KoBERT를 이용한 언론과 부동산시장의 관계 분석
Analysis of the relationship between the media and the real estate market using KoBERT

양 건 필* · 전 해 정**
Geonpil Yang · Haejung Chun

目 次

ABSTRACT

1. CONTENTS

(1) RESEARCH OBJECTIVES

The real estate market is of most interest, and for this reason, changes in real estate prices and research predicting them have been actively conducted. In tracking and predicting changes in real estate prices, various methodologies such as machine learning and deep learning have emerged due to recent technological advances, but these methodologies have many limitations due to the difficulty of securing data that is the source. Therefore, research is needed through securing increasing online data.

(2) RESEARCH METHOD

This study aims to analyze the impact of the media on real estate prices through the KoBERT model and Granger causality test.

* 주저자 : 상명대학교 일반대학원 부동산학과, 박사과정, jpfeel1986@gmail.com

** 교신저자 : 상명대학교 일반대학원 부동산학과 교수, hjchun6807@smu.ac.kr

(3) RESEARCH FINDINGS

As a result of classifying the polarity values (positive, neutral, and negative) of news articles by applying KoBERT, the classification rate was high with an accuracy of about 84%. As a result of performing Granger causality analysis by housing type, the media had different effects for each category.

2. RESULTS

The media has adopted a hypothesis that affects real estate prices, and has implications that it can be used as an important variable in predicting changes in real estate prices by housing type and area in the future.

3. KEY WORDS

・KoBERT, media, real estate price, Granger causality test, polarity value

국문초록

부동산 가격에 영향을 미칠 수 있는 뉴스 기사 중 6가지 카테고리(금융, 증권, 산업/재계, 부동산, 글로벌 경제, 경제 일반)별 뉴스 내용을 크롤링 하여 자연어 처리 모형인 KoBERT 모형을 적용해 극성 값을 분류하고 이를 한국부동산원의 유형별 주택 매매가격지수 및 주택 실거래가격지수와의 그랜저 인과 검정을 통해 실증분석하였다. 뉴스 기사의 시간적 범위는 2011년 7월부터 2021년 6월까지로 하였고 공간적 범위는 수도권, 지방권, 서울, 서울 강남, 서울 강북으로 구성하였고 내용적 범위는 아파트, 다세대주택, 단독주택으로 설정하였다. 뉴스 기사 중 부동산 카테고리의 경우 수도권, 서울, 서울 강남권의 아파트 매매 지수 및 실거래가격지수에 영향을 미치는 것으로 나타났다. 금융 관련 언론 지수도 수도권 및 서울 강남에 영향을 미쳤지만 주택 유형에 있어서는 단독주택 매매 지수에 영향을 주었다고 나타났다. 증권, 글로벌 경제, 경제 일반 카테고리의 언론 지수는 서울이나 수도권이 아닌 지방 도시의 주택 유형들에 영향을 준 것으로 나타났으며 산업 재계 카테고리의 경우 다른 카테고리와 달리 다세대주택의 매매 지수 및 실거래가격지수에만 영향을 주는 것으로 나타났다.

핵심어 : KoBERT, 언론, 부동산 가격, 그랜저 인과검정, 극성 값

Ⅰ. 서 론

국내 부동산 시장의 가장 큰 특징은 필수재이면서 투자재의 성격을 동시에 갖는다는 것이다.
부동산 시장은 대부분의 관심사이며 이로 인해 특히 부동산 가격의 변화와 이를 예측하는 연구가 활발히 이루어져왔다. 부동산 가격의 변화를 추적하고 예측하는 데 있어 최근 기술의 발전으로 머신러닝, 딥러닝 등 다양한 방법론들이 나왔지만 이러한 방법론들은 그 원천이 되는 데이터의 확보의 어려움으로 인해 많은 한계점들을 갖고 있다.

현재 부동산 가격 예측을 위해 활용되는 대부분의 데이터는 주택 변수(면적, 층, 건축년도 등), 주변 환경 변수(주변 대중교통시설과의 거리, 유통시설수, 공원 면적 등) 및 경제적 변수(금리, 물가, 통화량) 등 정량적 수치로 수집이 가능한 범위 내로 제한된다. 이를 극복하기 위해 정량적 데이터 외 온라인에서 다량으로 발생하고 있는 정성적 데이터를 수집 후 이를 정량화해 부동산 가격 예측에 활용하는 움직임이 최근 들어 점점 활발해지고 있다. 온라인에서 수집한 정성적 데이터의 정량화를 위해서 텍스트 마이닝 기술이 적용되고 있으며 그 중 감성분석이 대표적인 예이다. 감성 분석은 온라인의 텍스트들을 긍정과 부정, 중립으로 분류하는 것으로 다량의 텍스트들을 수집하는 것과 해당 텍스트를 정확하게 분류할 수 있는 알고리즘 적용이 가장 중요하다. 대다수의 연구들은 감성 분석을 문장의 형태소 분석을 통해 명사, 동사 등으로 추출한 뒤 문맥이 아닌 단어 기반으로 극성 값(긍정, 부정, 중립)을 구분한다. 문맥을 고려하지 못하는 단어 기반의 감성 분석은 해당 문장들의 뜻을 왜곡하는 결과를 낳아 부정확한 극성 값을 부여하게 된다.

