

아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석

– 온라인 뉴스 기사의 비정형 빅데이터를 활용한 감성분석 기법의 적용 –

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 –

박재수* · 이재수**

Park, Jae-Soo · Lee, Jae-Su

Abstract

The coming of the digital age has dramatically increased the information exchanges among the individuals. Before the digital age, the flow of information was confined to the smaller area around the individual. However, the introduction of smartphones and social media such as facebook and twitter have transformed this flow of information. News articles are one of the important sources from which the information flow originates and the real estate market which is one of the asset markets responds strongly to the news. The purpose of this study to analyze the relationship between sentiment index and real estate price by crawling news article and using topic analysis, cross correlation and Granger causality test. The result of this study showed that the news had significant leading as well as Granger causality relationships with the apartment sales price in small and small and medium apartment houses. This result could be used as evidence to infer the diminishing interest in large houses after the price fall in 2008 Financial Crisis and population aging.

키워드 주택매매가격, 빅데이터, 텍스트 마이닝, 감성분석, 그래저 인과관계 검정

Keywords Housing Sales Price, Big data, Text Mining, Sentiment Analysis, Granger Causality Test

1. 서론

1. 연구 배경 및 목적

디지털과 모바일 시대의 진입으로 세상은 디지털 세상으로 빠르게 변화하고 있다. 최근 사회 시스템 또한 스마트폰 등 스마트

기기들을 매개로 디지털과 모바일 환경으로 급속히 이동하고 있다. 사람 간의 정보교환 방식도 트위터, 페이스북, 카카오톡, 네이버 밴드와 같은 소셜미디어 및 메신저 플랫폼을 기반으로 변화되어 가고 있다. 이런 시대적 변화에 따라 부동산 분야에서도 통계학적 접근, 거시경제학적 접근, 생애주기적 접근 등과 같은 전통적 연구 및 분석 방법과 더불어 최근에는 빅데이터 등을 활용한

* Doctorate Candidate, Department of Real Estate, Kangwon National University(first author: okparkjaesoo@naver.com)

** Assistant Professor, Department of Real Estate, Kangwon National University(Corresponding author: jslee25@kangwon.ac.kr)

새로운 분석 방법의 개발 및 적용이 요구되고 있다.

부동산 시장에서 주택 등의 가격결정 메커니즘을 설명하고 예측하기 위해서 전통적인 연구방법은 센서스 자료, 공공 통계자료, 설문조사 자료 등 정형 데이터에 근거해 헤도닉 가격모형, 시계열 분석모형, 패널 분석모형 등을 활용하였다. Won and Lee(2018)는 서울시 소형주택의 가격결정 메커니즘을 설명하기 위해 특정 연도에 조사한 데이터를 활용한 헤도닉 가격모형을 추정하였다. 주택가격은 주택의 물리적 특성, 입지 특성, 근린환경 특성과 인구 사회적 특성 변수들에 의해 결정되며, 소형주택은 아파트의 가격결정 특성과 다름을 실증하였다. 권건우·진창하(2016)는 생애주기별 가구의 주거비 부담 결정요인의 분석을 통해 생애주기별 가구의 주거비 부담을 결정하는 요인이 상이하고 한 요인에 특정 세대가 더 민감하게 반응함을 주장하였다. 즉 가구의 생애주기에 따라 주택의 소비와 수요를 달리한다는 것으로 청년과 노년가구의 주거비 부담이 청년과 장년가구 보다 높다는 것과 청년과 장년가구의 주거비 부담요인으로 주택점유형태, 기타주택 대비 아파트 거주 면적, 가구원수, 근로형태가 있음을 통계적으로 유의미함을 밝혔다. 이를 통해 권역별 주택수요를 간접적으로 측정할 수 있을 것이다. 임대봉(2015)은 유동성이 주가 및 주택가격에 미치는 효과의 분석을 통해 유동성이 서울 및 인천 아파트에는 유의미한 영향을 미치지 않지만, 다른 지역에서는 영향이 미미하다고 주장하였다.

대부분의 선행연구는 부동산 구매자의 심리적인 관점이 배제된 상태에서 미시 또는 거시 변수들의 변화만으로 부동산 시장을 분석한 한계가 내재되어 있다. 그러나 부동산 시장의 다양한 현상을 분석하는 선행연구들은 부동산 수요자들의 기대 등 심리적 요인들을 충분히 고려하지 못하는 등 데이터 수집 및 변수 측정과 관련된 비판이 지속적으로 제기되었다. 최근 데이터의 양과 속도가 폭발적으로 증가하고 다양한 데이터의 활용이 가능해짐에 따라 이러한 전통적 접근방법의 한계를 극복할 수 있는 새로운 시도가 제시되고 있다. 박종영·서충원(2015)은 부동산 뉴스 기사를 수집하여 시기별로 주택시장의 변화를 설명할 수 있는 핵심 단어와 중요도의 계량화를 시도하였다. 김대원·유정석(2016)은 소셜 미디어인 트위터 단어의 빈도 변수와 아파트 매매가 및 전세가 지수의 관계를 분석하였다. 또한 경정익·이국철(2016)은 부동산 뉴스 기사를 수집하고 감성지수를 산출하여 부동산 시장 진단과 가격을 예측할 수 있는 모형을 제시하였다.

기존 선행연구는 다양한 미시 또는 거시 변수들의 변화에 따른 주택가격 결정 또는 예측모형을 제시하거나 생애주기에 따라 어떻게 이동하고 어떻게 주택을 소비하는지를 주로 다루었다. 하지만 정보의 유통속도가 빠른 현재는 기존의 변수 이외에 다양한 온라인 데이터를 활용한 변수들을 연구에 적용하기 위한 방법론의 필요성이 증대되고 있다. 최근 많은 연구 및 실무분야에서 정형 데이터가 아닌 비정형 데이터를 이용한 환경이 조성되고 있다.

이에 부동산 분야에서도 다양한 빅데이터를 활용하여 기존과는 다른 새로운 분석방법과 연구결과를 제시하는 시도가 필요하다고 판단된다.

본 연구는 서울시를 사례지역으로 주요 온라인 신문기사에서 추출한 아파트 매매와 관련된 감성지수가 아파트 매매가격과 시계열적으로 유의미한 영향 관계를 나타내는지 분석하고 시사점을 제시하는 것을 목적으로 한다. 온라인 신문기사의 텍스트 마이닝(Text Mining)을 위해 웹 크롤링(Web Crawling) 기술을 개발하고 적용하여 아파트 매매와 관련된 신문 기사를 수집한다. 수집된 기사는 토픽분석을 통해 연관된 주제로 분류하고 아파트 매매와 관련된 단어로 감성분석을 실시한다. 감성분석을 통해 산출된 월별 감성지수가 아파트 매매가격에 유의미한 선행관계와 인과관계를 형성하는지 교차상관분석과 인과관계 분석을 통해 검증한다. 서울시의 하위 주택시장을 고려하여 도심권, 서북권, 동북권, 서남권, 동남권의 5대 권역으로 구분하고, 아파트의 규모별로 5개 규모로 세분하여 분석하였다.

본 연구의 차별성은 분석대상 아파트시장을 서울시의 권역과 아파트의 규모에 따라 구분하고, 분석대상 빅데이터인 온라인 뉴스를 토픽모델링 분석을 통해 신문에 포함된 단어들을 체계적으로 분류하여 감성사전 및 감성지수를 만드는데 활용하였다는 것이다.

2. 연구 범위 및 내용

본 연구는 시간적으로 2010년 1월부터 2017년 12월까지 96개월을 분석기간으로 설정하고 분석의 시간 단위는 월단위를 기본으로 한다. 분석대상을 월단위 기준으로 한 이유는 다른 부동산 데이터와의 비교분석이 용이하기 때문이다. 대부분 부동산 관련 데이터는 주간 또는 월간으로 공표되기 때문이다. 공간적 분석범위는 (Fig. 1)에서와 같이 서울시 전체와 5대 권역(대생활권)으로 세분하였다.

아파트 실거래가격은 기준으로 환산하고 2010년 1월 가격을 100으로 하여 지수화 하였다. 아파트의 규모는 국토교통부의 실거래가 공개시스템에서 분류하는 기준에 따라 5개로 구분하였다. 소형은 60m² 이하, 중소형은 60~85m², 중형은 85~102m², 중대형은 102~135m², 대형은 135m² 이상으로 분류한다.

신문기사는 3개 일간지, 조선일보(www.chosun.com), 동아일보(www.donga.com), 중앙일보(www.joins.com)와 3개 경제전문지, 한국경제(www.hankyung.com), 매일경제(www.mk.com), 서울경제(www.sedaily.com)를 대상으로 한다. 이 신문사들을 선정할 이유는 국내에서 상대적으로 영향력이 큰 일간지와 경제지를 적절히 고려하기 위해서이다. 한국ABC 협회의 2017년도 일간신문 발행 유료부수 자료에 따르면, 유료부수를 기준으로 조선일보 1위, 동아일보 2위, 중앙일보 3위, 매일경제 4위

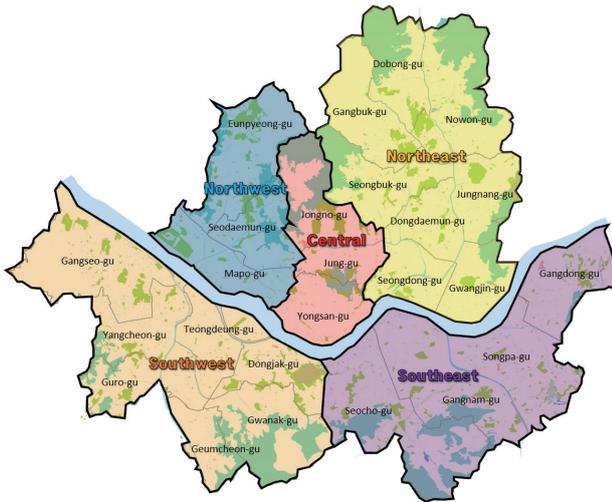


Fig. 1. Seoul Metropolitan Area and Subregions

Source: SMG(2014)

한국경제 5위, 서울경제 24위 순으로 나타났다. 해당 신문의 웹 페이지에서 분석대상 기간의 부동산 관련 뉴스를 웹 크롤링 기법을 이용하여 수집하고 감성지수를 측정하였다. 감성지수는 수집된 기사에 나타난 긍정적 표현과 부정적 표현의 빈도수 비율로 아파트 가격을 지수로 만든 것과 동일하게 2010년 1월을 100으로 하여 상대적 지수로 산출하였다.

II. 이론 및 선행연구 검토

1. 빅데이터의 개념

빅데이터에 대한 정의는 연구자들마다 다양하게 정의하고 있다. 가트너그룹(Gartner Group)의 애널리스트 레이니(Laney)는 데이터의 급성장에 따른 이슈와 기회를 데이터의 양(Volume), 데이터 입출력의 속도(Velocity), 데이터 종류의 다양성(Variety)이라는 세 개의 차원으로 정의하였다(Laney, 2001). 최근에는 빅데이터의 분석을 통해 가치(Value) 있는 정보를 추출하여 활용하는 시대로 4V로 표현한다(김정선 외, 2014). 또한 신뢰성과 분석할 만한 가치 등을 담보하기 위해 데이터의 새로운 속성인 정확성(Veracity)도 제시되고 있다. 따라서 가치(Value)와 정확성(Veracity)을 모두 포함하는 5V로 개념이 확장되고 있다.

