뉴스 빅데이터를 이용한 전세 가격 예측

- 토픽모형 분석을 중심으로 -

The Jeonse Price Forecasting used by News Big Data

- Focusing on Topic Modeling Analysis -

이 종 민* · 이 종 아** · 정 준 호***

Lee, Jong Min · Lee, Jong Ah · Jeong, Jun Ho

目 次

Ⅰ. 서론

Ⅱ. 선행 연구

Ⅲ. 데이터 수집 및 모형 설계
1. 데이터 수집
2. 모형 설계

Ⅳ. 예측력 검증
1. 기본적 검정
2. 기본모형 설정
3. 유용성 분석

Ⅴ. 결론 및 시사점

〈abstract〉

〈참고문헌〉

ABSTRACT

1. CONTENTS

(1) RESEARCH OBJECTIVES

The purpose of this study is to firstly analyze the patterns of news articles that have been analyzing the market trend relatively objectively, and to distinguish the key topics in the jeonse market by distinguishing meaningful topics in the jeonse and analyzing their meaning. Second, the extent to which the change in the proportion of these topics affects the improvement of the forecasting power of the jeonse market.

(2) RESEARCH METHOD

In this paper, we use the LDA algorithm, which is mainly used to find the hidden semantic structures in the text during topic modeling. The ARIMA model is used for

* 주 저 자: 강원대학교 부동산학과, 박사과정 수료, goodrealtor@korea.com

** 공동저자: KB금융지주, 선임연구위원, 부동산학박사, jongah01@hanmail.net

*** 교신저자: 강원대학교 부동산학과 교수, D.Phil., jhj33@kangwon.ac.kr

▷ 접수일(2017년 4월 20일), 수정일(1차:2017년 5월 3일), 게재확정일(2017년 5월 20일)

the 'Jeonse' price forecasting model.

(3) RESEARCH FINDINGS

The prediction accuracy of the ARIMA model, which is a model of the rent price change rate, is also analyzed. The model with topic 'loan' had 17.39% RMSE in time difference 10 and MAE was 15.72%, which showed the greatest predictive power improvement among the models, and it was analyzed that the prediction power was improved by 12.14~17.39% in the whole comparative model

2. RESULTS

Similar to the previous research that examined whether the topic weight of news articles is substitute for economic forecasting, the accuracy of the prediction of the rent price fluctuation rate is 12.14~17.39% when using the 'Jeonse'. It has been verified that there are improved results.

3. KEY WORDS

• ARIMA, Cross correlation, Granger causality test, Jeonse, Topic Modeling

국문초록

본 연구는 '전세'관련 뉴스기사 빅데이터의 월별 토픽 비중을 이용하여 전세가격변동률 예측력 향상에 관한 실증분석이다. 이를 위하여 KB국민은행의 전국 전세가격지수 자체를 활용한 아리마 기본모형과 여기에 토픽 비중을 포함하는 모형과의 RMSE 및 MAE를 비교하였다. 사전분석으로 진행한 교차상관분석에서는 '전세', '월세', '정책', '분양'의 토픽이 전세가격변동률에 선행성을 보였으며, 그랜저인과관계 분석 결과 월별 각 토픽 비중이 전세가격변동률에 직·간접적으로 영향을 미치는 것으로 분석되었다. 결론적으로 토픽 비중을 포함하는 모형의 예측력이 10~11개월의 시차에서 기본모형대비 최대의 향상률을 보였다. 이는 전세시장 분석시 토픽 비중을 활용한다면, 보다 정확한 가격예측이 가능함을 의미하며, 이러한 결과는 향후 전세 거주의 주 계층인 서민층의 주거안정에 선제적으로 대응하는 정책 입안 활용에 도움이 될 것으로 판단된다.

