JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 7, pp. 73-80, July 2024
https://doi.org/10.9708/jksci.2024.29.07.073

KOSPI index prediction using topic modeling and LSTM

Jin-Hyeon Joo*, Geun-Duk Park**

*Assist Professor, Dept. Convergence, Hoseo University, Asan-si, Korea
**Professor, Dept. of Computer Engineering, Hoseo University, Asan-si, Korea

[Abstract]

In this paper, we proposes a method to improve the accuracy of predicting the Korea Composite Stock Price Index (KOSPI) by combining topic modeling and Long Short-Term Memory (LSTM) neural networks. In this paper, we use the Latent Dirichlet Allocation (LDA) technique to extract ten major topics related to interest rate increases and decreases from financial news data. The extracted topics, along with historical KOSPI index data, are input into an LSTM model to predict the KOSPI index. The proposed model has the characteristic of predicting the KOSPI index by combining the time series prediction method by inputting the historical KOSPI index into the LSTM model and the topic modeling method by inputting news data. To verify the performance of the proposed model, this paper designs four models (LSTM_K model, LSTM_KNS model, LDA_K model, LDA_KNS model) based on the types of input data for the LSTM and presents the predictive performance of each model. The comparison of prediction performance results shows that the LSTM model (LDA_K model), which uses financial news topic data and historical KOSPI index data as inputs, recorded the lowest RMSE (Root Mean Square Error), demonstrating the best predictive performance.

▸ **Key words**: Topic Modeling, LSTM, Machine Learning, Predictive Modeling, LDA

[요 약]

본 연구는 토픽 모델링과 장단기 기억(LSTM) 신경망을 결합하여 한국 종합주가지수(KOSPI) 예측의 정확도를 향상하는 방법을 제안한다. 본 논문에서는 LDA(Latent Dirichlet Allocation) 기법을 이용해 금융 뉴스 데이터에서 금리 인상 및 인하와 관련된 10개의 주요 주제를 추출하고, 추출된 주제를 과거 KOSPI 지수와 함께 LSTM 모델에 입력하여 KOSPI 지수를 예측하는 모델을 제안한다. 제안된 모델은 과거 KOSPI 지수를 LSTM 모델에 입력하여 시계열 예측 방법과 뉴스 데이터를 입력하여 토픽 모델링 하는 방법을 결합하여 KOSPI 지수를 예측하는 특성을 가진다. 제안된 모델의 성능을 검증하기 위해, 본 논문에서는 LSTM의 입력 데이터의 종류에 따라 4개의 모델(LSTM_K 모델, LSTM_KNS 모델, LDA_K 모델, LDA_KNS 모델)을 설계하고 각 모델의 예측 성능을 제시하였다. 예측 성능을 비교한 결과, 금융 뉴스 주제 데이터와 과거 KOSPI 지수 데이터를 입력으로 하는 LSTM 모델(LDA_K 모델)이 가장 낮은 RMSE(Root Mean Square Error)를 기록하여 가장 좋은 예측 성능을 보였다.

▸ **주제어**: 토픽 모델링, LSTM, 머신러닝, 예측 모델링, LDA

• First Author: Jin-Hyeon Joo, Corresponding Author: Geun-Duk Park
*Jin-Hyeon Joo (joojin4381@hoseo.edu), Dept. Convergence, Hoseo University
**Geun-Duk Park (gdpark@hoseo.edu), Dept. of Computer Engineering, Hoseo University
• Received: 2024. 06. 11, Revised: 2024. 07. 01, Accepted: 2024. 07. 02.

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

I. Introduction

금융시장에서 주가를 예측하는 것은 투자자들의 투자 결정, 기업 가치 평가, 경제 분석 및 리스크 관리 등에 주가 정보를 사용하기 때문에 중요하다. 그러나 주가는 다양한 요인들로 결정되기 때문에 주가를 예측하기는 어려운 문제이다[9].

주가 예측은 많은 연구가 이루어졌는데, 과거에는 변동성 지수와 같은 통계적인 방법이 주가 예측 연구에 주로 사용되었다[7, 8]. 하지만, 하드웨어의 발달과 인공 지능에 관한 연구가 이루어지면서 다양한 신경망 모델을 이용한 주가 예측 연구가 활발하게 이루어졌다. 신경망을 이용한 주가 예측에는 주가가 시계열 데이터임을 주목하여 과거 주가를 학습시킨 후 미래의 주가를 예측하는 방법이 많이 사용되었다.