빅데이터는 정형화된 데이터와 비정형화된 데이터로 크게 구분할 수 있다. 먼저 정형화된 데이터는 통계청과 같은 정부기관이나 국가기관은 아니지만 통계를 작성 발표할 수 있는 지정기관에서 작성하고 있는 통계로 200여 종에 달하며 주로 숫자로 작성되고 있다. 반면 비정형데이터는 우리사회 곳곳에서 만들어지고 있는데, 특히 페이스북, 트위터, 카카오톡과 같은 소셜 네트워크 서비스 이용자들이 많은 데이터를 생산하고 있다(박원준, 2012). 비정형 데이터의 대표적인 형태는 숫자를 포함하여 문자, 사진,

이모티콘 등과 같은 형태의 데이터들이다.

2. 웹 크롤링과 감성분석

1) 웹 크롤링(Web Crawling)

웹 크롤링은 크롤러를 이용하여 접근할 수 있는 웹의 정보를 수집하는 기술을 의미한다. 웹크롤러는 로봇(robots), 스파이더(spiders) 또는 봇(bots)으로도 불리며, 이를 통하여 웹 페이지를 자동으로 내려 받아 빅데이터를 구축할 수 있게 해주는 기술을 말한다. 데이터 크롤링 방식은 다양하며 일반적으로 웹페이지의 각 링크를 일일이 따라가 정보를 얻는 작업을 대신하여 자동적으로 웹서버를 순회하며 웹페이지의 내용을 분석하고, 그 안에 포함되어 있는 수많은 텍스트·수치·그림·멀티미디어 정보 등을 수집하는 방식을 활용한다(통계청 통계개발원, 2015a).

웹 크롤링은 5가지 요소로 구성되어 있다. 첫째, 사이트주소 대기(URL Queue)로 사이트주소 필터링에서 매칭되지 않은 사이트주소, 응용 실시간 필터링에서 크롤링하기 위한 최초 사이트주소 리스트이다. 둘째, 다운로더(Downloader)로 사이트 주소 대기(URL Queue)에서 사이트주소를 가지고 와 해당 웹 서버에 접속하여 정상적으로 응답을 받으면 해당 사이트주소의 웹 페이지를 다운로드 받는다. 셋째, 파싱(Parsing)으로 다운로드 받은 웹 페이지의 소스에서 각 태그를 파싱하여 각 태그를 저장하고 형태에 따른 사이트주소를 추출한다. 넷째, 중복검사(Redundancy Check)로 추출된 태그별로 추출된 사이트주소를 저장 시 사이트주소의 끝대기열(URL End Queue)에 있는 리스트와 매칭하여 이미 분석된 사이트주소는 제외하고, 분석 및 크롤링 되지 않은 사이트주소만 저장한다. 다섯째, 사이트주소 끝대기열(URL End Queue)로 크롤링하여 분석된 사이트주소 리스트이다(정승일·김현우, 2014).

2) 감성분석(Sentiment Analysis)

오피니언 마이닝(Opinion mining)으로도 불리는 감성분석은 텍스트마이닝(Text Mining)기법으로 비정형 데이터인 문서나 문장에서 작성한 사람의 감정을 추출해 내는 기술이다. 이와 같이 문서에 기록되어 나타난 사람들의 의견, 성향, 태도와 같은 주관적인 데이터를 정제하고 분석하는 자연어 처리 기술인 감성분석을 수행하기 위해서는 다음과 같은 일련의 3단계 과정으로 이루어진다.

첫째, 데이터 수집단계는 감성분석을 하기 위한 핵심자료가 된다. 데이터 수집과정은 첫 번째 단계로서 비정형 대규모 텍스트 데이터를 수집하는 단계이다. 감성 분석 기술을 적용할 데이터를 수집하기 위해서는 검색 엔진을 활용한다. 검색 엔진은 사용자의 질의어를 입력 받아 질의어가 포함된 모든 문서, 각종 데이터를 수집한다. 둘째, 주관성 탐지단계로 연구에 필요한 텍스트를 수

집하고 난 후 감성 분석에 사용될 텍스트를 분리, 분류하는 작업이 필요하다. 문장의 단어, 구획 등을 분석해 전체를 대표하는 용어들을 추출하는 기법이다. 일반적으로 인터넷에서 수집된 신문 기사 중 '감성'과는 관련이 없다고 판단되는 부분은 제외시킨다. 셋째, 극성 탐지단계로 주어진 데이터가 '긍정'인지, 혹은 '부정'인지를 판단하는 극성분석(Polarity Detection) 작업이 이뤄진다. 텍스트 안에 있는 긍정적, 부정적인 단어를 구분 한 뒤 통계적 기법을 적용하여 '빈도'와 '속성'에 점수 가중치를 부여한 뒤, 각 단어가 나타내는 점수의 총합이나 평균을 구해 긍정적인지 혹은 부정적인지 알아내는 것이다(신수정, 2014).

3. 선행연구 검토

본 연구와 관련된 선행연구는 크게 3가지 유형으로 나뉜다. 첫째, 뉴스기사나 트위터와 같은 비정형 데이터가 부동산시장에 미치는 영향 연구. 둘째, 감성분석 모델을 제시하고 모델의 유용성을 부동산시장 분석을 통해 실증한 연구. 셋째, 감성분석을 주식시장이나 석유와 같은 상품시장에 적용하고 여러 분석방법의 유용성을 실증한 연구이다.

우선, 뉴스기사나 트위터 같은 소셜미디어를 중심으로 감성을 분석한 초기 연구로 김진유(2006)는 투기라는 단어를 통해 신문기사가 부동산 가격변동에 미치는 영향을 그랜저 인과관계를 분석하였다. 결과는 투기관련 기사의 양(기사건수)은 부동산가격과 양방향의 그랜저 인과관계가 있음을 밝혔다. 신문기사의 방향과 주택가격의 변화 사이에도 유의미한 인과관계가 있음을 밝혔다.

진창하·Gallimore(2012)는 미국 애틀란타 지역 부동산 시장에서 시장 참여자는 다른 자산시장과는 상이한 특징을 나타내고 있다고 주장하였다. 특히 수익형 부동산 시장과 달리 주택과 같은 주거용 부동산 시장은 비전문적인 시장참여자에 의해 시장의 형태가 영향을 받는다고 주장하였다. 공적분 검정, 그랜저 인과관계 검정과 오차수정모형을 이용하여 주택시장과 신문기사와의 관계를 분석하였다. 분석 결과, 긍정적인 용어의 사용 보다 부정적인 용어의 사용이 주택시장의 가격변화에 대해 연관성이 높은 것으로 나타났다.

최근 연구로 김기홍·김재태(2014)는 동남권의 3개구(강남구, 서초구, 송파구)의 아파트를 대상으로 공동주택공시가 대비 실거래가율, 공동주택가격 대비 경매 낙찰가율, 공동주택가격 대비 경매 감정가율에 대한 그랜저 인과관계 검정을 실시하였다. 공동주택가격 대비 경매 낙찰가율과 공동주택공시가 대비 실거래가율이 공동주택가격 대비 경매 감정가에 영향을 미치고, 공동주택가격 대비 경매 낙찰가율이 공동주택공시가 대비 실거래가율에 영향을 미치는 결과를 도출했다. 통계청 통계개발원(2015b)은 소셜미디어 데이터를 활용하여 공식통계와 비교·분석하였다. 감성지수 산출은 긍정표현의 빈도수를 긍정표현의 빈도수와 부정표

현의 빈도수 합계로 나누어 지수화 하였다. 연구 결과, 소셜미디어 기반 지표가 방법론적으로 안전하게 생산될 수 있다면 빅데이터 기반 통계들은 종래의 공식 통계보다 저렴하고 더 빠를 수 있는 잠재력이 있음을 확인하였다.

김대원·유정석(2016)은 소셜미디어와 주택시장 간의 동적 관계를 실증 분석하였다. 주택가격과 전세가격으로 검색된 트위터 자료를 수집하고 월별로 상승 및 하락 단어의 빈도를 산정하여 소셜 미디어의 대응변수로 사용하였다. 아파트 매매가격 지수 및 전세가격 지수를 로그 차분한 변동률을 주택가격의 대응변수로 사용하였다. 벡터자기회귀(VAR) 모형 분석에 따른 충격 반응분석을 실시하였고, 그랜저 인과관계 검정을 실시하여 변수들 간 인과성을 확인하였다. 분석 결과, 상승보다는 하락이 아파트 매매가 변동률에 더 많은 영향을 미친 것으로 확인하였다. 이종민 외(2017a)는 네이버와 구글에서 제공하는 검색지수 중에서 전세와 전세대출 검색지수를 이용하여 전세가격 예측의 유용성을 검증하였다. 그랜저 인과관계 분석에서 네이버와 구글의 검색 지수가 한쪽 방향으로 전세가격 변동률에 영향을 미치고 있음을 확인하였다. 교차상관분석에서는 네이버 검색 지수가 6-7개월의 시차를 두고 전세가격변동률과 동조하는 것을 확인하였다. 그리고 포털 검색 지수를 포함한 전세가격예측모형이 전세가격만 이용한 예측력을 보다 향상시키고 있음을 실증하였다.

둘째, 감성분석 모형개발과 관련된 연구로 경정익·이국철(2016)은 부동산과 관련된 비정형데이터를 활용하여 부동산시장을 분석 예측할 수 있는 빅데이터 감성분석 모형 개발을 시도하였다. 감성분석 모형에 사용될 감성지수는 수집된 뉴스를 지역별, 일자별, 매체, 뉴스 유형별로 분류한 후 뉴스 내용을 일련의 언어 처리 과정인 형태소 분석 및 의존 구문 분석을 통해 어휘와 의미의 특징을 도출하여 감성분석에 유용한 형태로 정규화하였다. 감성사전을 기반으로 각 문장에 극성 값을 태깅한 후 극성 값을 기간별 지역별로 구분하였다. 모형을 검정하기 위해 감성분석 데이터와 이와 관련된 부동산 정형데이터와 상관분석을 실시한 결과, 상관계수 0.5~0.6을 나타내 비교적 높은 상관관계가 나타났음을 밝혔다. 박종영·서충원(2015)은 부동산 뉴스기사를 수집하여 주택시장과 관련된 단어를 빈도수, 엠-그램(N-gram) 순열, 단어 빈도-역문서빈도(TF-IDF) 가중치를 구하고 시기별로 다루어지는 핵심 단어를 발견하여 당시의 주택시장 상황을 파악하는 연구에 적용된 분석기법을 소개하였다. 시기별로 빈도수가 많은 단어의 종류가 달라짐을 확인하고 이를 통해 주택시장의 분위기를 가늠할 수 있는 핵심 단어와 단어의 중요도를 계량화해 객관적인 정보를 찾고자 하였다.