2017년 이전까지 단어 수준으로 분석되었던 텍스트 처리 기술이 2018년 구글에서 개발한 BERT(Bidirectional Encoder Representations from Transformer) 모델의 출현으로 문장 수준에서의 텍스트 분석이 가능해졌다.[1] 이 모델은 최근까지 딥러닝 모델을 적용한 모든 자연어 처리 분야에서 좋은 성능을 보이고 있는 범용 언어 모델이다. BERT를 적용한 모델링 과정을 살펴보면 Pre-trained는 비지도 학습(Unsupervised Learning) 방식으로 진행되고 대량의 문장을 인코드(Encoder)가 임베딩하고, 이를 트렌스퍼(Transfer)하여 Fine-tuning을 통해 목적에 맞는 학습을 수행하여 과업을 수행하는 것이 특징이다.[2] 하지만, 범용 언어 모델이라도 한국어는 다른 언어와 달리 고유의 특성을 갖고 있으므로 이에 맞는 개선된 모델이 필요하다. 이에 BERT의 한국어 성능 한계를 극복한 KoBERT 모델을 2019년 SKT에서 개발해 공개하였다. 위키피디아 및 뉴스 등에서 수집한 수천만 개의 한국어 문장으로 이루어진 대규모말뭉치(corpus)를 학습하였으며, 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화(Tokenization) 기법을 적용하여 성능 향상을 이끌어 냈다. 이를 활용해 단어가 아닌 문장 수준에서의 텍스트 감성분석을 통해 보다 정확한 극성 분류가 가능해졌다고 할 수 있다.

이에 본 연구는 KoBERT 모형을 통해 뉴스 기사들의 극성 값을 분류하고 이를 그랜저 인과관계 모형으로 검정해 언론이 부동산 가격에 미치는 영향을 분석하고자 한다. 즉, 언론에서의 긍정 또는 부정적 여론이 부동산 주택 가격에 영향을 미치는지? 영향을 미친다면 어느 기간까지 그 영향을 미치는지?를 실증분석하고자 한다.

본 연구는 부동산 주택 가격에 영향을 미칠 수 있는 6가지 카테고리(금융, 증권, 산업/재계, 부동산, 글로벌 경제, 경제 일반)별 뉴스를 국내에서 가장 점유율이 높은 네이버 포털에서 웹 크롤링 하여 KoBERT 적용으로 극성을 분류하고 이를 한국부동산원의 유형별 주택 매매가격지수 및 주택 실거래 가격지수와의 그랜저 인과관계 분석을 실시하여 분석하고자 한다.

본 연구의 시간적 범위는 2011년 7월부터 2021년 6월까지로 하였고 공간적 범위는 전국을 한국부동산원 자료의 지역 분류체계를 따라 수도권, 지방권, 서울, 서울 강남, 서울 강북으로, 내용적 범위는 주택 유형 중 아파트, 다세대주택, 단독주택으로 설정하였다.

본 연구의 구성은 다음과 같다. 2장은 관련된 선행연구를 살펴본다. 3장은 KoBERT 모형과 그랜저 인과관계 모형에 대해 알아본다. 4장은 실증분석으로 텍스트 데이터 수집 및 구축 과정과 KoBERT 분석 결과 및 그랜저 인과관계 분석 결과에 대해 구술한다. 마지막은 결론으로 연구결과를 요약하고 시사점을 제시한다.

---

1) Wikipedia, "BERT(language model)"

2) 이제로, 박은환, 이재구, "BERT파생모델의 한국어에 대한 성능 비교", 학술대회논문집, 한국통신학회, 2020, pp.901-902.

Ⅱ. 선행연구

감성분석은 크게 감성 사전 구축을 통한 분석 방법, 기계 학습 기반의 분석 방법이 있으며 이 중 감성 사전 구축을 통한 분석 방법을 적용한 연구가 더 많다. 감성 사전 구축 분석 방법은 단어의 극성 값(긍정, 부정, 중립)을 연구자가 직접 부여한 감성 사전을 제작 후 새로운 문장 내 단어의 출현 빈도에 따라 극성 값을 판단하는 방법이다. 사전에 정의된 감성사전을 기반으로 문장의 극성을 판단하므로 개별 연구자에게 주어진 시간과 경험 및 주관에 따라 연구의 질이 차이가 날 수 있다. 만약 극성 값을 부여받지 않은 새로운 단어의 출현 시 이를 반영하여 감성사전을 계속 추가해야 한다는 어려움이 존재한다. 또한, 분야마다 동일한 단어도 다른 의미로 사용되는 경우가 존재해 각 분야마다 적합한 감성 사전이 구축되어야 한다는 한계점이 있다.