핵심어 : 교차상관분석, 아리마모형, 예측, 인과관계분석, 전세, 토픽

Ⅰ. 서 론

뉴스 기사는 그 속성상 사실을 기본으로 향후 시장에 미치는 영향에 대해 제시한다는 점에서 시장 예측에 매우 유의미한 정보이다.[1] 실제 뉴스가 주식이나 부동산 등 경제 시장에 영향을 미치고 있음은 이미 많은 연구를 통해 입증되고

1) 김유신·김남규·정승렬, "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형", 지능정보연구, 한국지능정보시스템학회, 2012, 제18권 제2호, pp.143~156.

있다[김진유(2006), 우윤석·이은정(2011), 송치영(2012) 등][2]. 다만 그동안의 뉴스를 활용한 시장 예측에 대한 연구는 기술 및 데이터 접근성의 한계로 일부 몇 개의 신문으로 국한하거나 일정기간 관련 기사의 빈도 측정 등을 활용한 연구로 제한되어 왔다.

그러나 최근 빅데이터 수집의 접근성이 크게 개선되었을 뿐만 아니라, 빅데이터 기술의 발전으로 인해 뉴스 등 다양한 비정형데이터에 포함된 단어들의 특성까지도 상세하게 분석할 수 있도록 진일보되어 위와 같은 연구의 한계를 극복하는 새로운 방법론이 제시되고 있다. 최근 뉴스의 단순한 빈도뿐만 아니라 감성(긍정/부정)적 특성이나 연관 키워드 분류 등에 대한 연구[Yoosin Kim et al.(2014), Min Jeong Park(2015), Ko Ichinose, Kazutaka Shinmada(2016), 김대원·유정석(2016)][3]등이 그러한 예이다.

한편, 뉴스에서 의미 있는 키워드를 분리하는 작업인 텍스트 마이닝(Text Mining)에 대한 기술적 발전도 뉴스 활용 연구 확대에 기여하고 있다[4]. 대표적인 방법은 문헌 클러스터링(Literature clustering)[5]과 토픽모형(Topic modeling)[6] 등이 있다. 이중 대용량의 문서에 포함된 숨겨진 토픽들을 찾아내는 통계추론 모델인 토픽모형은 신문 기사나 언론등이 가지는 토픽과 각 단어들이 토픽에 포함될 확률을 찾아내는데 유의미한 분석방법으로 대두되고 있으며, 박자현·송민(2013)[7], 현윤진 외(2015)[8], 김건·윤혜정(2016)[9] 등이 해당 방법론을 통해 시장흐름을 분석하는데 활용한 연구를 시도하였다.

본 연구는 빅데이터의 기술 발달과 시장의 예측에 대한 관심 확대 등의 변화를 감안하여 국내 자산 시장의 한 축인 주택 시장을 뉴스 빅데이터를 활용해 예측력 확대 가능성을 분석하고자 한다. 본 연구에서 진행하고자 하는 주택 시장은 상대적으로 주거 취약 계층인 일반 서민층에 관심이 높은 전세시장으로 정하였으며[10], 빅데이터 텍스트 마이닝 기법은 토픽모형을, 예측력 향상의 검증방법은 ARIMA 모형을 활용하고자 한다. 연구 범위를 전세시장으로 국한한 것은 전세라는 텍스트가 의미가 명확하여 동음 이의어로 인한 중복기사 수집의 오류가 다른 키워드 보다 적은 장점이 있기 때문이기도 하다.

본 연구의 목적은 첫째, 전세 관련 뉴스 빅

---

2) 김진유, "신문기사가 부동산가격변동에 미치는 영향", 주택연구, 한국주택학회, 2006, Vol.14, No.2, pp.39~63.
우윤석·이은정, "언론보도와 시계열 주택가격 간의 관계에 관한 연구", 주택연구, 한국주택학회, 2011, 제19권 제4호, pp.111~134.
송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구", 국제경제연구, 한국국제경제학회, 2012, 제8권 제3호, pp.1~34..

