國土計劃
pISSN: 1226-7147, eISSN: 2383-9171
https://doi.org/10.17208/jkpa.2021.08.56.4.99

Journal of Korea Planning Association Vol.56, No.4 (2021)
대한국토·도시계획학회지 「국토계획」 제56권 제4호 pp.99-111

부동산 감성지수의 주택가격 예측 유용성*
: 뉴스기사와 방송뉴스 빅데이터 활용 사례

Predictability of Housing Sales Prices Employing a Real Estate Sentiment Index
: Using Unstructured Big Data of Online Newspaper and TV Broadcast News

박재수** · **이재수*****
Park, Jaesoo · Lee, Jae-Su

Abstract

In recent years, computer hardware and software have improved in quality, and unstructured data analysis has become possible. Sentiment analysis, a technique extracting information such as opinions and attitudes from unstructured text, which is useful for grasping the sentiments of participants in the real estate market, is emerging as a viable alternative using Big Data. This study first aims to determine whether the new sentiment index is useful in improving the predictive power of the apartment sale price index autoregressive integrated moving average (ARIMA) model. Second, it strives to compare the degree of improvement in the predictive power of the sentiment index using the prediction error. The results showed that the national NPSI ARIMAX model including the newspaper sentiment index showed an improved prediction error of 7.90% for the root-mean-square error (RMSE) and 6.21% for the mean absolute error (MAE) compared to the national NAPI ARIMA model. Additionally, the national TVSI ARIMAX model, including the broadcast sentiment index, showed an improved prediction error of 5.05% for the RMSE and 5.42% for the MAE compared to the national NAPI ARIMA model. This confirmed that if the sentiment index were included, the predictability of the apartment sale price index could be improved. Moreover, the contribution of the newspaper sentiment index was higher than the broadcast sentiment index regarding improving the predictability of the apartment price index.

주제어 빅데이터, 머신러닝, 아파트 매매가격지수, 감성지수, ARIMAX
Keywords Big Data, Machine Learning, Apartment Sales Price Index, Sentiment Index, ARIMAX

Ⅰ. 서 론

1. 연구 배경 및 목적

주택시장은 정치·경제·사회적 변동요인에 의해 영향을 받음과 동시에 주택의 수요자와 공급자, 그리고 부동산시장을 통제·관리하려는 정부의 움직임에 의해 결정된다. 시장경제체제에서 가격은 일반적으로 수요와 공급에 의해 결정되지만, 시장이 비정상적으로 움직이면 정부는 시장에 개입하여 시장을 안정시키려고 노력한다(하성규, 2006). 이런 과정에서 시장 참여자들은 재화와 서비스의 거래에 영향을 받는다.

대다수의 주택 소비자들은 주택시장의 움직임을 방송뉴스나

* 이 논문은 박재수의 2020년 박사학위 논문 '주택시장 예측을 위한 부동산 감성지수 개발 연구' 결과의 일부를 정리하였음.

** Ph.D. of Real Estate, Kangwon National University, Managing Director of Hanil Networks Co. (First Author: okparkjaesoo@naver.com)
*** Associate Professor, Department of Real Estate, Kangwon National University (Corresponding Author: jslee25@kangwon.ac.kr)

신문기사를 통해 정보를 구득한다. 주식시장과 다르게 일반적으로 부동산 정보에는 비대칭성이 존재한다. 그래서 주택시장의 초기 변화는 많은 정보를 가진 전문가들에 의하여 가격이 견인된다. 주택가격의 움직임이 뚜렷해지기 시작하면 방송이나 신문사에서 이를 확인하고 기사 또는 방송을 통해 전달된다. '경제는 심리'라는 주장이 있듯이 주택시장도 일반대중의 심리를 얼마나 잘 이해하는지가 중요하다. 전통적 경제이론은 시장 참여자들이 합리적인 의사결정을 한다고 가정하고 가격결정 모형을 제시하고 있으나, 실제 자산시장에는 시장이 항상 효율적으로 움직이지 않는 다수의 연구결과가 제시되고 있다(Baker and Nofsinger, 2010).

우리나라에서도 부동산시장 참여자들의 심리를 측정하여 부동산시장의 모니터링 및 예측 지표로 활용하려는 시도가 나타나고 있다. 국토연구원과 KB국민은행은 설문 및 전화응답 조사자료를 통해 부동산 관련 소비자의 심리지수를 정기적으로 조사·발표하고 있다. 그러나 이들은 조사시점과 발표시점 사이의 시차, 조사 및 측정 기간의 유연성, 그리고 특정 이슈의 발생에 따른 심리지수의 변화 측정의 즉시성 문제가 제기되고 있다(송민채·신경식, 2017; 박재수, 2020).

부동산시장 참여자들의 심리는 부동산 시장의 가격을 결정하는 중요한 요소이다. 학계에서도 신문기사가 부동산 가격에 미치는 영향을 규명하는 시도가 최근 진행되고 있다. 박종영·서충원(2015)은 부동산 관련 기사를 시기별로 수집하여 주택시장의 변화를 설명할 수 있는 주요 단어와 중요도의 계량화를 시도하였고, 김대원·유정석(2016)은 소셜 미디어 중 하나인 트위터에서 부동산 관련 단어의 빈도 변수와 아파트 매매 및 전세가격 지수의 관계를 분석하였다. 경정익·이국철(2016)은 부동산 관련 신문기사를 통해 감성지수를 산출하여 부동산 시장을 예측할 수 있는 모형을 제안하였다.

이 연구의 목적은 신문기사와 방송뉴스에서 도출한 감성지수가 주택매매가격의 예측에 유용한지 실증하는 것이다. 예측 유용성 검증은 시계열 분석 모형인 ARIMA 모형과 감성지수가 추가된 ARIMAX 모형을 추정하고 계수의 통계적 유의성을 검증한다. 또한 예측오차인 평균 제곱근 오차(Root Mean Square Error, RMSE)와 평균 절대 오차(Mean Absolute Error, MAE) 값을 비교하여 감성지수가 주택매매가격의 예측을 향상시키는지 예측의 유용성을 분석한다.

연구목적을 달성하기 위해서 신뢰 있고 정확한 감성지수 산출이 필요하다. 이를 위해 기계학습(Machine Learning) 기법을 활용한 감성지수 산정 방법을 체계화하고, 부동산 관련 신문기사와 방송뉴스 텍스트 데이터를 수집하여 부동산시장에 대한 긍정과 부정 문장을 분석한다. 이를 통해 보다 개선되고 정확도 높은 부동산 감성지수를 산출한다. 감성지수 개발 방법의 정립은 본 연구에서는 상세히 다루지 않는다.[1)]

2. 연구 범위 및 방법

연구의 공간적 범위는 전국으로 설정하였다. 서울시 등 일부 대도시가 아닌 전국을 분석의 공간적 범위로 설정한 이유는 이 연구에서 활용되는 신문기사와 방송뉴스를 제공하는 언론사가 전국 부동산 뉴스를 다루는 중앙 일간지, 경제신문, 방송사이기 때문이다. 우윤석·이은정(2011)은 부동산 참여자들이 부동산 관련 경제정보를 언론보도를 통해 얻으며, 다른 지역의 주택가격이 오른다는 보도를 접하면 본인 거주지역의 주택가격도 상승할 것이라는 기대감을 갖는다고 주장하였다. 결국 언론매체는 전국 부동산시장에 영향력을 미치고 있으며, 일부 대도시에 관한 부동산 뉴스도 전국적인 파급효과를 갖는다는 것을 뒷받침한다.

분석대상 주택은 유형 측면에서는 아파트, 그리고 점유형태 측면에서는 매매로 한정한다. 아파트를 분석대상으로 선정한 이유는 전체 주택에서 아파트가 차지하는 비율이 약 60%, 공동주택에서 차지하는 비율이 약 75%로 우리나라의 대표적인 주택유형이기 때문이다. 또한 거래빈도가 타 주택유형보다 월등히 많고 구조 특성상 표준화·규격화되어 있어 가격 변동을 연구하는데 적합한 대상이다. 매매를 선택한 이유는 우리나라 가구의 자산 비중에서 부동산이 차지하는 비중이 높기 때문이다. 주택가격은 1998년 IMF 외환위기, 2008년 금융위기와 같은 큰 사건이 발생하기 전까지 지속적으로 상승했기 때문에 전세보다 매매를 통해 주택 소유 욕구가 강하게 형성되어 있는 것도 중요한 이유이다.

분석 자료는 비정형 빅데이터로서 텍스트 형태의 신문과 방송기사, 그리고 정형 시계열 데이터인 아파트 매매가격지수이다. 분석 기간은 2008년 미국 금융위기가 일단락되기 시작한 2012년 1월부터 2018년 12월까지의 기간이다. 신문기사는 중앙 일간지 3개(조선일보, 동아일보, 중앙일보), 경제지 3개(매일경제, 한국경제, 서울경제신문)에서, 그리고 방송뉴스는 지상파 3사(KBS, MBC, SBS)에서 아파트가 포함된 뉴스기사를 웹 크롤링(Web Crawling)을 통해 일간 단위로 수집하여 분석한다. 아파트 매매가격지수는 선행연구에서 활용된 KB국민은행의 월간 전국 아파트 매매가격지수를 활용한다. 수집된 신문과 방송 뉴스기사는 새로운 감성지수 개발을 위해 활용되며, 전국 아파트 매매가격지수는 개발된 감성지수와 함께 예측 유용성 분석을 위한 ARIMAX 모형의 종속변수로 활용된다.