경정익(2016)은 2013년 1월부터 2015년 8월까지 게재된 35,082개의 부동산 관련 뉴스 데이터를 수집하여 부동산 가격, 부동산 투자, 부동산 정책, 부동산 시장, 부동산 경기 등 5개의 대표 속성으로 구분하여 수집된 부동산 뉴스 텍스트를 감성분석하여 긍정과 부정의 감성을 표현하는 3,527개의 감성 표현 어휘로 작성하였다. 이를 통해 도출한 뉴스 기사의 감성분석지수와 실거래가 데이터가 높은 상관관계가 있음을 발견했다고 하였다.[3)]

박재수(2019)는 2010년 1월부터 2017년 12월까지 3개 일간지(조선일보, 동아일보, 중앙일보)와 3개 경제전문지(한국경제, 매일경제, 서울경제)의 신문기사를 크롤링 해 감성지수를 산출하였다. 분석 결과 온라인 신문기사와 관련된 긍정 감성지수는 소형 아파트의 매매가격지수에 1개월의 시차로 유의미한 영향을 주는 것으로 나타났다고 하였다.[4)]

서정석(2021)은 2011년 7월부터 2018년 12월까지 총 90개월을 시간적 범위로 설정하고 "주택", "부동산", "아파트", "집값", "청약", "분양", "재개발", "재건축" 등 총 8개의 검색 키워드가 포함된 신문기사 총 640,445건을 수집하여 이 중 4,000건의 표본을 추출한 뒤 감성사전을 구축하였다. 분석 결과 단기적으로는 언론 보도 논조 지수가 주택 매매시장 소비 심리 지수에 부정적인 영향을 미치는 것으로 나타났다고 하였다.[5)]

기계 학습 기반의 분석 방법은 문장 내 단어의 출현 빈도만으로 계산하는 방식, 문장 내 단어들의 관계 등의 언어 규칙을 기계 학습 방법을 적용하여 분류하는 방식이 존재한다.

이재수(2020)는 2012년 1월부터 2018년 12월까지 지상파 3사(KBS, MBC, SBS)의 8163개의 부동산 관련 방송 뉴스를 웹 크롤링을 통해 수집 및 분류하고 기계학습 기법을 활용하여 감성분석을 적용해 감성지수를 산출하였다. 분석 결과 부동산 관련 방송 감성지수는 아파트 매매가격지수와 양의 상관관계를 나타내며 감성지수가 아파트 매매가격지수에 영향을 미치는 후행 관계가 1개월 시차에서 상관성이 높다고 주장하였다.[6)]

Ⅲ. 분석모형

1. KoBERT

BERT[2]는 자연어를 양방향으로 사전 학습이 가능한 자연어 처리 모델이다.[7)] 모델은 대용량의 레이블이 없는 데이터로 모델을 미리 학습한 후, 레이블이 있는 데이터를 이용하여 전이 학습(Transfer Learning)을 수행한다. 이후 Sentence Embedding 혹은 Contextual Word Embedding 기법을 통하여 문장을 토큰 단위로 분할하여 네트워크에 입력하면 전체 문장에 대한 Vector와 문장을 구성하는 단어인 토큰 각각에 대응되는 Vector를 출력한다. 이를 기반으로 Text

3) 경정익, 이국철, "Textmining에 의한 부동산 빅데이터 감성분석 모형 개발", 주택연구, 한국주택학회, 2016, 제24권, 제4호, pp.115-136.

4) 박재수, 이재수, "아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석", 국토계획, 대한국토·도시계획학회, 2019, 제54권, 제1호, pp.131-147.

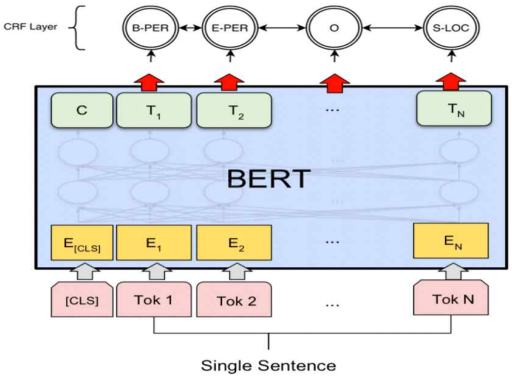
5) 서정석, 오지훈, 김정섭, "부동산 뉴스와 주택경기의 동적 관계에 대한 고찰: 언론보도논조지수 개발을 중심으로", 한국지역개발학회지, 한국지역개발학회, 2021, 제33권, 제1호, pp.89-112.

6) 이재수, 박재수, "방송뉴스 감성지수와 서울시 주택매매가격의 상관 및 인과관계 분석", 주택도시금융연구, 대한국토·도시계획학회, 2020, 제54권, 제1호, pp.53-68.

7) Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Under-standing", Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp.4171-4186.

Classification을 학습할 경우, 전체 네트워크가 많은 양의 문서로 Masked Language Model을 사전 학습하였기 때문에 향상된 성능을 얻을 수 있다. 길이가 TT 인 어절 단위로 토큰화된 문장 xx 가 KoBERT 모델에 입력으로 사용되고, KoBERT를 거쳐 입력 문장의 길이만큼 의미 정보가 벡터의 형태 (oo, $h1$, $h2$, ⋯ , hTT)로 출력된다.

