prices are influenced by a number of external factors, such as laws and trends, as well as number-based internal factors such as trading volume and closing prices. Since many factors affect stock prices, it is very difficult to accurately predict stock prices using only fragmentary stock data. In particular, since the value of a company is greatly affected by the perception of people who actually trade stocks, emotional information about a specific company is considered an important factor. In this paper, we propose a deep learning-based stock price prediction model using sentiment analysis with news data considering temporal characteristics. Stock and news data, two heterogeneous data with different characteristics, are integrated according to time scale and used as input to the model, and the effect of time scale and sentiment index on stock price prediction is finally compared and analyzed. Also, we verify that the accuracy of the proposed model is improved through comparative experiments with existing models.

▸ **Key words**: Stock Price Forecasting, LSTM, ResNet, Sentiment Analysis, Text Summarization

[요 약]

주가는 거래량, 종가 등과 같은 숫자 기반의 내부적인 요인뿐만 아니라 법, 유행 등 여러 외부 요인에 의해 영향을 받는다. 수많은 요인이 주가에 영향을 미치기 때문에 단편적인 주식 데이터만을 이용한 정확한 주가 예측은 매우 어려운 일이다. 특히 기업의 가치는 실제 주식을 거래하는 사람들의 인식에 영향을 많이 받기 때문에 특정 기업에 대한 감성 정보가 중요한 요인으로 여겨진다. 본 논문에서는 시간적 특성을 고려한 뉴스 데이터의 감성 분석을 이용한 딥러닝 기반 주가 예측 모델을 제안하고자 한다. 주식과 뉴스 데이터, 서로 다른 특성을 가진 2개의 이종 데이터를 시간 크기에 따라 통합하여 모델의 입력으로 사용하며, 시간 크기와 감성 지표가 주가 예측에 미치는 영향에 대해 최종적으로 비교 및 분석한다. 또한 우리는 기존 모델과의 비교 실험을 통해 제안 모델의 정확성이 개선되었음을 검증한다.

▸ **주제어**: 주가 예측 모델, LSTM, ResNet, 감정 분석, 텍스트 요약

• First Author: Doo-Won Kang, Co-Author: So-Yeop Yoo, Ha-Young Lee, Corresponding Author: Ok-Ran Jeong
*Doo-Won Kang (pch145@gachon.ac.kr), Dept. of Software, Gachon University
*So-Yeop Yoo (bbusso@gachon.ac.kr), Dept. of Software, Gachon University
*Ha-Young Lee (hhzet11@gachon.ac.kr), Dept. of Software, Gachon University
*Ok-Ran Jeong (orjeong@gachon.ac.kr), Dept. of Software, Gachon University
• Received: 2022. 07. 13, Revised: 2022. 07. 27, Accepted: 2022. 08. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

모든 기업의 수익원은 소비자들에게 제공하는 상품이다. 그러한 상품들의 형태는 공산품 혹은 무형의 서비스나 컨텐츠일 수도 있다. 그렇게 소비자들에게 제공하는 상품들로 쌓아온 기업의 평판은 그 기업의 정체성과 앞으로의 성장 가능성을 보여주는 거울이라고 할 수 있다. 즉, 기업 이익의 근본은 상품을 제공하는 대상인 소비자이고, 기업의 흥망성쇠는 소비자들의 기업에 대한 인식에 따라 결정되며, 이를 통하여 기업의 이익에 민감하게 반응하는 주식 시장에서의 기업의 가치를 결정할 수 있다는 것이다. 소비자들의 기업 인식률에 대한 영향력은 최근 ESG 경영이 기업의 중요한 투자 판단 가치가 됨에 따라서, 전통적인 정형적 데이터만을 보고 투자해왔던 대형 자산 운용기관의 판단 결정에도 큰 영향을 끼치고 있다.

최근 몇 년 동안 급속한 경제 발전으로 금융 활동의 수가 증가하고 있으며 그 변동추세 또한 점점 복잡해지고 있다. 금융 활동의 패턴을 이해하고 변화를 예측하는 것은 금융계에서 주요한 연구 중 하나이다. 하지만 주식 시장은 매우 복잡하고 다양한 정치, 경제적 요인, 정권 교체, 투자 심리 등 다양한 요인에 따라 변동성이 커서 예측이 어렵다. 효율적 시장 가설(Efficient Market Hypothesis)은 자산 가격이 금융 뉴스 기사, 소셜 미디어, 블로그 등과 같은 새로운 정보에 반응한다[1].

주식 시장에서 생성되는 데이터는 매우 방대하고 비선형적이다. 이러한 종류의 데이터를 모델링하기 위해서는 숨겨진 패턴을 분석할 수 있는 모델이 필요하다. 딥러닝 모델은 자체 학습 과정을 통해 데이터 속에 존재하는 상호작용 및 패턴을 식별해 낼 수 있다.

최근 컴퓨팅 기술의 발전으로 투자자의 투자 심리에 영향을 주는 방대한 SNS 데이터를 처리하여 감성을 추출할 수 있게 되었고, 이를 금융 시계열 데이터를 조합하여 주가를 예측하는 연구가 활발하게 진행되고 있다[2-6].

본 연구의 제안하는 영향력 예측 모델은 웹 크롤링으로 온라인상에서 소비자들의 기업에 대한 인식에 대한 데이터를 수집하여 감성 분석을 진행한다. 최근 크게 발전한 딥러닝 기술을 활용하여 분석하고, 이를 기존의 정형화된 수치만을 사용한 예측 모델과 비교함으로써 소비자들의 기업 활동에 대한 인식률에 따라 기업의 주가가 어떻게 변화하는지 예상하고자 한다.