3) Yoosin Kim, Seung Ryul Jeong, Imran Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction", *Int. J. Advance. Soft Comput. Appl*,. 2014, Vol.6, No.1, pp.1~13.
Min Jeong Park, Doo Sup Choi, Il Chul Moon, "Does news affect the housing market?", *Preprint submitted to Journal of Housing Economics*, 2015, pp.1~40.
Ko Ichinose, Kazutaka Shimada, "Stock market prediction from news on the Web and a new evaluation approach in trading", *2016 5th IIAI International Congress on Advanced Applied Informatics*, pp.77~81.
김대원·유정석, "뉴스기사가 아파트 매매·전세 가격 국면 지속기간에 미치는 영향 분석", 서울도시연구, 서울연구원, 2016, 제17권 제2호, pp.29~46.

4) 텍스트 마이닝이란 블러그나 댓글 뉴스 등 비정형 데이터를 통하여 유용한 정보를 분석하고 추출하는 것을 의미

5) 문헌클러스터링이란 밀접하게 관련된 문헌들은 동일한 정보요구에 대해 모두 적합하고, 적합한 문헌들과 부적합한 문헌들은 서로 다른 클러스터에 속할 것이라는 가설로 사전분류 체계없이 문헌간의 유사성을 따라 분류함.

6) 토픽모형이란 텍스트 마이닝의 방법중 하나로 비구조화된 텍스트 자료의 문서에서 주제를 추출하는 확률모형 알고리즘임.

7) 박자현·송민, "토픽모델링을 활용한 국내문헌정보학 연구동향분석", 정보관리학회지, 정보관리학회, 2013, 제30권 제1호, pp.7~32.

8) 현윤진·김남규·조윤호, "토픽 분석을 활용한 관심 기반 고객 세분화 방법론", *Journal of Information Technology Applications&Management*, 2015, 제22권 제1호, pp.77~93.

9) 김건·윤혜정, "토픽모델링을 활용한 서울지역 호텔서비스에 대한 고객인식의 변화 분석", 서비스경영학회지, 서비스경영학회, 2016, 제17권 제3호, pp.217~230.

10) 이윤홍, "전국 아파트 전세변동률 기준 시장 유형화와 유형별 전세가격 변동률 영향요인 분석", 부동산학보, 한국부동산학회, 2016, 64집, pp.212~226.

데이터의 주제어와 연관어 상의 패턴을 분석해 시장 내 유의미한 토픽들을 구분하고, 시간이 변화함에 따라 해당 토픽들의 영향이 시기적으로 어떻게 변화하였는지를 확인하는 것이다. 이는 당시의 시장의 흐름에 따라 관심과 중심이 되는 키워드들이 어떻게 달라졌는가를 확인할 수 있는 것에 유의미한 분석이다. 둘째는, 이러한 토픽 비중의 변화가 전세가격의 예측력 향상에 어느 정도 영향을 주고 있는 것을 밝히는 것으로 시장의 심리적 흐름 및 관심 이슈가 가격의 선행적 지표로써 활용 될 수 있음을 검증할 수 있다.

이러한 분석의 시도는 매일 무수히 발표되고 있는 뉴스기사를 활용해 향후 전세시장을 보다 정밀하게 예측함으로써 주거라는 본질적인 이용가치를 목적으로 하는 서민층의 주거안정에 선제적으로 대응하는 정책 입안과 전세 수요자의 최적 전세 구입 시기 선택에 도움이 될 것으로 판단된다.

Ⅱ. 선행 연구

뉴스의 빈도를 활용한 초기 예측 연구로는 안성원·조성배(2010)[11]는 주가의 예측력 향상을 위하여 Bag of Words 모델과 Naive Bayesian분류 기법을 사용하여, 뉴스만을 이용하여 주가를 예측하는 것에 비하여 주가의 시계열 데이터 분석 기법을 추가할 경우 평균 5.5%의 예측성능이 향상됨을 분석하였다. 변태우 외(2011)[12]은 텍스트 마이닝을 이용하여 뉴스의 단어빈도수와 환율을 함께 예측에 사용할 경우 환율과 뉴스의 단어 빈도를 각각 주가 예측에 사용할 때보다 11.5~12.5%의 예측성공률이 높아지는 것을 밝혔다.

