뉴스 빅데이터를 활용한 부동산시장의 인식에 대한 연구*
– 정책 기조 변화를 중심으로

Comparative A study on the perception of the real estate market using news big data
– Focusing on the change of policy

정 홍 ** 남 진 ***
Jung, Hong Nam, Jin

Abstract

Real estate is a complex interdisciplinary study, and it is very difficult to analyze and interpret the relationship between macroscopic and microscopic phenomena in society because they are interdependently connected. With the recent rapid development of big data technology and tools, text mining studies, which collect and analyze texts, which are unstructured data, to derive meaningful information are being used in real estate policy-related research. The real estate market reflects not only macroeconomic conditions such as income and interest rates, but also the movements of real estate market participants. When evaluating the effectiveness of the media, it is necessary to analyze the media as well.

Therefore, this study analyzes the difference in trends in the real estate market by dividing it into before and after Corona how news reports express social issues and market conditions on the housing market using text-mining, one of the unstructured data analysis methods. to check it. It is considered that this study can be used as basic data for research on the perception of real estate and housing markets.

색인어 : 텍스트마이닝, 부동산시장, 뉴스
Keyword : Textmining, Real-estate market, News

* 이 연구는 국토교통부의 「스마트시티 혁신인재육성사업('19-'23)」으로 지원되었습니다.
** 서울시립대 스마트시티학과 석사과정(주 저자 : eqnd777@naver.com)
*** 서울시립대 도시공학과 • 스마트시티학과 교수(공동저자 : jnam@uos.ac.kr)

Ⅰ. 서론

1. 연구의 배경 및 목적

급속한 기술 발전과 함께 디지털 혁명이 발생하고, 매일 생산되는 방대한 양의 지식과 정보에서 인사이트와 가치를 추출하는 과정이 중요해짐에 따라 빅데이터의 중요성이 대두되고 있다. 그리고 모든 것이 디지털 기술과 상호 유기적으로 되는 지능형 사회로 정의 내려지는 4차 산업혁명 사회에서 빅데이터의 가치는 계속 재고되고 있다.

많은 국가와 국내외 기업들이 빅데이터의 기술개발과 활용을 선점하기 위해 연구개발을 진행하고 있다. 빅데이터는 데이터 기반의 과학적·합리적 예측과 의사결정 지원을 통해 혁신의 기반으로 활용할 수 있는 기술로, 많은 국가와 국내외 기업이 기술개발과 활용을 선점하기 위해 연구개발을 진행하고 있다. 빅데이터는 상호의존성, 불확실성, 복잡성을 높이는 사회 발전을 예측하고 문제 해결을 위한 가치와 지식을 제공하기 위해 실시간으로 사용할 수 없는 데이터를 수집, 저장 및 분석해야 한다.

빅데이터는 데이터를 기반으로 논리적인 미래 예측 등을 통한 발전의 원동력으로 활용될 수 있는 혁신기술로 이미 많은 기업과 국가에서 혁신적인 기술을 선점하기 위해 연구 개발을 진행하고 있다. 빅데이터는 이전까지는 사용하기 어려웠던 다양한 비정형데이터 등을 기술의 발전으로 인해 즉각적으로 수집 및 분석이 가능해져 복잡하게 상호작용하고 있는 사회구조를 해석하고 더 나아가 예측하여 사회문제를 해결하고 있다.

데이터마이닝(data mining), 텍스트마이닝(text mining), 시스템다이나믹스(system dynamics) 등 빅데이터의 분석기법은 다양하게 포함되어있다. 이러한 분석기법 중에서도 텍스트마이닝은 대량의 텍스트 데이터에서 분석을 통해 가치와 인사이트를 추출하는 것으로. 관련 키워드에 대한 트렌드와 문제를 인식하는데 활용된다. 2017년 미국 대선 당시, 데이터 벤처기업 제닉AI는 대선을 열흘 앞둔 시점부터 트럼프가 당선될 것이라고 정확하게 예측하여 선거 직후 큰 화제가 되었다. 이들은 검색 엔진과 소셜미디어 등에서 2000만 건이 넘는 대량의 데이터를 수집하여 빅데이터 분석기법을 활용해 정확한 예측해 세계를 놀라게 한 사례이다.

부동산이라는 것은 학제 간 혼재된 복합적인 학문으로, 사회의 거시적·미시적 현상들이 상호의존적으로 연결되어 있어 서로의 관계를 분석하고 해석하는 것은 매우 어려운 일이다. 기존의 부동산 관련 선행연구는 구득할 수 있는 정형데이터를 가지고 부동산과 관련한 현상들을 해석해내고자 하였다. 최근 빅데이터의 기술 발전으로 인해 비정형데이터를 다룰 수 있게 되자 이 데이터를 사용하여 부동산 현상들을 해석

하는 연구가 많이 진행되고 있다. 신문과 뉴스기사와 같은 언론 비정형데이터를 대용량으로 수집 및 활용할 수 있는 기술과 도구들이 등장하여 뉴스 빅데이터가 내포하고 있는 가치와 정보를 해석하고자 주목하기 시작한 것이다(박대민, 2016).

이처럼 신문기사와 같은 비정형 데이터인 텍스트를 수집 및 분석하여 유의미한 정보를 도출하고자 하는 기법으로 텍스트 마이닝은 여러 분야에서 적용되고 있으며 부동산정책 관련 연구들 또한 이러한 기조에 맞춘 연구들이 진행되고 있다. 부동산시장은 소득, 이자율 등과 같은 거시경제의 상태뿐만 아니라 부동산시장 참여자들의 움직임도 반영하고 있으며, 부동산 수요자가 시장의 정보들을 크게 접할 수 있는 대표적 매체 중 하나가 뉴스 등 언론보도라고 할 수 있기 때문에 정책의 효과성을 평가할 때 언론에 대한 분석도 함께 이루어져야 할 필요성이 있다(박재수, 이재수, 2019).