감성지수 산정을 위해 토픽모형(Topic Model), 텍스트랭크(TextRank), TF-IDF(Term Frequency-Inverse Document Frequency), 그리고 나이브 베이즈(Naïve Bayes)와 같은 기계학습 기법을 활용하였다. 토픽모형을 통해 텍스트 데이터에서 의미있는 단어를 추출하고, 텍스트랭크 알고리즘을 이용해 추출한 단어의 중요도를 측정하여 이를 근거로 감성사전을 만든다. 감성사전과 TF-IDF를 활용하여 분석할 문장에서 특정 단어들의 중요도를 수

치화하고, 나이브 베이즈 분류 모델은 특정 단어들에 부여된 수치를 활용하여 분석 문장의 긍정 및 부정 극성과 극성에 따른 가중치를 계산한다. 이 가중치를 이용하여 월별 감성지수를 산출한다.

감성지수의 유용성은 시계열 자료인 아파트 매매가격지수의 예측오차를 얼마나 감소시키는지를 측정하여 확인한다. 우선, 감성지수와 아파트 매매가격지수 간의 상관성을 교차상관 분석을 통해 분석한다. ARIMAX 모형을 이용하여 감성지수가 아파트 매매가격 예측 모형에서 유의미한 변수인지 분석한다. 아파트 매매가격지수 단일변수 ARIMA 모형과 감성지수가 포함된 ARIAMX 모형에서 도출한 예측오차인 평균 제곱근 오차(RMSE)와 평균 절대 오차(MAE)를 비교하여 감성지수의 유용성을 분석한다. 감성분석은 파이썬(Python) 프로그램을 이용하고, 시계열 분석은 R 통계 프로그램(Version 3.4.1)과 Eviews(Student Version 11)를 활용한다.

II. 선행연구 검토

1. 심리지수를 이용한 부동산시장 연구

경제지표의 변화 예측에 시장 참여자들의 심리변수가 유용하다는 연구가 다양하게 제시되고 있다. 옥기율·김지수(2012)는 한국은행이 발표하는 소비자 심리지수가 주식시장의 수익률에 미치는 영향을 분석하였다. 소비자 심리지수 정보가 부정적으로 변화하면 주식시장은 과잉반응을, 반대로 긍정적으로 변화하면 과소반응을 나타내며 영향을 미치는 것으로 분석되었다. Li et al.(2016)은 로이터에서 수집한 석유 뉴스를 빅데이터 분석을 통해 서부 텍사스 중질유(West Texas Intermediate)의 가격 예측에 유용한지 실증하였다. 기사에서 도출한 감성지수가 원유가격과 동일한 방향성을 나타내고, 이들 사이에 3주 시차를 두고 유의미한 인과관계가 나타남을 실증하였다.

부동산시장 참여자들의 부동산시장의 변화에 대한 태도와 심리의 중요성이 강조된다. 국토연구원은 부동산 소비자의 행태 변화와 인지수준을 조사하고 부동산시장 변화 분석의 자료로 사용하기 위해 2013년 4월부터 부동산시장 소비자심리지수를 정기적으로 발표한다. 심리지수와 부동산시장의 관계를 분석한 주요 연구를 정리하면 다음과 같다.

최영걸 외(2004)는 서울시 주택시장의 가격에 대한 기대심리를 적응적 기대와 합리적 기대가설을 적용하여 실증하였다. 분석 결과, 서울시 주택시장은 적응적 기대가 지배하는 시장으로 기대심리에 의해 버블이 있을 수 있다고 주장하였다. 최희갑·임병준(2009)은 투자자의 부정적 태도가 부동산시장의 침체에 영향을 미칠 수 있는 가설을 검증하였다. 인과관계 검정과 오차수정모형을 통해 가격전망지수의 시차변수는 주택가격의 증가율에 대해 유의미한 설명력을 나타냈다.

김대원·유정석(2013)은 주택소비심리지수와 주택 거래량의 관계를 유한시차분포(FDL) 모형을 이용하여 분석하였다. 주택가격과 관련된 심리 변수인 주택소비심리지수는 일정 시차를 두고 주택 거래량 결정에 유의미한 영향을 미치는 것으로 분석되었다. 이는 우리나라 주택시장에서 소비심리지수가 주택시장의 변화를 예측할 수 있는 변수임을 시사한다. 김리영·안지아(2013)는 소비자의 주택가치 전망이 주택가격이나 거래에 미치는 영향을 분석하였다. 분석 결과, 주택시장에서 소비자의 주택가치에 대한 전망이 주택의 가격보다는 거래량에 영향을 미친다고 주장하였다.

조태진(2014)은 서울·부산·대구를 포함한 7개의 대도시를 중심으로 거시경제변수와 부동산심리지수가 주택시장에 주는 영향을 패널 데이터 모형으로 분석하였다. 연구 결과, 경제심리지수와 소비자심리지수는 통계적으로 유의하지 않으나, 부동산전망지수는 모든 시차에서 유의미하게 나타났으나, 분석 결과의 일관성은 나타나지 않았다. 유한수·정재호(2015)는 주택시장에서의 투자자 매매심리를 나타내는 주택매매시장 소비심리지수와 주택매매가격지수의 관계를 분석하였다. 분석 결과, 주택매매가격지수와 소비심리지수는 양방향의 인과관계를 나타냈다. 소비심리지수의 증가는 주택매매가격의 증가에 영향을 미치고, 주택가격의 증가는 다시 소비심리를 높이는 것으로 판단된다.

노민지·유선종(2016)은 주택시장의 수요와 공급에 의한 작동원리와 더불어 소비자의 심리가 주택가격의 변동을 설명하는 데 유용한지 분석하였다. 국토연구원이 발표하는 주택매매시장 소비자심리지수와 인터넷 검색량을 이용하여 아파트 실거래가격에 미치는 영향을 분석하였다. 분석 결과, 인터넷 검색량과 아파트 매매가격 변동률과는 동시성을 나타냈고, 소비자심리지수 변동률은 0개월 시차와 3개월 시차에서 아파트 매매가격 변동률과 양의 영향 관계를 나타냈다.

2. 뉴스기사를 이용한 부동산시장 연구

일반적으로 자산시장에서 가격은 시장에 도달하는 새로운 정보에 의하여 영향을 받는다. 부동산시장도 이에 예외가 될 수 없다. 이와 관련된 선행연구를 부동산시장 이외의 시장과 부동산시장으로 구분하여 정리하면 다음과 같다.

먼저 부동산 이외의 자산시장에 대한 연구로 송치영(2002)은 뉴스 발생과 주식 및 외환시장의 변동을 분석해 뉴스가 주가와 원-달러 환율에 유의미한 영향을 미치고, 영향력은 외환시장보다 주식시장에서 크게 나타난다고 주장하였다. 안희중 외(2010)는 남북관계 뉴스가 우리나라 주식시장에 미치는 영향을 분석하기 위하여 단변량 회귀분석을 활용하였다. 분석 결과, 긍정적 뉴스에는 양의 방향으로, 부정적 뉴스에는 음의 방향으로 주가가 반응하는 것으로 나타났다.

김유신 외(2012)는 빅데이터 감성분석을 이용해 지능형 투자의

사결정모형을 제안하였다. 연구를 통해 주식시장 개장 전 뉴스 콘텐츠의 감성지수와 주가지수의 등락이 통계적으로 유의미한 관계가 있다고 주장하였다. 김동영 외(2014)는 기업 주가의 변동을 뉴스기사와 SNS 데이터를 활용해 감성분석과 기계학습 방법으로 예측하는 연구를 수행하였다. 기계학습 기법을 적용하면 기존 연구보다 개선된 결과를 도출할 수 있음을 밝혔다.

뉴스기사가 부동산가격에 미치는 영향에 관한 연구로 Gayer and Viscusi(2002)는 유해 폐기물 지역에 입지한 주택의 가격과 관련 뉴스기사 사이의 관계를 분석하였다. 분석 결과, 정부의 유해물질 처리 기금과 관련된 기사 수의 증가가 주택가격에 긍정적인 영향을 미치는 것으로 나타났다. 김진유(2006)는 '투기'라는 단어가 들어간 신문기사가 부동산가격에 어떤 영향을 미치는지 분석하였다. 전국과 서울시의 아파트 가격과 투기가 포함된 부동산 기사의 수 사이에 양방향의 인과관계가 나타난 것으로 분석되었다. 이 연구는 신문기사의 방향성과 주택가격 변동 사이에 유의미한 인과관계가 있음을 밝혔다.

우윤석·이은정(2011)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. 또한 언론보도에 영향을 받은 시장 참여자의 부동산시장에 대한 기대가 정부의 부동산 정책의 효과성에도 영향을 준다고 주장하였다.

진창하·Gallimore(2012)는 객관적인 정보뿐만 아니라 직관과 투자심리가 부동산 시장 참여자들에게 영향을 준다고 주장하였다. 이를 분석하기 위해 애틀랜타 CMSA를 사례로 오차수정모형을 이용하여 신문기사의 내용과 부동산가격 사이의 관계를 분석하였다. 이 연구는 신문기사의 내용이 주택가격 변동에 유의미한 영향을 미친다는 결과를 제시하였다. 또한 부정적 용어의 사용이 긍정적 용어의 사용보다 신문기사와 주택가격의 변화와 더 높은 연관성이 있다고 주장하였다.