빅데이터를 이용해 본격적으로 예측을 시도한 한 연구는 공성배 외(2015)[13]가 기상요소, 사회적요소등으로 정확한 예측이 어려운 전력수요를 소셜네트워크, 블로그등의 키워드분석을 통하여 수집된 자료에서 더위와 관련된 언어의 언급된 횟수를 통하여 사람들의 심리상태를 분석하여 수요예측의 정확성 향상 방안에 대하여 분석한 연구가 있다. 또한 박민주(2015)[14]는 교통상황 예측을 기상, 사고, 공사시의 교통상황 Data를 활용한 단기/중기/장기의 예측정보를 구축하여 실시간 교통정보를 제공하는 사후적인 교통관리에서 벗어나 사전적 혼잡관리 개념을 도입하였다. 고숙자·정영호(2012)[15]는 보건의료·복지 통계 데이터를 활용하여 질병의 확산을 예측함으로써 국가차원의 질병의 예방과 조기 대응 및 예방이 가능하다고 주장하였다.

본 연구에서 적용한 토픽모형을 활용한 빅데이터 분석을 진행한 연구는 연구의 주제 흐름 변화 등을 파악하거나, 고객의 인식 변화 등을 파악하는데 주로 활용되었다. 박자현・송민(2013)[16]은 국내 문헌정보학의 연구동향을 LDA(Latent dirichlet allocation) 기반의 토픽 모형으로 분석하였으며, 나상태 외(2016)[17]는 학술진흥재단 등재학술지에 게재된 논문을 대상으로 토픽모형과 시계열 분석을 통하

11) 안성원·조성배, "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측", 한국컴퓨터종합학술대회 논문집, 한국정보과학회, 2010, Vol.37, No.1, pp.364~369.

12) 변태우·이성우·김재광·이지형, "텍스트마이닝과 환율데이터를 활용한 주가의 위험성 예측에 관한 연구", *Proceedings of KIIS Fall Conference*, KISS, 2011, Vol.21, No.2, pp.132~133.

13) 공성배·장민석·주성관, "빅데이터 분석을 이용한 단기전력 수요예측", 대한전기학회 전력기술부문회 전력계통연구회 춘계학술대회 논문집, 대한전기학회, 2015, pp.144~145.

14) 박민주, "빅데이터를 활용한 교통상황 예측기술개발", 월간교통, 한국교통연구원, 2015, Vol.212, pp.42~46.

15) 고숙자·정영호, "국민건강 미래예측 시스템 구축방안: 빅데이터를 활용한 건강위험 예측방안 모색을 중심으로", 보건복지포럼, 한국보건사회연구원, 2012, 통권193호, pp.43~52.

16) 박자현·송민, 전게논문, pp.7~32.

17) 나상태·김자희·정민호·안주연, "토픽 모델링을 이용한 시뮬레이션 연구 동향 분석", 한국시뮬레이션학회 논문지, 한국시뮬레이션학회, 2016, Vol.25, No.3, pp.107~116.

여 국내 시뮬레이션 연구 활동 동향을 확인하였다. 한편 차윤정 외(2015)[18]는 미국에서 판매되고 있는 스마트폰 6개 기종의 미국 트위터 데이터를 수집하여 각각의 기기에 대한 토픽을 비교 분석한 후 마케팅전략을 제안하였으며, 현윤진 외(2015)는[19] 고객이 일상에서 사용하는 인터넷 뉴스 기사 조회 기록에 대한 텍스트 분석을 통하여 고객의 관심 분야를 찾아내고 이를 이용하여 고객을 세분화하는 방안을 제시하였다. 그 외에도 김건·윤혜정(2016)[20]은 12년간의 호텔사용 후기 204,251문장을 토픽모형을 활용하여 서울지역 호텔에 대한 인식과 변화를 파악하여 정책적 시사점을 제시하였다.