강범일 외(2017)은 토픽 모델링 기법을 이용하여 18대 대선 후보 기사에서 주로 형성되는 주제들을 추출했다. 이 주제들이 매체별로 어떤 차이를 보이는지, 형성된 주제가 구성된 단어들의 내용에 무슨 차이가 있는지, 또한 시기별로 주제 분포 차이가 어

떻게 드러나는지 분석하였다. 분석 결과, 매체별, 시기별, 주제별로 주제의 분포가 상이함을 확인하였다. 또한 각 정파 진영에서는 후보 자신에게 비추어 유리한 이슈는 적극적으로 보도하고 불리한 이슈는 다소 소극적으로 보도하는 경향이 있는 것으로 나타났다. 감미아·송민(2012)은 텍스트마이닝 기법을 활용하여 단순 빈도 분석을 위한 신문기사의 키워드와 클러스터링, 분류 결과를 분석하여 국제, 정치, 경제, 사회, 문화 및 사설 분야에서 신문사간 차이점을 분석하였다. 수집된 신문기사들은 불용어 제거, 형태소 분석과 자연어 처리 후 키워드들 간 코사인 유사도를 분석하고 네트워크 지도를 만들어 단어들의 네트워크를 통해 클러스터 결과를 분석하였다. 분석 결과, 특정 기사에 대해서 신문 기사엔 논조의 차이가 있음을 발견하였다. 예를 들어, 4대강 사업 관련 신문기사에서 진보적인 신문은 특정 단어가 비슷한 순위에 위치한 반면, 보수적인 신문은 이와 다르게 구성되는 경향이 있어 현상을 다르게 바라보고 있다고 주장하였다.

셋째, 주식시장과 빅데이터와의 관계를 분석한 연구로 이득환 외(2013)는 빅데이터에 나타난 9가지 감성들의 특징을 다음소프트(Daum-soft)에서 제공받은 자료를 대상으로 자기상관분석, 주성분 분석, 벡터자기상관(VAR) 추정을 실시하여 감성이 가지고 있는 특징을 분석하였다. 분석 결과, 감성들은 일정한 패턴을 가지고 있고, 주성분 분석을 통해 긍정성, 부정성 형태의 9가지 감성들로 묶일 수 있음을 나타냈다. 이를 바탕으로 빅데이터에 나타난 주가 정보는 다양한 감성들을 담고 있으며 무작위적인 정보의 나열이 아니라 일정부분 주식시장과 흐름을 같이 하고 있으며 과거 값을 통해 예측이 가능함을 보였다. Li et al.(2016)은 로이터에서 수집한 석유 관련 뉴스를 감성분석과 빅데이터 분석도구를 사용하여 미국 텍사스 중질유 가격을 예측하는데 유용하다고 주장하였다. 기사에서 뽑아낸 지수가 실제 원유가격과 방향성이 동일함을 확인하였다. 또한 그랜저 인과관계 검증에서는 3주의 시차를 두고 기사와 가격 사이에 유의미한 인과관계가 나타남을 실증하였다.

이 연구는 텍스트마이닝을 활용한 감성분석 중 토픽분석을 통해 감성사전을 만들고 감성지수를 산출하였다. 토픽모형을 활용함으로써 신문기사에서 사용한 단어들을 보다 체계적으로 분석할 수 있고, 이를 통해 부동산 시장에 영향을 미치는 단어들을 추출할 수 있다. 이를 위해 국내에서 상대적으로 영향력이 큰 3개 일간지와 3개 경제지를 고려하였다. 또한 부동산 시장의 감성지수와 아파트 매매가격의 영향관계를 분석하기 위해 주택시장의 공간범위를 5대 권역으로 구분하고, 서울에 입지한 아파트를 국토교통부의 기준에 따라 소형, 중소형, 중형, 중대형 및 대형의 5개 규모로 세분하여 지역별 특성을 비교·분석하고자 한 점에서 연구의 차별성이 있다. 또한 많은 선행연구들이 활용한 주택가격지수 또는 아파트가격지수 대신 실거래가격을 단위면적당 매매가격으로 환산하고 2010년 1월 가격을 100으로 하여 지수화 하였다.

다. 시간적으로는 2010년 1월부터 2017년 12월까지 96개월을 분석기간으로 설정하여 최근 부동산 온라인 뉴스와 서울의 아파트 매매가격의 관계를 분석한 점에서도 연구의 차별성이 있다.

III. 분석틀 설정

1. 자료수집 및 분석절차

본 연구를 위해 신문기사와 아파트 실거래가격 자료를 수집하였다. 먼저 신문 기사를 수집하기 위해 일간지 3개(조선일보, 동아일보, 중앙일보)와 경제지 3개(매일경제, 한국경제, 서울경제)를 선정하였다. 기간은 2010년 1월부터 2017년 12월까지 96개월로 하였다. 본 연구는 자산시장의 급격한 가격 움직임을 분석에서 제거하기 위해 2008년 글로벌 금융위기 영향이 줄어들기 시작한 2010년을 분석의 기산점으로 정하였다.

해당 웹사이트에서 웹 크롤링을 이용해서 '아파트'를 키워드로 총 184,280건의 아파트 매매관련 기사를 수집하였다. 본 연구는 '주택'이라는 키워드 보다는 분석 목적과 대상에 실질적으로 부합할 수 있는 '아파트'를 키워드로 신문 기사를 수집하였다. 아파트 실거래가격은 국토교통부 실거래가 공개시스템(rtdown.molit.go.kr)에서 월별로 발표하는 자료를 활용하였고, 2010년 1월부터 2017년 12월까지 96개월간의 자료를 사용하였다.

아파트 매매가격과 감성지수 사이의 인과성 및 선행성을 검증하기 위해 <Fig. 2>와 같은 절차에 의해 연구를 진행하였다. 먼저 아파트 매매가격은 국토교통부 실거래가 공개시스템에서 수집하였고, 신문기사는 3개 일간지와 3개 경제지를 포함한 6개 신문사의 웹사이트에서 아파트를 키워드로 관련 기사를 웹 크롤링하여 수집하였다. 수집된 신문기사는 토픽분석을 통해 감성사전에 사용할 단어를 추출하고 단어와 관련된 긍정과 부정 서술어를 통해 감성사전을 만들었다. 구축된 감성사전을 이용하여 월별로 긍정과 부정의 단어개수를 계산하여 긍정 및 부정 감성지수를 산정하였다.

아파트 실거래가격을 분석에 이용하기 위해 수집된 원자료를 m²당 매매금액으로 환산한 후 2010년 1월을 기준월로 설정하고 실거래 가격지수를 100으로 설정한 후 매월 실거래가격을 모두 지수화 하였다. 아파트 매매가격을 기반으로 산정한 매매가격지수와 토픽분석과 감성사전을 활용하여 만든 감성지수를 이용하여 그랜저 인과관계와 교차상관분석을 실시하였다.

2. 분석방법

1) 토픽 모델링(Topic Modeling)

토픽 모델(Topic model)이란 문서 집합의 추상적인 '주제'를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의

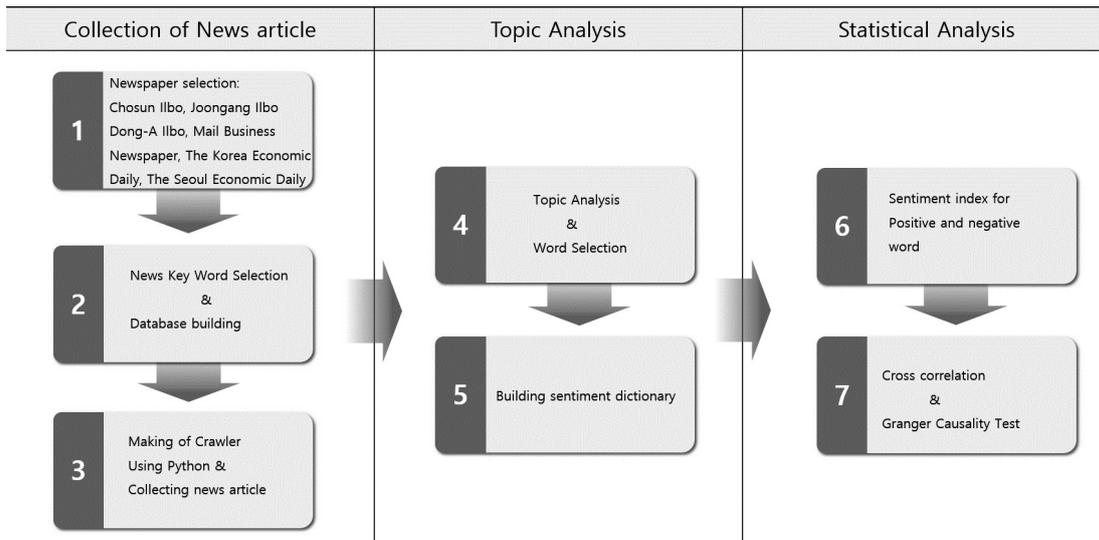


Fig. 2. Analysis Procedure

미구조를 발견하기 위해 사용되는 텍스트마이닝 기법 중 하나이다. 특정 주제에 관한 문헌에서는 그 주제에 관한 단어가 다른 단어들에 비해 더 자주 등장할 것이다. 이것이 토픽 모델의 개략적인 개념이다. 실제로 문헌 내에 어떤 주제가 들어있고, 주제 간의 비중이 어떤지는 문헌 집합 내의 단어 통계를 수학적으로 분석함으로써 알아낼 수 있다. 토픽 모델링은 문서를 이루고 있는 키워드들을 바탕으로 문서에서 주제를 찾아내기 위해 사용되는 방법론으로 대량의 문서 집합에 적용되며 다양한 종류의 데이터에 적용이 가능하다(신규식 외, 2015).

토픽 모델링의 가장 대표적인 방법론은 잠재디리클레할당(Latent Dirichlet Allocation: LDA)이다. LDA 알고리즘은 생성모델로서 문헌 내의 숨겨져 있는 주제들을 찾아내는 알고리즘이다. 따라서 문헌, 단어 등 관찰된 변수(Observed variable)를 통해 문헌의 구조와 같은 보이지 않는 변수(Hidden variable)를 추론하는 것을 목적으로 한다. 이 분석을 통해 전체 문서 집합의 주제들과 각 문서별 주제 비율, 각 주제에 포함될 단어들의 분포를 알 수 있다(Blei, 2012).

토픽 모델은 또한 확률적 토픽 모델이라고도 불리는데, 이는 광범위한 텍스트 본문의 잠재적 의미 구조를 발견하기 위한 통계적 알고리즘을 가리키는 의미로도 쓰인다. 정보화 시대가 도래하면서 매일 생성되는 텍스트는 인간이 직접 처리할 수 있는 양을 크게 넘어서는데, 토픽 모델은 자동적으로 비정형 텍스트의 집합을 이해하기 쉽도록 조직하고 정리하는 데에 쓰일 수 있다. 또한 토픽 모델은 원래 개발된 목적인 텍스트마이닝 분야 이외에도 유전자 정보, 이미지, 네트워크와 같은 자료에서 유의미한 구조를 발견하는데 유용하게 사용되고 있다(Blei, 2012).