김대원·유정석(2016)은 트위터 정보와 전국과 서울의 아파트 매매 및 전세가격 간 동적 관계를 분석하였다. 분석 결과, 이들 사이에는 유의미한 관계가 있는 것으로 나타났다. 아파트 매매가 변동에 더 큰 영향을 주는 트위터 내용은 상승보다 하락과 관련된 내용으로 나타났다. Sun et al.(2014)은 온라인 뉴스와 웹 검색 데이터를 이용하여 부동산가격 예측 모델을 제시하였다. 이를 통해 뉴스와 검색 데이터 변수를 추가한 모델이 예측오차를 줄일 수 있다고 주장하였다.

박재수·이재수(2019)는 온라인 신문기사 데이터를 이용하여 분석한 감성지수가 서울시 아파트 매매가격과 상관 및 인과관계가 있는지 분석하였다. 분석 결과, 소형아파트의 매매가격만 감성지수와 1개월 시차의 교차상관관계 및 인과관계를 나타냈다. 이 연구는 신문기사 내용 중 긍정적인 표현이 아파트 매매가격과 유의미한 영향 관계가 있으며, 이는 주로 소형아파트에서 나타난다고 주장하였다.

주택가격을 포함한 부동산가격의 예측을 위한 선행연구는 대부분 금리, 물가지수 등 다양한 거시경제지표를 활용한 모형을 추정하는 방향으로 지속되었다. 부동산시장 심리지수는 비교적 최근 연구에서 고려하기 시작하였다. 국토연구원에서 주기적으로 조사·분석을 통해 공개하는 부동산시장 소비자심리지수는 부동산시장의 심리지수의 측정 및 활용에 의의가 있다. 그러나 직접 조사를 통해 측정된 심리지수는 시장의 여건 변화를 즉각적으로 반영하기 어렵고, 가격 또는 거래 등 부동산시장의 주요 지표를 별도로 분리하지 않는다. 또한 월 단위 지수로써 부동산시장 변화를 시간적으로 바로 반영하기 위한 즉시성의 한계도 있다.

최근 거시경제변수만을 이용한 부동산시장 변화 예측의 한계를 극복하려는 새로운 노력이 진행되었다. 경제부문 참여자들이 실시간으로 생성하는 데이터를 수집·분석하여 부동산시장의 변화를 설명 또는 예측하려는 시도가 나타나고 있다. 신문 및 방송뉴스 텍스트 데이터 등 비정형 빅데이터를 이용하여 감성지수 등을 산정하고, 이를 주택가격 등 부동산시장의 변화를 설명 또는 예측하는 데 활용하는 연구가 제시되고 있다. 그러나 선행연구는 대부분 텍스트 데이터에 내재한 질적 내용과 논조보다는 기사 수나 특정 단어의 수 등 양적 변수를 활용하는 데 그치고 있다.

부동산과 관련된 비정형 빅데이터 분석기법을 적용하여 감성지수 산출하고, 이를 주택가격 예측에 활용한 국내 연구는 여전히 미흡하다. 이 연구는 신문기사와 방송뉴스 텍스트에서 도출한 감성지수를 이용하여 주택가격의 예측 유용성을 분석한 점에서 학술적 의의가 있다. 신문기사를 수집·분석하여 부동산가격의 설명 또는 예측에 활용한 선행연구는 최근 일부 제시되고 있다. 그러나 이 연구는 신문과 방송뉴스 데이터를 분석하여 신문 감성지수와 방송 감성지수를 산정하고, 매체 간 주택가격의 예측 유용성을 비교·분석한 최초의 연구인 점에서도 차별성이 있다.

이 연구는 통계학과 기계학습(Machine Learning)을 이용하여 n-gram 단어사전을 만들고 이를 기반으로 BERT모델[2]을 부동산 감성사전에 적용하여 성능이 뛰어나면서도 범용적인 AI모델을 만들었다. 본 논문은 여기에서 중요한 특징을 도출하고 감성사전을 구축하기 위하여 토픽모델링과 비지도학습인 텍스트랭크 알고리즘을 활용하였다. 또한 문장 단위의 극성을 분류하기 위한 모형을 만들기 위하여 나이브 베이즈 분류 모델을 적용하였다. 연구에서 개발하고자 하는 감성지수는 일반가구가 부동산시장에 대한 의견형성의 기초가 되는 신문기사와 방송뉴스의 문장에 대한 극성을 분석해서 지수를 도출한다. 이러한 방법은 월간, 주간 또는 일간 단위까지 분석기간을 유연하게 적용할 수 있는 장점도 있다.

Ⅲ. 분석자료와 방법

1. 분석자료

본 연구는 주택 매매가격지수로 KB국민은행에서 공개하는 전국 아파트 매매가격지수를 사용하였다. 아파트는 우리나라에서 가장 많은 주택유형이고 규격화되어 거래가 쉽고 빈번하기 때문이다. 통계청의 2015년 인구주택총조사에서 아파트가 차지하는 비중이 60%로 가장 높은 비중을 차지한다. 자료의 수집기간은 2008년 금융위기 국면이 진정되기 시작한 2012년부터 2018년 말까지 총 84개월이다(〈Table 1〉 참고).

이 연구는 유효부수가 상대적으로 많은 일간지와 경제지를 적절히 고려하여 주요 일간지와 경제지를 각각 3개씩 선정하였다. 일간지는 조선일보, 동아일보와 중앙일보이고, 경제지는 매일경제, 한국경제와 서울경제신문이다. 한국ABC협회(http://www. kabc.or.kr) 2017년도 자료에 따르면, 유료부수를 기준으로 조선일보 1위, 동아일보 2위, 중앙일보 3위, 매일경제 4위 한국경제 5위, 서울경제 24위로 제시되었다.

부동산 방송뉴스는 지상파 3사(KBS, MBC, SBS)에서 부동산이 포함된 뉴스를 추출하였다. 해당 사이트의 경제면에서 '부동산'과 관련된 기사 99,427건을 웹 크롤링(Web Crawling)한 후, 연구 목적에 맞게 '아파트'와 '매매'가 포함된 뉴스기사를 2차로 분류하였다. 수집된 신문 및 방송뉴스 기사 건수는 〈Table 2〉와 같다.

Table 1. Data description

Variable	Period	Region	From-To	Obs.
KB APT price index	Monthly	National	2012.1.-2018.12.	84

Table 2. The number of news articles and sentences

News -paper	No. of articles	No. of sentences	Broad -cast	No of articles	No. of sentences
Chosun Ilbo	3,129	67,162	KBS	3,276	42,460
Dong-A Ilbo	2,560	46,134	MBC	2,001	20,033
Joongang Ilbo	2,563	78,868	SBS	2,886	42,460
Korea Economic Daily	8,264	147,363	-	-	-
Maeil Business Newspaper	7,847	137,714	-	-	-
Seoul Economic Daily	4,031	75,096	-	-	-

2. 분석 방법

1) 분석절차

언론매체를 통한 부동산 관련 뉴스는 부동산시장 참여자에게 심리적 영향을 미친다. 하지만 이런 심리지수를 계량화하여 객관적인 지표로 나타내는 것은 어려운 문제이다. 최근에 부동산 시장 참여자들의 심리상태를 수치화하려는 시도가 지속되고 있다(박재수·이재수, 2019). 본 연구는 부동산시장 참여자들의 심리상태를 신문과 방송의 뉴스기사를 활용하여 지수화하고, 이 감성지수가 전국 아파트 매매가격지수의 예측에 유용한지 통계적 검증을 통해 분석하는 것이 목적이다.

우선, 신문 및 방송 뉴스기사를 활용하여 타당한 감성지수를 산출하는 것이 필요하다. 이를 위해 뉴스기사의 수집 및 전처리, 뉴스기사 데이터를 이용한 토픽분석을 실시하였다. 토픽분석으로 추출한 주요 주제와 단어가 포함된 문장을 추출하고 텍스트랭크알고리즘를 활용하여 감성사전을 만든다. 또한 TF-IDF 알고리즘과 나이브 베이즈 분류 모델을 적용하여 문장의 가중치를 산정하고 신문과 방송 감성지수를 산출한다.

둘째, 신문 및 방송 감성지수를 외생변수로 선정하고, 이 지수가 전국 아파트매매가격지수를 예측하는 데 유용한지 검증한다. 예측 유용성 분석은 시계열 분석에서 널리 이용되는 ARIMA와 ARIMAX 분석방법을 적용하였다. 전국 아파트 매매가격지수의 최적 ARIMA 모형을 추정하고, 모형의 예측 오차인 RMSE, MAE를 구한다. 그리고 신문 및 방송 감성지수의 선행시차 변수를 외생변수로 투입한 전국 아파트 매매가격지수의 ARIMAX 모형을 추정하고, 모형의 예측 오차인 RSME, MAE를 계산한다. 마지막으로, ARIMA 모형과 ARIMAX 모형의 예측 오차를 비교하여 신문 및 방송 감성지수의 유용성을 비교·분석한다.

2) 감성지수 산정 방법

감성지수를 산정하기 위해 토픽모델, 텍스트랭크, TF-IDF, 나이브 베이즈 방법론을 활용하였다. 〈Figure 1〉은 감성지수 분석을 위한 절차도이다.

첫 번째 단계로 부동산 관련 신문기사와 방송뉴스 텍스트 데이터를 월별로 수집하고, 문장 단위로 데이터를 재분류한 후 전처리 작업을 시행하였다. 두 번째 단계로 토픽분석를 실시하여 신문기사와 방송뉴스의 부동산 관련 토픽을 각각 8개씩 분류하고, 한 토픽에 30개 단어를 추출하였다. 각 단어가 포함된 총 9,600개 샘플 문장을 무작위로 선정하였다.