2장에서는 제안하는 주가 예측 모델과 관련된 연구들에 관해 설명하며, 3장에서는 본 논문에서 제안하는 모델에 대해 자세하게 소개하며 4장에서는 모델에 대한 평가 실험과 그 결과에 대해 서술한다. 마지막으로 5장에서는 본 논문의 결론에 관해 소개한다.

II. Related Works

1. Stock price forecasting

[7]에서는 시계열 데이터 분석을 위해 다양한 딥러닝 모델을 적용하였다. 신경망 모델을 사용한 금융 시계열 데이터 모델링은 [7]에서 처음 시도하였으며, 이는 IBM의 자산 가격 변동에서 비선형 규칙성을 찾아내기 위한 신경망을 모델링하였고 효율적 시장 가설에 대한 증거를 확립하는 데 기여하였다.

[8]에서는 과거 금융 정보를 사용하여 CNN기반 주가 예측 모델을 만들어 CNN기반 모델이 주가 예측에서 좋은 성능을 나타냄을 입증하였다. [9]에서는 CNN의 일종인 ResNet 구조를 주가 예측 모델에 적용하려는 시도가 있었으며, ResNet 구조를 포함한 모델이 더 높은 성능을 보인다는 결과를 도출하였다. [10]은 주가 예측을 위한 양방향 및 순방향 LSTM 모델의 성능을 평가하였고, 양방향 LSTM 모델이 순방향 LSTM 모델보다 주가 예측에서의 성능이 더 뛰어남을 입증하였다.

[11]은 암호화폐 가격의 sequence size를 다르게 하여 LSTM을 사용한 암호화폐 가격 예측 모델을 제작하였고, sequence size가 작을수록 더 작은 오차를 가지지만 이는 매번 오차를 재조정한 결과이며 실제 금융 시장에서는 sequence size가 10일 때 가장 큰 실용성을 갖는다는 것을 입증하였다.

2. Stock price forecasting based on sentiment analysis of news data

뉴스는 금융 시장 가격에 영향을 미치는 요소 중 하나이다. 최근 주가 예측 모델 연구에서는 뉴스 데이터를 인공지능에 접목하려는 여러 시도가 있었으며 기업과 관련된 뉴스 기사와 주가의 변동성 사이에 가한 상관관계가 있음이 드러났다. 뉴스 데이터를 주가 예측 모델에 적용하기 위해서 뉴스 데이터를 벡터화하는 다양한 방법들이 제시되었다[12-14].

[14]는 SVM(Support Vector Machine)에 입력 변수로 뉴스 데이터 전문을 벡터화하여 사용하는 것보다 summarization하여 사용하는 것이 더 좋은 성능을 끌어낸다는 것을 보였다.

[2]는 RNN을 사용하여 S&P 500 기업의 주가 데이터와

해당 기업에 대한 뉴스 기사 뉴스 본문 중 해당 기업명이 들어간 5문장을 추출하여 감성을 분석한 뒤, 이 값들의 평균을 변수로 하여 주가 예측 모델을 제작하였다. 다만 뉴스 본문 중 일부만을 사용하여 변수가 뉴스의 전체적인 감정을 다 내포하지 못하였고 뉴스 기사 빈도수에 따른 화제성을 고려하지 못했다는 한계가 존재한다.

III. The Proposed Scheme

1. Model structure

본 논문에서는 정형 데이터인 주가 정보 데이터와 비정형 데이터인 뉴스 데이터를 활용하여 딥러닝 모델 기반 주가 예측 모델에 감성 지표가 미치는 영향을 비교하고 그 결과를 분석하고자 한다. 주가 데이터와 뉴스 데이터를 수집하고, 수집된 데이터를 각각 전처리하여 감성 지표를 계산함으로써 주가 정보 예측이 가능한 모델을 구현한다.

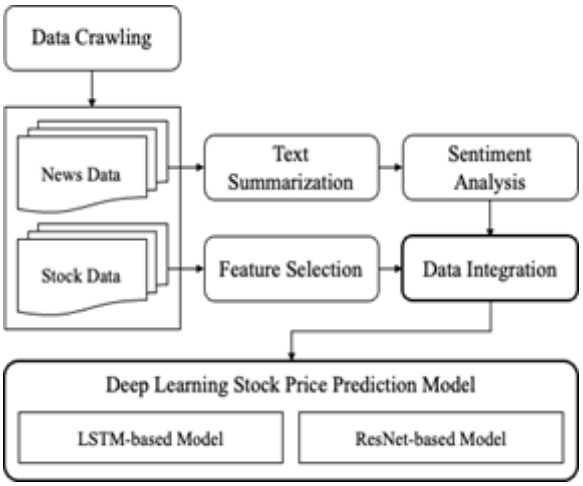


Fig. 1. Overall architecture

그림 1은 제안하는 모델의 전체적인 구조를 보여준다. 주가 데이터와 뉴스 데이터를 수집하는 데이터 크롤링, 모델에 학습시키기 위한 데이터 전처리, 그리고 감성 지표 분석과 이를 활용해 주가를 예측하는 딥러닝 모델로 구성된다. 다음에서는 각 구성요소에 대해 자세하게 설명하고자 한다.

2. Data crawling

본 논문에서는 감성 분석을 활용하여 주가의 변화를 예측하는 딥러닝 모델을 구축하고 감성 지표의 영향을 비교 및 분석한다. 주가 예측을 위해서는 반드시 주가를 예측하고자 하는 기업에 대한 주가 정보가 필요할 뿐만 아니라 관련된 다양한 정보들이 필요하다. 우리는 소비자들의 기업에 대한 인식과 같은 기업에 대한 감성을 분석하고 주가 예측 모델에 적용하고자 한다. 기업에 대한 주가 정보를 알 수 있는 주가 데이터와 기업에 대한 감성 지표를 알 수 있는 뉴스 데이터를 수집하고 활용한다.