따라서 이 연구의 목적은 비정형 데이터 분석방법 중 하나인 텍스트마이닝(text-mining)을 활용하여 뉴스보도가 주택시장에 대한 사회적 이슈와 시장 상황에 대해 어떻게 표현하고 있는지 코로나 전과 후로 나누어 부동산시장의 동향 차이를 분석하고자 확인하는 데 있다. 이러한 연구는 부동산 및 주택시장의 인식에 관한 연구의 기초자료로 활용할 수 있을 것으로 고려된다.

2. 연구의 범위 및 내용

이 연구는 뉴스기사를 텍스트마이닝을 활용하여 부동산시장의 인식이 정책 기조의 변화에 따라 어떻게 변하는지 분석하고자 한다. 연구의 범위와 구성은 다음과 같다. 시간적 범위를 문재인 대통령이 취임된 해인 2017년부터 2021년까지로 설정하였다. 수집자료는 네이버 포털에 올라온 뉴스기사를 각 1년씩 총 5개년의 뉴스기사를 수집하여 분석한다. 2장은 이론적 고찰 및 선행연구 검토로 빅데이터와 텍스트마이닝에 대한 고찰과 이를 활용한 선행연구를 살펴보고. 3장은 연구방법으로 분석자료 수집, 빈도분석, 전처리, N-Gram순열, TF-IDF(Term Frequency - Inverse Document Frequency) 분석과정에 대해 간략히 정리한다.

4장은 분석 결과로 가공한 데이터를 빈도 및 TF-IDF 분석과 N-gram 순열결과를 연도별로 비교하고 분석한다. 이후 5장은 결론으로 총 5개년의 연구결과를 요약하여 정리하고, 연구에 따른 정책적 시사점을 제시하고자 한다.

Ⅱ. 이론적 고찰 및 선행연구 검토

1. 텍스트 마이닝의 정의 및 개념

빅데이터는 일반적으로 Volume, Velocit, Variaty라고 불리는 3V의 기본 특성을 가진다. 즉, 막대한 데이터를 빠르게 수집하여 저장한 후 분석을 통해 새로운 가치를 창출할 수 있는 프로세스와 기술을 의미한다(한국정보화진흥원, 2013). 또한 LG경제연구원(2012)은 비정형 데이터에 집중하여 이전까지는 활용할 수 없었던 사람들의 일상생활이나 그 안에 포함된 감정이나 상황, 상태 등 비정형 형태의 데이터까지 분석함으로써 컴퓨터의 알고리즘에 의한 반사작용이 아닌 '가치'와 '생각'에 의한 반응을 추출해내는 기술이라 한다.

최근 빅데이터가 가지는 가치에 대해 기대가 상승한 것은 데이터의 특징인 대용량이라는 것보다는 이러한 대용량의 비정형데이터를 실시간으로 반영하여 분석하고 분석에 따라 예측도 가능하다는 점 때문이다.

빅데이터 기술의 특징을 살펴보면 첫째, 기존의 데이터 처리방식에 비교해보면 데이터 구축방식이 편리하여 비용을 절감할 수 있으며 효율성을 제공한다. 그리고 정형 및 비정형 데이터의 수집, 저장, 분류, 분석, 표현 등 빅데이터 처리의 모든 과정에서 다양한 기술 발전이 요구된다. 셋째, 빅데이터 기술은 데이터의 수집 양과 기술보다 그 데이터를 분석해서 추출할 수 있는 인사이트(Insight)를 찾아내는 것이 더욱 중요한 기술이다. 따라서 이에 해당하는 다양한 기술을 필요영역에 적정하게 적용하여 의미를 찾아내는 일련의 프로세스라는 것이다(김정선 외, 2014).

빅데이터의 대표적인 분석기법은 아래 <표 1>에서 보는 바와 같이 데이터마이닝, 텍스트마이닝, 오피니언 마이닝, 소셜분석, 클러스터 분석, 현실마이닝 등이 있다(한국정보화 진흥원, 2012; 이정미, 2013; 박상훈, 2018).

<표 1> 빅데이터 기술

구분	내용
데이터 마이닝 (Data Mining)	• 대용량의 데이터, 데이터 베이스 등에서 감춰진 지식, 기대하지 못했던 경향, 새로운 규칙 등의 유용한 정보를 발견하는 과정 • 데이터 마이닝을 통해 정보의 연관성(순차 패턴, 유사성 등)을 파악함으로써 가치 있는 정보를 만들어 의사결정에 적용
텍스트 마이닝 (Text Mining)	•자연어로 구성된 비정형 또는 반정형 텍스트 데이터에서 자연어 처리 기술에 기반해 관계 또는 패턴을 추출하여 가치와 의미있는 정보를 찾아내는 마이닝 기법
오피니언 마이닝 (Opinion Mining)	• 웹상의 데이터베이스에 저장되어 있는 어떤 주제 혹은 특정 대상자의 의견을 포함하고 있는 텍스트 속에서의 의미를 추출하여 감성(긍정, 부정 등)을 분석하는 기법
웹 마이닝 (Web Mining)	•인터넷 수집 정보를 데이터 마이닝 방법으로 분석하여 통합하는 기법 • 콘텐츠 마이닝(웹 검색, 수집 데이터), 구조 마이닝(웹 사이트 구조), 활용 마이닝(사용자 이용형태) 등으로 세분화 됨
클러스터 분석 (Cluster analysis)	• 통계기법에 의해 비슷한 특성이 있는 개체를 클러스터로 나누는 방법을 통해 유사성을 판단하는 기술로서 일명 군집분석이라 함
소셜 분석 (Social Mining)	• 소셜 미디어 글과 사용자를 분석해 소비자의 패턴 등을 분석하는 기법 • 마케팅 분야뿐만 아니라 사회의 흐름과 트렌드, 여론 변화 추이를 읽어내는 소셜 미디어 시대의 마이닝 기법
현실 마이닝 (Reality Mining)	• 사람의 행동 패턴을 예측하기 위해 사회적 행동과 관련된 정보를 기기(휴대폰, GPS 등)를 통해 얻고 분석하는 기법 • 휴대폰 등 모바일 기기들을 통해 현실에서 발생하는 정보를 기반으로 인간 관계와 행동 양태 등을 추론