세 번째 단계로 텍스크랭크를 이용하여 샘플 문장에 포함된 단어들의 관계를 분석하고, 감성사전을 만든다. 네 번째 단계로 TF-IDF 알고리즘을 이용하여 문장에 나온 단어들의 점수를 계산하고, 이를 나이브 베이즈 분류 모델에 투입한다.

다섯 번째 단계로 나이브 베이즈 모델을 이용하여 문장에 긍정

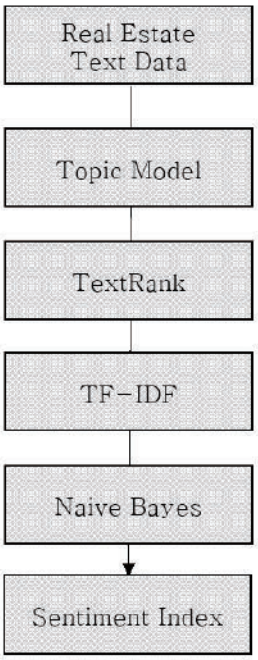

Figure 1. Flow of sentiment index

Source: Park, 2020

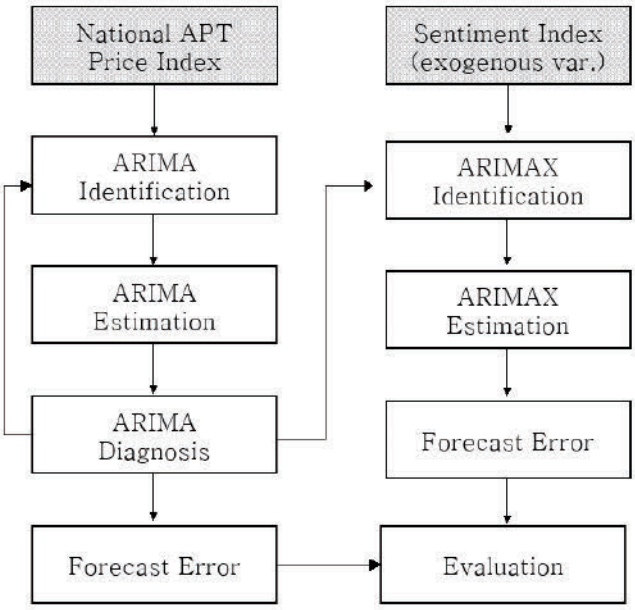

Figure 2. Flow of ARIMA and ARIMAX

Source: Park, 2020

및 부정 극성을 분류하고, 긍정 및 부정 가중치를 계산한다. 마지막으로 Eq. 1에 따라 월별 신문 감성지수(NPSI)와 방송 감성지수(TVSI)를 계산한다. 밑수 p는 긍정, n은 부정을 의미한다.

$$NPSI_i = \sum_{i=1}^{n} NPSI_p(i) - \sum_{i=1}^{n} NPSI_n(i) \quad (1)$$
$$TVSI_i = \sum_{i=1}^{n} TVSI_p(i) - \sum_{i=1}^{n} TVSI_n(i)$$

3) 예측 유용성 분석

감성지수의 예측 유용성 분석방법은 ARIMA와 ARIMAX 모형을 활용한다. ARIMA 모형은 자신의 시계열 자료를 근거로 가격 변동을 예측하며, 어떠한 시계열에도 적용이 가능하다. 특히 시간의 흐름에 따라 자료가 빠르게 변동하는 경우에도 시계열 예측이 가능한 장점으로 경제지표 예측에 많이 사용되고 있다(송경재·양희민, 2005).

〈Figure 2〉는 전국 아파트 매매가격지수에 대한 ARIMA 모형과 감성지수가 포함된 ARIMAX 모형을 추정하여 감성지수의 예측 유용성을 분석하는 흐름을 나타내고 있다. 시계열 자료의 예측에 이용하는 기본모형은 자기회귀요소를 이용하는 AR(Auto Regressive) 모형과 이동평균요소를 이용하는 MA(Moving Average) 모형이 있다. 그리고 이 두 요소를 동시에 고려하는 ARMA(Auto Regressive Moving Average: 자기회귀이동평균) 모형이 있다.

자기회귀모형인 AR모형은 현재 관측 값 Y_t가 과거 관측 값에 의해 설명되는 모형이다. AR(p)모형은 아래 식과 같다. ϕ_p는 y_{t-k}의 자기회귀계수, p는 자기회귀시차를 의미한다.

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (2)$$

이동평균모형인 MA모형은 현재 관측 값 Y_t가 과거 오차항의 선형결합으로 설명되는 모형으로 MA(q)모형은 아래 식과 같다. θ_q는 이동평균계수, q는 이동평균 시차를 나타낸다.

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots \theta_q \varepsilon_{t-q} \quad (3)$$

시계열 Y_t가 AR과 MA를 동시에 가지고 있으면 시계열 데이터는 자기회귀이동평균(ARMA) 모형을 따른다. ARMA(p, q)는 아래 식과 같다.

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-p} \quad (4)$$

그러나 대부분의 시계열 자료는 안정성을 확보하지 못하여 차분을 통해 안정적인 시계열로 변환한다. 불안정한 시계열 자료를 차분을 통해 안정적인 시계열 자료로 변환한 후 AR과 MA모형을 적합한 모형이 ARIMA 모형이다. ARIMA(p, d, q)모형에서 p는 AR모형의 차수, d는 차분 차수, q는 MA모형의 차수이다.

신문 및 방송 감성지수의 아파트매매가격지수 예측 유용성을 검증하기 위하여 독립변수를 고려한 시계열 모형인 ARIMAX-(Auto Regressive Integrated Moving Average with eXogenous variables) 모형을 적용한다. ARIMA 모형은 변수 자신의 과거 값만을 이용하기 때문에 회귀분석에서의 관심사인 특정 독립변수와의 관계를 분석하기는 어렵다. ARIMAX 모형은 다변

량 시계열 자료에서 종속변수에 영향을 미치는 독립변수인 시계열 변수 사이의 인과관계를 분석하는 대표적인 모형으로 단변량 ARIMA 모형을 확장한 것이다(이종민 외, 2017). 본 연구에서 사용한 감성지수가 포함된 ARIMAX 모형은 아래의 식과 같다.

$$\Delta y_t = \phi_0 + \phi_1 \Delta y_{t-1} + \ldots + e_t - \theta_0 - \theta_1 \varepsilon_{t-1} - \ldots + \alpha_k si_t \quad (5)$$

감성지수의 예측 유용성을 검증하기 위하여 변수의 오차 측정 지표 중 평균 제곱근 오차(RMSE)와 평균 절대 오차(MAE)를 사용한다. RMSE는 Eq. 6과 같이 모형의 예측 값과 실제 값의 차이를 의미하며, 예측 값의 크기에 의존한다. k는 자유도, Y_t는 실제 값, $\widehat{Y}_t$는 예측 값을 의미한다.

$$\sqrt{\frac{1}{n-k}\sum_{t}^{n}(Y_t - \widehat{Y}_t)^2} \quad (6)$$

MAE는 Eq. 7과 같이 실제 값과 예측 값의 차이를 양의 정수로 변환한 차이 값의 평균이다. MAE의 장점은 계산이 용이한 점이다. Y_t는 실제 값, $\widehat{Y}_t$는 예측 값을 의미한다.

$$\frac{1}{T}\sum_{t=1}^{T}|Y_t - \widehat{Y}_t| \quad (7)$$

두 지표 모두 낮은 예측 오차값이 더 좋은 결과를 의미한다. 그러나 RMSE는 오류가 평균화되기 전에 제곱으로 계산되기 때문에 큰 오류에 상대적으로 높은 가중치를 부여할 수 있다(이종민, 2018).

Ⅳ. 분석 결과

1. 기초통계 분석

1) 전국 아파트 매매가격지수 기초 통계량

KB국민은행의 전국 아파트 매매가격지수는 2012년 1월부터 2013년 중반까지는 하향 안정세를 나타냈다. 그러나 2013년 3분기부터 매매가격이 상승세를 보이면서 대부분 2018년 말까지 우상향하는 흐름을 보였다.

전국 아파트 매매가격지수의 시간적 변화를 나타내면 〈Figure 3〉과 같다. 2012년부터 2018년까지 정부의 부동산 대책발표는 4번이 있었다. 2008년 금융위기 이후 정부는 부동산시장 안정화 및 부양을 위한 여러 조치를 취했다. 2012년 '5.10 부동산 대책', 2014년 '7.24 부동산 대책' 등이 대표적이다.

KB 국민은행이 2019년 1월을 100으로 하여 산출한 전국 아파트 매매가격지수 수준변수와 계절조정 후 수준변수의 기초 통계량은 〈Table 3〉과 같다.