데이터의 수집 기간은 2021년 1월 1일부터 2022년 4월 30일까지 총 327일의 거래일을 기준으로 한다. 그림 2와 같이 2020년 코로나 팬데믹 이후 코스피 지수는 약 3개월간 폭락 기간을 거친 뒤 약 1년 동안 꾸준한 상승세를 보였다[15]. 예측 모델 학습에 영향을 주는 비대칭성 데이터[16]의 수집을 최소화하기 위해 데이터의 수집 기간을 위의 기간으로 제한하였다.

Fig. 2. KOSPI index chart (2020~2021)

국내에 존재하는 수많은 기업 중, 주가 데이터와 뉴스 데이터의 충분한 연관성을 포함하기 위해 2021년 한 해 동안 가장 언급이 많이 된 회사들을 선정하여 주가 예측 모델의 학습을 진행한다. 회사 선정을 위해 Requests와 BeautifulSoup 라이브러리와 함께 오픈소스인 KoreanNewsCrawler[17]를 활용하여 2021년 1년 동안의 네이버 경제 뉴스[18]에서 KOSPI 상위 50개 기업들이 제목에 언급된 뉴스를 모두 수집하였다. 그림 3은 제목에 기업이 언급된 뉴스의 수를 통계 낸 결과를 보여준다. SK 하이닉스가 30375개로 가장 많이 언급되었으며, 삼성전자, 카카오, LG전자 순으로 많이 언급되고 있음을 확인할 수 있다.

본 논문에서는 2021년 네이버 경제 뉴스에 가장 많이 언급된 상위 4개 기업인 SK 하이닉스, 삼성전자, 카카오, LG전자를 연구 대상으로 선정한다. 4개의 기업에 대해 2021년 1월 1일부터 2022년 4월 30일까지 뉴스 데이터와 주가 데이터를 크롤링하여 연구에 활용한다. 뉴스 데이터는 네이버 경제 뉴스에 게재된 본문을 Scrapy와 Newspaper 라이브러리를 사용하여 크롤링한다. 뉴스 크

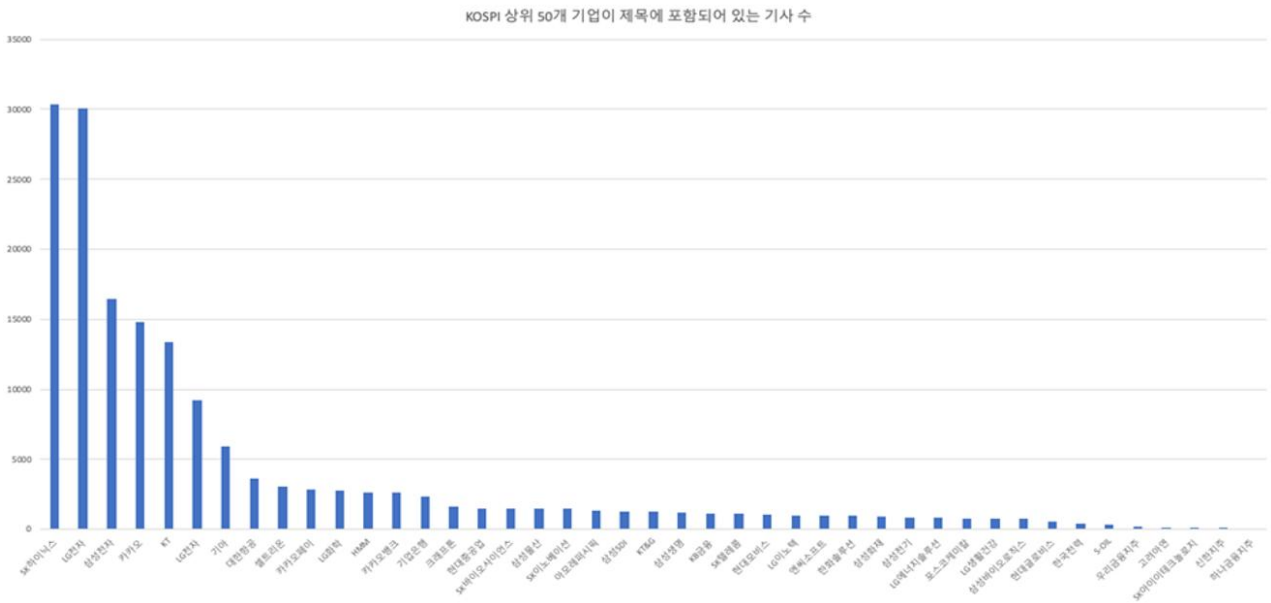


Fig. 3. Number of mentioned in title of Naver News by company as of 2021

롤링의 제한에 의해 각 기업별 1일 최대 120개의 뉴스 기사를 수집하도록 한정하여 최종적으로 총 134,036개의 뉴스 기사를 수집하였다.

주식 시장은 변동이 크고 다양한 요소들이 주식 가격 결정에 많은 영향을 미치기 때문에 정확한 예측이 어렵다. 주식에 영향을 미치는 여러 정보 중 당일 거래의 가장 마지막에 체결된 거래 가격인 종가와, 주식 시장에서 당일 주식이 거래된 총량을 의미하며 주가에 선행되는 특징이 있는 거래량 데이터를 주식 데이터로 활용하고자 한다. 뉴스 데이터를 수집한 기간과 동일한 총 327일 동안의 종가와 거래량 일별 데이터를 야후 파이낸스[19]에서 수집하고 모델 학습에 활용한다.

