The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.5, pp.209~216
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.5.209

Sentiment Analysis of News Based on Generative AI and Real Estate Price Prediction: Application of LSTM and VAR Models

Sua Kim† · Mi Ju Kwon†† · Hyon Hee Kim†††

ABSTRACT

Real estate market prices are determined by various factors, including macroeconomic variables, as well as the influence of a variety of unstructured text data such as news articles and social media. News articles are a crucial factor in predicting real estate transaction prices as they reflect the economic sentiment of the public. This study utilizes sentiment analysis on news articles to generate a News Sentiment Index score, which is then seamlessly integrated into a real estate price prediction model. To calculate the sentiment index, the content of the articles is first summarized. Then, using AI, the summaries are categorized into positive, negative, and neutral sentiments, and a total score is calculated. This score is then applied to the real estate price prediction model. The models used for real estate price prediction include the Multi-head attention LSTM model and the Vector Auto Regression model. The LSTM prediction model, without applying the News Sentiment Index (NSI), showed Root Mean Square Error (RMSE) values of 0.60, 0.872, and 1.117 for the 1-month, 2-month, and 3-month forecasts, respectively. With the NSI applied, the RMSE values were reduced to 0.40, 0.724, and 1.03 for the same forecast periods. Similarly, the VAR prediction model without the NSI showed RMSE values of 1.6484, 0.6254, and 0.9220 for the 1-month, 2-month, and 3-month forecasts, respectively, while applying the NSI led to RMSE values of 1.1315, 0.3413, and 1.6227 for these periods. These results demonstrate the effectiveness of the proposed model in predicting apartment transaction price index and its ability to forecast real estate market price fluctuations that reflect socio-economic trends.

Keywords : Generated AI, Prediction of Real Estate Price, News Sentiment Index, Multi-head attention LSTM, Vetor Auto Regression

생성 AI기반 뉴스 감성 분석과 부동산 가격 예측: LSTM과 VAR모델의 적용

김 수 아† · 권 미 주†† · 김 현 희†††

요 약

부동산 시장은 다양한 요인에 의해 가격이 결정되며 거시경제 변수뿐 만 아니라 뉴스 기사, SNS 등 다양한 텍스트 데이터의 영향을 받는다. 특히 뉴스 기사는 국민들이 느끼는 경제 심리를 반영하고 있으므로 부동산 매매 가격 예측에 있어 중요한 요인이다. 본 연구에서는 뉴스 기사를 감성 분석하여 그 결과를 뉴스 감성 지수로 점수화 한 후 부동산 가격 예측 모델에 적용하였다. 먼저 기사 본문을 요약 후 요약된 내용을 바탕으로 생성 AI를 활용하여 긍정, 부정, 중립으로 분류한 다음 총 점수를 산출하였고 이를 부동산 가격 예측 모델에 적용하였다. 부동산 가격 예측 모델로는 Multi-head attention LSTM 모델과 Vector Auto Regression 모델을 사용하였다. 제안하는 뉴스 감성 지수를 적용하지 않은 LSTM 예측 모델은 1개월, 2개월, 3개월 예측에서 각각 0.60, 0.872, 1.117의 Root Mean Square Error (RMSE)을 보였으며, 뉴스 감성 지수를 적용한 LSTM 예측 모델은 각각 0.40, 0.724, 1.03의 RMSE값을 나타낸다. 또한 뉴스 감성 지수를 적용하지 않은 Vector Auto Regression 예측 모델은 1개월, 2개월, 3개월 예측에서 각각 1.6484, 0.6254, 0.9220, 뉴스 감성 지수를 적용한 Vector Auto Regression 예측 모델은 각각 1.1315, 0.3413, 1.6227의 RMSE 값을 나타낸다. 앞선 아파트 매매가격지수 예측 모델을 통해 사회/경제적 동향을 반영한 부동산 시장 가격 변동을 예측할 수 있을 것으로 보인다.

키워드 : 생성 AI, 부동산 가격 예측, 뉴스 감성 지수, Multi-head Attention LSTM, Vector Auto Regression

※ 이 논문은 2022년도 동덕여자대학교 동덕여자대학교 학술 연구비 지원에 의하여 수행된 것임.
※ 이 논문은 2023년 ACK 2023의 우수논문으로 “생성 AI 기반 뉴스 기사 심리지수를 활용한 부동산 가격 예측 모델”의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 서강대학교 인공지능학과 석사과정
†† 준 회 원 : 동덕여자대학교 정보통계학과 학사과정
††† 종신회원 : 동덕여자대학교 정보통계학과 부교수
Manuscript Received : December 27, 2023
First Revision : March 12, 2024
Accepted : April 6, 2024
* Corresponding Author : Hyon Hee Kim(heekim@dongduk.ac.kr)

1. 서 론

부동산 매매가격지수를 예측하기 위해 전통적으로 금리, GDP, 대출 그림, 소비자 물가 지수 등과 같은 거시 경제적 지표들을 주로 사용해왔다[1]. 2006년 경제 위기 이후로, 소비 심리가 경제 정보의 중요한 부분으로 인식되기 시작했으며, 이는 사람들의 구매 결정에 매우 중요한 요인으로 밝혀졌다[2]. 그럼에도 불구하고, 이와 같은 거시 경제 지표들은 개인의 소비 습관이나 정책적인 요인들을 충분히 반영하지 못한다는 한계점을 가지고 있다. 부동산을 매매하는 사람들의 심리는 시장의 수요와 공급에 큰 영향을 미친다. 따라서 정부정책과 심리 요소들을 반영하고 보다 정확한 예측을 위해서는 정성적인 데이터에 대한 반영이 반드시 필요하다[3].