〈그림 1〉 사전 학습기반의 언어모델(BERT)



대다수의 기존 연구들은 문장 처리 시 뉴스 기사의 제목을 기준으로 자연어를 처리하고 극성 값을 부여했다. 뉴스 기사의 제목은 전체의 내용을 함축하고 있지만 중립적인 내용의 제목이 많아 정확한 극성을 판별하기 위해서는 내용 전체의 문맥을 고려해야 한다. 따라서, 본 연구에서는 뉴스 기사의 제목이 아닌 내용 전체를 KoBERT 모형에 적용하여 분석을 수행하였다.

2. 그랜저 인과관계

두 개의 시계열 데이터에서 한 변수의 과거 데이터와 다른 한 변수의 과거 데이터의 결합으로 그 변수를 선형 예측했을 때 통계적으로 유의미하게 영향력이 있다면 그랜저 인과(Granger Caisality)가 있다고 말한다. 그랜저 인과관계를 확인하기 위해서는 두 개의 시계열의 데이터 모두 정상성(Stationary)이 전제되어야 하며 시차(lag)를 파라미터로 넣어 주어야 한다. 시차의 적절한 값을 찾는 것은 연구자의 판단에 의해서 결정되며 본 연구에서는 언론 기사의 특성을 고려하여 시차(lag)를 1부터 3까지로 정의하였다.

$$Y_t = \mu + \sum_{i=1}^{k} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j Y_{t-j} + e_{1t} \quad (1)$$

$$X_t = \mu' + \sum_{i=1}^{m} \lambda_i X_{t-i} + \sum_{j=1}^{n} \delta_j Y_{t-j} + e_{2t} \quad (2)$$

식(2)에서 X_t는 X의 t기 변동률, Y_t는 t기 변동률이며, e_{1t}와 e_{2t}는 시계열간에 상관관계가 없는 잔차항(residual)을 나타낸다.

만약 식(1)에서 모든 α_i값이 0이라는 가설이 기각되지 않으면 X의 변화가 X변화의 원인이라고 할 수 있으며, 반대로 식(2)에서 모든 δ_j값이 0이라는 가설이 기각되지 않으면 X의 변화가 Y의 변화의 원인이라고 할 수 있다. 두 가설이 모두 기각되지 않는다면, 이때는 X와 X의 변화 상호 간에 영향을 주고받는다고 결론을 내린다.

특히, 분석 결과를 해석하는 데 주의를 기울여야 한다. 그랜저 인과관계의 의미로 변수 X가 Y의 원인이라고 판정되더라도 이는 X가 Y와 어떤 일정한 관계를 가지며 선행하므로 Y의 예측에 있어서 X의 자료가 도움이 된다는 의미일 뿐, X가 Y의 충분조건이라거나 X를 조작함으로서 Y에 관련된 일정 목표를 달성할 수 있다고 하는 의미를 갖는 것은 아니기 때문이다.[8)9)]

Ⅳ. 실증분석

1. 변수설명

본 연구는 언론이 부동산 가격에 미치는 영향을 공간적 범위는 수도권, 지방권, 서울, 서울 강남, 서울 강북으로 내용적 범위는 주택 중 아파트, 다세대주택, 단독주택으로 시간적 범위는 2011년 7월부터 2020년 6월까지로 설정 후 KoBERT 및 그

8) 손재영, "지가와 거시경제변수간의 인과관계에 관한 실증분석", 한국개발연구, 한국개발연구원, 1991, 제13권, 제3호, pp.57-58.

9) 박재수, 이재수, "아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석", 국토계획, 대한국토·도시계획학회, 2019, 제54권, 제1호, pp.137-138.

랜저 인과관계 모형을 이용해 실증분석하였다.

〈표 1〉 변수설명 : 카테고리별 뉴스기사수

구 분	카테고리	뉴스기사수(개)
언론정보	금융	591,315
	증권	557,413
	산업/재계	622,862
	부동산	538,131
	글로벌 경제	408,469
	경제 일반	676,028
	전체	3,394,218

자료 : 네이버포털뉴스

주요 데이터인 뉴스 기사는 다양한 언론사 뉴스들을 카테고리별로 제공하는 네이버 포털을 웹 크롤링 해 수집하였다.
〈표1〉은 전체 기사 수와 카테고리별 기사 수를 나타낸다. 전체 기사 수는 3,394,218개로 관련 연구들 중 가장 많은 데이터를 확보하였다. 이는 언론정보의 긍정, 중립, 부정의 극성 값을 일반화하기해서는 가능한 많은 뉴스 기사들의 표본들을 확보
하는 게 중요하기 때문이다. 카테고리별 전체 기사 수는 경제 일반, 산업/재계, 금융, 증권, 부동산, 글로벌 경제순으로 나타났다.
〈표2〉는 유형 및 공간 단위별 부동산 지수 기초통계량으로 실거래가격지수 및 매매 지수 모두 수도권의 아파트 단위에서 가장 높게 나타났으며, 지역적으로 지방보다 수도권 또는 서울의 실거래가격지수와 매매 지수가 높게 나타났다.