3. Sentiment analysis of news data

주가 정보의 변동은 다양한 요소에 영향을 받게 된다. 거래량, 종가 등을 포함한 주식 자체의 내부적인 데이터도 다음 주가 예측에 중요한 요소가 되지만, 내부적인 요소들 이외에 많은 외부적인 요인들에도 영향을 받는다. 특히 주식 거래는 사람이 관여하기 때문에 특정 기업에 대한 사람들의 인식과 같은 외부 요인들이 존재할 수 있다. 본 논문에서는 주식 데이터의 내부 요소들뿐만 아니라 외부 요소로 감성 지표를 주가 예측에 활용하고자 한다.

감성 지표의 분석은 외부 데이터로 수집한 뉴스 데이터를 사용한다. 뉴스 데이터는 기사의 제목과 본문 등으로 이루어져 있다. 경제 분야의 기사들은 기업에 대한 단순한 정보를 포함하기도 하지만 특정 기업에 대한 대중들의 인식과 같은 감정적인 정보를 포함하고 있다. 기사의 제목은 주로 내용 전체를 아우를 수 있는 핵심적인 내용을 담고 있는 텍스트 데이터이지만, 짧은 문장으로 구성되는 특징 때문에 감성을 파악하기에는 부족하다. 본문 내용의 경우는 모든 정보가 포함되어 있어 상대적으로 긴 텍스트 데이터로 나타나며, 감성 정보를 파악할 수 있는 다양한 정보들이 포함되어 있다. 하지만 부가적인 정보를 많이 포함하고 있어 핵심이 되는 내용만 찾아 감성 정보를 파악하기에는 어려움이 존재한다. 본 논문에서는 이러한 문제점을 해결하고 적절한 감성 지표를 추출하기 위해 기사의 본문을 요약하고, 요약된 핵심 문장들로부터 감성을 추출한다.

뉴스 데이터의 제목과 본문 내용을 요약하고 감성을 추출하기 위해 KakaoBrain의 Pororo[20] 모델을 사용한다. Pororo는 Platform Of neuRal mOdels for natuRal language prOcessing의 약자로 자연어처리 오픈 소스이며, 자연어 처리와 음성에 관련된 여러 태스크를 수행한다. Pororo의 요약 기능을 활용하여 기사를 요약하고, 요약된 정보를 기반으로 감성 분석 기능을 통해 수치화된 감성 정보를 추출한다. 감성 분석의 결과는 긍정(positive)과 부정(negative)으로 나타나며 긍정과 부정의 수치 합은 1이다. 즉 positive는 0과 1 사이의 값을 가지며, negative는 1-positive로 표현이 가능하다.

그림 4는 Pororo를 사용하여 뉴스 데이터를 요약하고 감성 분석을 수행한 결과 예시를 보여준다. 수집된 뉴스 데이터의 제목과 본문을 연결하여 News Content로 사용하였으며 요약을 위해 Pororo를 사용한다. 입력값에 대한 요약을 수행한 결과는 Summarize에 해당하며, 이 요약된 결과를 이용해 긍정 혹은 부정을 계산할 수 있는 감성 분

News Content

하반기 디지털 손해보험사 출범 목표보험업계 게임체인저 될까.
카카오뱅크와 카카오페이가 금융 소비자들의 손에 익어가자 카카오가 다음 금융 플랫폼 타깃으로 보험을 겨냥했다. 하반기 빅테크와 보험이 결합한 디지털 손해보험사가 약 4500만명에 달하는 카카오톡이 카카오의 보험영업을 뒷받침해줄 전망이다. 또 코로나19 사태로 증폭된 시대적 변화를 등에 업고 정부가 추진하고 있는 마이데이터 사업이 잘개를 달아주며 전통 보험사들을 긴장하게 할 것으로 보인다. 이미 업계에서는 카카오톡을 무기로 카카오뱅크를 은행권 메기로 성공시킨 카카오의 보험사 설립을 주시하는 분위기다. 카카오뱅크는 불과 2년 만인 2019년에는 연간 기준 흑자전환에도 성공했고 은행업을 바라보는 소비자의 인식을 전환하는 데도 기여했다는 평가를 받기 때문이다. 카카오페이 역시 가입자 수 3500만명 MAU 2000만명을 넘어섰고 2020년 3분기까지 누적거래액 47조원으로 2019년 연간 거래액을 3분기만에 달성했다. 카카오의 새로운 금융 플랫폼이 될 보험은 자동차보험과 단기소액보험 분야부터 사업을 전개할 것으로 예상된다. 전통 보험사를 카카오페이가 단숨에 따라잡기는 어렵지만 고객 접근성을 바탕으로 20,30대부터 차근차근 세력을 넓혀나갈 것으로 보인다. 금융권 관계자는 카카오뱅크가 은행권의 디지털화에 방아쇠를 당겼다고 얘기할 정도로 카카오의 금융 서비스는 영향력이 크다며 번뜩이는 상품을 내놓는 것도 카카오 금융 서비스의 강점으로 작용할 것이다고 말했다. 권지예 기자 kwon.jiyejoongang.co.kr

Summarize

카카오뱅크와 카카오페이가 금융 소비자들의 손에 익어가자 카카오가 다음 금융 플랫폼 타깃으로 보험을 겨냥해 디지털 손해보험사가 약 4500만명에 달하는 카카오톡이 카카오의 보험 영여업을 뒷받침해줄 전망이며 카카오의 새로운 금융 플랫폼이 될 보험은 자동차보험과 단기소액보험 분야부터 사업을 전개할 것으로 예상되어 전통 보험사들을 긴장하게 할 것으로 보인다.

Sentiment Analysis

{ 'positive' : 0.8008, 'negative' : 0.1992}

Fig. 4. Examples of news data summary and sentiment analysis results using Pororo

석을 수행한다.