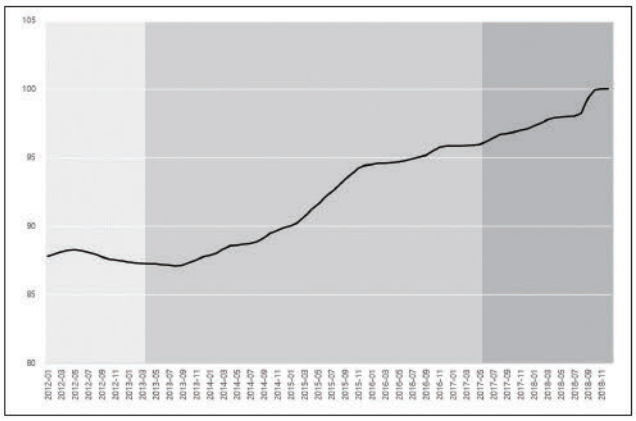

Figure 3. National APT price index

Source: KB Kookmin Bank, https://onland.kbstar.com/

Table 3. Descriptive statistics of national APT

	Mean	Median	S.D.	Min	Max
Level variable	92.31	92.27	4.218	87.09	100.03
Seasonal adjustment	92.30	92.34	4.206	87.12	99.93

2) 감성지수 기초 통계량

신문과 방송뉴스의 부동산 관련 기사에서 산출한 신문 및 방송 감성지수의 기초 통계량은 〈Table 4〉와 같다. 매체별 감성지수는 나이브 베이즈 분류기법을 활용하여 긍정과 부정을 판별하고, 기준 값인 0을 중심으로 긍정 감성이 높으면 양의 값을, 부정 감성이 높으면 음의 값을 나타낸다. 평균과 중위수 모두 0보다 큰 이유는 분석기간 중 부정적 기사보다 긍정적 기사가 많았기 때문이다. 2012년부터 2016년 말까지 부동산시장 부양을 통해 경제를 활성화하려는 정부 정책의 영향으로 판단된다.

언론매체별 감성지수의 수준변수와 계절조정 수치의 변화를 나타내면 〈Figure 4〉, 〈Figure 5〉와 같다. 2012년 중반부터 2013년 중반, 그리고 2017년 초반과 중반을 제외하고 신문과 방송 모두 전반적으로 부동산시장에 대한 긍정 감성지수가 높게 나타난다.

또한 신문 감성지수의 변동이 방송 감성지수의 변동보다 적다.

Table 4. Descriptive statistics of sentiment index

		Mean	Median	S.D.	Min	Max
News paper	Level var.	0.119	0.1448	0.140	-0.239	0.393
	Seasonal adj.	0.119	0.1582	0.136	-0.263	0.337
TV	Level var.	0.111	0.1559	0.176	-0.315	0.479
	Seasonal adj.	0.111	0.1294	0.168	-0.381	0.432

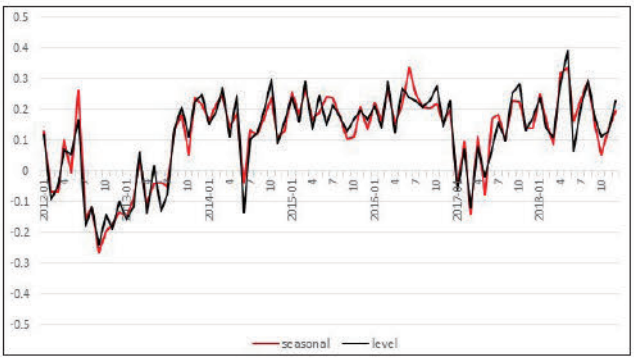

Figure 4. Newspaper sentiment index

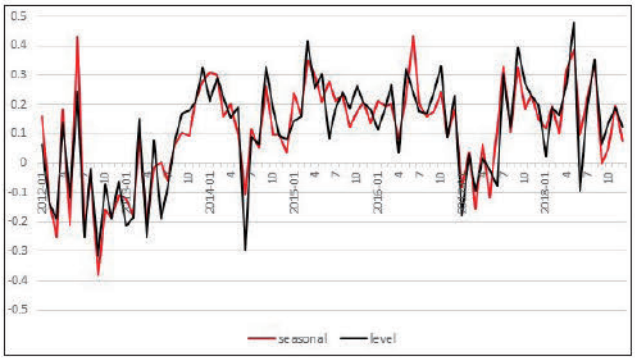

Figure 5. TV sentiment index

신문기사는 방송뉴스와 달리 부동산시장에 대해 다양한 기사를 생산할 수 있어 긍정과 부정 기사가 고르게 보도되기 때문이다. 방송뉴스는 부동산시장에 대한 보도량이 신문보다 적고 방송의 시간적 제약도 있다. 따라서 부동산시장이 호황이면 호황에 대한 뉴스 위주로, 불황이면 불황에 대한 뉴스 위주로 전달하는 경향이 있다. 하지만 신문기사는 부동산경기에 대한 기사뿐만 아니라 아파트 분양과 관련된 기사내용도 포함되어 있다. 신문기사는 양비론적 논조의 특성도 있어 부동산시장이 불황일 때도 긍정적인 면을 강조하는 경향이 있어 감성지수 변동이 상대적으로 적은 것으로 판단된다.

2. 단위근 검정과 교차상관 분석

1) 단위근 검정

시계열 자료에 단위근이 존재하면 추세를 포함하기 때문에 추세를 제거하여 안정적인 시계열로 만들어야 한다. 단위근 검정은 ADF(Augmented Dickey-Fuller), PP(Phillios -Perron) 검정법을 사용하였다. ADF는 검정 조건으로 절편만을 방정식에 포함하고, 시차는 최대 11차까지 SC(Schwatz Criterion)를 기준으로 검정하였다. PP도 검정 조건으로 절편만을 방정식에 포함하고, Bandwidth는 Newey-West Bandwidth을 사용하였다.

전국 아파트 매매가격지수의 단위근 검정 결과는 〈Table 5〉와 같다. 단위근에 대한 귀무가설을 기각하지 못해 수준변수가 단위근을 가지고 있는 것으로 분석되었다. 지수 데이터의 안정성을 확보하기 위해서 1차 차분하여 추세를 제거하고 정상 시계열로 변환하였다. 매체별 감성지수는 수준변수에서 단위근을 가지고 있지 않은 것으로 나타나 수준변수를 그대로 사용하였다.

Table 5. Result of unit root test

Variable	Level		1st difference	
	ADF	PP	ADF	PP
NAPI	0.6014	1.2156	-3.9088***	-3.8003***
NPSI	-2.6270*	-4.5233***	-	-
TVSI	-3.0119**	-6.7480***	-	-

Significance level: * < 0.10, ** < 0.05, *** < 0.01
NAPI: National Apartment Price Index
NPSI: NewsPaper Sentiment Index
TVSI: TeleVision Sentiment Index

2) 교차상관 분석

전국 아파트 매매가격지수와 감성지수 간 교차상관관계를 확인하기 위하여 ±24개월의 시차로 교차상관 분석을 실시하였다. 분석결과는 〈Table 6〉과 같다. 시차 0을 중심으로 양의 상관관계를 나타내고 있고, 시차 1에서 상관계수가 가장 크다. 부동산가격 상승과 관련된 긍정적 뉴스의 증가가 시차를 두고 부동산가격에 양의 관계를 나타내는 것이다. 이종민 외(2017)도 전세가격지수와 전세 토픽 및 검색지수 간 교차상관 분석에서 시차 0을 중심으로 대칭을 나타낸다.

시차 1에서 전국 아파트 매매가격지수와 신문 감성지수 간의 상관이 가장 강하게 나타난 것은 아파트 매매가격의 상승(하락) 초기에 이와 관련된 뉴스를 언론사들이 사전에 기사화하기 때문이다(김대원·유정원 2016). 감성지수가 전국 아파트 매매가격지수에 선행성을 나타낸 것은 이 지수가 전국 아파트 매매가격지수 ARIMA 모형의 외생변수로 활용될 수 있음을 의미한다.

Table 6. Result of cross correlation

Lag	NPSI-NAPI		TVSI-NAPI	
	Lead	Lag	Lead	Lag
0	0.4321		0.3666	
1	0.4997*	0.4021	0.4566*	0.3930
2	0.4811	0.4237	0.4399	0.3949
3	0.3742	0.3490	0.3126	0.2961
4	0.4428	0.3110	0.3989	0.2608
5	0.4221	0.2375	0.3937	0.1961
6	0.3201	0.2088	0.2935	0.1696

NAPI: National Apartment Price Index
NPSI: NewsPaper Sentiment Index
TVSI: TeleVision Sentiment Index

3. 예측 유용성 분석

1) 나이브 베이즈 분류 결과

본 연구에서 사용한 나이브 베이즈 모형의 목표 값은 예측해야 할 부동산 뉴스의 긍정 또는 부정 극성이며, 특정 값은 나이브 베이즈 모델에 투입되는 특정 단어이다. 또한 나이브 베이즈의 결과로 도출된 값은 나이브 베이즈 모델에서 개별 단어로 구성된 문장이 긍정 또는 부정에 영향을 미치는 정도를 의미한다.

예를 들어 '분양시장의 열기는 계속되고 있다'라는 문장은 '분양', '시장', '열기', '계속'의 단어로 분리된 후 각 단어가 나이브 베이즈 분류 모델에 투입된다. 투입된 단어들은 나이브 베이즈 모델에 의해 긍정 또는 부정이라는 극성이 부여됨과 동시에 해당 극성의 강도를 숫자로 표시한다. 이와 같이 각 문장에 대한 극성 및 극성의 강도를 긍정과 부정으로 분류한 후 긍정과 부정 값의 월별 합계 차이를 〈Table 7〉과 같이 얻었다.