하지만, 주가는 앞서 언급한 바와 같이 다양한 요인들에 의하여 결정되기 때문에 과거 주가 데이터만 가지고 미래의 주가 데이터를 예측하기는 쉬운 작업이 아니다. 이를 해결하기 위하여 승현수(2021)는 LSTM, GRU(Gated Recurrent Unit)에 CNN(Convolution Neural Network)을 결합하는 모델을 제안하여 예측 정확도를 높였다[1]. 이강훈(2003)은 뉴스 기사를 기반으로 경제 심리를 지수화한 NSI(News Sentiment Index), 기업경기실사지수와 소비자 동향 지수를 합성한 ESI(Economic Sentiment Index) 그리고 변동성 지수인 V-KOSPI(KOSPI Volatility Index)를 Lasso 회귀 모델과 결합하여 KOSPI 지수를 예측하는 연구를 진행하였다[2]. 김성곤(2023)는 KRX 증시 Brief에 반복적으로 등장하는 시작, 출발, 마감과 같은 시점 관련 키워드를 중심으로 시점을 구분하고, 단어의 출현 빈도와 KOSPI 지수 등락 여부를 기반으로 시점별 감성 사전을 구축하여 각 Brief의 시점별 감성지수를 도출하여 다음 날 KOSPI 지수를 예측하였다[3]. 장성희(2023)은 KOSPI 지수 방향성 예측의 정확도 향상을 위하여 자연어 처리 기술인 KoBERT를 적용한 BERT 모델을 사용하여 한국어 뉴스의 감성을 분석하고, LSTM 모델을 적용함으로써 텍스트 감성 분석과 자연어 처리 기술인 KoBERT를 결합하여 KOSPI 지수 방향 예측력을 향상할 수 있다는 사실을 확인하였다[4]. 강두원(2022)은 웹 크롤링을 통해 온라인상에서 소비자들의 기업에 대한 인식에 관한 데이터 수집 후 감성 분석 단계를 거친 후, 해당 데이터에 딥러닝 기술을 적용하여 소비자들의 기업 활동에 대한 인식에 따라 기업의 주가가 어떻게 변화하는지 예측하고자 했다[5].

기존 연구들은 주로 뉴스 데이터로부터 감성 데이터를 얻어오는 방법을 활용하고 있다. 경제적인 상황이 좋으면 긍정적인 뉴스가 증가하고, 경제 상황이 나쁘면 부정적인 뉴스가 늘어나는 것에 기반한 방법이다. 하지만, 언론은 경제가 나빠지는 상황에서 부정적인 경제 뉴스를 지나치게 과장하는 경향이 있다[6]. 따라서, 감성 데이터를 이용하여 KOSPI 지수를 예측하면 부정적인 부분이 강조될 가능성이 크다. 이에 본 연구에서는 감성 분석 대신에 문서의 주제 빈도수만을 입력 데이터로 사용하였다.

또한, 본 연구에서는 뉴스 데이터에 대해 필터링을 적용하였다. 뉴스 데이터에는 다양한 분야가 존재하고, 각 분야에 해당하는 이슈에 민감하게 반응한다. 따라서, 주가 예측을 하는 데 있어 주가와 관련된 뉴스 데이터를 활용하면 주가를 예측하는 데 큰 도움을 받을 수 있지만 주가와 상관없는 뉴스 데이터를 사용하면 주가를 예측하는 데 오히려 방해될 수 있다.

본 연구에서는 다음과 같은 조건으로 뉴스 데이터를 필터링하여 사용하였다. 1) 금리와 관련되어 있을 것 2) 언론사는 경제와 관련되어 있을 것. 첫 번째 조건의 경우 금리는 주가와 밀접한 관련이 있어서 금리와 관련된 뉴스를 선택하였다. 그리고 제목이나 내용 내에 “금리 인상”, “금리 인하”가 포함된 뉴스 데이터를 사용하였으며, 언론사는 매일경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제 이상 8개 언론사의 뉴스 데이터를 사용하였다.