최종적으로 추출된 감성은 긍정과 부정에 대한 확률값으로 0과 1 사이의 값을 가지게 된다. 이 수치를 딥러닝 모델에 사용하게 되면 유의미 한 정보를 얻기 어렵기 때문에 감성 지표로 변환하는 과정이 필요하다. 가장 먼저 확률값의 범위를 0~1이 아닌 -∞~∞로 변환하기 위해 수식 1을 사용한다. P는 Pororo를 통해 계산된 긍정(positive) 값을, N은 부정(negative)값을 나타내며, 딥러닝에서 자주 사용되는 시그모이드 함수의 역수인 로짓(logit) 함수를 사용한다.

$$logit = \ln(\frac{P}{N}) \tag{1}$$

$$intensity_t = \ln\left(1 + \frac{n_t}{avg\ of\ N}\right) \tag{2}$$

주식 데이터는 시간 정보를 포함하고 있기 때문에 뉴스 데이터에서 감성을 분석할 때 단순하게 특정 기사의 감성만 추출하는 것이 아니라 시간 정보를 기반으로 한 분석이 필요하다. 특정 기업에 대한 기사가 많이 나타나는 일자는 해당 일자에 이슈가 있었음을 의미하며, 이러한 이슈는 주가 변동에 많은 영향을 미치게 된다. 감성 지표는 일자별 중요도를 계산하고 반영해야 한다. 이를 위해 수식 2를 사용한다. 특정 일자 t의 중요도인 $intensity$를 계산하기 위해 2021년 1월 1일부터 2022년 4월 30일까지 전체 기간에 작성된 뉴스 수의 평균인 $avg\ of\ N$과 해당 일자에 작성된 뉴스 수인 N_t를 활용한다.

$$sentiment\_index_t = logit * intensity_t \tag{3}$$

최종적으로 기업의 특정 일자에 대한 감성 지표 $sentiment_index_t$는 수식 3과 같이 표현할 수 있다. 수식 1로 계산된 감성 점수에 수식 2에서 계산한 특정 일자에 대한 중요도를 곱함으로써 감성 지표의 계산이 가능하다. 최종 계산된 감성 지표는 딥러닝 주가 예측 모델의 입력값으로 활용된다.

4. Deep learning stock price forecasting model

시간의 흐름에 따라 구성된 주식 데이터와 감성 분석을 위해 사용된 뉴스 데이터는 서로 다른 형태의 데이터이기 때문에 딥러닝 주가 예측 모델을 학습시키기 위한 입력으로 넣기 위해서는 반드시 2개의 데이터를 통합하는 과정이 필요하다. 본 논문에서는 감성 지표를 이용하여 타임 스탬프 t에 대한 주가를 예측하기 위하여 활용할 이전 데이터의 양에 따라 예측 모델의 성능이 어떻게 달라지는지에 대해 연구한다.

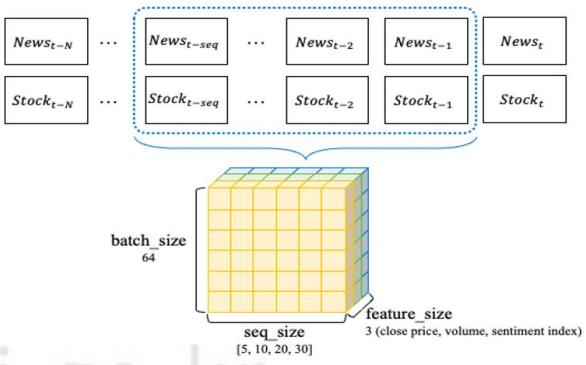


Fig. 5. Data integration based on seq_size

그림 5는 데이터 통합을 거친 후의 입력 데이터의 형태를 보여준다. 데이터 통합과 주가 예측 모델에 대한 연구를 위해 사용하는 seq_size는 예측하고자 하는 t일자를 기준으로 며칠 전까지의 데이터를 사용할지를 결정한다. seq_size에 따라 t 이전의 뉴스 데이터와 주식 데이터를 각각 가져온 후 모델에서 사용할 3가지 요소를 기반으로 통합한다. 이때 3가지 요소는 주식 데이터의 종가, 거래량과 뉴스 데이터의 감성 지표이다.

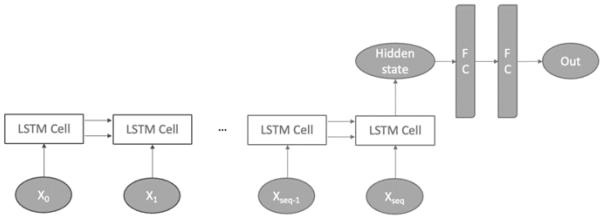


Fig. 6. Stock price prediction model architecture based on LSTM

본 논문에서는 LSTM과 ResNet 모델을 이용해 감성 지표가 주가에 미치는 영향에 대해 분석한다. LSTM(Long Short-Term Memory) 모델은 은닉 계층에 과거 데이터에 대한 정보를 담고 있어 시계열 데이터의 학습에 적합한 모델인 RNN 모델의 한 종류로 주가 예측에 많이 활용된다. 본 논문에서 활용한 LSTM은 그림 6과 같다. 기존 LSTM과 같은 구조이지만 출력값에 Fully Connected Layer(FC)를 추가하여 학습 매개변수를 증가시켰다. CNN 계열 모델은 주로 이미지의 패턴을 파악하는 데 사용된다. 기술적 분석의 전략은 차트를 통해 대중의 매수, 매도 심리를 역이용하여 주식시장에서 수익을 얻는 것이다[21]. 본 논문에서는 CNN계열 모델을 사용하여 차트를 통해 주가의 변동 패턴을 찾아내고자 하였고 CNN 계열 모델로는 ResNet을 사용한다.