자료: 한국정보화진흥원(2012), 이정미(2013), 박상훈(2018) 재구성

2. 부동산시장 뉴스기사와 심리에 관한 연구

다른 시장과 다르게 주택시장은 비전문적인 시장 참여자들의 영향을 더욱 많이 받게 된다. 다시 말해 주택시장은 객관적인 사실만이 아니라 시장 참여하는 모든 사람의 투자기대심리와 생각 등 심리라는 주관적인 부분도 어느 정도 받게 된다는 것이다(진창하, 2012). 이와 같이 많은 선행연구를 살펴보면 시장 참여자는 직접관찰이나 직접경험 또는 사회의 비공식 커뮤니케이션을 통해 경제 상황에 대해 대략의 판단을 하게 된다(Weatherford, 1983). 특히 매스미디어인 경제뉴스는 잠재적으로 소비자 심리의 장기적인 경제상황에 널리 영향을 미친다는 사실이 밝혀지고 있다.

김대원·유정석(2013)은 주택매매량에 주택가격에 대한 심리가 미치는 영향을 분석하였고, 주택시장을 예측하기 위한 변수로 소비 심리 지수가 활용될 수 있고 시장 참여자의 주택가격에 대한 심리는 주택가격의 기대치가 매매량 결정에 영향을 미친다고 설명하였다.

박재수·이재수(2017)는 신문기사를 활용하여 아파트 가격의 동향을 텍스트마이닝 및 감성분석으로 분석했다. 분석 결과를 보면 강남의 아파트 매매가격이 변화한 뒤 이러한 내용이 기사에 실리면 서남권의 아파트 매매가격이 후행한다고 했다.

전해정(2019)은 부동산시장의 동향을 텍스트마이닝을 활용하여 확인하고자 했다. 시간적 범위는 2016년 8월부터 2017년 8월까지이며 이 기간의 뉴스를 분석한 결과를 보면 재건축, 강남 등과 관련된 키워드가 다수 출현하여, 강남의 재건축에 대한 시장의 기대가 많은 것으로 보았다.

장몽형·김한수(2021)는 텍스트마이닝을 활용하여 주택가격변동을 분석했다. 분석 결과를 보면 부동산 뉴스기사에 다수 출현된 키워드가 지역별 아파트 매매가격변동에 영향을 준다고 하였다. 그중에서도 재건축 키워드가 주택매매가격과 상당히 관련성이 있다고 하였다.

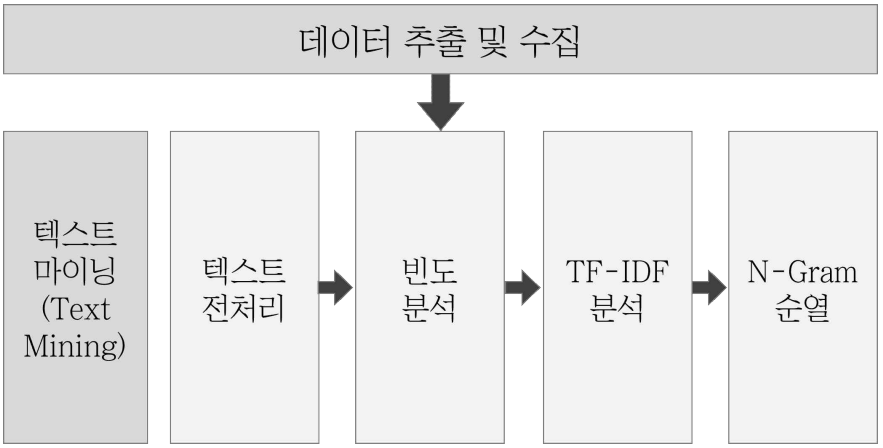
선행연구를 정리 및 비교하여 이 연구가 가지는 차별성은 다음과 같다. 첫째, 비정형데이터인 뉴스기사를 텍스트마이닝으로 수집하고 계량화 분석을 하여 보다 객관성을 가지며, 이는 부동산시장에 대한 심리적 태도 변화를 파악하기에 용이하다. 둘째, 부동산시장의 인식 및 동향을 즉각적으로 확인하다는 점이다. 다수의 선행연구는 주택가격과의 관계를 확인하기 위해 뉴스 기사를 가공하여 지수화하였다. 가공된 뉴스기사 지수는 부동산시장에 대한 인식을 즉각적으로 확인하기에는 다소 어려움이 있다.

Ⅲ. 연구방법

1. 분석 모형

이 연구는 부동산시장 동향에 대한 인터넷 뉴스기사를 대상으로 텍스트마이닝을 활용하여 부동산시장의 인식에 대해 살펴보고자 한다. 연구의 순서는 <그림 1>과 같다. 인터넷 뉴스기사에서 수집된 비정형데이터에서 의미를 가지지 않는 숫자, 특수문자, 접속사 등은 제거하여 텍스트 전처리(preprocessing process)를 진행했다. 이후 출현빈도가 높은 키워드를 순서대로 추출하여 빈도분석과 TF-IDF분석을 진행하였으며, N-Gram순열 분석을 진행하여 어떤 단어와 단어가 묶여서 출현하는지를 확인하였다.