본 연구에서는 앞서 언급한 점들을 고려하여 기존의 거시 경제 변수들과 더불어 뉴스 기사의 감성 분석을 토대로 도출된 심리 지수를 새로운 변수로 활용하였다. 최근에는 신문과 뉴스 기사의 감성 분석을 통해 감성 지수를 계산하여 시장 변화 예측에 사용하는 연구가 나타나고 있다[4]. 감성 분석은 주관적 태도나 감정을 텍스트로부터 추출하는데 널리 사용되는 방법으로 해당 기사가 긍정적인지 부정적인지 분류를 할 수 있다. 그러나 한국어의 특성상 전체 어휘를 통한 감성의 의미 파악에는 어려움이 있으며, 특히 부동산과 같이 전문적인 지식이 필요한 분야에서 뉴스 해석이 달라질 수 있어, 일반적인 감성 사전의 적용에 한계가 있다. 이러한 문제점을 해결하기 위해 본 연구에서는 생성 AI의 프롬프트 엔지니어링을 활용하여 뉴스 기사의 긍정 및 부정을 수치화하고, 이를 예측 모델의 새로운 입력 변수로 활용하였다.

예측 모델로는 시계열 데이터에 적합한 Long Short-Term Memory (LSTM) 모델과 Vector Auto Regression (VAR) 모델[5]이 사용되었다. 최근LSTM이라는 인공신경망의 진화된 형태가 주목받고 있으며, 이는 특히 주택가격 예측 분야에서 딥러닝의 뛰어난 예측 능력을 입증하고 있다[6]. 또한, LSTM과 Attention 매커니즘을 결합한 모델은 주가 예측과 같은 분야에서 상당한 연구가 진행되고 있다[7]. 아파트매매가격지수와 같은 다변량 시계열 데이터의 예측에서는 과거 데이터와 변수 간 상호 작용을 고려하기 위해 VAR 모델이 사용되고 있다[8]. VAR 모델은 특정 이론에 기반한 가설 설정이 필요 없어 내생변수와 외생 변수 간의 구분 없이 연구자의 선입견을 최소화하는 장점이 있다[9].

실험 결과 제안한 뉴스 감성 지수를 포함한 Multi-head attention LSTM과 VAR 모델이 포함하지 않은 모델보다 높은 정확성을 보였다. Multi-head attention LSTM 모델의 경우 평가지표인 Root Mean Square Error (RMSE) 평균 3개월 기준으로 뉴스 감성 지수를 사용하지 않은 모델은 0.8630, 뉴스 감성 지수를 사용한 모델은 0.7180의 성능을 보였다. VAR 모델의 경우 뉴스 감성 지수를 사용한 모델은 1.0318의 성능을 보였다. 따라서 부동산 매매가격지수와 같이 소비자 심리에 의해 가격이 결정되는 다양한 분야에서 뉴스 감성 지수를 활용하면 보다 성능이 뛰어난 예측 모델을 만들 수 있을 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 제2장에서는 부동산 매매 예측에 관한 선행 연구들을 살펴보고 제3장에서 제안하는 부동산 가격 예측 모델을 자세히 설명한다. 제4장에서 실험 설계 및 성능 평가 결과를 서술하고 마지막으로 제5장에서 결론 및 향후 연구를 제시한다.

2. 선행 연구

2.1 부동산 시장 분석을 위한 시계열 분석 기법

주택가격을 예측하기 위해서는 시계열 모델이 주로 적용되고 있다. LSTM은 금융, 자연어 처리, 시계열 예측 등 다양한 분야에서 활용되고 있으며 특히 시계열 분야에서 높은 성과를 확인할 수 있다. Park et al.[10]은 주식 예측을 위해 LSTM을 사용하였으며 우수한 성과를 확인하였다. Li et al.[11]은 주가 데이터에 상관관계 데이터를 결합하였고 이를 LSTM과 Attention을 결합한 모형의 입력값으로 사용하여 비교 모형보다 좋은 성과를 내었다. Attention의 성능을 향상하기 위해 제안된 Multi-head attention은 여러 개의 Attention을 병렬적으로 사용하고 각각의 결과를 결합하는 과정이다.

박재수·이재수[12]는 부동산 시장과 관련된 온라인 신문 기사의 비정형 빅데이터를 수집하고 이를 통해 산출된 부동산 감성지수와 실거래가 기반의 아파트 매매가격지수의 동태적 관계를 시계열 벡터자기회귀모형을 적용하여 분석하였다. 충격반응분석 결과, 감성지수는 일부 권역을 제외하고 정(+)의 반응을 나타냈으며 반응의 크기는 2개월에서 가장 크게 형성되었음을 확인하였다.