ResNet모델은 CNN모델의 한 종류로 convolution layer를 깊게 쌓았을 때 발생하는 기울기 소실 현상을 해결하기 위해 residual block 구조를 도입한 모델이다. Residual block은 인접하지 않은 다음 convolution layer에 직접 학습 데이터를 넘겨주는 구조로 연산량 증가 없이 convolution layer를 더 많이 쌓기 위해 제안되었고, 많은 이미지 처리 분야에서 좋은 결과를 얻었다. 본 논문에서는 그림 7과 같이 2개의 convolution layer뒤에 3개의 residual block을 추가한 모델을 사용하였다.

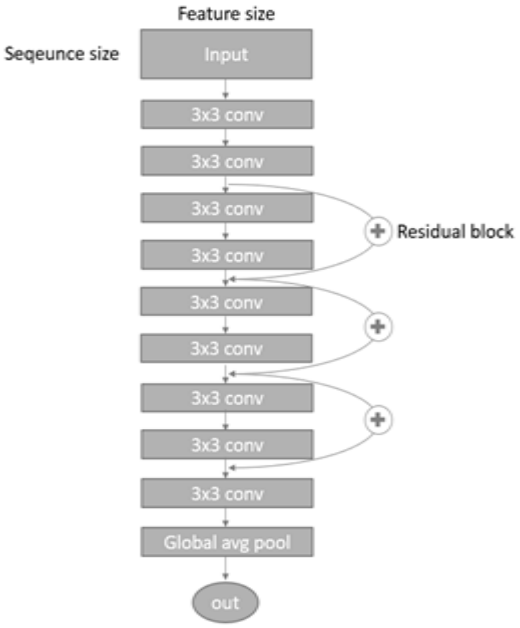


Fig. 7. Stock price prediction model architecture based on ResNet

IV. Experiment

1. Dataset

감성 지표를 이용한 딥러닝 주가 예측 분석 모델에 대한 연구를 위해 뉴스 데이터와 주가 데이터를 모두 사용하였다. 뉴스 데이터와 주가 데이터 모두 2021년 1월 1일부터 2022년 4월 30일까지 총 327일에 해당하는 데이터를 수집하였으며, 뉴스 데이터는 네이버 경제 뉴스에서 KOSPI 상위 50개 기업들에 대한 기사들을 수집하였다. 이 중 가장 많이 언급된 카카오, 삼성전자, LG전자, SK하이닉스를 선정하였다. 주가 데이터는 야후 파이낸스에서 동일한 기간에 대한 일별 데이터를 수집하였다.

Table 1. Statistics of data crawled by company

	Kakao (카카오)	Samsung Electronics (삼성 전자)	LG Electronics (LG 전자)	SK hynix (SK 하이닉스)
Number of articles	39235	35354	27018	26704
Number of stock datas	327	327	327	327
Total	39562	35681	27345	27031

표 1은 기업별로 수집된 데이터에 대한 통계를 보여준다. 실험을 위해 총 128,311개의 뉴스 데이터와 1,308개의 주가 데이터를 활용한다. 실험에 사용되는 데이터의 수가 한정되어 있기 때문에 K-fold 교차검증을 사용하며, 이때 k 값은 4로 결정하여 진행한다. 전체 데이터를 k개, 즉 4개로 나누어 한 번씩 검증 데이터로 사용하고 검증 데이터에 사용되지 않는 나머지 데이터를 훈련 데이터로 사용하여 모델을 검증한다.

2. Experiment environment

감성 지표를 이용한 딥러닝 기반 주가 예측 모델을 연구하기 위해 LSTM과 ResNet을 사용한다. 모델의 입력 변수로는 종가, 거래량의 주가 지표 2개와 감성 지표 1개가 사용된다. LSTM 모델의 경우 2개의 layer를 사용하였으며, dropout 비율은 0.2로, 1000 epoch을 통해 훈련을 진행하며, ResNet 모델은 500epoch를 사용하였다. 2개의 모델을 사용한 주가 예측 모델은 모두 평균 제곱근 편차(RMSE: Root Mean Square Error)를 손실 함수로 사용하였으며, 정규화를 위해 1e-5의 L2 정규화를 적용한다. 또한 learning rate scheduler를 사용하여 매 epoch마다 0.99만큼 감소할 수 있도록 진행한다. 특히 본 논문은 감성 지표를 이용한 딥러닝 기반 주가 예측에 있어서 영향을 미치는 데이터의 범위를 파악하기 위해 seq_size를 5, 10, 20, 30으로 설정하여 각 모델에 대한 실험을 진행하였다.

3. Results

표 2는 sequence size에 따른 모델별 loss와 정확도를 비교한 결과이다. 주가 지표만을 사용한 모델의 정확도는 LSTM이 평균 약 49.6%, ResNet이 약 50.5%로 ResNet의 정확도가 약 0.9%만큼 더 높게 나타났다. 감정 지표를 사용한 모델의 정확도 또한 LSTM이 평균 약 49.6%, ResNet이 약 51.7%로 ResNet의 정확도가 더 높게 나타났다. 반면 loss 값의 경우 LSTM이 감정 지표를 사용하지 않았을 때와 감정 지표를 사용하였을 때를 비교해보면 각각 2.13과 2.12이며, ResNet의 경우 2.25와 2.36으로 [11]에서 언급했듯이 금융 시장에서는 loss의 감소가 실용성의 증가, 즉 정확도의 향상으로 이어지지 않음을 확인할 수 있다.