(Fig. 3)은 LDA 알고리즘에서 사용되는 확률 계산을 도식화한 것으로 α 는 각 문서별 토픽의 디리클레 사전확률을 의미하며, β 는 각각의 토픽별 단어의 분포를 디리클레 사후 확률로 나타낸

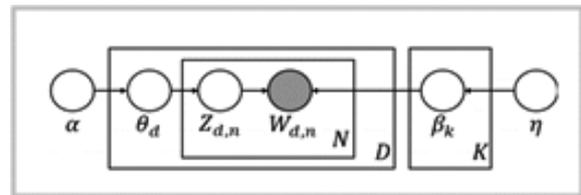


Fig. 3. Plate Notion for Smoothed LDA

다. θ 는 각 문서에 대한 토픽 분포를 의미하며, Z 는 각 문서에서 출현하는 단어의 토픽을 나타내며, ω 는 토픽에 포함된 단어를 의미한다. 잠재변수 α 와 β , 토픽 결합분포 θ , 단어의 토픽 Z , 토픽에 포함된 ω 가 주어졌을 때 확률분포는 (equation 1)같이 계산된다(이종민 외, 2017a).

$$p(\beta_{1:k}, \theta_{1:D}, z_{1:D}, \omega_{1:D}) = \prod_{i=1}^k p(\beta_i) \prod_{d=1}^D p(\theta_d) \left(\prod_{i=1}^k p(z_{d,n} | \theta_d) p(\omega_{d,n} | \beta_{1:k}, z_{d:n}) \right) \quad (1)$$

요약하면, 어떤 문서에 대해 파라미터 θ 인 주제 벡터가 있고 앞서부터 단어를 하나씩 θ 로부터 하나의 주제를 선택하고, 다시 그 주제로부터 단어를 선택하는 방식으로 문서 생성과정을 모델링하는 것이다. 문서1과 문서2가 있을 때, 주제는 비슷하더라도 각 문서에 등장하는 단어의 종류나 빈도는 다를 수 있기 때문에 단순한 키워드 기반의 모델로는 유사도를 계산하거나 주제를 분류하는 데에 한계가 있다. 그러나 이미 보유한 많은 텍스트에 기초하여 α 와 β 를 알고 개별 문서의 θ 를 계산할 수 있다면, 이 θ 를 갖고 유사도 계산이나 분류 작업을 훨씬 쉽고도 정확하게 해낼 수 있다.

2) 감성분석

감성분석은 형태소 분석, 감성사전 구축, 감성 판별의 과정을 통해 이루어진다. 형태소란 더 이상 분석하면 뜻을 잃어버리는 말의 최소 단위이다. 한국어는 문법 형태소가 어휘 형태소와 결합하여 문장 성분을 구성하고, 이것이 문장 내에서 단위 의미를 하는 특성이 있다(심광섭, 2016). 형태소 분석이란 어절을 형태소 단위로 분리하고 각 형태소에 적절한 품사 태그를 부여하는 과정이다(이재성, 2011). 과거에는 사람의 손으로 만든 사전을 이용하여 절차적으로 형태소 분석을 하는 방안들이 주로 제안되었다. 그런데 최근에는 품사 부착 말뭉치로부터 획득한 정보를 이용하는 확률 모델기반의 한국어 형태소 분석 방법론들이 제시되고 있다(심광섭, 2014). 형태소 분석 내용은 예를 들어, '아파트/nng(명사)가격/nng(명사)이/ec(어미)상승/nng(명사)하/xsv(미사)다/ec(어미)'로 하게 된다(경정익·이국철, 2016).

감성사전은 긍정 또는 부정과 같은 감성을 포함한 형태소들을 모아 놓은 집합체다. 분석대상이 되는 문장의 긍정 및 부정을 판별하기 위해서 형태소 분석이 선행되어야 하며 감성을 가지는 형태소를 제외한 나머지 형태소들을 삭제하는 문장 처리단계가 필요하다. 감성사전은 이 단계에서 문장을 구성하는 각 형태소의 삭제 및 유지 여부를 판단하기 위한 도구로 사용된다. 따라서 감성 분석을 위해서는 감성사전의 구축이 필요하다.

감성사전을 구축하기 위해서는 분석 문장들 내에 존재하는 감성을 포함하는 형태소들을 1차적으로 수집하게 된다. 다음으로 감성을 표현하는데, 보편적으로 사용되는 형태소만을 남기는 필터링 과정이 뒤따르게 된다. 모든 과정을 거친 후에 남아있는 형태소들이 최종적으로 감성 사전에 등재되게 된다(홍두표 외, 2017).

감성분석을 수행하기 위해서는 감성사전에 등록된 형태소별 긍정 혹은 부정을 나타내는 가중치 정보가 필요하다. 감성 판별은 가중치 산출 방법에 따라 크게 범용적인 의미에 따라 가중치를 부여하는 방법과 주제별 가중치를 계산하는 방법으로 나눌 수 있다.

범용적인 의미에 따라 가중치를 부여하는 방법은 부정을 의미하는 단어에는 -1, 긍정을 의미하는 단어에는 +1 값을 분석가가 주관적으로 부여한 후 가중치의 합이 0보다 클 경우에는 긍정, 작을 경우에는 부정, 같을 경우에는 중립으로 판단하며, 주제별 가중치를 계산하는 방법은 특정 단어가 긍정문장에 포함되어 있을 경우에는 +1, 반대로 특정 단어가 부정문장에 포함되어 있을 경우에는 -1로 각 단어에 가중치를 부여한다(홍두표 외, 2017).

3) 교차상관분석

교차상관 검정(cross-correlation test)은 두 시계열 사이에 존재하는 발생 시간의 시차관계를 이용하여 두 시계열 사이에 존재하는 동행성, 선행성 또는 후행성의 관계를 규명하기 위해 활용되는 통계적 검정 기법이다. 교차상관분석은 두 시계열의 교차상관함수(cross-correlation function)에 기초한다. 이는 두 시계

열 사이에서 발생 시간의 시차에 따라 변화하는 상관관계의 정도 및 방향을 측정하는 척도로 활용되며 다음과 같이 계산된다(한국 건설산업연구원, 2004).

동행성이란 두 시계열 사이의 통계적 유의미한 값이 시차가 0인 곳에서 발생한다는 의미로 두 시계열 사이에 시차가 없이 동시에 움직인다는 의미이며, 선행성은 두 시계열 사이의 통계적 유의미한 값이 시차가 마이너스(-)인 곳에서 발생한다는 의미로 두 시계열 중 독립변수로 사용된 시계열의 움직임이 종속변수로 사용된 시계열의 움직임 보다 먼저 발생했다는 의미이고, 후행성은 두 시계열 사이의 통계적 유의미한 값이 시차가 플러스(+)-인 곳에서 발생한다는 의미로 두 시계열 중 독립변수로 사용된 시계열의 움직임이 종속변수로 사용된 시계열의 움직임 보다 나중에 발생했다는 의미이다.

두 시계열 $\{x_t\}$ 와 $\{y_t\}$ 가 결합 정상적이면 $\{x_t\}$ 와 $\{y_t\}$ 사이에 교차공분산함수는 <equation 2>와 같이 정의된다.

$$Cov_{xy}(k) = E[(x_{t+k} - \mu_x)(y_t - \mu_y)], \quad (2)$$

$$k = 0, \pm 1, \pm 2, \dots$$

여기에서 μ_x 와 μ_y 는 각각 시계열 $\{x_t\}$ 와 $\{y_t\}$ 의 모집단 평균이다. 두 시계열 $\{x_t\}$ 와 $\{y_t\}$ 의 교차상관함수는 <equation 3>과 같은 공식에 의해 정의된다.

$$Cov_{xy}(k) = \frac{Cov_{xy}(k)}{\sqrt{Cov_{xx}(0)}\sqrt{Cov_{yy}(0)}}, \quad (3)$$

$$k = 0, \pm 1, \pm 2, \dots$$

4) 그랜저 인과관계 분석

Granger 인과관계는 쉽게 표현하면 어떤 변수 Y 를 다른 변수 X 를 제외한 모든 정보로써 예측하는 경우보다 추가적으로 변수 X 에 관한 정보를 이용할 때 예측력이 향상되는 경우, X 가 Y 의 원인이라고 정의할 수 있다. Y 에 대해 X 를 그 원인 또는 결과로 생각되는 다른 변수라 할 때, <equation 4>와 <equation 5>를 추정하여 그 계수 값에 대해서 검증하는 것이다.

$$G(Y_t) = b_0 + \sum_{j=1}^k a_j G(X_{t-j}) + \sum_{j=1}^l b_j G(Y_{t-j}) + U_t \quad (4)$$

$$G(X_t) = c_0 + \sum_{j=1}^m c_j G(X_{t-j}) + \sum_{j=1}^n d_j G(Y_{t-j}) + V_t \quad (5)$$

여기서 $G(X_t)$ 는 X 의 t 기 변동률, $G(Y_t)$ 는 Y 의 t 기 변동률이며, U_t 와 V_t 는 시계열간에 상관관계가 없는 잔차항(residual)을 나타낸다.

만약 식(4)에서 모든 a_i 값이 0이라는 가설이 기각되지 않으면 X 의 변화가 X 변화의 원인이라고 할 수 있으며, 반대로 식(5)에서 모든 d_i 값이 0이라는 가설이 기각되지 않으면 X 의 변화가 X 의 변화의 원인이라고 할 수 있다. 두 가설이 모두 기각되지 않는다면, 이 때는 X 와 X 의 변화 상호간에 영향을 주고받는다라고 결론을 내린다.

특히, 분석결과를 해석하는데 주의를 기울여야 한다. Granger 인과관계의 의미로 변수 X 가 Y 의 원인이라고 판정되더라도 이는 X 가 Y 와 어떤 일정한 관계를 가지며 선행하므로 Y 의 예측에 있어 X 의 자료가 도움이 된다는 의미일 뿐, X 가 Y 의 충분조건이라거나 X 를 조작함으로써 Y 에 관련된 일정 목표를 달성할 수 있거나 하는 의미를 갖는 것은 아니기 때문이다(손재영, 1991).

IV. 실증분석

1. 토픽 및 감성분석 결과

1) 토픽분석 결과

‘아파트’를 키워드로 웹문서를 크롤링을 이용하여 분석기간에 해당하는 신문기사를 수집하였다. 수집된 신문기사의 내용을 개별적으로 모두 확인 하는 것은 물리적으로 불가능한 작업이다. 이를 극복하기 위해 토픽분석을 이용하여 신문기사가 포함하고 있는 단어를 비슷한 주제별로 분류하였다. 본 연구에서는 8개의 토픽으로 구분하였다. 또한 형태소 분석을 위하여 세종 패키지(Sejong Package)를 이용하여 데이터에서 명사를 추출하였다. 글자 수는 2자에서 4자까지로 지정하였다. 글자 수를 2자에서 4자로 지정 한 이유는 신문기사에서 단어를 축약해서 쓰는 경향이 일반적으로 나타나기 때문에 글자 수를 4자까지만 지정하여도 분석에 어려움이 없다. 또한 숫자나 문장기호, 영문자 등은 불용어 처리하였다.

토픽분석 결과를 정리하면 <Table 1>과 같다. 각 토픽의 내용을 구성하는 단어들을 분석 및 종합하여 개별 토픽의 주제를 정하였다. Topic 1은 아파트 매매, 전세 그리고 주로 가격의 방향성을 알려주는 단어들로 구성되어 있었다. Topic 2는 재개발, 재건축이 진행되고 있는 지역명 단어들이, Topic 3은 분양, Topic 4는 상권, Topic 5는 아파트 구조, Topic 6은 투자, Topic 7은 청약, Topic 8은 정확히 규정하기 어려운 기타 단어들이었다.

이 중에서 아파트 매매가격의 변동과 관련성이 높은 단어를 포함한 Topic 1과 Topic 2에 감성사전에 포함할 단어를 선택하였다. 포함된 단어는 총 21개로 가격, 집값, 대출, 거래, 매매, 정부, 규제, 대책, 완화, 세금, 경매, 낙찰, 전세값, 수요, 투자, 매물, 전망, 완화, 중소, 재개발, 재건축이 해당된다. 선택된 단어들은 신문기사에서 부동산 시장의 현재 상황을 설명하거나 향후 전망을 언급하는 단어들이 거의 포함되어 있다고 할 수 있다.