필터링된 뉴스 데이터는 금리 인상과 금리 인하, 두 가지 유형으로 나누어 토픽 모델링(Topic Modeling)으로 주제(Topic)를 추출해 LSTM(Long Short Term Memory)을 이용하여 주가 예측에 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련된 연구에 관하여 기술하고, 3장에는 제안된 모델의 세부 구현에 관해 설명한다. 4장에는 제안된 모델의 실험 결과를 통해 성능 분석을 한 후 5장에서 결론을 맺는다.

II. Preliminaries

1. Related works

1.1 Topic Modeling

모든 문서에는 표현하고자 하는 것이 있고, 표현하고자 하는 것과 관련된 단어는 문서에서 관련 없는 단어들보다 빈번하게 등장한다. 토픽 모델링은 문서에서 자주 등장하는 단어들을 추출하여 문서에서 표현하고자 하는 것, 즉 주제(Topic)를 추측하는 비지도 학습 분류 방법론이다.

Papadimitriou, Raghavan, Tamaki, Vempala(1998)

는 문서와 단어 간 행렬(Matrix)을 통해 잠재 의미를 도출하기 위하여 LSI(Latent Semantic Indexing) 알고리즘을 개발하였고, 추후 Blei, Ng, Jordan(2003)가 현재 많이 사용되고 있는 LDA(Latent Dirichlet Allocation)을 제시하였다.

그 외에도 문서 내 응답 변수(Response Variable)를 통해 문서를 분류할 수 있도록 하는 sLDA(supervised LDA), 문서 내 메타데이터(meta-data)를 활용하여 메타데이터와 토픽 간의 상관관계 추정 및 토픽 간 관계를 구분하여 해석할 수 있는 STM(Structural Topic Model) 그리고 LSI와 LDA와 달리 시간의 흐름에 따라 토픽의 내용 변화 파악에 활용되는 DTM(Dynamic Topic Model)이 있다.

1.2 LDA(Latent Dirichlet)

LDA는 잠재 딜리클레(Latent Dirichlet) 확률 기반의 비지도 학습을 통해 문서의 주제를 파악하여 문서를 분류할 수 있는 비지도 학습 방법론이다.

LDA는 문서 내 단어의 빈도수를 표현하는 데이터 BoW(Bag of Word) 혹은 TF-IDF를 사용하여 전체 문서 내에서 각 문서가 나타내는 주제를 딜리클레 분포를 이용해서 알아내게 된다.

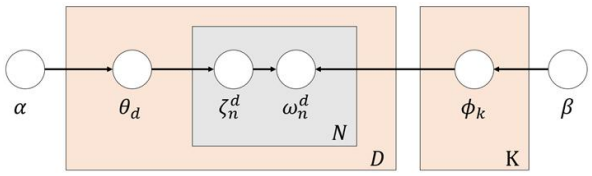

Fig. 1. The process of selecting topics using LDA (Latent Dirichlet Allocation)

<Fig. 1>의 그림은 LDA에서 주제를 선정하는 방법에 대한 것으로, 문서는 N개 단어의 연속을 나타내며 $w=(\omega_1,\omega_2,...,\omega_N)$으로 표기한다. 전체 문서 M은 문서의 집합을 나타내며 $D=(w_1, w_2, \dots, w_M)$으로 나타낸다. 모든 문서는 각각 k개의 주제에 속할 때, α는 k 차원 딜리클레 분포의 매개 변수이고, β는 $k\times V$ 크기의 매개 변수로 β_{ij}는 i번째 주제가 단어집(V)의 j번째 단어를 생성할 확률을 나타낸다.

주어진 α에 대해서 딜리클레 분포 $Dir(\alpha)$에서 d번째 문서에 대한 주제 가중치 θ_d를 무작위로 추출하고, 주어진 β에 대해서 딜리클레 분포 $Dir(\beta)$에서 k번째 주제에 대한 가중치 ϕ_k를 추출한다. ζ_n은 k 차원 벡터로 ζ_n^d는 단어 ω_n이 d번째 주제에 속할 확률 분포를 나타낸다.