2. KoBERT 분류 결과

본 연구는 언론이 부동산 가격에 미치는 영향을 공간적 범위는 수도권, KoBERT 적용을 통해 수집한 뉴스 기사들의 극성 값을 산출하기 위해서는 극성 값이 라벨링 된 데이터가 필요하다. 라벨링 된 데이터의 확보 방법은 크게 2가지인데 하나는 연구자가 문장을 직접 보고 극성 값을 부여하는 방법과 다른 하나는 라벨링 되어 있는 데이터를 민간/공공기관에서 수급하는 방법이다. 전자의 경우 시간적 한계로 인해 많은 문장들의 극성 값을 부여하지 못하고 상대적으로 소수의 데이터만을 확보할 수 있다는데 단점이 있다. 후자의 경우 한국어 문장들을 라벨링 한 데이터의 부재로 인해 데이터 확보가 어려웠지만 2019년 NIA 한국지능정보사회진흥원에서 AI 개발 및 AI 데이터 기반을 조성하기 위해 구축한 AI HUB에서 데이터 셋을 공개해 이를 활용하였다. 2021년을 기준으로 총 전문 작업자들의 수작업을 통해 133,013개의 한국어 문장의 극성 값이 분류되어 있으며 매년 데이터들이 추가로 갱신된다는 측면에 있어 향후 활용성이 높을 것으로 예상된다.

〈표 3〉 문장 극성값 분류 결과

Model	Accuary(%)		
Epoch(#)	3	5	10
KoBERT	0.84337	0.84643	0.84475

확보한 133,013개의 데이터들을 7.5 : 2.5의 비율로 각각 학습 데이터 셋, 테스트 데이터 셋을 만들었고 이를 반복횟수를 변경하며 모델링해 결과를 산출하였다. 〈표3〉은 반복 횟수에 따른 KoBERT 모형의 극성 값 분류 정확도를 나타낸다. 반복 횟수가 5회일 때 가장 높은 정확도를 나타냈으며 정확도 약 84%로 높은 분류율을 나타냈다.

〈표 4〉 KoBERT 극성값 분류 결과

구분	카테고리	전체 기사(개)	긍정 기사(개)	중립 기사(개)	부정 기사(개)
전체	금융	591,315	76,144	396,290	84,979
	증권	557,413	66,532	441,057	83,726
	산업/재계	622,862	84,094	453,981	84,787
	부동산	538,131	49,015	430,753	58,363
	글로벌 경제	408,469	27,938	288,379	92,152
	경제 일반	676,028	69,062	450,651	156,315
	총계	3,394,218	372,785	2,461,111	560,322

〈표 2〉 변수설명 : 유형 및 공간단위별 부동산지수 기초통계량

유형	공간	평균	표준편차	최소값	최대값
다세대 실거래가격지수	수도권	97.7125	77.37673	86.8	122.4
	지방	94.87167	57.27583	78.5	110.9
	서울	97.41667	141.0165	83.6	131
아파트전세 실거래가격지수	수도권	90.96373	108.0699	71.498	109.531
	지방	95.15977	27.97619	80.343	104.804
	서울	91.33325	116.0211	71.209	107.846
아파트매매 실거래가격지수	수도권	98.53649	278.4683	81.017	155.961
	지방	94.93908	49.7553	82.213	115.661
	서울	99.85968	655.3092	74.794	171.16
다세대매매지수	수도권	99.74306	4.725386	96.474	104.907
	지방권	97.11354	5.244834	90.004	100.112
	서울	99.5949	10.09478	95.33	106.612
	서울 강북	99.72005	12.94939	95.279	107.806
	서울 강남	99.48843	7.965115	95.302	105.49
아파트매매지수	수도권	97.82508	65.59191	87.265	122.981
	지방권	96.54502	14.00052	85.599	105.26
	서울	97.56248	95.38598	85.115	117.068
	서울강북	97.99926	69.17811	87.13	115.473
	서울강남	97.22744	120.5792	83.394	118.427
단독주택매매지수	수도권	101.0843	36.26181	96.108	115.84
	지방권	99.1614	23.22937	92.879	109.46
	서울	100.9355	58.25549	93.988	118.876
	서울강북	101.0651	57.50486	94.511	118.972
	서울강남	100.7365	59.55146	93.053	118.728

〈표4〉는 네이버 포털에서 수집한 3,394,218개의 뉴스 기사의 극성 값을 구축된 모형으로 분류한 결과이다. 극성 값을 기준으로 카테고리를 살펴보면 중립 값의 기사들이 모든 카테고리에서 많이 나타났으며 부정기사의 비율이 긍정 기사의 비율보다 높게 나타났다. 그중 경제 일반, 글로벌 경제 카테고리의 부정기사의 비율이 가장 높게 나타났다.