<그림 1> 분석 모형



1) 분석자료 수집

부동산시장의 인식을 살펴보기 위해 분석 포털 사이트를 국내 인터넷 포털 점유율 1위인 네이버 뉴스에서 텍스트 크롤링을 진행하였다. 연구를 위한 키워드로는 “부동산”, “시장”, “전망” 총 3가지 키워드로 뉴스기사 데이터를 수집하였다. 데이터 수집 기간은 문재인 대통령의 취임 해인 2017년 1월 1일부터 2021년 12월 31일까지로 설정하여 텍스트 데이터를 수집하였다. 수집된 데이터 건수는 2017년 총 11,204건, 2018년 총 11,058건, 2019년 총 9,947건, 2020년 총 11,078건, 2021년이 총 11,233건으로 5년간 약 55,000건에 달하는 기사를 추출했다.

수집된 원문을 보면서 광고기사 등 관련이 없거나 결과에 왜곡되게 할 수 있는 기사들은 수작업으로 삭제하였으며, 이후 형태소 분석을 실시하여 정제 작업을 하였다. 최종 정제된 텍스트 데이터로 분석을 진행하였다. 다음 <표 2>는 각 수집된 데이터의 개요이다.

<표 2> 네이버 뉴스 데이터 수집 개요

구분	내 용
수집 단어	“부동산”, “시장”, “전망”
수집 기간	2017년 / 2018년 / 2019년 / 2020년 / 2021년
수집 용량	11,204건 / 11,058건 / 9,947건 / 11,078건 / 11,233건
수집 매체	네이버 뉴스
수집 도구	텍스톰(Textom)

2) 전처리

전처리(Preprocessing)는 텍스트 데이터를 적절히 분석하고자 진행하는 정제 과정이다. 데이터를 수집하는 과정에서 광고와 같이 필요하지 않은 기사들을 제거하지만 그럼에도 불필요한 텍스트와 기사가 포함되는 경우가 많다. 그래서 의미를 가지지 않는 어미, 조사, 접속사 등의 불용어를 제거하여 단어를 정제해야 한다. 또한 문서에서 문맥상 유사한 의미를 가지는 단어는 하나의 형태로 묶어 대표성을 가지게 해야 한다. 유사한 의미를 내포하는 단어가 따로 묶여 있을 경우, 문서의 분석 결과값이 과소 또는 과대 평가될 수 있기 때문이다.

3) 빈도 분석

비정형데이터인 뉴스기사 등을 분석하는 방법으로 많이 사용되는 코퍼스(corpus) 언어분석은 언어학의 한 분야로 언어를 통계학적으로 분석하여 매우 중요한 분석이다. 대표적인 분석 방법이 빈도 분석인데 동일한 의미를 표현하더라도 빈도가 높은 단어는 빈도가 낮은 단어에 비해서 기억을 더 잘하는 심리적 연관성과 관련이 있으며 이를 빈도효과(frequency effect)라고 한다. 예를 들어, ‘구매’ 라는 단어보다는 ‘매수’ 라는 단어를 더 잘 기억한다. ‘방안’ 보다는 ‘대책’, ‘양도소득세’ 보다는 ‘양도세’, ‘버블’ 보다는 ‘거품’이라는 표현에 더 익숙하다. 따라서, 빈도수는 해당 문서에서 표현하고자 하는 메시지를 함축적으로 전달하므로 연구자가 적절한 단어를 선택하여야 한다.

언어자료는 빈도가 높은 단어와 낮은 단어 간의 차이가 커서 대표적인 멱함수 분포를 보이며 표본에 대한 평균과 표준편차는 적용될 수 없다. 결국, 언어자료는 수집한 데이터의 전부를 전수 조사하는 것을 원칙으로 한다.

4) TF-IDF 분석

TF-IDF는 용어빈도-역문서빈도(Term-Frequency-Inverse Document Frequency)의 준말로서 주어진 문서에서 용어가 갖고 있는 정보의 양을 정량화 시키는 숫자를 사용해 텍스트 형식을 행렬 형식(행-열/테이블 형식)을 표현하는 분석 모형이다. TF-IDF모형은 토큰 발생빈도가 문서에 대한 정보의 양을 완전히 나타내지 못한다는 단점을 보완하기 위해 쓰이며 자주 출현하지 않는 용어가 문서에 담긴 훨씬 많은 정보를 전달할 수 있다는 점을 가정한다.

문서 i에서 주어진 용어 j에 대한 용어 빈도(TF)는 용어 j가 문서 i에서 나타나는 횟수와 같다. 드물게 발생하는 용어는 자주 발생하는 일반 용어보다 더 많은 정보를 담고 있다. 이를 설명하기 위해 또 다른 값 T를 곱해야 한다. 이 요소 T는 해당 용어가 주어진 문서에 얼마나 특별한 용어인지를 나타낸다. 이것을 IDF(Inverse Document Frequency)라고 한다. 주어진 용어의 IDF는 다음과 같이 나타낼 수 있다.

$$term\,j(IDF_j) = \log_{10}(N/DF_j) \qquad (1)$$

여기서 DF_j는 용어 j를 갖는 문서 수를 나타낸다. N은 총 문서 개수다. 따라서 문서 i에서 용어 j의 TF-IDF는 다음과 같이 나타낼 수 있다.

$$a_{ij} = TF-IDF_j = TF_{ij} \times IDF_j = TF_{ij} \times \log_{10}(N/DF_j) \qquad (2)$$

TF-IDF 가중치는 하나의 문서에서 TF가 크고 전체 문서에서 DF가 작을수록 커지며 값이 클수록 상대적으로 문서 내에서 핵심적인 단어임을 나타낸다.