2.2 뉴스 기사를 이용한 부동산 가격 예측

최근에는 거시경제지표만을 반영하는 부동산 시장 예측 모델의 한계를 극복하기 위해 뉴스 기사, 소셜 미디어, 검색어 데이터 등을 활용하는 연구가 나오고 있다. 우윤석·이은정[13]은 언론보도가 부동산시장의 참여자의 기대심리에 영향을 미친다고 주장하였다. 언론보도의 수가 아파트 가격 변화에 미치는 영향을 분석한 결과, 서울 강남의 아파트 가격 상승과 관련된 언론기사 수가 시차를 두고 기타 서울지역의 아파트 가격 상승을 이끈다는 점을 확인하였다. 장몽현·김한수[14]는 뉴스 단어의 빈도와 아파트 가격의 변동에 관한 관계 분석 결과, 벡터자기회귀(VAR) 모형을 통해 뉴스 기사 속 자주 언급되는 일부 단어가 아파트 가격변동과 통계적으로 유의미한 관계가 있다고 주장하였다. 박재수·이재수[15]는 신문 및 방송 감정지수 산출을 통해 아파트 매매가격지수에 예측 모형을 활용하였고 변수의 통계적 유의성, 설명력과 오차 측면에서 모형의 설명 및 예측에 유용하다고 주장하였다.

Table 1. Variables for Real Estate Price Prediction Using News Articles

Authors	Predictor variable
Shin and Kim and Hong [3] (2022)	Economic Indicators: Housing Permits, Construction Investment, Private Consumption, Equipment Investment Indicators, Housing Loans, Household Loans, GDP, Real GDP Growth Rate, Economic Growth Rate, Consumer Price Index, Composite Index of Coincident Indicators, Composite Index of Leading Economic Indicators, CD (91-Day Maturity), 3-Year Treasury Bond Yield, Time Deposit Interest Rate, Lending Interest Rate, Composite Stock Price Index, Exchange Rate, News Data.
Park and Lee [15] (2020)	broadcast news Data
Baek and Song and Cho [16] (2023)	Composite Consumer Sentiment Index, Naver Search Query Data, Mortgage Interest Rate, KOSPI Stock Price Index
Seo and Kim [17] (2021)	Search Volume, Employment Rate, Base Rate, Real Estate Consumer Sentiment Index, Regulatory Tightening, Tax Increase, Loan Reduction, Supply Increase

이러한 연구들을 통해서 뉴스 데이터를 분석하여 긍정과 부정을 판별 후 뉴스 심리지수를 산정하여 부동산 시장 가격 예측에 이용하는 것이 효과적인 접근 방법임을 알 수 있다. 따라서 본 연구에서는 생성AI를 활용하여 뉴스 기사의 감성분석을 진행 후 수치화하여 전국 아파트 매매가격지수 예측에 유용한지 검증하고자 한다.

부동산 가격은 다양한 사회/경제적 요인에 의해 결정되며, Table 1은 부동산 가격 예측에 뉴스 기사를 활용한 연구에서 사용된 변수들을 요약한 것이다.

3. NSI 기반 부동산 가격 예측 모델

3.1 부동산 가격 예측 모델 개요

본 연구에서 제안하는 부동산 가격 예측 모델의 개요는 Fig. 1과 같다. 제1단계로 KB 부동산 데이터 허브, 한국은행, 한국부동산원, 그리고 부동산 관련 뉴스를 크롤링하고 전처리를 실시하였다. 제2단계에서는 수집한 뉴스 기사를 바탕으로 생성 AI 프롬프트 엔지니어링을 통하여 뉴스 감성 지수 (News Sentiment Index, NSI)를 산출하였다. 마지막으로 제3단계에서 Multi-head attention based LSTM과 VAR 모델에 뉴스 감성 지수를 변수로 적용하여 성능 평가를 실시하였다.

3.2 데이터 수집 및 전처리

본 연구에서는 매매가격지수를 예측하기 위해 2019년 1월부터 2023년 3월까지의 아파트 매매가격지수와 부동산 관련 거시경제 지표를 활용하였다. 종속변수로 이용되는 아파트 매매가격지수는 KB 부동산 데이터 허브에서 제공하는 월 간 전국 아파트 매매가격지수를 사용하였다. 240 개의 구시군에 대한 월간 매매가격지수를 활용하였으며 독립변수로는 거시경제 지표인 기준금리, 명목 GDP, 대출금리 및 전국 아파트 매매거래량으로 구성하였다. 명목 GDP는 분기별 자료이기 때문에 종속변수와 데이터의 구조가 맞지 않다. 월자료로 변환하기 위해 데이터의 결측값을 보간하는 것이 필수적이며 직전의 값으로 대체하는 전방 채우기[18]를 통해 월별 데이터로

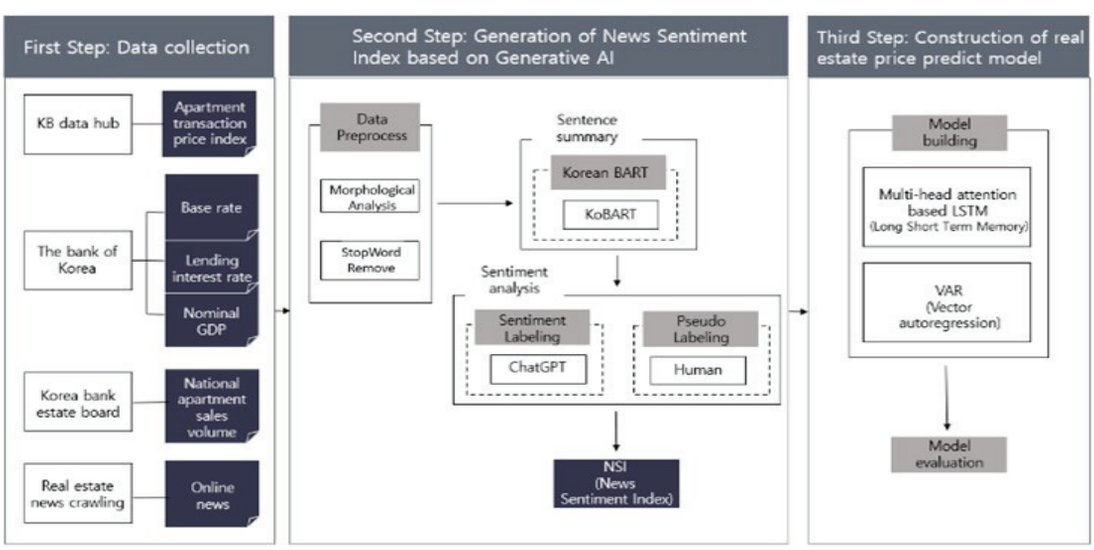

Fig. 1. Overview of the Real Estate Price Prediction

변환 후 사용하였다.