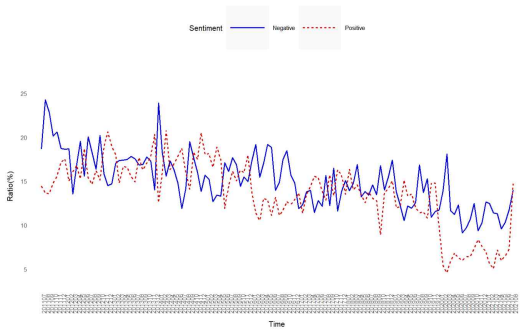
〈그림2〉는 카테고리별 긍정, 부정 비율을 나타내는 그래프로 2020년 2월 이후 언론의 긍정적 논조가 모든 카테고리에서 급격하게 감소하는 현상을 관찰할 수 있었다. 반면에 부정적 논조는 긍정적 논조와 달리 상대적으로 영향을 받지 않은 것으로 관찰되었다. 이는, 글로벌 이슈였던 COVID-19가 언론의 긍정적 보도에 영향을 크게 준 것으로 추정할 수 있다. 카테고리로 살펴보면 금융, 증권 및 산업/재계에 가장 큰 영향을 준 것으로 보이며 그 다음으로 부동산, 글로벌 경제, 경제 일반 분야의 긍정적 논조에 영향을 준 것으로 나타났다.

일부 뉴스 카테고리를 살펴보면 부동산 카테고리의 경우 2013년 이후 감소했던 부정적 논조가 2018년 말부터 다시 상승하는 경향을 보이고 긍정적 논조가 적어지는 현상을 나타냈다. 산업 재계 카

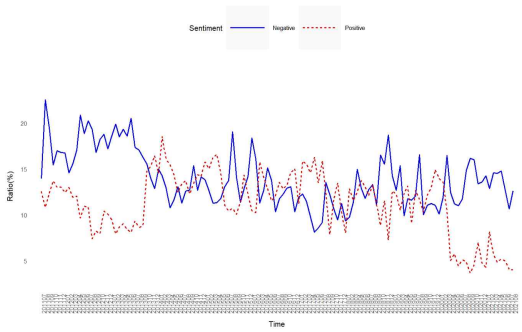
테고리의 경우 2015년 이전까지 긍정적 논조가 더 많았던 반면에 2018년 이후에는 부정적 논조가 강해진 것을 확인할 수 있었다.

〈그림 2〉 카테고리별 긍부정 기사수비율(%)

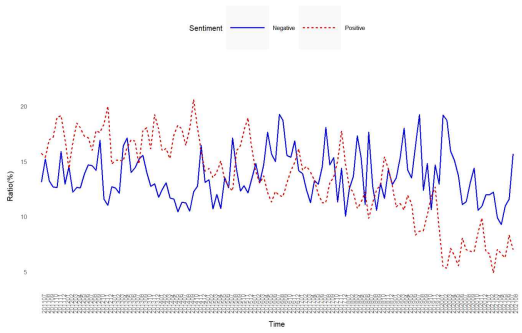
금융



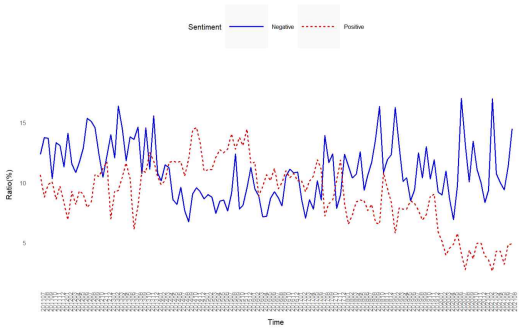
증권



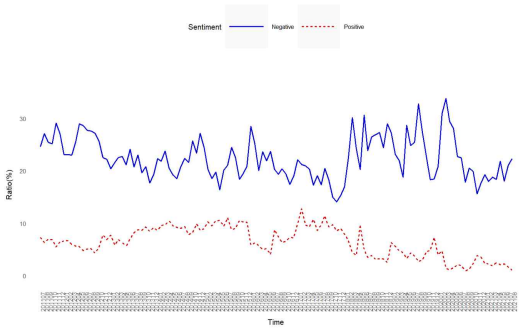
산업/재계



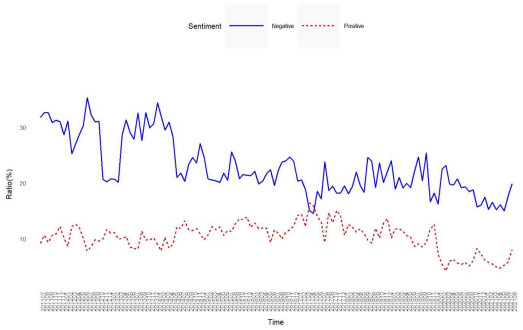
부동산



글로벌 경제



경제 일반



3. 그랜저 인과관계 검정결과

그랜저 인과관계는 시계열의 정상성이 전제되어야 하므로 ADF(augmented dickey-fuller)방법을 이용하고, 유의수준은 5%를 적용하여 수행하였다.