끝으로 본 연구와 같이 토픽모형을 활용한 시장을 예측한 연구로는 Dick Stevens (2014)[21]가 회귀모형을 활용한 부동산가격예측에 텍스트 마이닝의 tf-idf를 사용하였으며, 실증분석 결과 기준선보다 최대 38%의 예측력 향상이 있음을 증명한 사례와 Hannes Mueller, Christopher Rauh (2016)[22]가 New York Times등 신문기사 약700,000건을 수집하여 분석한 토픽을 이용할 경우 분쟁의 시점을 예측할 수 있고, 분쟁이 발생하기 이전의 문서상의 주제 비중 변화를 추정할 수 있음을 실증 분석하여 제시한 것이 대표적이다. 한편 국내 연구로는 Min Jeong Park et. al(2015)[23]이 09~14년까지의 연합뉴스기사 46,631개를 수집하여, 10개로 분류한 토픽을 포함한 전세가격 회귀모형과 미포함 회귀모형의 R^2값을 비교한 결과 토픽을 포함했을 경우 모형의 설명력이 15.3%가 향상됨을 실증한 연구가 존재한다.

위에서 살펴 본 것처럼 토픽을 활용한 시장 예측을 시도한 연구는 실질적으로 국내에서는 거의 없다. 또한 본 연구와 같이 토픽 모형과 ARIMA 모형을 결합하여 실증분석을 시도한 것은 국내에서는 유일한 시도로 보여진다. Min Jeong Park et. al(2015)는 연구의 범위를 연합뉴스 하나로 한정하였으나, 본 연구는 16개 언론사 뉴스로 확대한 후 중복 기사의 제거 등의 전처리 작업 등을 진행한 바, 이는 빅데이터 분석의 본질적인 측면에서 기존 연구보다 보다 발전적 접근으로 판단되어진다.

Ⅲ. 데이터 수집 및 모형 설계

1. 데이터 수집

뉴스 기사의 빅데이터는 한국언론진흥재단에서 2016년부터 서비스를 개시한 빅데이터분석시스템인 BIGKINDS를 통하여 수집하였으며, 해당 시스템에서는 국내 주요 언론사에서 제공하는 뉴스를 DB로 제공하고 있다.

본 연구에서는 분석 대상 데이터의 신뢰성과 객관성을 확보하기 위하여 전국을 대상으로 기사를 제공하는 언론사 16개[24]에서 유형분류 '경제'에 검색어 '전세'가 포함된 뉴스 전체중 중복뉴스를 제외한 42,474개를 수집하여 텍스트 기반의 빅데이터로 선정하였다.[25]

---

18) 차윤정·이지혜·최지은·김희웅, "소셜미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원", 지식경영연구, 한국지식경영학회, 2015, 제16권 4호, pp.69~87.

19) 현윤진.김남규.조윤호, 전게논문, pp.77~93.

20) 김건·윤혜정, 전게논문, pp.217~230.

21) Dick Stevens, "Predicting real estate price using Text Mining", *HAIT Master Thesis,* Tilburg University School of Humanities, 2014, pp.1~70.

22) Hannes Mueller, Christopher Rauh, "Reading between the lines: Prediction of political violence using newspaper text", *Cambridge working paper economics,* Cambridge-INET Institute, 2016, 1630, pp.1~51.

23) Min Jeong Park, Doo Sup Choi, Il Chul Moon, *op.cit,* pp.1~40.