2) 감성분석 결과

토픽분석을 통해 선정된 주요 단어들을 이용하여 감성사전을 구축하고 감성지수 산출을 위해 주요 단어별 긍정 및 부정 표현을 구체화 하였다. 감성사전의 구축을 위해서 아파트 매매가격의 변화와 관련이 깊은 Topic 1과 재건축, 재개발이 들어간 Topic 2에서 총 21개의 단어를 선정하여 활용하였다. 선택된 명사 단어가 표현할 수 있는 긍정 표현과 부정 표현을 정의하여 <Table 2>와 같이 감성사전을 구축하였다.

각 단어에 해당하는 긍정 표현과 부정 표현은 예시에서 제시한 것 보다 많이 수집하여 분석에 적용하였다. 예를 들면, 아파트 가격 또는 집값이 상승하거나 하락하는 것을 설명하는 신문기사의 표현방식은 다양하다. 집값 또는 아파트 가격이 상승하는 경우, ‘상승하다’라는 표현 이외에 ‘오르다’, ‘올랐다’, ‘오르고 있다’, ‘급등하다’, ‘올라갈 것으로 예상된다’ 등으로 표현될 수 있다. 또한 집값 또는 아파트 가격이 하락하는 경우, ‘하락하다’라는 표현 이외에 ‘내렸다’, ‘내리고 있다’, ‘급락했다’, ‘하락할 것이다’ 등의 여러 가지 형태로 표현될 수 있다.

감성사전을 <Table 2>와 같이 정의한 후 2010년 1월부터 월단위로 수집된 신문기사 대해 감성사전에 언급된 단어들의 긍정 표현과 부정 표현들의 개수를 수집하여 긍정 감성지수와 부정 감성지수를 만들었다. 긍정 단어개수와 부정 단어개수의 2010년 1월 개수를 100으로 하여 시계열 기준 대비 움직임을 감성지수로 정의하였다. 이렇게 함으로써 아파트 가격이 긍정감성지수와 부정 감성지수에 대해 각각 어떻게 반응하는지 파악할 수 있기 때문이다.

<Fig. 4>와 <Fig. 5>는 감성지수에 대한 기초 통계량과 시계열 그래프를 나타낸 것이다. 2010년부터 2016년까지는 비슷한 움직임을 보였다. 즉 긍정 단어 개수와 부정 단어 개수의 비율의 증감이 같은 방향성을 나타냈다. 그러나 아파트 가격의 상승세가 굳어지기 시작한 2017년도 들어와서 긍정단어 개수 증가는 가속화된 반면 부정 단어 개수는 반대로 감소하고 있다. 본 연구에서는 긍정 감성지수와 부정 감성지수 중에서 아파트 매매가격과 흐름을 같이 하는 긍정 감성지수를 변수로 선정하여 통계분석을 진행하였다. 아파트 매매가격과 같은 흐름을 나타내는 변수를 모형에 투입해야 같은 흐름을 나타내는 두 변수 사이의 통계적 유의성을 관찰할 수 있기 때문이다.

2. 아파트 매매가격지수 및 단위근 검정

1) 아파트 매매가격지수

서울의 5개 권역(도심권, 서북권, 동북권, 서남권, 동남권)별 아파트의 월매매가격은 당 거래금액의 월별 총합을 거래건수로 나누어 단순 평균가격으로 계산하였다. 지역별 아파트가격의 수준이 상이함을 보정하기 위해 단순평균 월별매매가격을 2010년

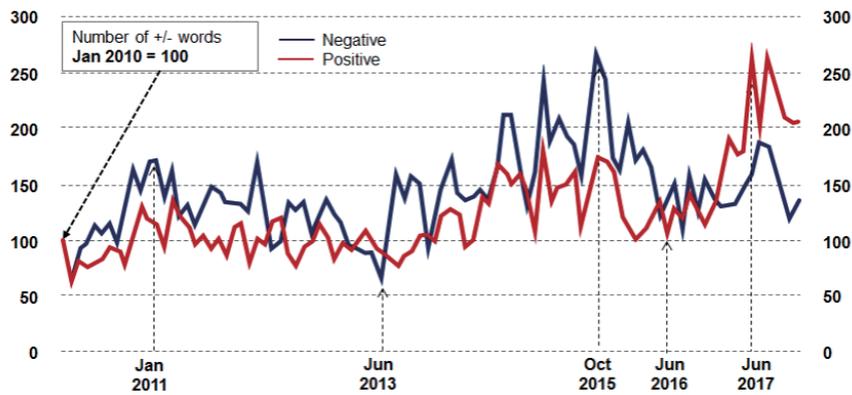
Table 1. Result of Topic Analysis

Topic1 (Sale)	Topic2 (Area)	Topic3 (Parcel Out)	Topic4 (Commercial District)	Topic5 (Structure)	Topic6 (Invest)	Topic7 (Tender)	Topic8 (Others)
housing	redevelop*	parcel out	store	drawing	deposit	tender	housing
price*	Seocho-gu	lot	profit	system	lend license	offer	company
jeonse	Yeongdeungpo-gu	park	invest	family	financing	house	develop
loan*	Yeoksam-dong	near	officetel	energy	house	lot	service
transaction*	Songpa-gu	size	building	living room	quick sale	quantity	telecommunication
rise	Mapo-gu	develop	ROI	community	studio	rank	retail
10 billion Won	Seocho-dong	form	commercial district	design	near	competition rate	Korea
sale*	Guro-gu	area	shop	area	Sejong City	block	customer
fall	Geumcheon-gu	above ground	subway station zone	kitchen	Cheonan	average	Guro-dong
auction*	consulting	New Town	Gangnam	plane	Chungnam	area	construction
government*	Gangnam	neighborhood	company office	structure	owner	contract	domestic
interest rate*	partnership	underground	commerce	housing	monthly rent	un-parceled out	group
demand*	Yeouido-dong	demand	company office	household	family	business cycle	Gasam-dong
monthly rent	redevelop*	complex	commerce fee	underground	walk	win	City of Seoul
investment*	City of Seoul	around	liquid	life	rent fee	housing lot	financing
solution*	Yongsan-gu	life	investor	storage	hospital	close	land
housing price*	Seongdong-gu	housing	demand	in door	profit	receive	invest
tax*	Jongno-gu	Gangnam	hotel	balcony	new city	demand	house
regulation*	partner	highway	item on sale	parking lot	Daejeon	volume	evaluation
earning	Gangdong-gu	environment	business cycle	type	market price	new	operation
win bid*	New Town	education	behind	inside	adjacent	gather	government
jeonse price*	Seodaemun-gu	view	broker	eco-friendly	ground	advance sale	insurance
collateral	Seongbuk-gu	town	lease	landscaping	land	new town	solution
item on sale*	propel	officetel	shop	model	fund	Innovation city	existing
outlook*	Dongjak-gu	location	fund	expand	population	Wilye New City	partial
lessen*	Nowon-gu	walk	parking lot	heating	recommend	Dongtan	startup
small/medium*	brokerage	area	rent fee	terrace	invest	Prugio	cost
recession	Samseong-dong	culture	area	daylighting	responsibility	Raemian	modern
for sale	Gwanak-gu	inquiry	business type	house	Kyungbuk	interest	item
fund	Seungsu-dong	condition	parcel out	environment	local	small/medium	work

* Words that are included in the sentiment dictionary because they are relevant to the fluctuation of the apartment sale prices

Table 2. Example of Sentiment Dictionary

Word	Positive Expression	Negative Expression
housing price	go up etc.	go down etc.
sales	go up etc.	go down etc.
trade volume	go up, increase etc.	go down, decrease etc.
interest rate	go up, go higher, increase etc.	go down, decrease etc.
loan	increase etc.	decrease etc.
investment	active etc.	inactive etc.
regulation	loose etc.	tighten etc.
tax	increase, go up etc.	decrease, go down etc.
auction volume	increase etc.	decrease etc.
auction price	go up, jump up etc.	go down, drop etc.
forecast	bright, good etc.	dark, bad etc.



	Mean	Std. dev.	Min.	Max.	Median
Positive Sentiment Index	124.07	41.12	63.50	267.00	113.70
Negative Sentiment Index	143.40	37.85	64.70	266.50	137.10

Fig. 4. Sentiment Index Statistic (2010.1.~2017.12.)

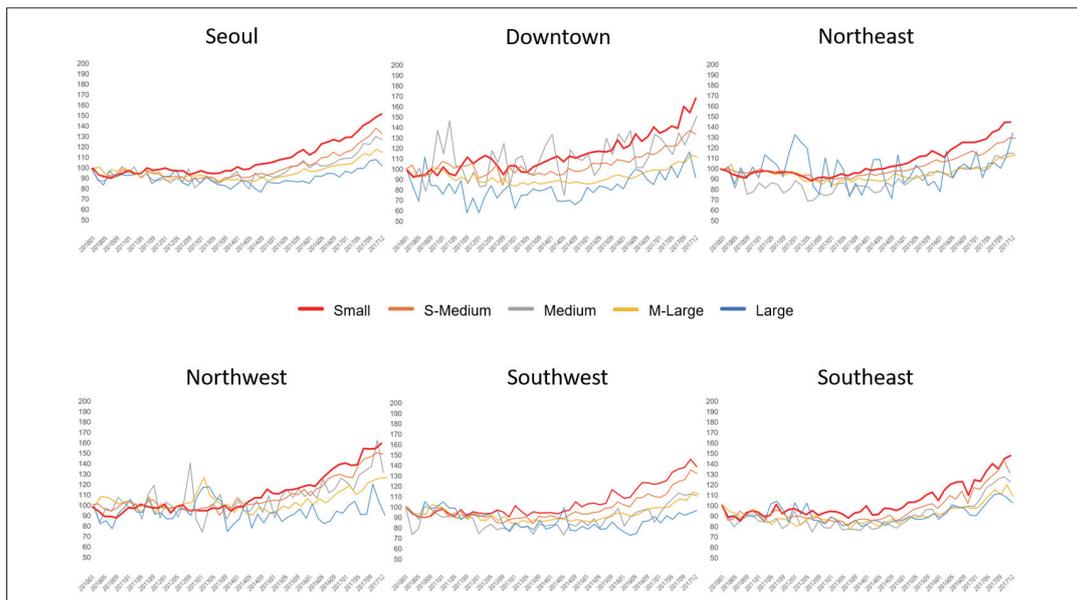


Fig. 5. A Time Series of Apartment House Sale Prices in Seoul(2010. 1~2017. 12.)

1월을 기준(100)으로 하여 매매가격지수로 변경하여 분석에 적용했다.

〈Table 3〉에서 서울지역 아파트 매매가격지수는 서울 전체와 권역별로 중위 값이 평균값보다 낮게 나타났다. 매매 가격지수의 중위 값을 기준으로 도심권과 서북권의 소형부터 중형 아파트 매매가격지수가 100 이상으로 다른 권역의 가격지수보다 높게 나타났다. 아파트의 절대가격은 동남권이 가장 높지만, 도심권과 서북권의 소형, 중소형 및 중형 아파트의 매매가 상승률이 높게 나타나고 있다.