〈표 5〉 그랜저 인과관계 검정결과

카테고리	GrangerTest			lag	Pr.F	lag	Pr.F	lag	Pr.F	영향지속 기간(개월)
증권	언론긍정지수	→	지방권아파트전세실거래지수	1	0.0981*	2	0.1470	3	0.0806*	1
증권	언론긍정지수	→	지방권아파트매매지수	1	0.0077***	2	0.0191**	3	0.0069***	3
금융	언론긍정지수	→	수도권단독매매지수	1	0.0620*	2	0.1511	3	0.1474	1
금융	언론긍정지수	→	서울강남단독매매지수	1	0.0606*	2	0.1812	3	0.4320	1
부동산	언론긍정지수	→	수도권아파트매매실거래지수	1	0.0603**	2	0.0276**	3	0.1498	2
부동산	언론긍정지수	→	서울아파트매매지수	1	0.0302**	2	0.0507*	3	0.0770*	3
부동산	언론긍정지수	→	서울강남아파트매매지수	1	0.0169**	2	0.0410**	3	0.0793*	3
부동산	서울아파트매매실거래지수	→	언론긍정지수	1	0.0293**	2	0.0072***	3	0.0326**	3
산업재계	언론긍정지수	→	지방권다세대실거래지수	1	0.0199**	2	0.0913*	3	0.1045	2
산업재계	언론긍정지수	→	서울다세대매매지수	1	0.0710*	2	0.2528	3	0.1355	1
산업재계	언론긍정지수	→	서울강남다세대매매지수	1	0.0335**	2	0.1294	3	0.2594	1
글로벌경제	언론긍정지수	→	지방권아파트전세실거래지수	1	0.0087***	2	0.0062***	3	0.0200**	3
글로벌경제	언론긍정지수	→	지방권아파트매매지수	1	0.0140**	2	0.0118**	3	0.0313**	3
경제일반	언론긍정지수	→	지방권다세대실거래지수	1	0.0984*	2	0.1788	3	0.2318	1
경제일반	언론긍정지수	→	지방권아파트전세실거래지수	1	0.0141**	2	0.0335**	3	0.0373**	3
경제일반	언론긍정지수	→	지방권아파트매매지수	1	0.0203**	2	0.0583*	3	0.1087	2

사용되는 변수 중 5% 유의수준에서 단위근을 갖는다는 귀무가설을 기각하지 못한 변수의 경우 1차 차분하여 단위근이 없는 정상 시계열로 변환하여 검정을 수행하였다.

〈표5〉는 뉴스 기사 카테고리별로 주택매매 지수 및 실거래가격지수에 유의한 영향을 준 값들을 나타낸 것이며 유의수준은 90%신뢰도(α=0.1)로 판단하였다. 카테고리별로 결과를 보면 증권, 글로벌 경제, 경제 일반 카테고리의 언론 긍정 지수는 서울 또는 수도권이 아닌 지방권에 영향을 주는 것으로 나타났다. 특히, 지방권 아파트 전세 실거래가격지수 및 지방권 아파트 매매지수는 글로벌 경제 및 경제 일반 카테고리의 언론 긍정 지수의 영향을 2개월 이상의 상대적으로 긴 시간 동안 받는 것으로

나타났다. 반면에 금융 및 부동산 카테고리의 언론 긍정 지수는 수도권, 서울 및 서울 강남에 영향을 미치는 것으로 나타나며 금융은 단독주택 매매 지수에 부동산은 아파트 매매 지수 및 실거래가격지수에 영향을 미치는 것으로 나타났다. 산업 재계 카테고리의 언론 긍정 지수는 주택 유형 중 다세대주택에 영향을 주며 1개월의 시차로 다세대 매매 지수 및 실거래가격지수에 영향을 미치는 것으로 나타났다. 부동산 카테고리를 제외한 언론 긍정 지수와 부동산 가격지수는 양방향성이 아닌 일방향 상관관계 및 인과관계를 나타냈지만 부동산 카테고리의 경우 서울 아파트 매매실거래가격지수가 언론 긍정 지수에 지속적으로 영향을 주는 것으로 나타났다. 유의수준을 95%신뢰도(α=0.05)로 좀 더 엄격하게 판단하면 증권, 부동산, 산업 재계, 글로벌 경제, 경제 일반에서 수도권, 지방권, 서울 및 서울 강남의 아파트 관련 지수와 서울 강남 다세대 매매지수에만 영향을 준다고 나타났다.

Ⅴ. 결론

본 연구는 언론이 부동산 가격에 미치는 영향을 시공간적으로 나누어 실증분석하였다. 언론정보를 정량화하기 위해 KoBERT 모형을 AI HUB 감정 자연어 데이터 셋에 적용해 극성 값(긍정, 중립, 부정)분류 모델을 구축하였고 이를 크롤링을 통해 네이버 포털에서 수집한 6개 카테고리(금융, 증권, 산업/재계, 부동산, 글로벌 경제, 경제 일반)별 총 3,394,218개 뉴스 기사 내용에 적용해 월별 언론 논조의 극성 값을 부여하였다. 모형의 정확도는 약 84%로 비교적 높게 나타났다. 뉴스 기사의 중립 비율은 약 72%로 가장 높았고 부정 비율은 약 16%, 긍정 비율은 약 10%로 모든 카테고리에서 부정적 논조가 긍정적 논조보다 많았다.