본 연구의 기간은 2019년 1월부터 2023년 3월이며, 학습 세트는 2019년 1월부터 2022년 12월까지의 총 48개월로 설정하였다. 테스트 기간은 2023년 1월부터 2023년 3월까지이다.

LSTM 모델의 경우 단위가 다른 데이터를 정규화하기 위해 기준 금리를 제외한 모든 변수에 최대최소 정규화를 실시하였다. VAR 모델의 경우 표준화 및 정규화는 진행하지 않았으며 변수들의 시계열 정상성을 확보하기 위해 ADF (augmented dickey-fuller) 검정법을 이용하여 단위근 검정 후 차분을 실시하였다.

3.3 뉴스 감성 지수 생성

부동산 가격은 정부 정책과 외부 환경 요소에 크게 영향을 받으며 부동산 뉴스 기사는 이러한 시장 심리를 반영한다. 따라서 아파트 매매 가격 지수를 예측할 때, 뉴스 기사를 활용하는 것이 필요하다. 이를 위해 국내 포털 사이트인 네이버에서 51개월 동안 월별로 약 600개의 뉴스 기사를 크롤링하여 총 30,600개의 뉴스 기사를 수집하였다.

먼저, 뉴스 감성 지수를 산정하기 위해 뉴스 전문을 읽고 감성 분석하는 것은 비효율적이기 때문에 정보 처리의 효율성과 정확성을 높이기 위해 텍스트 요약을 수행하였다.

대량의 텍스트를 요약하는 방법은 크게 통계 기반의 추출적 요약과 신경망을 이용한 신경망을 이용한 생성 요약 방법으로 분류할 수 있다. 추출적 요약 방법으로는 text rank 기법[19]을 활용한 summarize 함수가 대표적으로 사용되며 이를 사용하면 주어진 문장 집합에서 중요한 문장을 추출하여 요약한다. 이 방법은 첫 번째 문장을 주로 핵심 문장으로 선정하는 경향이 있고 원문의 문장만을 요약문에 포함하기 때문에 빈도가 높은 불필요한 단어를 추출하는 문제가 있다. 반면 생성적 요약 방법은 의미 있는 문장을 비교적 정확하게 요약할 수 있는 것으로 알려져 있어 본 연구에서는 KoBART라는 한국어 BART 모델을 사용하였다[20]. 이 모델은 BART에서 소개된 “Text Infilling” 노이즈 함수를 사용하여 40 GB 이상의 한국어 텍스트에 대해 학습된 한국어 언어 모델이다. 따라서 KoBART를 활용한 생성적 요약을 통해 뉴스 기사를 효과적으로 요약하였다.

다음으로, 요약된 텍스트를 이용해 ChatGPT를 활용한 데이터 라벨링을 수행하였다. ChatGPT는 사용자가 텍스트 프롬프트에 입력 값을 주면 입력된 프롬프트를 이해하고 적절한 응답을 생성하는 과정을 통해 대화가 이루어진다. GPT 모델 중 2020년에 발표된 모델인 ‘text-davinchi-002’를 사용하였으며 이는 175억 개의 파라미터를 가지며 대화형 AI, 문장 생성, 질의응답에서 높은 성능을 나타낸다. 각 뉴스 기사 요약문에 대해 감성 분석 후 긍정 1, 중립 0, 부정 -1로 수치화하였다. GPT를 통한 자동 라벨링 결과 중 라벨링 되지 않은 부분과 중립으로 라벨링된 내용에 대해서는 사람이 직접 개입하는 반자동 라벨링을 통해 정확한 라벨링을 수행하였다. Algorithm 1은 뉴스 감성 지수 산출을 위한 유사 코드를 나타낸다.

Algorithm 1. Computation of News Sentiment Index

```
PreprocessedNews <- preprocess(news)
Summary <- kobart_summarize(PreprocessedNews)
Sentiment <- chatgpt_sentiment(Summary)
if Sentiment is 'Positive' then
    SentimentScore = 1
else if Sentiment is 'Negative' then
    SentimentScore = -1
else
    SentimentScore = 0
end if
    return SentimenScore
```

마지막으로, 월별로 감정분석 수치의 총합을 뉴스 심리지수라 하고 다음 Equation (1)과 정의하였다.

$$\text{News Sentiment Index (NSI)} = \sum (\text{Pos(i)}\times 1 + \text{Neu(j)}\times 0 + \text{Neg(k)}\times -1) \quad (1)$$

Pos(i): 긍정으로 판별된 i개 뉴스
Neu(j): 중립으로 판별된 j개 뉴스
Neg(k): 부정으로 판별된 k개 뉴스

라고 할 때, NSI는 각각의 판별된 뉴스 기사에 해당되는 점수의 총합으로 정의된다. 이와 같이 산출된 NSI 지수는 예측 모델의 새로운 변수로 사용된다. 다른 거시경제 지표와의 비교를 위해 본 연구에서는 월별 기준에 근거해 산출하였으나 일별, 주별 등 특정 목적에 따라 다양하게 확장이 가능하다.