분석기간인 2010년부터 2017년 말까지 서울시의 중형, 중대형과 대형 아파트의 매매가격지수는 100보다 낮고 중형에서 대형으

로 규모가 커질수록 매매가격지수 평균은 감소하는 경향이 있다. 이러한 경향은 권역별로도 나타나는데, 서남권과 동남권은 중소형부터 대형 아파트의 매매가격지수가 100 미만으로 나타나고 있다. 도심권은 중대형과 대형 아파트, 동북권은 중형과 중대형 아파트, 서북권은 대형 아파트의 매매가격지수가 100 미만으로 나타났다.

2) 단위근 검정 결과

단위근 검정은 시계열 자료의 안정성에 관한 검정방법으로 시계열이 불안정하면 변수 내에 단위근이 존재하게 된다. 단위근이 존재하게 되면 시장에 무작위적 충격이 올 경우 그 충격이 미래의

Table 3. Descriptive Statistics of Seoul Apartment House Sales Price Index

Region	APT size	Mean	Median	sd	Minimum	Maximum
Seoul	Small	106.95	100.35	14.08	91.10	148.20
	Small-medium	100.23	95.95	10.75	87.60	135.20
	Medium	97.91	96.00	10.58	81.60	128.60
	Medium-large	95.81	94.00	7.08	85.50	117.30
	Large	91.75	90.70	6.63	79.00	118.60
Central	Small	113.90	110.50	15.66	89.10	163.60
	Small-medium	103.88	102.70	9.45	85.60	133.40
	Medium	111.11	110.95	18.14	63.40	151.10
	Medium-large	97.71	97.10	8.67	82.00	117.60
	Large	86.09	84.60	14.43	61.90	167.20
Northeast	Small	105.29	99.10	13.27	89.80	141.50
	Small-medium	100.92	97.90	9.34	89.20	126.50
	Medium	91.04	88.80	10.41	68.30	115.20
	Medium-large	95.95	95.00	6.09	84.40	113.40
	Large	100.02	100.00	13.34	74.70	143.30
Northwest	Small	110.80	102.40	16.88	90.30	155.40
	Small-medium	106.50	100.90	13.43	90.80	144.50
	Medium	108.54	106.45	18.01	77.00	186.70
	Medium-large	104.90	103.00	9.02	88.40	130.90
	Large	95.83	94.70	10.95	65.30	120.10
Southwest	Small	104.30	99.00	12.96	91.00	142.40
	Small-medium	97.86	93.75	10.24	84.90	131.80
	Medium	91.52	90.90	8.91	72.90	114.40
	Medium-large	92.56	90.90	6.49	84.00	112.80
	Large	88.35	87.60	7.41	73.70	106.50
Southeast	Small	102.70	97.50	13.65	86.10	144.20
	Small-medium	95.62	90.80	11.98	81.20	137.90
	Medium	92.01	88.40	12.00	73.80	127.40
	Medium-large	91.36	88.30	8.60	78.70	119.30
	Large	92.28	90.50	7.49	78.30	111.40

수치에 연속적으로 영향을 주게 된다. 또한 시계열이 안정적인 추세궤도에서 벗어나게 되어 가성회귀(spurious regression)의 문제가 생기고 이는 추정의 신뢰성을 떨어뜨리게 된다(Woo, 2005).

서울시 5개 권역과 아파트 규모별 매매가격지수와 감성지수의 단위근 검정 결과는 <Table 4>와 같다. 아파트 매매가격지수에 대한 1% 유의수준의 단위근 검정 결과, 도심권의 중형 아파트를 제외한 모든 분석대상에서 귀무가설을 기각하지 못하였다.

따라서 수준변수를 1차 차분하여 추가 단위근 검정을 실시한 결과, 유의수준 1%에서 서울시 및 모든 권역에서 규모별 아파트 매매가격지수의 단위근이 존재하지 않음을 확인하였다. 긍정 감성지수에 대해서도 같은 절차에 따라 단위근 검정을 실시하였다. 분석 결과, 수준변수에서 단위근이 존재하는 것으로 나타났고, 1차 차분변수에서는 단위근이 존재하는 않음을 확인하였다.

Table 4. Result of Root Unit Test

Region	APT size	Level Variables		Difference Variables	
		Dickey-Fuller	p-value	Dickey-Fuller	p-value
Seoul	Small	1.1407	0.9900	-5.2180	0.01***
	Small-medium	1.3357	0.9900	-4.7371	0.01***
	Medium	-0.8261	0.9563	-5.932	0.01***
	Medium-large	0.7152	0.9900	-5.0857	0.01***
	Large	-0.7977	0.9588	-8.1751	0.01***
Central	Small	-0.4466	0.9825	-5.9356	0.01***
	Small-medium	0.7793	0.9900	-5.6686	0.01***
	Medium	-4.2095	0.01***	-5.7023	0.01***
	Medium-large	-0.5716	0.9770	-6.0244	0.01***
Northeast	Large	-2.3118	0.4476	-7.3330	0.01***
	Small	-0.2104	0.9900	-5.2559	0.01***
	Small-medium	-0.0937	0.9900	-5.2435	0.01***
	Medium	-2.0370	0.5611	-6.9789	0.01***
Northwest	Medium-large	-0.0261	0.9900	-5.1129	0.01***
	Large	-3.7742	0.0232	-5.7725	0.01***
	Small	0.4340	0.9900	-5.6393	0.01***
	Small-medium	0.0457	0.9900	-5.1794	0.01***
Southwest	Medium	-0.7409	0.9639	-8.1963	0.01***
	Medium-large	-1.4540	0.8020	-5.4669	0.01***
	Large	-2.9101	0.2004	-6.5394	0.01***
	Small	-0.0909	0.9900	-6.0840	0.01***
Southeast	Small-medium	0.5147	0.9900	-5.5241	0.01***
	Medium	-0.7266	0.9652	-6.0092	0.01***
	Medium-large	0.3237	0.9900	-5.7774	0.01***
	Large	-1.3216	0.8568	-7.4486	0.01***
positive sentiment index	Small	-0.3749	0.9856	-5.4569	0.01***
	Small-medium	1.2293	0.9900	-6.1992	0.01***
	Medium	0.7285	0.9900	-6.5404	0.01***
	Medium-large	-1.1905	0.9041	-6.1369	0.01***
	Large	-1.3791	0.8330	-6.3599	0.01***
		-1.9840	0.5831	-5.4801	0.01***

- confidence level: * 0.10 lower, ** 0.05 lower, *** 0.01 lower

3. 교차상관 및 인과관계 분석 결과

1) 교차상관 분석 결과

2010년 1월부터 2017년 말까지 서울시 및 5개 권역별 아파트 실거래 매매가격지수와 주요 온라인 신문기사를 이용해 산출한 긍정감성지수의 교차상관관계 분석 결과는 <Table 5>와 같다.

우선, 이 기간 서울시는 오직 소형아파트 매매가격지수만 긍정감성지수와 유의미한 교차상관관계를 나타냈다. 긍정감성지수가 1개월의 시차를 두고 소형아파트 매매가격지수에 선행성을 보이는 것으로 분석된다. 서울시는 아파트 매매와 관련된 온라인 뉴스기사의 긍정감성지수가 소형아파트 매매가격과 상당히 빠른 선행성 교차상관관계를 보이는 것으로 해석된다.

둘째, 서울시 5개 권역과 주택규모를 고려한 세부 분석 결과, 총 25개 유형 중 13개 유형에서 유의미한 교차상관관계가 나타났다. 이 중 8개 유형에서 긍정감성지수가 아파트 매매가격지수에 선행성을 보이는 것으로 분석되었다. 특히 긍정감성지수가 유의미한 선행성을 나타내는 8개 유형은 모두 3개월 이하의 시차를 보이고 있다. 이는 긍정감성지수와 아파트 매매가격이 유의미한 상관관계를 나타내면 긍정감성지수는 비교적 빨리 아파트 매매가격에 영향을 미치고 있음을 시사한다.

셋째, 서울시 권역별 분석 결과를 살펴보면, 모든 권역에서 긍정감성지수는 소형아파트의 매매가격과 유의미한 교차상관관계를 보이는 것으로 나타났다. 도심권, 서북권, 서남권은 온라인 뉴스기사의 긍정감성지수가 소형아파트의 매매가격과 1개월의 시차를 두고 선행성을 보이는 것으로 분석되었다. 동남권은 2개월, 그리고 동북권은 3개월의 시차를 두고 긍정감성지수가 소형아파트의 매매가격에 선행성을 보이는 것으로 나타났다. 서북권

과 서남권은 온라인 뉴스기사의 긍정감성지수가 소형아파트 매매가격에 1개월의 시차를 두고 선행성을 나타냄과 동시에 이보다 긴 시차를 두고 긍정감성지수가 소형아파트 매매가격에 후행성도 보이고 있다.

넷째, 동북권, 서북권과 서남권에서는 온라인 뉴스기사의 긍정감성지수가 소형아파트 이외의 아파트의 매매가격에 선행성을 보이는 것으로 분석되었다. 동북권은 긍정감성지수가 중형아파트의 매매가격에 2개월의 시차를 두고 선행성을 나타냈다. 서북권은 긍정감성지수가 대형아파트의 매매가격에 3개월의 시차를 두고 선행성을 보였다. 서남권은 긍정감성지수가 중소형아파트의 매매가격에 1개월의 시차를 두고 선행성을 나타냈다. 소형이외의 다른 규모의 아파트 유형에서도 긍정감성지수가 아파트 매매가격에 선행성을 보이는 것은 아파트 규모별 재고 및 신규 공급, 거래건수, 주택수요 특성 등 권역별 주택시장의 특성과 관련이 있다.

분석 결과를 종합하면, 서울시와 모든 권역에서 공통적으로 온라인 뉴스기사의 긍정감성지수는 소형아파트의 매매가격에 유의미한 교차상관관계와 선행성을 보인다. 긍정감성지수는 3개월 이내의 선행성을 보이는 특징이 있다. 또한 권역별 주택시장의 특성에 따라 다른 규모의 아파트 유형에서도 일부 유의미한 교차상관관계와 선행성이 나타났다.

2) 인과관계 분석 결과

2010년 1월부터 2017년 12월까지 서울시와 5개 권역별 아파트 실거래 매매가격지수와 주요 온라인 신문기사에 대한 긍정감성지수의 그랜저 인과관계를 분석한 결과를 정리하면, <Table 6>과 같다.