Table 2-1. Loss and accuracy by sequence size at LSTM

seq_size	value	Kakao (카카오)	Samsung Electronics (삼성 전자)	LG Electronics (LG 전자)	SK hynix (SK 하이닉스)
5	loss	2.4372	1.3591	2.5914	2.2903
	loss + sent	2.4455	1.3313	2.5881	2.2875
	acc	0.4412	0.5218	0.5001	0.531
	acc + sent	0.4412	0.5373	0.5001	0.5032
10	loss	2.4385	1.2991	2.6097	2.2815
	loss + sent	2.4431	1.2798	2.6036	2.2693
	acc	0.4292	0.4859	0.5206	0.486
	acc + sent	0.4292	0.5046	0.5331	0.486
20	loss	2.3904	1.3043	2.4274	2.2546
	loss + sent	2.3874	1.278	2.4191	2.2404
	acc	0.4949	0.4853	0.482	0.5013
	acc + sent	0.4949	0.495	0.4852	0.5013
30	loss	2.4145	1.3319	2.4162	2.2711
	loss + sent	2.4072	1.2888	2.4094	2.247
	acc	0.5419	0.4882	0.505	0.5353
	acc + sent	0.5419	0.4882	0.505	0.5053

Table 2-2. Loss and accuracy by sequence size at ResNet

seq_size	value	Kakao (카카오)	Samsung Electronics (삼성 전자)	LG Electronics (LG 전자)	SK hynix (SK 하이닉스)
5	loss	2.6438	1.4579	2.7204	2.3841
	loss + sent	2.5355	1.5318	3.1424	2.9762
	acc	0.5161	0.5216	0.5218	0.4877
	acc + sent	0.5527	0.5311	0.531	0.509
10	loss	2.5995	1.2817	2.7107	2.4698
	loss + sent	2.5837	1.2819	2.7937	2.6765
	acc	0.4858	0.4951	0.5553	0.5078
	acc + sent	0.5488	0.5046	0.5237	0.4889
20	loss	2.8084	1.3087	2.5362	2.367
	loss + sent	2.6822	1.3284	2.7009	2.4095
	acc	0.5538	0.4427	0.4335	0.5377
	acc + sent	0.5505	0.4883	0.4657	0.5538
30	loss	2.6231	1.2617	2.4348	2.3957
	loss + sent	2.5649	1.4601	2.4249	2.7539
	acc	0.5252	0.5253	0.5151	0.458
	acc + sent	0.5487	0.4578	0.5253	0.5016

이어서 표 3은 감정 지표를 사용했을 때의 loss와 정확도 증가율을 비교한 표이다. 감성 지표를 변수로 포함하여 모델을 훈련시켰을 경우, LSTM의 평균 loss는 약 0.7% 감소하였으며 평균 정확도는 약 0.06% 증가하였다. ResNet의 경우 평균 loss가 약 5% 증가하였지만, 평균 정확도는 약 2.7% 증가하였음을 확인할 수 있다. 이는 주가와 감정 지표 사이의 복잡한 비선형적 패턴을 LSTM이 충분히 추출하지 못했고, ResNet이 금융 시장의 패턴 추출에 더 뛰어난 성능을 가진다고 생각된다.

Table 3. Loss and Accuracy Growth

	Model	Seq_size	Kakao (카카오)	Samsung Electronics (삼성 전자)	LG Electronics (LG 전자)	SK hynix (SK 하이닉스)	Avg
Increase of Loss(%)	LSTM	5	0.341	-2.045	-0.127	-0.122	-0.489
		10	0.189	-1.486	-0.234	-0.535	-0.516
		20	-0.126	-2.016	-0.342	-0.630	-0.778
		30	-0.302	-3.236	-0.281	-1.061	-1.220
		Avg	0.025	-2.196	-0.246	-0.587	-0.751
	ResNet	5	-4.096	5.069	15.512	24.835	10.330
		10	-0.608	0.016	3.062	8.369	2.710
		20	-4.494	1.505	6.494	1.796	1.325
		30	-2.219	15.725	-0.407	14.952	7.013
		Avg	-2.854	5.579	6.165	12.488	5.344
Increase of Accuracy (%)	LSTM	5	0.000	2.970	0.000	-5.235	-0.566
		10	0.000	3.849	2.401	0.000	1.562
		20	0.000	1.999	0.664	0.000	0.666
		30	0.000	0.000	0.000	-5.604	-1.401
		Avg	0.000	2.204	0.766	-2.710	0.065
	ResNet	5	7.092	1.821	1.763	4.367	3.761
		10	12.968	1.919	-5.691	-3.722	1.369
		20	-0.596	10.300	7.428	2.994	5.032
		30	4.474	-12.850	1.980	9.520	0.781
		Avg	5.985	0.298	1.370	3.290	2.736

또한 LSTM의 경우 sequence size가 10일 때 약 1.56%, ResNet의 경우 sequence size가 20일 때 약 5%로 정확도 상승률이 가장 크게 나타났음을 확인할 수 있다. 이는 모델의 구조에 따라 알맞은 sequence size가 다름을 확인할 수 있다.

V. Conclusions

본 논문은 주가 등락률 예측 향상을 위해 딥러닝의 대표적인 모델인 RNN과 CNN계열 모델을 사용하여 주가 지표와 뉴스 감성 지표와의 효용성과 함께 효율적인 sequence size를 제시하였다. KOSPI 상위 50개 기업 중 네이버 경제 뉴스 기사 수가 가장 많았던 4개 기업을 대상으로 모델을 제작하였으며, 주가 지표는 일일 종가와 거래량을 사용하였다. 감성 지표는 4개 기업을 키워드로 네이버 뉴스 본문을 크롤링한 뒤 한국어 BERT 모델인 Pororo를 사용하여 Summarization 작업을 수행함으로써 본문을 정제해주고, 이어서 Pororo의 감성 분석 작업을 통해 감성 정보를 추출하였다. 그리고 추출한 감성 정보와 크롤링한 뉴스 기사 수를 사용하여 감성 지표를 계산하였다. 최종적으로 감성 지표를 변수로 추가하였을 경우 LSTM의 sequence size가 10일 때, 그리고 ResNet의 sequence size가 20일 때, 가장 큰 정확도 상승률을 보였다.