위와 같은 과정을 여러 문서를 거쳐 진행하면서 확률이 안정적으로 변하게 되면 이를 통해 각 문서가 어떤 주제를 담고 있는지 확률적으로 추정할 수 있게 된다.

1.3 RNN(Recurrent Neural Network)

신경망은 생물학적 뇌의 작동 원리에서 영감을 받아 만들어진 컴퓨터 모델로, 데이터 처리 및 학습에 사용된다. 이 모델은 인공 신경망(ANN)이라고도 불리며, 기계 학습과 인공 지능 분야에서 주요 기술로 사용된다.

인공 신경망은 여러 개의 뉴런(또는 노드)으로 구성된 계층적인 네트워크로 각 뉴런은 입력받은 가중치와 함께 연산을 수행하고, 그 결과를 다음 뉴런으로 전달한다. 각각의 계층은 입력층, 은닉층, 출력층으로 나눌 수 있다. 입력층은 데이터를 받고, 출력층은 최종 결과를 내보내며, 은닉층은 입력과 출력 사이의 중간 처리를 담당한다.

신경망은 학습 데이터를 통해 가중치를 조정하여 원하는 작업을 수행할 수 있다. 주요 학습 방법으로는 신경망의 출력과 실제 값 사이의 오차를 계산하고, 이 오차를 줄이기 위해 가중치를 조정하는 과정을 반복하여 학습 데이터에서 패턴을 학습하고, 이를 새로운 데이터에 일반화하는 역전파(backpropagation) 방법이 있다.

순환 신경망(Recurrent Neural Network, RNN)은 문장, 음성, 주가 시계열 데이터 등의 입력 순서나 시간적 의존성을 가지는 데이터인 시퀀스 데이터를 처리하는 데 사용되는 인공 신경망의 한 유형이다.

RNN은 순환 구조를 하고 있어서 이전 단계의 출력을 현재 단계의 입력으로 사용하기 때문에 이전 정보를 기억하고 새로운 입력에 대해 이전 상태를 고려하여 출력을 생성한다. 따라서 RNN은 시퀀스 데이터의 패턴 및 의미를 파악하는 데 효과적이다.

1.4 LSTM(Long Short Term Memory)

RNN은 문장 생성, 기계 번역, 감성 분석, 주식 시장 예측, 날씨 예측 등 다양한 작업에 사용된다. 그러나 RNN에는 장기 의존성(Long Term Dependencies) 문제가 있어서 시간이 오래 지난 이전 정보를 기억하는 데 어려움을 겪는 경우가 있다.

위와 같은 문제를 해결하기 위해 LSTM(Long Short Term Memory)과 GRU(Gated Recurrent Unit)와 같은 변형된 RNN 아키텍처가 개발되었다.

LSTM(Long Short Term Memory)은 순환 신경망(RNN)의 한 종류로, 장기 의존성 문제를 해결하기 위해 제안된 구조이다. LSTM은 RNN의 은닉 상태에 게이트 메

커니즘(gate mechanism)을 도입하여 장기 및 단기 메모리를 관리한다.

LSTM은 RNN의 기본 셀(cell)에 망각, 입력, 출력 3개의 게이트를 하나의 셀을 이루는 구조를 갖는다. 1) 입력 게이트(Input Gate)는 현재 입력이 얼마나 중요한지를 결정한다. 시그모이드 함수를 사용하여 0과 1 사이의 값을 출력한다. 2) 망각 게이트(Forget Gate)는 이전 메모리 상태의 어느 부분을 잊을지를 결정한다. 시그모이드 함수를 사용하여 0과 1 사이의 값을 출력하며, 0은 해당 정보를 완전히 잊는 것을 의미한다. 3) 출력 게이트(Output Gate)눈 현재 셀 상태가 다음 은닉 상태에 어떻게 반영될지를 결정한다. 시그모이드 함수를 사용하여 0과 1 사이의 값을 출력하며, 하이퍼볼릭 탄젠트 함수를 사용하여 현재 셀 상태를 -1과 1 사이의 값으로 변환한 후 이와 출력 게이트의 결과를 곱하여 최종 출력을 생성한다.