5) N-Gram 순열

N-gram은 문자열에서 N개의 연속된 요소를 추출하는 방법이며 확률학, 통신이론, 컴퓨터언어학 및 데이터압축 영역에서 많이 사용된다. 빈도분석을 통해 나온 단

어들은 원래의 자기 자리에 있던 순서를 완전히 무시하고 처리되므로 단어 자체가 가지는 의미를 제대로 표현하지 못할 수 있다. 이를 보완하기 위해 문서에서 인접한 단어들을 쌍으로 묶어 표현하는 N-gram 순열 기법을 적용한다.

대표적으로 '상승'과 '증가'라는 단어는 '집값-상승' 또는 '금리-상승' 그리고 '거래량-증가' 또는 '세부담-증가'와 같이 긍정과 부정의 의미를 모두 포함하고 있기 때문에 N-gram 순열은 문서의 의미를 정확히 전달하는 중요한 기법이라고 할 수 있다. N-gram 순열에서 요약된 단어들이 해당 문서들의 내용을 함축적으로 표현해 주기 때문에 이를 기반으로 TF-IDF의 빈도 단어 선정에 상당 부분 도움이 된다.

Ⅳ. 분석 결과

1. 2019년 이전 분석결과

1) 빈도 분석

<표 3> 2017년 추출 키워드 빈도분석표

	단어	빈도수	백분율(%)		단어	빈도수	백분율(%)
1	부동산	3243	4.16%	26	2018	175	0.22%
2	시장	1289	1.65%	27	오피스텔	165	0.21%
3	집값	697	0.89%	28	상승	165	0.21%
4	서울	590	0.76%	29	전문가	160	0.21%
5	전망	544	0.70%	30	종부세	159	0.20%
6	분양	496	0.64%	31	금리	154	0.20%
7	아파트	465	0.60%	32	주택	149	0.19%
8	내년	334	0.43%	33	상가	149	0.19%
9	보유세	315	0.40%	34	기준금리	149	0.19%
10	올해	284	0.36%	35	한은	145	0.19%
11	하반기	283	0.36%	36	거래	145	0.19%
12	재건축	278	0.36%	37	인기	139	0.18%
13	규제	271	0.35%	38	1년	138	0.18%
14	정부	252	0.32%	39	우려	134	0.17%
15	강남	249	0.32%	40	강화	128	0.16%
16	인상	232	0.30%	41	개발	127	0.16%
17	하락	230	0.30%	42	금리인상	125	0.16%
18	주택시장	224	0.29%	43	분양시장	122	0.16%
19	주목	213	0.27%	44	지방	122	0.16%
20	대책	212	0.27%	45	최대	122	0.16%
21	투자	210	0.27%	46	경제	117	0.15%

22	개최	201	0.26%	47	집	115	0.15%
23	공급	194	0.25%	48	경기	114	0.15%
24	수도권	190	0.24%	49	전국	114	0.15%
25	2018	175	0.22%	50	정책	113	0.14%

<표 4> 2018년 추출 키워드 빈도분석표

	단어	빈도수	백분율(%)		단어	빈도수	백분율(%)
1	부동산	2643	3.59%	26	상승	161	0.22%
2	시장	803	1.09%	27	종부세	158	0.21%
3	집값	680	0.92%	28	금리	154	0.21%
4	서울	583	0.79%	29	상가	150	0.20%
5	분양	479	0.65%	30	전문가	149	0.20%
6	전망	467	0.63%	31	기준금리	146	0.20%
7	아파트	451	0.61%	32	주택	145	0.20%
8	부동산시장	400	0.54%	33	2018	145	0.20%
9	내년	316	0.43%	34	눈길	144	0.20%
10	보유세	312	0.42%	35	거래	144	0.20%
11	재건축	275	0.37%	36	한은	140	0.19%
12	올해	271	0.37%	37	1년	134	0.18%
13	규제	270	0.37%	38	우려	133	0.18%
14	하반기	256	0.35%	39	강화	124	0.17%
15	정부	251	0.34%	40	지방	122	0.17%
16	강남	247	0.34%	41	금리인상	122	0.17%
17	인상	230	0.31%	42	분양시장	120	0.16%
18	하락	226	0.31%	43	최대	120	0.16%
19	주택시장	212	0.29%	44	개발	114	0.15%
20	대책	212	0.29%	45	경제	114	0.15%
21	주목	210	0.28%	46	발표	112	0.15%
22	공급	190	0.26%	47	경기	111	0.15%
23	투자	190	0.26%	48	전국	111	0.15%
24	수도권	188	0.26%	49	정책	111	0.15%
25	오피스텔	167	0.23%	50	양극화	107	0.15%

2) TF-IDF 분석

<표 5> 2017년 추출 키워드 TF-IDF표

	단어	TF-IDF		단어	TF-IDF
1	부동산	3928.182	16	인상	887.1458
2	시장	2719.422	17	하락	880.5002
3	집값	1899.529	18	주택시장	862.4756
4	서울	1702.301	19	대책	834.0308
5	전망	1615.905	20	주목	830.8473

6	분양	1523.547	21	투자	824.1336
7	아파트	1453.785	22	개최	795.6944
8	내년	1154.587	23	공급	774.8603
9	보유세	1109.49	24	수도권	764.8529
10	올해	1026.095	25	2018	717.0096
11	재건축	1023.667	26	오피스텔	688.7739
12	하반기	1023.48	27	상승	685.7463
13	규제	993.8308	28	전문가	669.8896
14	정부	943.6217	29	종부세	667.7028
15	강남	937.439	30	금리	653.6844