Table 5. Result of Cross Correlation Analysis

Region	leading/lagging	APT size				
		Small	Small-medium	Medium	Medium-large	Large
Seoul	leading	lag 1 (0.215)	-	-	-	-
	lagging	-	-	-	-	-
Central	leading	lag 1 (-0.230)	-	-	-	-
	lagging	-	-	-	-	lag 12 (-0.234)
Northeast	leading	lag 3 (0.205)	-	lag 2 (-0.272)	-	-
	lagging	-	-	lag 11 (0.215)	-	-
Northwest	leading	lag 1 (0.304)	-	-	-	lag 3 (0.275)
	lagging	lag 3 (-0.209)	-	lag 2 (0.244)	-	-
Southwest	leading	lag 1 (0.330)	lag 1 (0.204)	-	-	-
	lagging	lag10 (0.209)	-	-	-	-
Southeast	leading	lag 2 (-0.264)	-	-	-	-
	lagging	-	-	-	-	-

-(ACF value) reported by ccf function in R

Table 6. Result of Granger Causality Test

Region	GrangerTest	APT size				
		Small	Small-medium	Medium	Medium-large	Large
Seoul	p-sent-index → price	lag 1 * (3.8259 0.0535)	-	lag 1 * (3.0288 0.0851)	-	-
	price → p-sent-index	-	-	lag 2 * (2.6418 0.0768)	시차 1 ** (4.2846 0.0412)	-
Central	p-sent-index → price	-	-	-	-	-
	price → p-sent-index	-	-	-	-	-
North east	p-sent-index → price	lag 4 *** (5.0337 0.0011)	lag 1 ** (3.9572 0.0496)	-	-	-
	price → p-sent-index	-	-	-	-	-
North west	p-sent-index → price	lag 2 *** (6.9991 0.0015)	-	lag 5 ** (3.0505 0.0143)	-	lag 3 ** (3.0546 0.0327)
	price → p-sent-index	-	-	-	-	-
South west	p-sent-index → price	lag 3 ** (2.6691 0.0527)	lag 1 ** (4.6584 0.0335)	lag 1 ** (5.7137 0.0188)	-	-
	price → p-sent-index	lag 4 ** (2.7988 0.0311)	-	-	-	-
South east	p-sent-index → price	lag 3 ** (3.7713 0.0135)	-	-	-	lag 2 ** (4.2532 0.0172)
	price → p-sent-index	lag 2 ** (3.7466 0.0274)	-	-	lag 1 * (3.0868 0.0822)	-

- confidence level: * 0.10 lower, ** 0.05 lower, *** 0.01 lower (F-value, P-value)
 - p-sent-index : positive sentiment index

서울시 전체를 살펴보면, 온라인 신문기사와 관련된 긍정 감성 지수는 소형아파트의 매매가격지수에 영향을 미치는 유의미한 인과관계를 나타냈다. 시차는 1개월로 긍정 감성지수는 소형아파트 수요자와 투자자의 심리에 영향을 주고 이는 상당히 빨리 매매 가격에 영향을 미치는 것으로 해석된다. 이는 앞선 교차상관분석의 결과와 일관성을 보이고 있다. 이와 함께 중형 아파트와 중대형 아파트에서도 유의미한 인과관계가 나타났다. 서울시의 중형 아파트에서는 온라인 뉴스기사의 긍정 감성지수가 매매가격에 영향을 미치고, 매매가격은 다시 긍정 감성지수의 형성에 영향을 미치는 양방향 인과관계를 보이고 있다. 중대형 아파트에서는 오히려 매매가격이 온라인 뉴스기사의 긍정 감성지수에 영향을 미치는 인과관계가 나타났다. 서울시 아파트는 규모에 따라 긍정 감성지수와 매매가격의 영향 관계는 다르게 나타나고 있음을 알 수 있다. 소형 아파트의 매매가격은 긍정 감성지수에 비교적 민감한 반면, 중형 이상 아파트로 규모가 증가할수록 주택매매가격이 변동하면 이와 관련한 뉴스기사가 전개되는 방향으로 전환되는 것으로 해석된다. 아파트 규모에 따라 상호 영향관계가 다르게 나타나는 것으로 아파트의 규모별 특성과 해당 아파트의 수요자 및 투자자의 특성에 기인한 것으로 판단된다.

서울시 5개 권역과 주택규모를 고려한 세부 분석 결과, 총 25개 유형 중 13개 유형에서 인과관계가 유의미한 것으로 분석되었다.

이 중 10개 유형에서 긍정 감성지수가 아파트 매매가격에 영향을 미치는 관계를 보이고, 3개 유형에서는 매매가격이 긍정 감성지수에 영향을 미치는 관계를 보이고 있다.

서울시 권역과 주택규모를 고려한 분석 결과를 정리하면 다음과 같다. 첫째, 도심권에서는 어떤 아파트 규모에서도 긍정 감성지수와 매매가격 사이에 유의미한 인과관계가 나타나지 않았다. 이는 종로구, 중구, 용산구를 포함하는 도심권이 갖고 있는 특성에 기인한 것으로 판단된다. 도심권은 서울시 전체 아파트 재고의 4.1%를 차지하여 주택재고가 다른 권역에 비해 현저히 적다. 도심권의 사대문안은 서울의 전통적인 중심지로 상업지역으로 지정되어 있고, 권역 내에 역사건축물 등 보존지역이 많아 아파트 등 주택개발이 제한되기 때문이다. 따라서 아파트 매매거래가 활발하지 않아 인과관계가 분명치 않을 수 있다.

둘째, 도심권을 제외하고 서울시 권역에서 공통적으로 나타나는 사실은 전용면적 60m² 이하 소형 아파트에서만 온라인 뉴스기사의 긍정 감성지수가 매매가격에 유의미한 영향을 미치고 있다는 점이다. 이 사실은 소형 아파트의 특성과 해당 아파트의 수요자 및 투자자의 특성에 기인한 것으로 판단된다. 소형 아파트를 포함한 소형주택은 중대형 주택과 비교하여 단위면적당 매매가는 높지만, 호당 매매가격은 상대적으로 저렴하다. 따라서 거래가 비교적 용이하며, 이 연구의 시간적 범위에서 가장 많이 거래

가 된 유형에 해당한다.

소형 아파트의 수요자 특성도 긍정 감성지수가 매매가격에 미치는 영향 관계에 중요한 역할을 하는 것으로 판단된다. 소형 주택의 주요 수요계층은 1·2인 가구이면서 주택을 소유하지 않을 가능성이 높은 20-30대 계층이다. 2017년 현재, 서울시의 1·2인 가구 비율은 약 56.1%로 가장 높은 비율을 차지하며, 도심권을 제외한 모든 권역에서 1·2인 가구의 비율은 가장 높다. 이들은 주로 학업과 직장 등의 이유로 서울로 거주하며, 온라인 검색과 소셜 미디어 등의 이용이 활발하여 이에 비교적 민감하게 반응하는 특성이 있다. 소형 아파트의 투자자 또한 잠재 수요계층인 젊은 1·2 가구의 주택수요 및 선호에 민감하다. 따라서 투자자들도 온라인 뉴스 등을 통해 정보를 얻고 주택시장의 변화에 대응하는 경향이 있다. 이처럼 소형 아파트의 특성과 수요자 및 투자자의 특성이 온라인 뉴스기사의 긍정 감성지수가 소형 아파트 매매가격에 영향을 미치는 것으로 설명될 수 있다.

셋째, 소형 아파트 이외에도 권역별로 온라인 뉴스기사의 긍정 감성지수가 아파트 매매가격에 유의미한 영향을 미치는 인과관계가 분석되었다. 이는 아파트 재고 및 공급, 주택수요 특성 등 권역별 주택시장의 특성과 관련이 있는 것으로 판단된다. 동북권에서는 긍정 감성지수가 중소형 아파트의 매매가격에 1개월의 시차를 두고 유의미한 인과관계를 나타냈다. 서남권에서는 긍정 감성지수와 중소형 아파트뿐만 아니라 중형 아파트의 매매가격에 1개월의 시차를 두고 유의미한 영향을 미치는 것으로 분석되었다. 서북권에서는 중형과 대형 아파트의 매매가격에 온라인 뉴스의 긍정 감성지수가 영향을 미치고, 동남권에서는 긍정 감성지수가 대형 아파트의 매매가격에 영향을 미치는 것으로 나타났다.

V. 결론 및 시사점

부동산시장 거래자들이 정보를 얻는 경로는 다양하다. 빅데이터 시대에는 단순히 대용량의 데이터를 보유하는 것을 넘어서 어떻게 하면 그 안에서 가치 있는 정보를 정제해낼 수 있을지에 더 많은 관심이 집중되고 있다. 감성분석은 서로 다른 종류의 데이터를 마이닝 도구를 통하여 그 의미를 유추함으로써 콘텐츠에 내재하는 생각이나 감정을 재구성한다.

이 연구는 서울시를 대상으로 주요 온라인 신문기사에서 추출한 아파트 매매와 관련된 긍정 감성지수가 아파트 매매가격과 유의미한 영향 관계를 나타내는지 분석하는데 목적이 있다. 이를 위해 국토교통부가 제공하는 서울시 아파트 실거래 데이터를 2010년 1월부터 2017년 12월까지 96개월간 월 단위로 수집하였다. 또한 서울시를 5개 권역으로 구분하고 아파트도 5개 규모로 나누어 분석내용을 구체화하였다. 96개월의 감성지수를 산출하기 위해 3개 주요 일간지와 3개 주요 경제지의 모든 온라인 기사를 웹 크롤링하여 184,280개의 관련 기사를 추출하였다.

주요 분석 결과를 요약하면 다음과 같다. 첫째, 토픽분석 결과, 아파트 매매와 관련된 온라인 뉴스기사는 총 8개의 토픽으로 분석되었다. 이 중 아파트 매매가격과 관련된 단어들을 포함한 Topic 1(매매)과 Topic 2(지역)의 21개 단어들을 감성사전에 포함하였다. 이 단어들의 긍정 표현과 부정 표현들을 수집하여 감성지수를 산출한 결과, 2010년부터 2016년까지는 전반적으로 긍정 및 부정 감성지수가 유사한 패턴으로 증가하다가 2017년에 들어서는 긍정 감성지수만 크게 증가하는 패턴을 나타냈다.

둘째, 교차상관분석 결과, 서울시는 오직 소형아파트 매매가격만 긍정 감성지수와 1개월 시차의 유의미한 관계를 나타냈다. 서울시 5개 권역과 주택규모를 고려하면, 13개 유형에서 유의미한 교차상관관계가 나타났고, 이 중 8개 유형에서 긍정 감성지수가 아파트 매매가격에 선행성을 나타냈다. 서울시의 모든 권역에서 긍정 감성지수는 소형 아파트의 매매가격과 유의미한 교차상관관계를 보이는 것으로 분석되었다.

셋째, 온라인 뉴스기사의 긍정 감성지수와 아파트 매매가격 사이의 인과관계 분석 결과, 서울시는 긍정 감성지수가 소형아파트의 매매가격에 유의미한 영향을 미치는 인과관계가 나타났다. 1개월의 시차로 긍정 감성지수는 소유자와 투자자의 심리를 통해 소형아파트의 매매가격에 영향을 미치는 것으로 해석된다.

넷째, 권역과 주택규모를 고려하면, 13개 유형에서 유의미한 인과관계가 나타났고, 이 중 10개 유형에서 긍정 감성지수가 아파트 매매가격에 영향을 미치는 것으로 분석되었다. 도심권에서는 어떤 규모의 아파트에서도 이들 상호간의 유의미한 인과관계가 나타나지 않았다. 도심권을 제외한 서울시 모든 권역에서 소형 아파트에서만 온라인 뉴스기사의 긍정 감성지수가 매매가격에 유의미한 영향을 미치는 것으로 분석되었다. 또한 인과관계 분석에서 긍정 감성지수와 아파트매매가격 사이의 시차는 대부분 3개월 이내로 나타나 긍정 감성지수가 아파트 매매가격이 영향을 미치는 시간이 상당히 빠르다는 것을 알 수 있다.

이 연구는 최근에 활용가능성이 높아지고 있는 빅데이터를 이용하여 부동산시장을 설명 및 예측할 수 있는 가능성을 제시했다는 점에서 의미가 있다. 온라인 상에서 비정형 빅데이터를 수집하여 토픽분석을 적용함으로써 연구자의 주관적 개입가능성을 최대한 제거하고 객관적인 감성사전을 구축하는 방법론을 적용하였다. 감성사전을 활용하여 감성지수를 산출하는 시도도 방법론적 의미가 있다.