2) ARIMA 분석 결과

매체별 감성지수와 전국 아파트 매매가격지수 간의 교차상관 분석에서 감성지수의 선행성을 확인하였다. 전국 아파트 매매가격지수를 예측하기 위해 감성지수를 외생변수로 활용할 수 있다. 우선, 전국 아파트 매매가격지수(NAPI)에 대한 적정 ARIMA 모형을 Box-Jenkins 방법에 따라 선정하였다. 분석 결과, 전국 아파트 매매가격지수의 ACF (AutoCorrelation Function) 값은 서서히 감소하고, PACF (Partial AutoCorrelation Function) 값은 첫 번째에서 스파이크가 나타났다. ACF 값이 서서히 줄어든다는 것은 수준변수가 추세를 가지고 있음을 의미한다.

수준변수에서 추세를 제거한 1차 차분 변수를 이용하면 ACF 값은 지수적으로 감소하고, PACF 값은 첫 번째에서 스파이크를 나타낸다. 1차 차분한 매매가격지수의 PACF 값이 유의수준을 넘어서고 있어 전국 아파트 매매가격지수 ARIMA(p, d, q) 모형을 ARIMA(1,1,0)으로 결정하고 잔차에 대한 Q 통계량 검증을 실시하였다. 그러나 Q 통계량의 p값이 유의수준보다 작게 나타나 모형의 잔차(Residual)에 자기상관이 존재하는 것으로 분석되었다.

Table 7. Sample of result of naive bayes classification

Date	NPSI	TVSI
2012-01	0.1184	0.0636
2012-02	-0.0911	-0.1375
2012-03	-0.0441	-0.1874
2012-04	0.0175	0.1386
2012-05	0.0533	-0.1176
2012-06	0.1691	0.2407
2012-07	-0.1758	-0.2508

NPSI: NewsPaper Sentiment Index
TVSI: TeleVision Sentiment Index

잔차의 자기상관을 제거하기 위하여 AR(2)를 추가하여 새로운 전국 아파트 매매가격지수 ARIMA(2,1,0) 모형을 추정하였다. 이 모형에서 추정 계수는 유의수준 5% 이내이고 F-통계량 또한 유의하게 나타났다. 잔차에 대한 Q 통계량 검증 결과, 전국 아파트 매매가격지수 ARIMA(1,1,0)에서 발생했던 유의수준을 넘어서는 자기상관 값이나 편자기상관 값이 나타나지 않았다. 최종적으로 모형의 모든 추정 계수에 대한 유의수준이 5% 이내이고 잔차의 자기상관이 나타나지 않은 전국 아파트 매매가격지수 ARIMA(2,1,0) 모형을 감성지수 예측 유용성 측정을 위한 최종 준거모형으로 선정하였다. 전국 아파트 매매가격지수 ARIMA(2,1,0) 모형의 adj. R^2는 0.455이고 AIC(Akaike Information Criterion) 값은 -0.9853이다. 분석 결과는 〈Table 8〉과 같다.

3) ARIMAX 분석 결과

신문 감성지수를 포함한 전국 아파트 매매가격지수 모형(NPSI ARIMAX)은 신문 감성지수의 선행 시차를 -1에서 -4까지 시차별로 전국 아파트 매매가격지수(NAPI) ARIMA 모형에 투입하여 Q 통계량의 p값이 0.05 이상으로 잔차에 자기상관이 없고, 신문 감성지수(NPSI)의 p값이 0.1 이하인 모형으로 선정한다.

신문 감성지수의 선행시차에 따른 NPSI ARIMAX 모형의 추정 결과는 〈Table 9〉와 같다. 앞서 제시한 모형 선정기준에 부합하고 모든 변수의 계수가 통계적 유의성이 있는 모형은 신문 감성지수의 선행시차 -2를 포함한 모형으로 나타났다. 또한 시차 -2를 적용한 NPSI ARIMAX 모형의 설명력(adj. R^2)이 0.480으로 가장 높다. 분석 결과, 신문 감성지수는 전국 아파트 매매가격지수의 변동을 설명 또는 예측하는 데 유용한 것으로 나타났다.

방송 감성지수를 포함한 전국 아파트 매매가격지수 모형(TVSI ARIMAX)도 NPSI ARIMAX와 동일한 선정기준을 적용하였다. 방송 감성지수의 선행시차에 따른 TVSI ARIMAX 모형의 추정 결과는 〈Table 10〉과 같다. TVSI ARIMAX 모형의 선정기준에 부합하고 모형 내 모든 변수가 통계적으로 유의미한 모형은 시차 -1을 적용한 시계열 모형이다. 방송 감성지수도 신문 감성지수와 유사하게 전국 아파트 매매가격지수의 변동을 설명 또는 예측하는 데 유용한 것으로 분석되었다.

Table 8. Result of ARIMA (1,1,0) and ARIMA (2,1,0)

Model	Var.	Coef.	Std. err.	t	F	adj. R^2	AIC
ARIMA (1,1,0)	C	0.145	0.064	2.257**	33.54***	0.42	-0.974
	AR (1)	0.669	0.076	8.755***			
	-	-	-	-			
ARIMA (2,1,0)	C	0.146	0.053	2.741***	23.87***	0.45	-0.985
	AR (1)	0.795	0.071	11.16***			
	AR (2)	-0.180	0.077	-2.402**			

Significance level: * < 0.10, ** < 0.05, *** < 0.01

Table 9. Result of NPSI ARIMAX

Ex. Var.	Var.	Coef.	Std. err.	t	F	adj. R^2
NPSI	C	0.115	0.053	2.153**	19.07	0.468
	NPSI (-1)	0.252	0.162	1.556		
	AR (1)	0.729	0.094	7.760***		
	AR (2)	-0.180	0.086	-2.092**		
	C	0.108	0.047	2.258**	19.70	0.480
	NPSI (-2)	0.314	0.188	1.670*		
	AR (1)	0.747	0.079	9.426***		
	AR (2)	-0.232	0.088	-2.628**		
	C	0.166	0.068	2.437**	18.00	0.459
	NPSI (-3)	-0.186	0.158	-1.175		
	AR (1)	0.849	0.068	12.35***		
	AR (2)	-0.181	0.077	-2.337**		
	C	0.108	0.054	1.997**	19.09	0.478
	NPSI (-4)	0.296	0.183	1.618		
	AR (1)	0.761	0.091	8.345***		
	AR (2)	-0.223	0.085	-2.612**		

Significance level: * < 0.10, ** < 0.05, *** < 0.01

Table 10. Result of TVSI ARIMAX

Ex. var.	Var.	Coef.	Std. err.	t	F	adj. R^2
TVSI	C	0.129	0.056	2.301**	18.86	0.465
	TVSI (-1)	0.143	0.085	1.671*		
	AR (1)	0.759	0.095	7.934***		
	AR (2)	-0.185	0.082	-2.261**		
	C	0.127	0.050	2.529**	18.88	0.469
	TVSI (-2)	0.156	0.107	0.146		
	AR (1)	0.769	0.076	10.108***		
	AR (2)	-0.207	0.092	-2.233**		
	C	0.158	0.063	2.477**	18.22	0.462
	TVSI (-3)	-0.130	0.094	-1.379		
	AR (1)	0.840	0.079	-2.280**		
	AR (2)	-0.181	0.079	-2.280**		
	C	0.126	0.054	2.313**	18.37	0.468
	TVSI (-4)	0.155	0.105	1.468		
	AR (1)	0.778	0.090	8.555***		
	AR (2)	-0.208	0.084	-2.461**		

Significance level: * < 0.10, ** < 0.05, *** < 0.01

NPSI ARIMAX 모형과 TVSI ARIMAX 모형의 예측 유용성은 추정된 모형의 예측 오차가 얼마나 향상되는지 비교함으로써 파악할 수 있다. 모형의 예측력을 분석하는 지표로 제곱근 오차인 RMSE와 절댓값 평균인 MAE를 활용하여 예측의 정확도를 분석하였다. RMSE와 MAE는 값이 작을수록 예측에 대한 정확도가 높은 것을 의미한다.

Table 11. Comparison of prediction errors

Model	Exogenous variable	Prediction error	
		RMSE	MAE
NAPI ARIMA (2,1,0)	-	0.1959	0.1400
NPSI ARIMAX (2,1,0)	-	0.1959	0.1400
	NPSI (-1)	0.1809	0.1287
	NPSI (-2)	0.1803	0.1313
	NPSI (-3)	0.2069	0.1484
	NPSI (-4)	0.1826	0.1300
TVSI ARIMAX (2,1,0)	-	0.1959	0.1400
	TVSI (-1)	0.1860	0.1324
	TVSI (-2)	0.1871	0.1354
	TVSI (-3)	0.2039	0.1469
	TVSI (-4)	0.1874	0.1342

Significance Level: * < 0.10, ** < 0.05, *** < 0.01

분석 결과는 〈Table 11〉과 같다. 신문과 방송 감성지수 모두 전국 아파트 매매가격지수 변화 예측의 정확도를 향상시키는 것으로 나타났다. 전국 아파트 매매가격지수(NAPI) ARIMA 모형의 RMSE와 MAE는 각각 0.1959와 0.1400이다. 신문 감성지수(NPSI) ARIMAX 모형의 RMSE, MAE는 각각 0.1803, 0.1313으로 예측 오차가 각각 7.90%, 6.21% 향상된 것으로 분석되었다. 한편, 방송 감성지수(TVSI) ARIMAX 모형의 RMSE, MAE는 각각 0.1860, 0.1324로 예측 오차가 각각 5.05%, 5.42% 향상된 것으로 나타났다.