본 논문에서 진행한 연구를 통해 포괄적으로 대중의 인식을 반영하는 감정 지표를 제시하였다는 점과 주가 지표와 감정 지표에서 가장 효과적으로 패턴을 추출할 수 있는 sequence size를 제시했다는 것이다. 하지만, 실험에서 4개 기업만을 사용하여 상관관계를 추종하였기 때문에 통계적 신뢰도가 떨어진다는 점이 있으며, 다양한 주가 지표 중 종가와 거래량만을 사용하였기 때문에 국제 금값, 유가, 환율 등과 같은 다른 거시적 지표는 고려하지 못하였다는 한계점이 존재한다.

따라서 향후 연구로 더 다양한 기업들과 지표들을 추가한 모델을 분석함으로써 본 논문의 한계점을 보완해 나갈 예정이다.

ACKNOWLEDGEMENT

This work was supported by Gachon University research fund 2022 (GCU-202206850001)

REFERENCES

[1] B. G. Malkiel, “Efficient market hypothesis,” The New Palgrave: Finance. Norton, New York, pp. 127-134, 1989.

[2] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), pp.

205-208, 2019.

[3] Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N. “Explainable stock prices prediction from financial news articles using sentiment analysis,” PeerJ Computer Science 7:e340, 2021.

[4] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li and F. L. Wang, "Does summarization help stock prediction? a news impact analysis," IEEE Intelligent Systems, Vol. 30, No. 3, pp. 26-34, 2015.

[5] D. Duong, T. Nguyen and M. Dang, "Stock market prediction using financial news articles on ho chi minh stock exchange," Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication ser. IMCOM ‘16, pp. 71:1-71:6, 2016.

[6] Y. Wang, D. Seyler, S. K. K. Santu and C. Zhai, "A study of feature construction for text-based forecasting of time series variables," Proceedings of the 2017 ACM on Conference on Information and Knowledge Management ser. CIKM ‘17, pp. 2347-2350, 2017.

[7] H. White, "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, ser. Discussion paper-" Department of Economics University of California San Diego. Department of Economics, University of California, 1988.

[8] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," Expert Systems with Applications, Vol. 129, pp. 273-285, 2019.

[9] Guo, Baicun. "Research on Stock Price Forecast based on Resnet and LSTM," Frontiers in Economics and Management, Vol. 3, No. 5, pp. 161-167, 2022.

[10] K. A. Althelaya, E.-S. M. El-Alfy, and S. Mohammed, "Evaluation of Bidirectional LSTM for Short and Long-Term Stock Market Prediction," Proc. of the 9th International Conference on Information and Communication Systems, pp. 151- 156, 2018.

[11] Yao, Yifan, and Lina Wang. "Combination of window-sliding and prediction range method based on LSTM model for predicting cryptocurrency," arXiv preprint arXiv:2102.05448, 2021.

[12] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Transactions on Information Systems (TOIS), Vol.27, No. 2, pp. 1-19, 2009.

[13] Tetlock, Paul C., Maytal Saar‐Tsechansky, and Sofus Macskassy. "More than words: Quantifying language to measure firms' fundamentals," *The journal of finance* Vol. 63, No. 3, pp. 1437-1467. 2008.

[14] Li, Xiaodong, et al. "Does summarization help stock prediction? A news impact analysis," *IEEE intelligent systems* Vol. 30, No. 3, pp. 26-34, 2015.

[15] Google Finance, *https://www.google.com/finance/quote/KOSPI:KRX?hl=ko*

[16] N. V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets," ACM SIGKDD Explor. *Newslett*., Vol. 6, No. 1, pp. 1-6, 2004.

[17] KoreanNewsCrawler Open source, *https://github.com/lumyjuwon/KoreaNewsCrawler*

[18] Naver Finance News, *https://news.naver.com/main/main.naver?mode=LSD&mid=shm&sid1=101*

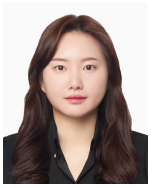
[19] Yahoo Finance, *https://finance.yahoo.com/*

[20] Pororo, *https://github.com/kakaobrain/pororo*

[21] Seungwoo Kim, "Can You Beat the Market? Characterizing the Market and a History of Investment Methods in the 20th Century," Critical Review of History, No. 138, pp. 8-35, 2022.

Authors

Doo-Won Kang is an undergraduate student of School of Computing at Gachon University, Seongnam, Korea. He is interested in machine learning and deep learning.



So-Yeop Yoo received the B.S. M.S. and Ph.D. degrees in Software from Gachon University, Korea, in 2014, 2016, and 2021, respectively. Dr. Yoo joined the faculty of the School of Computing at Gachon University, Seongnam, Korea, in 2021. She is currently a visiting professor in the School of Computing, Gachon University. She is interested in conversational AI, NLP, knowledge graph and emotional AI.

Ha-Young Lee received a B.S. degree in the School of Computing from Gachon University, Korea in 2022. She is currently an M.S. student in the School of Computing at Gachon University. She is interested in natural language processing, chat-bot and conversational AI.

Ok-Ran Jeong received Ph.D. degrees in Computer Science and Engineering from Ewha Womans University, Korea, in 2005. She was a postdoctoral researcher at the University of Illinois at Urbana-Champaign, USA and Seoul National University, Korea. Dr. Jeong joined the faculty of the Department of Software Design & Management at Gachon University, Seongnam, Korea, in 2009. She is currently a Professor in the School of Computing, Gachon University. She is interested in big data mining, machine learning, deep learning and applications of artificial intelligence.