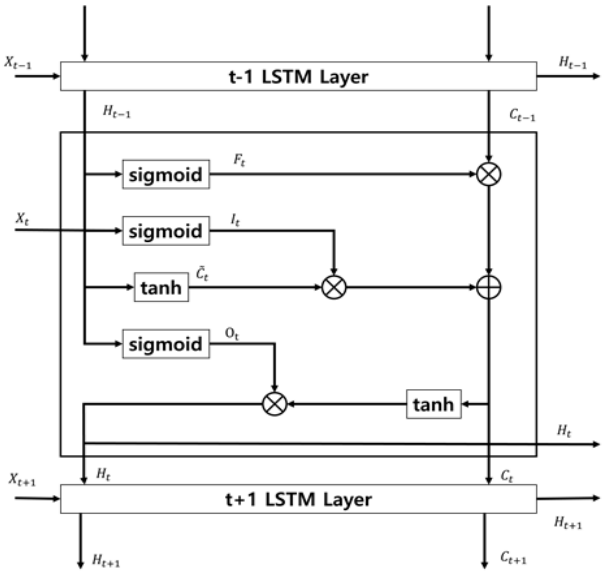

Fig. 2. Outline of the LSTM (Long Short-Term Memory) architecture

<Fig. 2>에서 C는 셀 상태(Cell state)로 현재 메모리 상태를 나타내며 입력 게이트와 망각 게이트를 사용하여 셀 상태가 업데이트된다. LSTM의 첫 과정으로 망각 게이트$F_t = \sigma(H_{t-1}, x_t)$의 값에 의해 이전 셀 상태의 망각 여부를 결정한다. 다음으로 입력 게이트 $I_t = \sigma(H_{t-1}, x_t)$와 새로운 셀 상태인 $\tilde{C} = \tanh(H_{t-1}, x_t)$을 사용하여 셀 상태를 업데이트한다. 마지막 과정으로 출력 게이트 $O_t = \sigma(H_{t-1}, x_t)$를 이용하여 $H_t = O_t \times \tanh(C_t)$를 결정해서 다음 LSTM 단계로 H_t와 C_t 값을 넘긴다.

III. The Proposed Scheme

본 연구에서는 KOSPI200 주가 예측의 정확성 향상을 위하여 기존에 KOSPI 주가 데이터만 LSTM의 입력 데이터로 사용하였던 것에 LDA를 사용하여 뉴스 데이터로부터 주제를 추출하고, 추출된 주제 데이터를 LSTM의 추가적인 입력 데이터로 사용하였다. 사용된 입력 데이터는 KOSPI200의 과거 주가 데이터와 금리 인상과 관련된 뉴스 데이터, 그리고 금리 인하와 관련된 뉴스 데이터가 사용되었다.

뉴스 데이터는 빅카인즈(Big Kinds) 사이트에서 제공하는 데이터를 사용하였으며, 약 2019년 3월부터 2024년 3월까지 약 5년간의 데이터를 연구에 사용하였다. 뉴스 데이터는 주가와 관련성이 높은 경제지 8개 (매일경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제)에 쓰인 뉴스 데이터를 선정하였다. 또한 주가와의 관련성을 강화하기 위하여 뉴스 데이터를 검색할 때 "금리 인상" 과 "금리 인하" 키워드를 사용하여 주가와 연관성이 높은 금리와 관련된 뉴스 데이터를 뽑아낼 수 있도록 하였다.

본 연구의 진행 순서는 한국어 전처리, LDA를 사용한 주제 추출, LSTM 학습, 주가 예측 순으로 진행되었다. 한국어 전처리에는 KoNLPy가 사용되었으며, 너무 짧거나 명사가 아닌 단어는 무의미하므로, 의미가 있는 2글자 이상의 명사만을 추출하였다. 추출된 명사를 LDA에 사용하여 주제를 추출한다. 이때, 각 주제가 뜻하는 바를 명확하게 하려고 금리 인상, 금리 인하와 각각 연관된 5개의 주제로 분류하여 총 10개의 주제로 분류한다. 분류된 주제는 금리 인상 관련 주제는 KOSPI 주가와 양의 관계를 맺고, 반대로 금리 인하 관련 주제는 KOSPI 주가와 음의 상관관계를 가진다.