<표 6> 2018년 추출 키워드 TF-IDF표

	단어	TF-IDF		단어	TF-IDF
1	부동산	3624.109	16	강남	920.8358
2	시장	2037.547	17	인상	871.1647
3	집값	1839.585	18	하락	859.0074
4	서울	1662.872	19	대책	824.4981
5	분양	1465.759	20	주택시장	818.4115
6	전망	1438.029	21	주목	812.6812
7	아파트	1403.613	22	투자	755.3015
8	부동산시장	1290.221	23	공급	754.2989
9	내년	1095.77	24	수도권	750.359
10	보유세	1087.918	25	오피스텔	687.5643
11	재건축	1003.385	26	상승	665.8338
12	올해	979.6377	27	종부세	657.402
13	규제	979.0284	28	금리	646.7597
14	하반기	939.9912	29	상가	638.126
15	정부	929.6009	30	전문가	627.7477

3) N-Gram 순열

<표 7> 2017년 추출 키워드 N-Gram표

	단어1	단어2	빈도		단어1	단어2	빈도
1	부동산	시장	979	16	부동산	전문가	50
2	부동산	대책	132	17	서울	아파트값	50
3	서울	집값	111	18	기준금리	인상	49
4	부동산	규제	72	19	지방	부동산	46
5	하반기	부동산	71	20	똘똘한	한채	45
6	서울	아파트	69	21	강남	재건축	45
7	보유세	개편안	67	22	집값	하락	41
8	부동산	투자	60	23	집값	상승	41
9	부동산	정책	59	24	보유세	인상	40

10	서울	부동산	58	25	하반기	주택시장	39
11	양도세	중과	57	26	9·13	부동산	35
12	2018	부동산	57	27	정부	부동산	35
13	부동산	보유세	56	28	강남	집값	33
14	시장	전망	56	29	아파트	시장	33
15	보유세	개편	54	30	금리	인상	31

<표 8> 2018년 추출 키워드 N-Gram표

	단어1	단어2	빈도		단어1	단어2	빈도
1	부동산	시장	517	16	기준금리	인상	47
2	부동산	대책	132	17	똘똘한	한채	45
3	서울	집값	110	18	강남	재건축	44
4	부동산	규제	71	19	서울	부동산	43
5	수익형	부동산	71	20	하반기	부동산	43
6	서울	아파트	69	21	집값	하락	41
7	보유세	개편안	67	22	보유세	인상	40
8	양도세	중과	57	23	집값	상승	40
9	부동산	정책	57	24	지방	부동산	40
10	부동산	보유세	54	25	하반기	주택시장	37
11	보유세	개편	51	26	9·13	부동산	35
12	2018	부동산	51	27	정부	부동산	35
13	서울	아파트값	50	28	상업용	부동산	35
14	부동산	전문가	49	29	강남	집값	33
15	부동산	투자	49	30	아파트	시장	33

2. 2019년 이후 분석결과

1) 빈도 분석

<표 9> 2019년 추출 키워드 빈도분석표

	단어	빈도수	백분율(%)		단어	빈도수	백분율(%)
1	부동산	2721	3.95%	26	경제	138	0.20%
2	시장	1136	1.65%	27	오피스텔	138	0.20%
3	집값	650	0.94%	28	강남	135	0.20%
4	서울	649	0.94%	29	인하	133	0.19%
5	분양	518	0.75%	30	분양시장	132	0.19%

6	전망	470	0.68%	31	수익형	127	0.18%
7	아파트	433	0.63%	32	지방	125	0.18%
8	분양가상한제	352	0.51%	33	한은	123	0.18%
9	올해	260	0.38%	34	인기	114	0.17%
10	투자	241	0.35%	35	이후	114	0.17%
11	하반기	234	0.34%	36	대책	113	0.16%
12	상승	229	0.33%	37	우려	112	0.16%
13	내년	220	0.32%	38	금리	111	0.16%
14	하락	216	0.31%	39	관심	110	0.16%
15	2019년	214	0.31%	40	인천	109	0.16%
16	규제	203	0.29%	41	개발	104	0.15%
17	전국	190	0.28%	42	1년	104	0.15%
18	주목	160	0.23%	43	눈길	100	0.15%
19	부산	155	0.22%	44	들썩	100	0.15%
20	정부	154	0.22%	45	거래	100	0.15%
21	수도권	150	0.22%	46	대전	99	0.14%
22	공급	147	0.21%	47	상반기	96	0.14%
23	기준금리	143	0.21%	48	주택	95	0.14%
24	전문가	142	0.21%	49	세미나	94	0.14%
25	주택시장	139	0.20%	50	최대	94	0.14%

<표 10> 2020년 추출 키워드 빈도분석표

	단어	빈도수	백분율(%)		단어	빈도수	백분율(%)
1	부동산	3895	0.0459	26	동결	162	0.19%
2	시장	932	1.10%	27	2020	161	0.19%
3	집값	738	0.87%	28	발표	159	0.19%
4	코로나	618	0.73%	29	부동산대책	153	0.18%
5	전망	593	0.70%	30	전국	147	0.17%
6	분양	561	0.66%	31	수익형	146	0.17%
7	아파트	507	0.60%	32	강남	142	0.17%
8	서울	494	0.58%	33	분양시장	142	0.17%
9	대책	468	0.55%	34	오피스텔	142	0.17%
10	규제	428	0.50%	35	주택시장	140	0.16%
11	부동산시장	393	0.46%	36	시대	139	0.16%
12	정부	383	0.45%	37	집	135	0.16%
13	올해	307	0.36%	38	종합	132	0.16%
14	수도권	265	0.31%	39	풍선효과	130	0.15%
15	한은	245	0.29%	40	우려	126	0.15%
16	공급	245	0.29%	41	경제	124	0.15%
17	홍남기	229	0.27%	42	추가	121	0.14%
18	상승	221	0.26%	43	최대	119	0.14%
19	기준금리	209	0.25%	44	하반기	117	0.14%
20	주목	204	0.24%	45	아파트값	113	0.13%
21	내년	186	0.22%	46	금리	108	0.13%
22	정책	184	0.22%	47	효과	108	0.13%