연구 결과, 아파트 관련 뉴스기사에 대한 긍정적인 표현의 증가와 아파트 매매가격 간에 유의미한 영향 관계가 존재함을 실증하였다. 이 결과는 신문기사에서 긍정표현이 늘어나는 추세가 나타나면 앞으로 아파트 매매가격이 상승하는 신호가 될 수 있다는 것을 의미한다. 이는 온라인 신문기사에 대한 지속적인 모니터링과 감성지수의 산출은 아파트 매매가격을 예측하는데 중요한 변수로 활용될 수 있음을 시사한다. 아파트 매매가격과 관련 변수

의 정보를 구득하고 이를 바탕으로 향후 매매가격의 추세를 지속적으로 예측하기는 매우 어렵다. 그러나 매일 발행되는 신문기사에서 실시간으로 매매 등 부동산과 관련된 정보를 수집하고 분석함으로써 미래 주택가격의 흐름을 예측하는데 효율적으로 활용할 수 있는 가능성이 있다.

이 연구는 다음과 같은 한계도 있다. 첫째, 분석기간을 2010년부터 2017년 말까지 한정함으로써 2008년 금융위기 이후 아파트 가격의 보합세와 상승기만을 대상으로 하고 있다는 것이다. 아파트 매매가격이 하락하는 시기에는 감성지수와 아파트 매매가격 사이의 관계가 어떤지 알 수 없어 종합적인 분석에 한계가 있다. 향후 분석기간을 연장하거나 주택가격의 하락기를 포함한 연구가 필요하며, 감성지수를 활용한 주택가격 예측모형의 개발을 위한 연구가 요구된다. 둘째, 아파트 가격에 영향을 미치는 경제지표를 분석 모델에 포함하는 연구가 필요하다. 이번 연구에서는 감성지수와 아파트 매매가격 사이의 관계를 중심으로 분석하는데 초점을 맞추었기 때문에 경제변수를 모형에 포함시키지는 않았다. 향후 심리지표인 감성지수와 경제변수를 통합하여 분석하는 방법론을 제시하는 연구가 요구된다. 셋째, 본 연구에서 감성지수를 도출하기 위해 선택된 신문사가 보수적 언론위주로 구성되어 있다는 것이다. 유료부수를 기준으로 언론사를 선정할 결과 우연히도 보수적 언론사가 다수 포함되는 결과가 나왔지만, 보수적 언론사와 진보적 언론사의 비율을 고려한 언론사 선정도 추후 연구에서 고려해야할 사항으로 여겨진다.

인용문헌 References

1. 감미아·송민, 2012. “텍스트 마이닝을 활용한 신문에 따른 내용 및 논조 차이점 분석”, 『지능정보연구』, 18(3): 53-77.
Kam, M.A. and Song, M., 2012 “A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis”, *Journal of Intelligence and Information Systems*, 18(3): 53-77.
2. 강범일·송민·조화순, 2013. “토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구”, 『한국문헌정보학회지』, 47(4): 315-334.
Kang, B.I., Song, M. and Jho, W.S., 2013. “A Study on Opinion Mining of Newspaper Texts based on Topic Modeling”, *Journal of the Korean Society for Library and Information Science*, 47(4): 315-334.
3. 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”, *Housing Studies Review*, 24(4): 115-136.
4. 권건우·진창하, 2016. “생애주기별 가구의 주거비 부담 결정요인에 관한 연구”, 『주택연구』, 24(3): 49-69.
Kwon, G.W. and Jin, C.H., 2016. “A Study on the Determinants of Housing Expenditure Burden Considering Family Life Cycle”, *Housing Studies Review*, 24(3): 49-69.
5. 김기홍·김재태, 2014. “아파트가격 유형별 상관관계분석을 통한 가격 간 선·후행 관계분석”, 『주거환경』, 12(2): 183-197.
Kim, K.H. and Kim, J.T., 2014. “A Relationship Analysis of the Leading or Lagging through the Correlation Analysis among Apartment Price Types”, *Journal of The Residential Environment Institute of Korea*, 12(2): 183-197.
6. 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 Chonseil Prices”, *Journal of the Korean Urban Management Association*, 29(1): 1-33.
7. 김정선·권은주·송태민, 2014. “소셜 빅데이터를 활용한 국내 빅데이터 수요공급 예측 : 데이터마이닝 의사결정나무 적용”, 한국경영정보학회 춘계공동학술대회, 천안시: 상록리조트.
Kim, J.S., Kwon, E.J. and Song, T.M., 2014, “Prediction for Demand of Big Data Using Social Big Data : Data Mining Decision Tree”, Paper presented at the annual meeting for The Korea Society of Management Information System Autumn Seminar, Cheonan: Sangnok resort.
8. 김진유, 2006. “신문기사가 부동산가격변동에 미치는 영향 - 투기가 포함된 신문기사와 주택가격간의 그랜저인과관계분석을 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. 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.
10. 박원준, 2012. “빅데이터 활용에 대한 기대와 우려”, 『방송통신전파저널』, 51(7): 28-47.
Park, W.J., 2012. “Expectation and concern about big data use”, *Journal of Communication & Radio Spectrum*, 51(7): 28-47.
11. 서울특별시, 2014. 『2030 서울플랜』, 서울.
Seoul Metropolitan Government, 2014. *2030 Seoul Plan*. Seoul.
12. 손재영, 1991. “지가와 거시경제변수간의 인과관계에 관한 실증분석”, 『한국개발연구』, 13(3): 57-58.
Son, J.Y., 1991. “Causality between Land Price Increase and Macroeconomic Variables”, *KDI Journal of Economic Policy*, 13(3): 57-58.
13. 신규식·최희련·이홍철, 2015. “신재생에너지 동향 파악을 위한 토픽 모형 분석”, 『한국산학기술학회논문지』, 16(9): 6411-6418.
Shin, K.S., Choi, H.R. and Lee, H.C., 2015. “Topic Model Analysis of Research Trend on Renewable Energy”, *Journal of the Korea Academia-Industrial Cooperation Society*, 16(9): 6411-6418.
14. 신수정, 2014.6.11 “글에서 감정을 읽다 - 감성분석의 이해”, 『IT WORLD』. <http://www.itworld.co.kr/techlibrary/87940>
Shin, S.J., 2014.6.11 “Read the feeling: Understanding of Sentiment Analysis”, *IT WORLD*. <http://www.itworld.co.kr/techlibrary/87940>
15. 심광섭, 2014. “한국어 형태소 분석을 위한 음절 단위 확률 모델”, 『정보과학회논문지』, 41(9): 642-651.
Shim, K.S., 2014. “Syllable-based Probabilistic Models for Korean Morphological Analysis”, *Journal of KIISE*, 41(9): 642-651.
16. 심광섭, 2016. “통계 기반 한국어 형태소 분석기의 성능 개선”,

- 「인문과학연구」, 34: 285-316.
Shim, K.S., 2016. "Improving the Performance of Statistical Korean Morphological Analyzer", *Journal of Humanities*, 34: 285-316.
17. 한국건설산업연구원, 2004. 「주택가격의 지역간 상관관계 분석 연구 - 수도권 아파트 가격을 중심으로」, 서울.
Construction & Economy Research Institute of Korea, 2004. *A Research on the Relationship between Housing Prices in the Metropolitan Area*, Seoul.
18. 이득환·강형구·김수현·이창민, 2013. "빅데이터에 나타난 감정 분석", 「금융공학연구」, 12(2): 79-96.
Lee, D.H., Kang, H.G., Kim, S.H. and Lee, C.M., 2013. "Autocorrelation Analysis of the Sentiment with Stock Information Appearing on Big-Data", *The Korean Journal of Financial Engineering*, 12(2): 79-96.
19. 이재성, 2011. "한국어 형태소 분석을 위한 3단계 확률 모델", 「정보과학회논문지: 소프트웨어 및 응용」, 38(5): 257-268.
Lee, J.S., 2011. "Three-Step Probabilistic Model for Korean Morphological Analyzer", *Journal of KIISE: Software and Applications*, 38(5): 257-268.
20. 이종민·이종아·정준호, 2017a. "포털 검색지수를 활용한 전세가격 예측 - 네이버, 구글을 중심으로 -", 「한국부동산학회 부동산학보」, 68: 134-148.
Lee, J.M., Lee, J.A. and Jeong, J.H., 2017a. "The Jeonse Price Forecasting Used by Portal Search Index : Focusing on Naver, Google Trend", *Korea Real Estate Academy Review*, 68: 134-148.
21. 이종민·이종아·정준호, 2017b. "뉴스 빅데이터를 이용한 전세가격 예측 - 토픽모형 분석을 중심으로", 「한국부동산학회 부동산학보」, 69: 43-57.
Lee, J.M., Lee, J.A. and Jeong, J.H., 2017b. "The Jeonse Price Forecasting used by News Big Data; Focusing on Topic Modeling Analysis", *Korea Real Estate Academy Review*, 69: 43-57.
22. 임대봉, 2015. "유동성이 주가 및 주택가격에 대한 파급효과 분석", 「한국부동산학회 부동산학보」, 61: 80-93.
Lim, D.B., 2015. "Analysis on Effect of Liquidity on Stock Prices and Housing Prices", *Korea Real Estate Academy Review*, 61: 80-93.
23. 정승일·김현우, 2014. "악성코드 탐지 시스템 Web-Anti-Malware", 「한국컴퓨터정보학회 학술발표논문집」, 22(2): 365-367.
Jung, S.I. and Kim, H.W., 2014. "Web-Anti-Malware Detection System", *Proceedings of the Korean Society of Computer Information Conference*, 22(2): 365-367.
24. 진창하·Paul Gallimore, 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 Analysis Association*, 18(2): 125-142.
25. 통계청 통계개발원, 2015a. 「빅데이터의 국가통계 활용을 위한 기초연구」, 대전.
Korea National Statistical Office, 2015a. *A Basic Study on Utilization of National Statistics of Big Data*, Daejeon.
26. 통계청 통계개발원, 2015b. 「SNS 데이터를 활용한 소비자성향 분석」, 대전.
Korea National Statistical Office, 2015b. *An Analysis of Consumer Trends Using SNS Data*, Daejeon.
27. 홍두표·정하림·박상민·한음·김홍희·윤일수, 2017. "SNS감성 분석을 이용한 정보 추출 방법론에 대한 연구", 「한국ITS학회 논문지」, 16(6): 141-155.
Hong, D.P., Jeong, H.R., Park, S.M., Han, E., Kim, H.H. and Yin, I.S., 2017. "Study on the Methodology for Extracting Information from SNS Using a Sentiment Analysis", *Journal of Korea Inst. Intell. Transp. Syst*, 16(6): 141-155.
28. Blei, D.M., 2012. "Probabilistic Topic Models", *Communications of the ACM*, 55(4): 77-84.
29. Laney, D., 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety", *META Group Research Note*, 6.
30. 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.
31. Woo, J.C., 2005. "An Analysis of the Determinants of Changes in the Land Prices in Korea", Master's Thesis, Yonsei University.
32. Won, J. and Lee, J.S., 2018. Investigating How the Rents of Small Urban Houses are Determined: Using Spatial Hedonic Modeling for Urban Residential Housing in Seoul. *Sustainability*, 10(1): 31.

Date Received 2018-12-17
Reviewed(1st) 2019-01-17
Date Accepted 2019-01-17
Date Revised 2019-01-24
Final Received 2019-01-24