매체별 예측 유용성을 비교하면, 방송 감성지수보다는 신문 감성지수의 예측 정확도가 더 높다. 그 이유는 매체별로 정보전달의 양과 제약이 다르고, 부동산 관련 기사에 대한 논조와 보도 특성이 다르기 때문인 것으로 판단된다. 신문은 방송보다 매체의 수가 많아 정보의 양이 많고, 정해진 시간에 화면을 통해 정보를 전달하는 방송뉴스보다 정보전달의 시간적 제약이 적다. 방송 뉴스도 부동산 대책과 정부 정책의 비중이 높아 양비론적인 신문보도와 대조적인 성향을 보인다(김수영·박승관 2017). 또한 이 연구에서 선정한 신문사는 시장 친화적이고 보수적인 데 비해 방송사는 중립적이고 사실보도에 중점을 두는 특성이 있다.

V. 결 론

이 연구는 비정형 빅데이터를 이용하여 부동산시장 참여자들의 심리지수를 측정하고, 이 지수가 부동산 가격의 설명 및 예측에 유용한지를 검증하는 데 초점을 둔다. 이를 위해 2012년 1월부

터 2018년 12월까지 수집한 신문기사와 방송뉴스의 부동산 관련 기사를 이용해 감성지수를 산출하였다. 이 감성지수를 포함한 전국 아파트 매매가격지수 모형이 유용한지, 그리고 어떤 매체의 감성지수가 더 유용한지 비교·분석하는 것이 연구의 목적이다.

주요 분석결과를 정리하면 다음과 같다. 우선, 전국 아파트 매매가격지수와 신문 및 방송 감성지수 간 교차상관분석을 통해 감성지수가 아파트 매매가격지수에 선행성을 나타내고, 시차 1에서 상관계수가 가장 크다. 주택가격 상승(하락) 초기에 이와 관련된 긍정(부정)적 뉴스를 보도하고, 뉴스기사의 증가가 시차를 두고 실제 주택가격에 양의 영향을 미치는 것으로 해석된다. 신문 및 방송 감성지수의 선행성은 감성지수가 전국 아파트 매매가격지수 예측 모형의 외생변수로 고려될 수 있음을 시사한다.

둘째, 전국 아파트 매매가격지수의 ARIMA 모형에 신문 및 방송 감성지수를 외생변수로 투입하여 ARIMAX 모형을 추정하였다. 전국 아파트 매매가격지수 모형을 ARIMA(2,1,0)로 결정하고 감성지수를 포함한 모형의 예측 유용성 분석을 위한 준거모형으로 설정하였다. 수정 모형은 기존 모형에 비해 자기상관이 없고 설명력도 개선되었다.

셋째, 전국 아파트 매매가격지수 ARIMA(2,1,0) 모형에 뉴스 또는 방송 감성지수를 각각 외생변수로 투입하여 ARIMAX 모형을 추정하였다. 신문 감성지수를 투입한 전국 아파트 매매가격지수 ARIMAX 모형은 선행시차 2의 신문 감성지수를 외생변수로 한 ARIMA(2,1,0) 모형으로 분석되었다. 방송 감성지수를 투입한 전국 아파트 매매가격지수 ARIMAX 모형은 선행시차 1의 방송 감성지수를 외생변수로 한 ARIMA(2,1,0) 모형으로 분석되었다. 추정된 모형에서 감성지수는 모두 통계적으로 아파트 매매가격지수 예측에 유의미하며, 설명력 또한 상대적으로 높은 것으로 분석되었다.

마지막으로, 전국 아파트 매매가격지수 ARIMA 모형과 신문 및 방송 감성지수가 외생변수로 포함된 개별 ARIMAX 모형의 예측오차를 분석하여 감성지수의 예측 유용성을 비교하였다. 분석 결과, 신문 감성지수(NPSI) ARIMAX 모형의 오차인 RMSE, MAE가 각각 7.90%, 6.21% 감소하여 예측오차가 개선되었다. 방송 감성지수(TVSI) ARIMAX 모형은 RMSE, MAE가 각각 5.05%, 5.42% 감소하여 예측오차가 향상된 것으로 나타났다. 신문과 방송 감성지수의 예측 유용성을 비교하면, 신문 감성지수의 예측력이 방송 감성지수보다 높다.

종합하면, 신문기사와 방송뉴스 텍스트 빅데이터에서 도출한 감성지수가 전국 아파트 매매가격지수의 예측 모형에 외생변수로 활용될 수 있고, 변수의 통계적 유의성, 설명력과 오차 측면에서 모형의 설명 및 예측에 유용하다. 또한 신문 감성지수가 방송 감성지수보다 예측 유용성이 더 높은 것으로 나타났다.

최근 국토연구원 등 우리나라에서도 부동산시장에 내재된 심리지수를 측정하여 모니터링 및 예측 자료로 활용하는 노력이 전개되고 있다. 그러나 직접 조사를 통해 취득한 자료를 이용한 부동산시장의 심리지수는 지수의 조사와 발표시기의 시차로 인한 즉시성 문제, 월 단위 이하 측정 기간에 대한 유연성 문제가 제기되고 있다. 이를 보완하기 위해 부동산의 다양한 이슈를 중심으로 비정형 텍스트 빅데이터를 수집하고 감성지수 등 심리지수를 산출하는 방법과 시스템을 마련하는 방안이 필요하다.

신문과 방송 감성지수의 예측 유용성을 비교한 결과, 일반적인 예상과는 달리 화면과 음성에 기반한 방송뉴스보다 텍스트와 그래픽에 기반한 신문 감성지수가 높은 성과를 보였다. 이는 언론매체의 정보전달 특성과 분석대상 언론매체의 속성에 기인하는 것으로 판단된다. 신문과 방송은 부동산과 관련된 정보전달의 양에 차이가 크다. 신문기사는 부동산시장에 대한 다양한 정보를 생산하고 방송뉴스에 비해 매체의 수와 보도량이 매우 많다. 또한 신문기사는 온라인으로 유통되어 거의 실시간으로 보도되고 업데이트된다. 그러나 방송뉴스는 여전히 정해진 시간에 보도되어 시간적 제약도 크다. 최근에는 인터넷과 SNS 등을 통해 시간적 제약은 감소하고 있으나, 아직까지는 시간적 제약의 차이는 유효한 것으로 보인다.

언론매체의 속성도 신문과 방송의 예측 유용성에 영향을 미칠 것으로 판단된다. 방송뉴스는 부동산시장이 호황이면 호황에 대한 뉴스를 위주로 보도하는 경향이 있다. 신문기사는 양비론적 보도 특성이 있어 부동산시장이 호황일 때에도 우려를 나타내는 경향이 있다. 또한 연구의 분석대상인 신문사는 시장 친화적이고 보수적인 데 비해 지상파 방송사는 사실보도에 중점을 두는 특성이 있다.

이 연구는 신문기사와 방송뉴스 빅데이터에서 감성지수를 산출하여 주택가격의 예측 유용성을 분석하고, 매체 간 예측력을 비교·분석한 점에서 학술적 의의가 있다. 그러나 유효부수를 기준으로 분석대상 신문사를 선정한 결과, 시장 친화적이고 보수적인 일간지와 경제지를 위주로 분석하였다. 진보적 성격의 일간지와 지상파 3사 이외에 케이블 등 방송뉴스를 고려하지 못한 점에 한계가 있다. 또한 분석 기간을 확장하였을 때에도 유사한 결과를 담보할 수 있는지 검증을 할 수 없다는 점도 연구의 한계이다. 이 연구의 한계는 후속 연구로 남긴다.

주1. 이 연구는 감성지수 개발 방법의 정립과 평가에 초점을 둔 것이 아니라 주택매매가격에 대한 감성지수의 예측 유용성 분석에 초점을 둔다. 기계학습 기법을 적용한 감성지수 개발 방법에 관한 자세한 내용은 별도의 연구논문으로 제시하므로 이 논문에서는 이를 간략히 다룬다.

주2. BERT(Bidirectional Encoder Representations Form Transformer)는 구글이 공개한 인공지능(AI) 언어모델로 일부 성능 평가에서 인간보다 더 높은 정확도를 보이며 현재 2020년까지 자연 언어처리(NLP) AI의 최첨단 딥러닝 모델로 평가받고 있다.