23	투자	175	0.21%	48	성장률	108	0.13%
24	주택	174	0.21%	49	상업용	103	0.12%
25	하락	170	0.20%	50	거래	103	0.12%

<표 11> 2021년 추출 키워드 빈도분석표

	단어	빈도수	백분율(%)		단어	빈도수	백분율(%)
1	부동산	3332	3.83%	26	인상	158	0.18%
2	시장	1229	1.41%	27	오세훈	158	0.18%
3	집값	964	1.11%	28	하반기	155	0.18%
4	전망	483	0.55%	29	투자	152	0.17%
5	아파트	442	0.51%	30	기준금리	151	0.17%
6	분양	418	0.48%	31	종합	150	0.17%
7	서울	406	0.47%	32	대출	147	0.17%
8	홍남기	360	0.41%	33	2021	142	0.16%
9	공급	353	0.41%	34	상승세	140	0.16%
10	정부	334	0.38%	35	아파트값	138	0.16%
11	상승	277	0.32%	36	발표	137	0.16%
12	규제	275	0.32%	37	경제	136	0.16%
13	올해	273	0.31%	38	거래	136	0.16%
14	내년	271	0.31%	39	안정	132	0.15%
15	정책	233	0.27%	40	우려	132	0.15%
16	금리	192	0.22%	41	LH	132	0.15%
17	헝다	179	0.21%	42	가격	124	0.14%
18	대책	175	0.20%	43	최대	124	0.14%
19	중국	174	0.20%	44	코로나	121	0.14%
20	종부세	174	0.20%	45	시대	121	0.14%
21	주택	174	0.20%	46	한은	120	0.14%
22	하락	171	0.20%	47	주택시장	118	0.14%
23	완화	166	0.19%	48	재건축	115	0.13%
24	수도권	164	0.19%	49	확대	112	0.13%
25	주목	162	0.19%	50	與	112	0.13%

2) TF-IDF 분석

<표 12> 2019년 추출 키워드 TF-IDF표

	단어	TF-IDF		단어	TF-IDF
1	부동산	3401.707	16	규제	775.0066
2	시장	2390.764	17	전국	738.9528
3	집값	1733.15	18	주목	648.9278
4	서울	1725.448	19	부산	636.5992
5	분양	1501.436	20	정부	630.4791
6	전망	1402.782	21	수도권	619.0539
7	아파트	1331.124	22	공급	608.6594

8	분양가상한제	1150.104	23	기준금리	596.0423
9	올해	928.2756	24	전문가	592.8706
10	투자	880.7366	25	주택시장	583.3133
11	하반기	860.1024	26	경제	583.1463
12	상승	849.6902	27	오피스텔	581.1168
13	내년	824.225	28	강남	572.4841
14	하락	816.2887	29	인하	566.0181
15	2019년	806.7116	30	분양시장	560.7585

<표 13> 2020년 추출 키워드 TF-IDF표

	단어	TF-IDF		단어	TF-IDF
1	부동산	4161.077	16	한은	931.7864
2	시장	2307.405	17	홍남기	886.4008
3	코로나	2208.329	18	상승	865.3026
4	집값	1998.004	19	기준금리	828.086
5	전망	1735.143	20	주목	813.215
6	분양	1686.096	21	내년	758.6422
7	아파트	1567.574	22	정책	753.4767
8	서울	1535.364	23	투자	726.4559
9	대책	1491.222	24	주택	721.3018
10	규제	1393.026	25	하락	710.6857
11	부동산시장	1308.949	26	동결	685.1457
12	정부	1286.515	27	2020	680.9164
13	올해	1100.336	28	발표	673.4546
14	수도권	988.0575	29	부동산대책	653.9266
15	공급	935.8194	30	전국	635.1666

<표 14> 2021년 추출 키워드 TF-IDF표

	단어	TF-IDF		단어	TF-IDF
1	부동산	4074.067	16	금리	779.3276
2	시장	2705.51	17	헝다	740.1867
3	집값	2362.386	18	대책	729.6813
4	전망	1512.396	19	중국	724.4971
5	아파트	1430.279	20	종부세	724.4971
6	분양	1373.299	21	주택	722.4855
7	서울	1342.798	22	하락	714.0058
8	홍남기	1233.06	23	완화	697.081
9	공급	1222.067	24	수도권	690.6704
10	정부	1169.043	25	주목	684.2353
11	상승	1022.371	26	인상	672.294
12	규제	1022.054	27	오세훈	672.294
13	올해	1010.591	28	하반기	661.5161
14	내년	1006.182	29	투자	654.7134