3.4 부동산 가격 예측 모델 구축

부동산 매매가격지수 예측을 위해 Multi-head attention based LSTM 알고리즘을 사용하였다. 입력 데이터에 대해 attention을 적용하는 부분에 LSTM을 사용함으로써 key, query, value를 생성하는 과정에서 입력 데이터의 중요성을 예측 모델에 강조하였다. 제안하는 예측 모델의 첫 번째 LSTM 레이어는 5개의 뉴런을 갖고 있으며, 두 번째 LSTM 레이어는 64개의 은닉 레이어를 가지고 있다. 이어서, 0.2의 드롭아웃 비율을 가진 추가적인 드롭아웃 레이어가 적용되었으며 최종적으로 LSTM 레이어는 32개의 은닉 레이어를 포함하고 있다. 마지막으로 n_heads가 5인 Multi-head attention 레이어를 사용하였다.

VAR 모델은 시계열의 정상성이 먼저 전제되어야 하므로 ADF 방법을 이용하고, 유의수준은 5%를 적용하여 수행하였다. 사용되는 변수 중 5% 유의수준에서 단위근을 갖는다는 귀무가설을 기각하지 못한 변수의 경우 1차 차분을 진행하였고, 1차 차분 후에도 단위근이 존재하는 경우 2차 차분을 통해 단위근이 없는 정상 시계열로 변환하여 검정을 수행하였다.

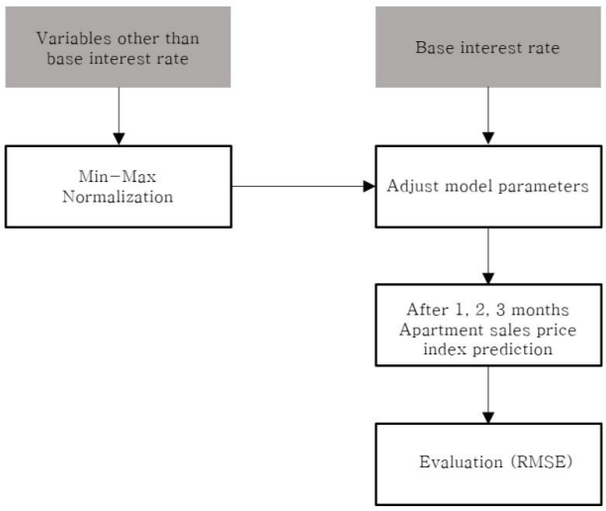

Fig. 2. Multi-head Attention based LSTM Model Framework

4. 실험 설계 및 성능 평가

예측 모델을 구축하고 실험을 진행하기 위해 Google Colab 환경에서 Python으로 개발하였다. 데이터 수집을 위해서 Beautifulsoup4 웹 크롤러를 사용하였으며 사전 학습된 KoBART모델과 open AI 라이브러리를 통해 개발하였다.

4.1 Multi-head attention based LSTM

Fig. 2는 Multi-head attention based LSTM 모델 구축 및 성능 평가 프레임워크를 나타낸다. 먼저, 기준금리를 제외한 모든 변수에 최대 최소 정규화를 진행하였다. Multi-head attention에 사용할 파라미터인 Attention의 개수를 정하기 위해 Attention의 개수를 5개에서 16개까지 각각 NSI 미반영 모델과 NSI 반영 모델에 적용하여 실험하였으며, 그 결과는 Table 2와 같다.

Table 2의 결과에 따라 미래 3개월의 손실 함수인 RMSE의 평균이 두 모델 모두에서 개수가 5일 경우 낮은 값을 보였기 때문에 Attention의 개수를 5개로 선정하였다. 학습에 사용된 파라미터는 window_size=3, n_heads=5, hidden_layer는 각각 64, 32이다. 손실 함수로는 MSE, 평가 지표는 RMSE, 최적화 알고리즘은 Adam을 사용하였으며, 학습률은 1e-3을 사용하였다. 1, 2, 3개월 후 가격 예측에 대한 성능 평가 결과는 Table 3과 같다.

Fig. 3은 예측 모델의 결과로 나타난 예측값과 실제값의 비교를 나타낸다. Fig. 3은 변수 NSI를 제외한 아파트 매매가격지수 예측 모델이고, Fig. 4는 변수 NSI를 포함한 아파트 매매가격지수 예측 모델이다. 미래 예측 성능은 NSI 반영 모델의 스케일링 이전의 1, 2, 3개월의 평균 RMSE가 0.341로 나타났다. 이를 통해 본 연구에서 가장 주요하게 보는 뉴스 감성 지수가 포함된 모델이 감성 지수가 포함되지 않은 모델보다 아파트 매매가격지수의 예측 성능이 높은 것으로 나타났다. 본

Table 2. 3-month RMSE Average of Models by Number of Attentions

Number of Attention	Prediction model without NSI	Prediction model with NSI
5	**0.2256**	**0.185**
6	0.247	0.202
7	0.272	0.206
8	0.184	0.231
9	0.30	0.268
10	0.532	0.432
11	0.20	0.388
12	0.47	0.734
13	0.172	0.265
14	0.30	0.273
15	0.664	0.662
16	0.187	0.227