24) 방송사 3개 MBC, SBS, YTN, 중앙일간지 8개 국민일보, 경향신문, 내일신문, 문화일보, 서울신문, 세계일보, 한겨레, 한국일보, 경제지 5개 매일경제, 서울경제, 파이넨셜뉴스, 한국경제, 헤럴드경제

한편, 전세가격지수는 KB국민은행이 86년 1월부터 주택매매 가격종합지수와 전세 가격종합지수를 전국과 서울, 수도권 그리고 각 지역별로 월별, 주간별로 제공하고 있다. 본 연구에서는 분석의 안정성 및 지역 대표성 등을 고려하여 전국 월간 전세가격지수를 가격 예측모형분석의 기본 데이터로 선정하였다.

데이터 분석기간은 '전세'관련 뉴스가 증가하고[26] 스마트폰의 이용자가 확대되어 스마트폰을 통해 다양한 매체의 뉴스를 접하는[27] 사람이 늘어나기 시작한 2012년 1월부터 2016년 12월까지 60개월로 확정하였으며, 데이터는 해당 기간 내의 전세가격지수와 뉴스 기사를 사용하였다.

〈표 1〉 휴대폰 및 스마트폰 보급률

년 도	2011	2012	2013	2014	2015	2016
휴 대 폰	89.2%	91.2%	91.8%	92.4%	93.2%	94.0%
스마트폰	21.6%	49.3%	66.2%	73.4%	77.6%	81.6%

자료:'16한국미디어패널조사(정보통신정책연구원, Vol.16-21)

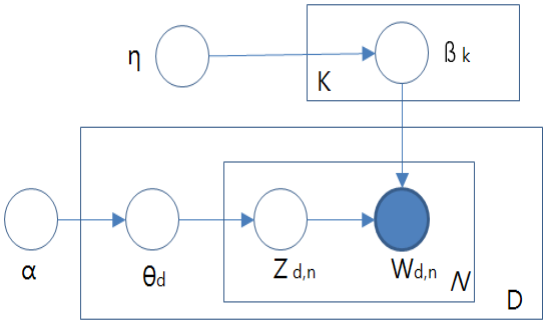
2. 모형 설계

본 논문의 빅데이터 분석을 위해 데이터 마이닝 기법 중에 하나인 토픽모형을 활용하였으며, 전세가격 예측을 위해 ARIMA 모형을 적용하였다

1) 토픽모형

토픽모형은 용량의 규모가 크고 구조화 되지 않은 문서의 주제를 이해하고 체계화하는데 도움이 된다. 토픽모형은 유전자 데이터, 이미지, 소셜네트워크 패턴의 발견에 사용되어 왔으며, 현재는 인간이 직관적으로 해석하기에 불가능한 다양한 전자적인 산출물을 요약하고 조직하는데 활용되고 있다.[28] 토픽모형의 대표적인 알고리즘으로 사용하는 LDA는 주어진 문서에 대하여 잠재적으로 갖는 주제들을 추론하는 확률모형 연구방법론이다.[29] LDA에서는 각각의 문서는 여러개의 주제를 가지고 있으며, 주제들은 Dirichlet 분포를 따르고 있다고 가정한다.

〈그림 1〉 Plate notation for smoothed LDA



〈그림 1〉은 LDA 알고리즘에서 사용하는 확률 계산방식을 도식화한 것으로 α는 각 문서별 토픽의 Dirichlet 사전확률을 의미하며, β는 각각의 토픽별 단어의 분포를 Dirichlet 사후 확률로 나타낸다. θ는 각 문서에 대한 토픽 분포를 의미하며, z는 각 문서에서 출현하는 단어의 토픽을 나타내며, w는 토픽에 포함된 단어를 의미한다. 잠재변수 α와 β,

25) 기존 연구는 주로 4개 전후(중앙일간지2개, 경제지2개) 신문에서 특정한 검색 키워드나 기사의 유형분류 없이 전체 기사를 빅데이터로 사용함.

26) '전세'관련 뉴스가 1999~2011년까지 월평균298.0건에서 12년 562건, 13년 798건, 14년 753건, 15년 864건, 16년 561건으로 큰 폭으로 늘어남.