카테고리별 긍정, 부정 비율은 시계열적으로 살펴본 결과, 대체적으로 긍정 비율은 감소하는 반면 부정 비율은 증감의 추세가 크게 나타나지 않았다. 특히, 2020년 2월 COVID-19의 영향으로 긍정 비율이 모든 카테고리에서 감소하는 일관적인 양상을 보였다. 이는 COVID-19가 언론에 영향을 크게 미친 것으로 추정해 볼 수 있다.

뉴스 기사의 카테고리별 극성 값과 부동산 주택 가격지수를 그랜저 인과관계 모형으로 분석한 결과 각 카테고리마다 영향을 미치는 종류와 범위가 상이하게 나타났다. 부동산 카테고리의 경우 수도권, 서울, 서울 강남권의 아파트 매매 지수 및 실거래가격지수에 영향을 미치는 것으로 나타났는데 이는 부동산 관련 언론 지수가 투자성이 가장 높은 수도권 내 아파트 가격에 민감하게 영향을 주었다고 볼 수 있다. 금융 관련 언론 지수도 수도권 및 서울 강남에 영향을 미쳤지만 주택 유형에 있어서는 단독주택 매매 지수에 영향을 주었다고 나타났다. 증권, 글로벌 경제, 경제 일반 카테고리의 언론 지수는 서울이나 수도권이 아닌 지방 도시의 주택 유형들에 영향을 준 것으로 나타났는데 특히 글로벌 경제, 경제 일반 카테고리는 단기적인 투자 관점보다 장기적인 수요 관점에서의 영향을 줄 수 있어 수도권보다 지방권의 주택 유형에 영향을 주었다고 판단된다. 이외에 산업 재계 카테고리의 경우 다른 카테고리와 달리 다세대주택의 매매 지수 및 실거래가격지수에만 영향을 주는 것으로 나타나 이에 대한 해석 및 연구가 더 필요할 것으로 보인다.

본 연구결과에 따르는 시사점은 언론이 부동산 가격에 영향을 미친다는 가설을 채택한다는 면에서 주택 유형 및 지역별 부동산 가격의 변화를 예측하는 데 있어 온라인 데이터가 중요한 변수로 활용될 수 있다는 점에 있다. 오프라인의 정보보다 온라인상의 정보가 기하급수적으로 증가하고 있는 빅데이터 시대에 온라인 정보를 유의미하게 가공 후 활용할 수 있다는 가능성을 보여준 점에서 해당 연구의 필요성이 존재한다고 판단된다.

데이터의 측면에서 네이버 포털 내 다양한 신문사들의 뉴스 기사를 수집해 활용했지만 언론의 논조에 있어서 편향성이 존재할 수 있으므로 이를 해결하기 위해 더 다양한 출처의 뉴스 기사 수집으로 이를 해결하는 것과 시간적인 측면에서 월 단위보다 주간 또는 일 단위의 보다 세분화된 단위로 언론과 부동산 가격의 인과관계를 밝히는 것은 추후 연구로 남긴다.

參考文獻

이제로, 박은환, 이재구, “BERT파생모델의 한국어에 대한 성능 비교”, 학술대회논문집, 한국통신학회, 2020, pp.901-902.

경정익, 이국철, “Textmining에 의한 부동산 빅데이터 감성분석 모형 개발”, 주택연구, 한국주택학회, 2016, 제24권, 제4호, pp.115-136.

박재수, 이재수, “아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석”, 국토계획, 대한국토·도시계획학회, 2019, 제54권, 제1호, pp.131-147.

서정석, 오지훈, 김정섭, “부동산 뉴스와 주택경기의 동적 관계에 대한 고찰: 언론보도논조지수 개발을 중심으로”, 한국지역개발학회지, 한국지역개발학회, 2021, 제33권, 제1호, pp.89-112.

이재수, 박재수, “방송뉴스 감성지수와 서울시 주택매매가격의 상관 및 인과관계 분석”, 주택도시금융연구, 대한국토·도시계획학회, 2020, 제54권, 제1호, pp.53-68.

Devlin, J., et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Under-standing”, Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp.4171-4186.

손재영, “지가와 거시경제변수간의 인과관계에 관한 실증분석”, 한국개발연구, 한국개발연구원, 1991, 제13권, 제3호, pp.57-58.

박재수, 이재수, “아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석”, 국토계획, 대한국토·도시계획학회, 2019, 제54권, 제1호, pp.137-138.

박종영, 서충원, “TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석”, 부동산학보, 한국부동산학회, 2015, 제63권, 제63호, pp.46-58.

전해정, 빅데이터와 텍스트마이닝을 이용한 부동산시장 동향분석, 디지털융복합연구, 한국디지털정책학회, 2019, 제17권, 제4호, pp.49-55.