Table 3. Evaluation of Prediction Performance

	1 month	2 month	3 month
Prediction model without NSI	0.60	0.872	1.117
Prediction model with NSI	0.40	0.724	1.03
Expert-knowledge based LSTM[3]	0.181	0.174	0.168

Fig. 3. Comparison of Predicted Values of Apartment Transaction Price Index Without NSI

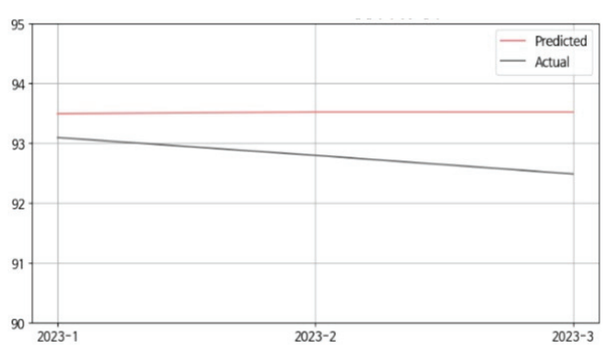

Fig. 4. Comparison of Predicted Values of Apartment Transaction Price Index With NSI

연구는 TF-IDF에 의해 선정된 주요 단어에 대해 상승, 보합, 하락을 평가한 선행연구[4]와 달리 문장 전체의 문맥을 파악한 후 긍정 및 부정 라벨링을 수행하였다. 선행 연구에 비해 성능 향상의 폭이 적은 이유는 선행 연구에서 더 큰 규모의 데이터셋을 이용해 모델을 학습시켰기 때문인 것으로 보인다.

4.2 Vector autoregression

Fig. 5는 VAR 모델 적용 프레임워크를 나타낸다. 변수들 간의 관계를 살펴보고 시계열 데이터의 정상성을 확보하기 위해 ADF 검정법을 이용하여 단위근 검정 후 차분을 실시하였다. 그 후 모델 파라미터로 적정 시차를 찾아주고 오차항의 독립성 검정을 거치는 과정이다.

앞선 LSTM과 마찬가지로 각각 4개, 5개 변수로 구성된 VAR 모델을 분석하기 위한 시차를 선택하기 위해 Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) 등을 사용하는 정보기준 방법을 적용하였다. 적정시차를 분석한 결과는 Table 4와 Table 5에 나타나 있다. 표에서 4개 변수로 구성된 VAR 모델과 5개 변수로 구성된 VAR 모델 모두 가장 최소화되는 AIC값을 갖는 시차 5를 적정시차로 결정하였다.

또한 5개 변수를 갖는 VAR 모델의 진단을 위해 오차항의 자기상관 존재에 대한 지표인 더븐-왓슨 통계량을 통해 오차항에 대한 자기상관을 평가하였다. 더븐-왓슨 통계량 d값은 0~4 사이의 값을 가지며 2에 가까울수록 자기상관이 없어 독립이라고 판단한다. 검정 결과 4개 변수로 구성된 모델의 경우 매매가격지수 2.29, 금리 2.03, 국내총생산 1.72, 대출금리 1.64, 아파트 거래량 2.2로 나타났다. 심리지수가 포함된 5개 변수로 구성된 VAR 모델의 경우도 심리지수 1.66, 매매가격지수 2.22, 명목 GDP 1.37, 금리 2.06, 대출금리 1.45, 아파트 거래량 1.75 로 나타났다. 이상의 검정결과에 따르면 5개 변수를 갖는 VAR 모델은 예측에 적합한 것으로 보인다.

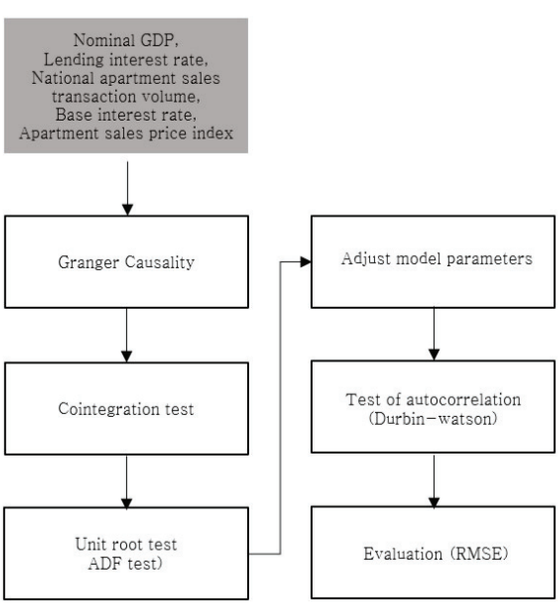

Fig. 5. VAR Model Framework

Table 4. Model Appropriate Time-lag Test Result Without NSI

		Time-lag				
		1	2	3	4	5
4 variables	AIC	30.33	30.11	29.86	28.59	28.24
	BIC	31.60	32.43	33.23	33.03	33.73

Table 5. Model Appropriate Time-lag Test Result with NSI

		Time-lag				
		1	2	3	4	5
5 variables	AIC	39.78	39.70	39.09	37.38	32.79
	BIC	41.53	42.96	43.86	43.65	40.57

Table 6. Evaluation of Prediction Performance

	1 month	2 month	3 month
Prediction model without NSI	1.6484	0.6254	0.9220
Prediction model with NSI	1.1315	0.3413	1.6227