<Table 1>은 금리 인상 관련 주제에 포함된 명사 10개이다. 각 주제는 “한국의 부동산 시장 전망”, “대출 및 금융 지원 정책에 대한 전망”, “국내 및 미국 시장에 대한 투자 전망과 채권 및 증권 투자 관련 상황”, “미국 연방준비제도의 기준금리 인하 및 인상 전망”, “코스피와 뉴욕증시의 지수 상승과 하락에 대한 거래 포인트 및 전망”을 나타낸다.

Table 1. Nouns on topics related to interest rate hikes

주제-1	주제-2	주제-3	주제-4	주제-5
전망	대출	투자	금리	지수
부동산	금리	시장	미국	증시
서울	금융	국내	인하	상승
아파트	은행	미국	연방	하락
금리	지원	채권	인상	코스피
올해	신용	증권	준비	거래
시장	이자	금리	현지	포인트
주택	인하	올해	제도	미국
정부	자금	펀드	시간	마감
한국 은행	담보	상장	기준 금리	뉴욕 증시

<Table 2>는 금리 인하 관련 주제에 포함된 명사 10개이다. 각 주제는 “서울 부동산 시장과 아파트 거래에 관한 정부 정책 및 가격 전망”, “금리와 지수에 따른 주가 상승과 하락에 대한 거래 포인트 및 코스피 마감에 관한 은행의 전망”, “한국경제의 물가와 경기에 따른 한국은행의 기준금리 인상 또는 동결에 대한 전망”, “미국 연방준비제도의 시간에 따른 중앙은행 금리 인상과 시장의 준비에 관한 현지 전망”, “국내 기업의 금융 투자와 대출에 관한 은행의 금리 변동에 대한 전망”을 나타낸다.

Table 2. Nouns on topics related to interest rate cuts

주제-1	주제-2	주제-3	주제-4	주제-5
부동산	금리	금리	미국	투자
주택	지수	물가	금리	금융
서울	하락	경제	연방	대출
시장	증시	전망	시간	기업
아파트	상승	한국 은행	현지	금리
가격	거래	인상	인상	은행
분양	포인트	기준 금리	준비	올해
건설	코스피	동결	제도	지난해
정부	은행	경기	중앙 은행	자금
거래	마감	올해	시장	국내

<Table 1>과 <Table 2>에 표현된 바와 같이 주가와 관련이 깊은 부동산, 금리, 한국은행, 미국 기준금리 등이 포함되어 있으며, 이를 통해 각 주제에 따른 뉴스 분류가 주가 예측에 큰 도움이 된다는 것을 알 수 있다.

추출된 주제에 따라 각각의 뉴스 데이터가 어느 주제에 속하는지 분류하고, 주제당 분류된 뉴스 데이터의 개수 그리고 이전 KOSPI 지수를 LSTM의 입력 데이터로 이용한다. 입력 데이터의 window size는 30으로 설정하였으며 epochs는 1000으로 설정하였다. 여기서 window size는 시계열 데이터나 다른 유형의 데이터 스트림에서 사용되는 데이터 포인트의 그룹으로 정의한다.

IV. Experiments

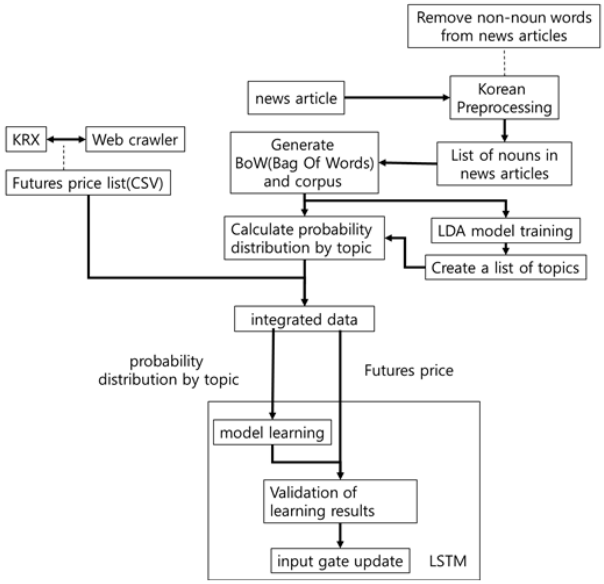

Fig. 3. Proposed Model

본 연구를 진행하기에 앞서 먼저 어느 정도의 window size를 사용하는 것이 적절한지 테스트를 진행하였다. 테스트에는 각 10, 20, 30, 40, 50 window size로 총 7차례 테스트를 진행하였다. 테스트 진행 결과는 <Table 3>과 같다. 평균적인 값은 window size 10을 제외하고는 근소한 차이를 가졌기 때문에 최소값을 기준으로 window size 30이 2.73의 가장 작은 값을 가졌다.