인용문헌
References

1. 경정익·이국철, 2016. "텍스트 마이닝에 의한 부동산 빅데이터 감성분석 모형 개발", 「주택연구」, 24(4): 115-136.
   Kyung, J.I. and Lee, K.C., 2016. "Development of Sentiment Analysis of Real Estate Big Data by Using Textmining", *Hosing Studies Review*, 24(4): 115-136.
2. 김대원·유정석, 2013. "주택가격에 대한 심리적 태도가 주택 매매거래량에 미치는 영향 분석", 「주택연구」, 21(2): 73-92.
   Kim, D.W. and Yu, J.S., 2013. "An Analysis on How Psychological Attitudes on the House Price Affect the Trading Volume", *Housing Studies Review*, 21(2): 73-92.
3. 김대원·유정석, 2016. "트위터 정보와 아파트 매매 및 전세 가격 간 동적 관계 분석", 「도시행정학보」, 29(1): 1-33.
   Kim, D.W. and Yu, J.S., 2016. "The Dynamic Relationship between Twitter Information and Apartment Sale and Jeonse Prices", *Journal of the Korean Urban Management Association*, 29(1): 1-33.
4. 김동영·박제원·최재현, 2014. "SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구", 「한국IT서비스학회지」, 13(3): 221-233.
   Kim, D.Y., Park, J.W., and Choi, J.H., 2014. "A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles", *Journal of Information Technology Service*, 13(3): 221-233.
5. 김리영·안지아, 2013. "소비자의 주택가치 전망이 가격 및 거래에 미치는 영향", 「국토계획」, 48(3): 403-417.
   Kim, L.Y. and An, J.A., 2013. "The Effect of the Consumers' Housing Value Expectation on the Prices and Trade Volume in the Seoul Metropolitan Region", *Journal of Korea Planning Association*, 48(3): 403-417.
6. 김수영·박승관, 2017. "KBS의 공보 방송 모형적 성격에 관한 연구: 부동산 뉴스 생산 과정을 중심으로", 「한국언론정보학보」, 81: 225-271.
   Kim, S.Y. and Park, S.K., 2017. "Public Broadcasting or Publicity Broadcasting?: An Analysis of KBS News Coverage of the Korean Housing Market", *Korean Journal of Communication & Information*, 81: 225-271.
7. 김유신·김남규·정승렬, 2012. "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형", 「지능정보연구」, 18(2): 143-156.
   Kim, Y.S., Kim, N.G., and Jeong, S.R., 2012. "Stock-Index Invest Model Using News Big Data Opinion Mining", *Journal of Intelligence and Information Systems*, 8(2): 143-156.
8. 김진유, 2006. "신문기사가 부동산가격변동에 미치는 영향 –'투기'가 포함된 신문기사와 주택가격간의 그랜져인과관계분석을 중심으로–", 「주택연구」, 14(2): 39-63.
   Kim, J.Y., 2006. "Influence of Newspaper Article on Real Estate Market", *Housing Studies Review*, 14(2): 39-63.
9. 노민지·유선종, 2016. "소비자 심리와 아파트 실거래가격 간 관계분석: 인터넷 검색량 및 국토연구원 주택매매시장 소비심리지수를 중심으로", 「국토연구」, 89: 3-13.
   Noh, M.J. and Yoo, S.J., 2016. "A Relationship between Sales Prices of APT and Comsumer Sentiment", *The Korea Spatial Planning Review*, 89: 3-13.
10. 박재수, 2020. "주택시장 예측을 위한 부동산 감성지수 개발 연구: 뉴스와 방송 빅데이터에 대한 AI 기술 적용", 강원대학교 대학원 박사학위논문.
    Park, J., 2020. "A Study of the Development of Real Estate Sentiment Index for Housing Market Prediction: AI Technologies Applied to the News and TV Broadcast Big Data", Ph.D. Dissertation, Kangwon National University.
11. 박재수·이재수, 2019. "아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석–온라인 뉴스 기사의 비정형 빅데이터를 활용한 감성분석 기법의 적용–", 「국토계획」, 54(1): 131-147.
    Park, J. and Lee, J.S., 2019, "An Investigation into the Causal Relationship and the Cross Correlation between Apartment House Sales Prices and Real Estate Online News–An Approach to the Sentiment Analysis Using Unstructured Big Data of Online News Articles–", *Journal of Korea Planning Association*, 54(1): 131-147.
12. 박종영·서충원, 2015. "TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석", 「부동산학보」, 63: 46-58.
    Park, J.Y. and Suh, C.W., 2015. "Analysis of Changes in the Housing Market Using TF-IDF Weight Model", *Korea Real Estate Academy Review*, 63: 46-58.
13. 송경재·양희민, 2005. "시계열 분석에 의한 국제유가 예측; Nymex-WTI 선물가격을 중심으로", 「통계연구」, 10(1): 62-81.
    Song, K.J. and Yang, H.M., 2005. "A Study on the Nymex WTI Prices Forecasting Using Time Series Analysis", *Journal of Korean Official Statistics*, 10(1): 62-81.
14. 송민채·신경식, 2017. "뉴스기사를 이용한 소비자의 경기심리지수", 「지능정보연구」, 23: 1-27.
    Song, M.C. and Shin, K.S., 2017. "Construction of Consumer Confidence Index Based on Sentiment Analysis Using News Articles", *Journal of Intelligence and Information Systems*, 23: 1-27.
15. 송치영, 2002. "뉴스가 금융시장에 미치는 영향에 관한 연구", 「국제경제연구」, 8(3): 1-34.
    Song, C.Y., 2002. "News and Financial Prices", *International Economic Journal*, 8(3): 1-34.
16. 안희중·전승표·최종범, 2010. "남북관계 관련 뉴스가 주식시장에 미치는 영향", 「한국경제의 분석」, 16(2): 119-231.
    Ahn, H.J., Jeon, S.P., and Chay, J.B., 2010. "The Effects of the News Related to the North-South Korean Relationship on the Korean Stock Markets", *Journal of Korean Economic Analysis (JKEA)*, 16(2): 119-231.
17. 옥기율·김지수, 2012. "소비자 심리지수가 KOSPI 수익률에 미치는 비대칭적 영향에 대한 연구", 「금융공학연구」, 11(1): 17-37.
    Ohk, K.Y. and Kim, J.S., 2012. "An Empirical Study between Consumer Sentiment Index and KOSPI's Return: Negativity Effect", *The Korean Journal of Financial Engineering*, 11(1): 17-37.
18. 우윤석·이은정, 2011. "언론보도와 시계열 주택가격 간의 관계에 관한 연구", 「주택연구」, 19(4): 111-134.
    Woo, Y.S. and Lee, E.J., 2011. "An Analysis on the Relationship between Media Coverage and Time-series Housing Prices", *Housing Studies Review*, 19(4): 111-134.

19. 유한수·정재호, 2015. "주택시장에서의 매매가격지수와 소비심리지수의 관계", 「부동산연구」, 25(4): 49-61.
Yoo, H.S. and Chung, J.H., 2015. "The Lead-Lag Relationship between Housing Purchase Price Index and Consumer Sentiment Index", *Korea Real Estate Review*, 25(4): 49-61.

20. 이종민, 2018. "비정형 빅데이터의 전세가격 예측 유용성 연구", 강원대학교 대학원 박사학위논문.
Lee, J.M., 2018. "An Empirical Study on the Availability of Unstructured Big Data in Jeonse Price Prediction", Ph.D. Dissertation, Kangwon National University.

21. 이종민·이종아·정준호, 2017. "뉴스 빅데이터를 이용한 전세가격 예측-토픽모형 분석을 중심으로-", 「부동산학보」, 69: 43-57.
Lee, J.M., Lee, J.A., and Jeong, J.H., 2017. "The Jeonse Price Forecasting Used by News Big Data-Focusing on Topic Modeling Analysis-", *Korea Real Estate Academy Review*, 69: 43-57.

22. 조태진, 2014. "심리지수가 주택시장에 미치는 영향에 관한 연구", 「주택연구」, 22(3): 25-48.
Cho, T.J., 2014. "A Study on the Effect of the Sentiment Index to the Housing Market", *Housing Studies Review*, 22(3): 25-48.

23. 진창하·Gallimore, P., 2012. "신문기사 내용과 주택가격: 인식, 사유, 그리고 투자심리", 「부동산학연구」, 18(2): 125-142.
Jin, C.H. and Gallimore, P., 2012. "Newspaper Content and Home Prices: Perception, Reasoning and Affect", *Journal of the Korea Real Estate Analysts Association*, 18(2): 125-142.

24. 최영걸·이창무·최막중, 2004. "서울시 주택시장에서 작동하는 가격기대심리에 관한 실증연구-적응적 기대와 합리적 기대를 중심으로", 「국토계획」, 39(2): 131-141.
Choi, Y.G., Lee, C.M., and Choi, M.J., 2004. "Relationship between the Present Price and Expectations on Future Capital Gains in the Housing Market: Adaptive Expectation and Rational Expectation Hypotheses", *Journal of Korea Planning Association*, 39(2): 131-141.

25. 최희갑·임병준, 2009. "주택가격 전망이 주택가격 및 경기에 미치는 영향", 「국토연구」, 63: 141-158.
Choi, H.G. and Rhim, B.J., 2009. "Role of the Housing Price Forecast in Housing Price and Business Cycle", *The Korea Spatial Planning Review*, 63: 141-158.

26. 하성규, 2006. 「주택정책론」, 서울: 박영사.
Ha, S.K., 2006. *Housing Policy Theory*, Seoul: Parkyoungsa.

27. Baker, H.K. and Nofsinger, J.R., 2010. *Behavior Finance: Investors, Corporations, and Markets*, Hoboken, New Jersey: John Wiley & Sons.

28. Gayer, T. and Viscusi, W.K., 2002. "Housing Price Responses to Newspaper Publicity of Hazardous Waste Sites", *Resource and Energy Economics*, 24(1-2): 33-51.

29. Li, J., Xu, Z., Yu, L., and Tang, L., 2016. "Forecasting Oil Price Trends with Sentiment of Online News Articles", *Procedia Computer Science*, 91: 1081-1087.

30. Sun, D., Zhang, C., Xu, W., Zou, M., Zhou, J., and Du, Y., 2014. "Does Web News Media Have Opinions? Evidence from Real Estate Market Prediction", Paper presented at the 18th Pacific Asia Conference on Information Systems, 2014.

Date Received 2021-02-15
Date Reviewed 2021-03-22
Date Accepted 2021-03-22
Date Revised 2021-05-20
Final Received 2021-05-20