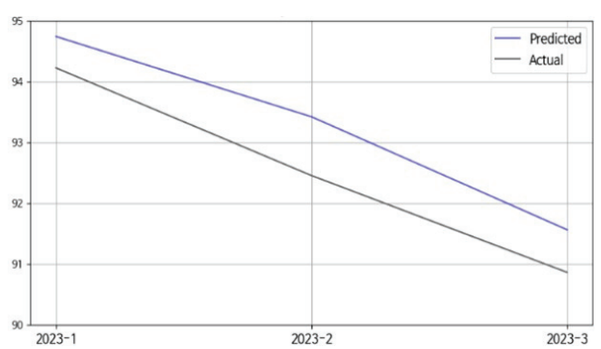

Fig. 6. Comparison of Predicted Values of Apartment Transaction Price Index Without NSI

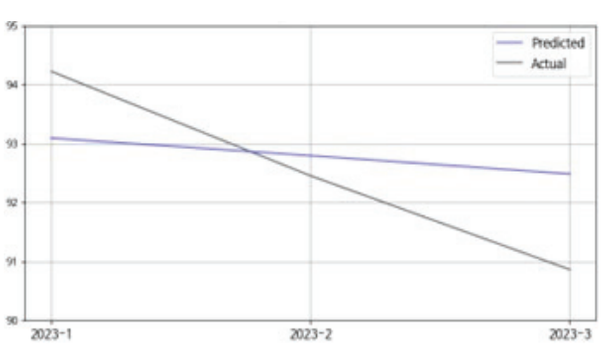

Fig. 7. Comparison of Predicted Values of Apartment Transaction Price Index with NSI

1, 2, 3개월 후 가격 예측도는 RMSE로 나타내었으며 Table 6과 같다. Table 6의 실험 결과에서 볼 수 있는 바와 같이 제

안하는 NSI를 활용한 예측 모델에서 성능이 향상됨을 알 수 있다. 또한 3개월 후의 예측 성과보다 1, 2개월의 단기 예측에서의 성과가 보다 우수함을 확인할 수 있었다.

Fig. 7은 예측 모델의 결과로 나타난 예측값과 실제값의 비교를 나타낸다. 미래 예측 성능은 NSI 반영 모델의 1, 2, 3개월의 평균 RMSE가 1.1927로 나타났다. 이를 통해 연구에서 가장 주요하게 보는 뉴스 감성 지수가 포함된 모델 이 뉴스 감성 지수가 포함되지 않은 모델보다 아파트 매매 가격지수의 예측 성능이 높은 것으로 나타났다.

5. 결 론

본 연구에서는 생성형 AI를 활용하여 뉴스 기사의 감성 분석을 실시하고 감성 분석 결과를 뉴스 감성 지수 (News Sentiment Index, NSI)로 정의하여 아파트 매매가격지수 예측 모델에 적용하였다. 시계열 예측 모델로는 Multi-head attention LSTM 모델과 Vector Auto Regression 모델을 사용하였다. 실험 결과를 통해 Multi-head attention LSTM 모델에서 제안하는 뉴스 감성 지수를 사용할 경우 아파트 매매가격지수의 예측력이 향상되었다. 또한, Vector Auto regression 모델에서도 NSI 지수를 활용한 경우 단기 1, 2개월의 미래 예측값이 향상된 것을 확인할 수 있다.

이러한 개선된 예측력은 부동산 뉴스와 같은 비정형 데이터가 부동산 시장과 가격에 유의미한 영향을 미친다는 것을 보여주며 금융데이터 예측 향상에 큰 공헌을 할 수 있을 것으로 기대된다. 또한, 부동산에 특화된 감성 사전이 존재하지 않는 상황에서 Chat GPT를 활용한 감성 분석 라벨링을 수행한 것에 의의가 있다. 라벨링 작업에서의 소요 시간, 비용 및 반복 작업 측면에서 ChatGPT가 갖는 유용성을 확인할 수 있으며 프롬프트 정보 구조화를 통해 신속한 데이터 구축의 용이함을 얻을 수 있음을 알 수 있다.

제약점으로는 감성분석이 정량적 평가가 아니라 정성적 평가로 이루어지기 때문에 ChatGPT를 이용한 감성분석 라벨링이 제대로 수행되었는지에 대한 명확한 평가가 어렵다는 한계가 존재한다. 따라서 향후 연구에서는 부동산 관련 전문가를 상대로 인터뷰를 진행하여 특정 단어들에 대해 긍부정 라벨링을 수행한 후 이를 ChatGPT 프롬프트에 시스템 메시지 또는 퓨샷 러닝의 예시로 추가한다면 좀 더 명확한 라벨링이 될 것으로 기대된다. 또한 사용된 데이터의 양이 상대적으로 적다는 한계점이 있다. 이로 인해 예측 모델이 데이터 패턴을 충분히 학습하지 못해 예측값에 큰 변동이 발생하지 않았을 가능성이 있다. 향후 연구에서는 훨씬 더 많은 양의 데이터를 수집하여 예측 모델의 성능을 평가한다면 한층 더 신뢰성 있는 결과를 얻을 수 있을 것으로 기대된다.

본 연구에서 뉴스 기사를 통해 산출한 감성 지수만을 활용하였다면 향후 연구에서는 부동산 카페와 같은 대중의 실질적 여론을 포함한 SNS 정보를 활용하고자 한다. 이러한 정보들을 수치화 하여 변수로 사용한다면 보다 더 다양한 정보의 데이터셋을 사용할 수 있을 것으로 기대된다.