Table 3. Test result by window size

window size	10	20	30	40	50
1st	4.03	2.85	2.79	3.52	3.10
2nd	4.67	3.17	2.73	2.87	4.29
3rd	7.22	3.56	3.12	3.55	3.18
4th	4.21	3.92	3.14	3.45	3.23
5th	2.95	3.57	4.53	3.31	3.05
6th	11.66	2.80	4.10	3.58	3.16
7th	5.22	2.80	3.01	3.36	3.18
Average	5.71	3.24	3.35	3.38	3.31

본 연구에서는 window size 30을 기준으로 총 4가지 모델에 대해서 비교 분석을 진행하였다. 첫 번째 모델은 KOSPI 데이터만을 입력 데이터로 사용하여 LSTM을 통해 KOSPI 데이터를 예측하였다(LSTM_K 모델). 두 번째는 KOSPI와 연관성이 깊은 KOSPI 데이터와 함께 NDX(나스닥-100) 지수와 SPX(S&P-500) 지수 데이터를 입력 데이터로 사용하여 LSTM을 통해 KOSPI 데이터를 예측하였다

Fig. 4. Comparing the predicted values of the KOSPI index between the LSTM_K and LSTM_KNS models

(LSTM_KNS 모델). 세 번째는 LDA로 추출한 주제 데이터와 KOSPI를 같이 입력 데이터로 사용하여 KOSPI 데이터를 예측하였다(LDA_K 모델). 마지막으로 LDA, KOSPI, NDX 그리고 SPX를 입력 데이터로 활용하여 LSTM을 통해 KOSPI 지수를 예측하였다(LDA_KNS 모델).

각 모델의 성능을 평가하기 위해 RMSE(Root Mean Square Error) 값을 기준으로 고찰하면 다음과 같다. 첫 번째 모델인 LSTM_K 모델의 경우 RMSE이 4.60으로 계산되었으며, <Fig. 4> 그래프와 같이 예측값이 실제 값과 비슷한 경향을 보이지만, 4개 모델 중 오차가 가장 크다. 두 번째 모델인 KOSPI_KNS 모델은 RMSE이 3.64로 KOSPI 데이터만 입력 데이터로 사용한 모델보다 좋은 결과 값을 보였지만 여전히 오차가 큰 편이다. 제안된 모델인 LDA_K 모델은 RMSE 값이 2.75이며, LSTM_K 모델과 LSTM_KNS 모델보다 우수한 성능을 보였다. 이는 LDA로 추출된 주제가 KOSPI 지수 예측 성능 향상에 기여함을 의미한다. 마지막 모델인 LDA_KNS 모델은 RMSE가 2.78이다. LDA만 사용했을 때와 비슷한 성능이 나왔는데, 이는 LDA로부터 추출한 주제 데이터를 사용할 때 NDX 및 SPX 입력 데이터는 예측 성능에 큰 영향을 미치지 못하는 것으로 판단된다.

V. Conclusions

본 연구에서는 LDA를 통해 토픽 모델링을 하고, 이를 KOSPI 지수와 함께 LSTM의 입력 데이터로 활용하여 주가를 예측하는 모델(LDA_K 모델)을 제시하였다. 또한, 본 논문에서는 KOSPI 지수에 영향을 미치는 NDX와 SPX 같은 해외 주가지수를 입력 데이터로 하는 다양한 모델(LSTM_KNS 모델, LDA_KNS 모델)을 제안하고 각 모델의 예측 실험 결과를 제시하였다. 실험 결과 LSTM_K 모델의 예측 성능이 가장 좋은 것으로 평가되었다. 이는 과거 KOSPI 지수만을 입력 데이터로 활용하거나 토픽 모델링만을 입력 데이터로 활용하는 기존 연구에 비해 더 많은 정보를 활용하여 KOSPI 지수를 예측하는 본 모델의 특성에서 기인한 것으로 판단된다.