15	정책	901.4428	30	기준금리	649.396

3) N－Gram 순열

<표 15> 2019년 추출 키워드 N-Gram표

	단어1	단어2	빈도		단어1	단어2	빈도
1	부동산	시장	874	16	서울	아파트값	43
2	서울	집값	178	17	기준금리	인하	41
3	부동산	투자	80	18	집값	하락	37
4	부동산	전문가	75	19	하반기	부동산	37
5	1년	뒤	58	20	지방	부동산	36
6	서울	아파트	56	21	2020	부동산	35
7	부동산	대책	56	22	민간택지	분양가상한제	34
8	시장	전망	53	23	부동산	정책	34
9	부동산	규제	53	24	올해	전국	33
10	서울	부동산	49	25	금리	인하	32
11	2019년	부동산	47	26	대전	부동산	32
12	10대	건설사	46	27	상업용	부동산	31
13	하락	전망	45	28	한국감정원	올해	30
14	부산	부동산	43	29	한은	기준금리	29
15	집값	상승	43	30	내집	마련	27

<표 16> 2020년 추출 키워드 N-Gram표

	단어1	단어2	빈도		단어1	단어2	빈도
1	부동산	시장	638	16	기준금리	동결	42
2	부동산	대책	268	17	가구	분양	42
3	부동산	규제	126	18	2020	부동산	39
4	부동산	정책	126	19	대책	발표	39
5	수익형	부동산	100	20	서울	아파트값	38
6	한은	기준금리	90	21	제로금리	시대	38
7	상업용	부동산	85	22	성장률	전망	37
8	정부	부동산	76	23	수도권	집값	36
9	서울	집값	69	24	분양	예정	36
10	부동산	투자	62	25	6·17	부동산	36
11	서울	아파트	60	26	포스트	코로나	35
12	집값	상승	57	27	부동산	전망	35
13	부동산	거래	48	28	부산	부동산	34
14	문	대통령	48	29	부동산	감독기구	33
15	지분적립형	주택	48	30	부동산시장	전망	32

<표 17> 2021년 추출 키워드 N-Gram표

	단어1	단어2	빈도		단어1	단어2	빈도
1	부동산	시장	883	16	문	대통령	40
2	부동산	정책	150	17	규제	완화	39
3	서울	아파트	67	18	내집	마련	37
4	상업용	부동산	67	19	시장	안정	37
5	수익형	부동산	62	20	집값	오른다	37
6	금리	인상	59	21	부동산	규제	36
7	부동산	대책	57	22	집값	안정	35
8	홍남기	부동산	56	23	한은	기준금리	35
9	서울	집값	51	24	올해	부동산	35
10	시장	전망	48	25	서울	부동산	34
11	집값	상승	46	26	주택	공급	34
12	부동산	투자	46	27	집값	상승세	34
13	서울	아파트값	43	28	2021년	부동산	34
14	집값	하락	42	29	대출	규제	33
15	기준금리	인상	41	30	부동산	거래	32

Ⅴ. 결론

본 연구는 다수의 뉴스 기사를 통해 텍스트 마이닝 기법을 이용하여 주택시장의 분위기를 가늠할 수 있는 키워드와 단어의 중요성을 정량화하여 객관적인 정보를 찾았다는 점에서 의의가 있다. 부동산 관련 소셜미디어, 뉴스 등 비정형 데이터의 수집 및 패턴을 실시간으로 모니터링하여 정책 데이터로 활용할 것을 제안한다.

또한 부동산 시장 동향은 빅데이터를 활용하여 빠르게 파악될 수 있으므로 정책입안자들은 부동산 빅데이터 관련 기술을 개발하고 정책에 적용할 수 있는 다양한 지원방안을 수립·시행해야 한다. 이를 통해 부동산 정보의 선진화에 기여하고 부동산 시장을 보다 체계적이고 빠르게 알 수 있을 것으로 기대된다.

본 연구를 통해 부동산시장의 현황을 직관적으로 확인할 수 있었으며, 향후 후속 연구를 진행하여 부동산시장 관련 키워드를 통한 주택가격변동과의 관계를 확인하고자 한다.

참고문헌

1. 경정익, & 이국철. (2016). Textmining 에 의한 부동산 빅데이터 감성분석 모형 개발. 주택연구, 24(4), 115-136.
2. 김대원, & 유정석. (2013). 주택가격에 대한 심리적 태도가 주택 매매 거래량에 미치는 영향 분석. 주택연구, 21(2), 73-92.
3. 박재수, & 이재수. (2019). 벡터자기회귀모형을 이용한 온라인 뉴스기사와 아파트 매매가격의 동태적 관계 연구: 비정형 빅데이터를 활용한 감성분석 기법의 적용. 감정평가학논집, 18(2), 83-113.
4. 박재수, & 이재수. (2021). 부동산 감성지수의 주택가격 예측 유용성: 뉴스기사와 방송뉴스 빅데이터 활용 사례. 국토계획, 56(4), 99-111.
5. 박종영, & 서충원. (2015). TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석. 부동산학보, 63(63), 46-58.
6. 이정미. (2013). 빅데이터의 이해와 도서관 정보서비스에의 활용. Journal of the Korean BIBLIA Society for library and Information Science, 24(4), 53-73.
7. 전해정. (2019). 빅데이터와 텍스트마이닝을 이용한 부동산시장 동향분석. 디지털융복합연구, 17(4), 49-55.
8. 진창하. (2012). Newspaper content and home prices: Perception, reasoning and affect. 부동산학연구, 18(2), 125-142.
9. 한국정보화진흥원. (2012). 성공적인 빅데이터 활용을 위한 3대 요소 : 자원, 기술, 인력, 빅데이터국가전략포럼.
10. LG경제연구원. (2012). 빅데이터 시대의 한국 갈라파고스가 되지 않으려면. LG 경제연구소 LG Business Insight, 2-6.
11. Weatherford, M. S. (1983). Economic voting and the “symbolic politics” argument: A reinterpretation and synthesis. American Political Science Review, 77(1), 158-174.