References

[1] J. Y. Park, “Consumer psychology and psychology,” Korea Development Bank, 2004.

[2] J.-Y. Ham and J.-y. Son, “Applying the ba yesian vector autoregressive model in house price pre diction,” *Review of Real Estate and Urban Studies*, Vol.8, No.2, pp.25-38, 2016.

[3] E. Shin, E. Kim, and T. Hong, “The prediction of real estate price based on deep learning using news sentiment and expert knowledge,” *The Journal of Internet Electronic Commerce Resarch*, Vol.22, No.3, pp.61-73, 2022.

[4] E. K. Shin, E. Kim, and T, Hong, “The prediction of real estate price based on deep learning using news sentiment and expert knowledge,” *The Journal of Internet Electronic Commerce Research*, Vol.22, No.3, pp.61-73, 2022.

[5] H. J. Chun, “Prediction of housing price using time series analysis and machine learning methods,” *Residential Environmental: Journal of the Residential Environment Institute of Korea*, Vol.18, No.1, pp.49-65, 2020.

[6] M. Kwon and J. Kim, “Forecasting Seoul Apart ment Price Index based on a Deep Learning Model,” in *Proceedings of Journal of Korean Institute of Industrial Engineers*, 2020.

[7] X. Zhang, X. Liang, A. Zhiyuli, S. Zhang, R. Xu, and B. Wu, “AT-LSTM: An attention-based LSTM model for financial time series prediction,” In *IOP Conference Series: Materials Science and Engineering*, Vol.569, No.5, pp.052037, IOP Publishing, 2019.

[8] H. M. Lee, and H. J. Chun, “Dynamic character istics of housing price in Seoul using panel VAR model,” *Residential Environment: Journal of the Reside ntial Environment Institute of Korea*, Vol.18, No.2, pp.27-42, 2020.

[9] G. P. Yang and H. J. Chun, “A study on the correlation between media and housing prices using a deep learn ing-based language model,” *Appraisal Studie*, Vol.20, No.3, pp.109-134, 2021.

[10] H. J. Park, Y. Kim, and H. Y. Kim, “Stock market foreca sting using a multi-task approach integrating long-short-term memory and the random forest framework,” *Applied Soft Computing*, Vol.114, pp.108106, 2022.

[11] H. Li, Y. Shen, and Y. Zhu, "Stock price prediction using attention-based multi-input LSTM," In *Asian Conference on Machine Learning*, pp.454-469, PMLR, 2018.

[12] J. S. Park and J.-S. Lee, "A study of dynamic relationship between apartment price and real estate online news using VAR model: An approach to the sen timent analysis using unstructured big data," *Apprai Sal Studies*, Vol.18, No.2, pp.83-113, 2019.

[13] Y. S. Woo and E. J. Lee, "An analysis on the relationship between media coverage and time-series housin ng prices," *Housing Study*, Vol.19, No.4, pp.111-134, 2011.

[14] M.-H. Jang and H.-S. Kim, "A research on fluctuations of housing prices using text mining," *Journal of the Korean Housing Association*, Vol.30, No.2, pp.35-42, 2019.

[15] J. Park and J.-S. Lee, "Predictability of housing sales prices employing a real estate sentiment index: Using unstructured big data of online newspaper and TV broadcast news," *Journal of Korea Planning Association*, Vol.56, No.4, pp.99-111, 2021.

[16] S. Baek, M. Cho, and M. Kang, "Comparative analysis of predictive power of apartment sales price index according to psychological variables: Focusing on the sentiment index using the survey and the online search engine," *Journal of Korea Planning Association*, Vol.58, No.2, pp.81-91, 2023.

[17] Y. Seo and K. Kim, "Development of an artificial intelligence model for predicting the policy and the environment affecting the public interest on the real estate," *The Journal of Korean Institute of Information Technology*, Vol.19, No.12, pp.135-141, 2021.

[18] H.-J. Choi, H.-W. Park, and S.-K. Ko, "Optinal data generation strategy for training RNN-based time-series data imputation models." In *Proceedings of Journal of Computing Science and Engineering*, 2022.

[19] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," *Proc. of the EMNLP*, Vol.4, pp.404-411, 2004.

[20] J. W. Lee, H. Kim, H. Jung, and S. Park, "Natural language processing-based korean summary system considering linguistic feature," *Journal of Knowledge Information Technolog y and Systems*, Vol.18, No.2, pp.389-398, 2023.

김 수 아

https://orcid.org/0009-0002-6400-3158
e-mail : rlatndkextra247@gmail.com
2023년 동덕여자대학교 정보통계학과(학사)
2024년 ~ 현 재 서강대학교 인공지능학과 석사과정
관심분야 : Large Language Model, Responsible AI

권 미 주

https://orcid.org/0009-0005-1043-4501
e-mail : miju9599@gmail.com
2020년 ~ 현 재 동덕여자대학교 정보통계학과 학사과정
관심분야 : Large Language Model, Responsible AI

김 현 희

https://orcid.org/0000-0002-7507-8342
e-mail : heekim@dongduk.ac.kr
1996년 이화여자대학교 컴퓨터학과(학사)
1998년 이화여자대학교 컴퓨터학과(석사)
2005년 이화여자대학교 컴퓨터공학과(박사)
2005년 ~ 2006년 LG전자 디지털미디어 연구소 선임연구원
2006년 ~ 현 재 동덕여자대학교 정보통계학과 부교수
관심분야 : Machine Learning, Deep Learning, Big Data Analysis