하지만 LSTM_K 모델도 급격한 지수 변동 상황에서는 여전히 예측 오차가 큰 경향을 보였다. 이러한 오차는 과거 지수 데이터와 뉴스 데이터만으로 주가를 예측하기 때문에 발생하는 한계로 생각된다. 또한, 투자자들의 심리적인 부분이 주가 예측 모델에 반영되지 않은 것이 급격한 변동성을 따라가지 못하는 원인의 하나로 판단된다. 따라서, 추후 연구에는 SNS 등의 다양한 채널을 통해 더 많은 정보로부터 데이터를 추출하는 연구와 투자자들의 심리 예측 부분을 데이터로 만들어 입력 데이터로 추가하는 연구가 진행될 필요가 있다고 생각된다.

Fig. 5. Comparing the predicted values of KOSPI index between the LDA_K model and LDA_KNS model

ACKNOWLEDGEMENT

This research was supported by the Hoseo University research grant in 2021.

REFERENCES

[1] Hyeon Su Seung, Jin Young Yang, Jung Hwa Kang, and Jae Hyun Kim, "Time series data through stock price prediction improving prediction accuracy," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 985-986, Jeollanam-do, 2021.

[2] Lee Kanghoon, and Lee Minhyeok, "A Study on the Prediction of KOSPI Index Direction using Investor Sentiment Indicators and Lasso Model," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 885-886, Gyeongbuk, 2023.

[3] Kim SeongGon, " A study on KOSPI direction forecasting using the combination of time separation and sentiment analysis", The Graduate School of Yonsei University, 2023

[4] Jang SeoHee, "KOSPI Index Movement Prediction Combining BERT sentiment analysis and LSTM neural networks", The Graduate School of Ewha Womans University, 2023

[5] Doo-Won Kang, So-Yeop Yoo, Ha-Young Lee, and Ok-Ran Jeong, "A study on Deep Learning-based Stock Price Prediction using News Sentiment Analysis," Journal of the Korea Society of Computer and Information , Vol. 27, No. 8, pp. 31-39, 2022.8.

[6] Kwangyee Ko, Seungwon Oh, and Jangsun Baek, "Development of economic fluctuation topic indices and topic indices regression model for KOSPI200 index," Journal of the Korean Data And Information Science Society, Vol. 31, No. 4, pp. 579-594, 2020.7. DOI: 10.7465/jkdi.2020.31.4.579

[7] Kim, Sun Woong, Choi, Heung Sik, & Oh, Jeong Hwan. Development of Options Trading System using KOSPI 200 Volatility Index. Journal of Information Technology Services, 13(2), 151-161. 2014.2 https://doi.org/10.9716/KITS.2014.13.2.151

[8] Sohee Shin, Hayoung Oh, and Jang Hyun Kim, "Estimation of KOSPI200 Index option volatility using Artificial Intelligence," Journal of the Korea Institute of Information and Communication Engineering, Vol. 26, No. 10, pp. 1423-1431, 2022.10.

[9] Nak Young Lee, and Kyong Joo Oh, "KOSPI200 futures index prediction using denoising filter and LSTM," Journal of the Korean Data And Information Science Society, Vol. 30, No. 3, pp. 645-654, 2019.5 DOI: 10.7465/jkdi.2019.30.3.645

Authors

Jin-Hyeon Joo received the B.S., M.S. degrees in Computer Science and Engineering from Hoseo University, Korea, in 2011, 2013 respectively, and He had worked INKA Entworks company for research about Media contents protection. He is currently a Assistance Professor in the Department of Convergence at Hoseo University. He is interested in Topic Modeling, Deep learning, Software Engineering, IoT.

Geun-Duk Park received M.S. and Ph.D. degrees in Computer Engineering from Seoul National University in 1997 and 2005 respectively, and B.S. degree in Computer Science and Statistics from Seoul National University in 1993. He is currently a professor of the department of computer engineering at Hoseo university. His current research interests include software engineering, topic modeling, and semantic web technology.