27) 2016 인터넷이용실태조사(한국인터넷진흥원, 16.12월)에서 스마트폰 이용자가 주로 이용하는 애플리케이션 종류는 '채팅, SNS 78.3%', '사진, 동영상 62.2%', '뉴스 54.4%', '게임 22.7%' 순으로 조사됐다.

28) David M. Blei, "Probabilistic Topic Models", *Communication of the ACM*, 2012, vol.55(4), p.77.

29) David M. Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation", *the Journal of machine Learning research*, 2003, 3, pp.993~1022.

토픽의 결합분포 θ, 단어의 토픽 z, 토픽에 포함된 w가 주어졌을 때 확률분포는 다음과 같이 구해진다.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n})) \quad \text{식 (1)}$$

토픽모형 분석은 대량의 디지털화된 텍스트 자료의 내용을 분석하는 데 유용한 도구로 활용되고 있다.[30] 본 연구에서는 LDA 알고리즘을 통계프로그램 R을 이용하여 실행하였다. 적정 토픽수를 통계 프로그램 R의 'Topicmodels'패키지[31]를 활용하여 월별 샘플 뉴스의 토픽수를 분석한 결과 10~17개의 토픽수가 적정하다고 확인되었으며, 이중 카테고리 별로 토픽이 명확하게 구분되고 해석이 가능한 12개를 최종 토픽수로 정하였다. 토픽의 비중을 구하기 위해서는 데이터 전처리 과정이 필요하다. 우선 형태소 분석을 위하여 'Sejong' Package를 이용하여 데이터에서 명사를 추출하였으며, 글자수는 2글자에서 6글자 사이로 지정하고 물음표, 느낌표, 따움표등의 문장 부호와 영문, 숫자등은 제거하였다. collapsed.gibbs.sampler를 이용하였으며, 설정은 k를12, alpha 0.1, 각1,000회의 샘플링을 반복하여 12개의 토픽과 상위 20개의 단어를 추출하였다. LDA 토픽 모델링 결과인 12개의 토픽 중 전세가격지수와 연관성이 없는 4개를 제외한 8개를 전세가격지수의 예측력 향상 실증검증에 활용하였다[32].

〈표 2〉는 2012년~2016년까지 '전세' 뉴스빅데이터를 LDA 알고리즘으로 월별로 분류한 토픽의 전체기간 비중과 각 토픽에 포함된 주요 단어들이며, 각 토픽의 명칭은 토픽에 포함된 주요 단어중 상위에 위치한 단어에 따라 분류하였다.

2) ARIMA 예측 모형

ARIMA 모형은 연속적인 시간에 대한 한 변수의 관측치에 대한 미래예측을 수행하는 시

〈표 2〉 뉴스 '전세'의 토픽 및 구성 단어 일부

Topic1(전세) 0.12533	Topic2(월세) 0.08707	Topic3(대출) 0.09192	Topic4(정책) 0.09765	Topic5(경기) 0.09614	Topic6(거래) 0.07993	Topic7(분양) 0.10325	Topic8(공급) 0.08557
전세	월세	대출	정책	시장	거래	분양	공급
이사철	보증금	은행	정부	하락	공인중개사	청약	입주
전세가격	주거비	금리	규제	상승	부동산	미분양	임대주택
매물	비율	담보대출	전망	전망	계약	전매	공공
비수기	전환	전세자금	대책	증가	등록	아파트	임대
재건축	부담	가계부채	집값	통계	서비스	전용	행복
아파트	오피스텔	상환	과열	수요	체결	단지	개발
거래량	물건	자금	시장	전국	신고	예정	사업
상승	세입	아파트	부동산	수도권	실거래	분양가	주거
강남	수익률	금액	가계부채	기록	정보	경쟁률	소득

30) 남춘호, "일기자료 연구에서 토픽모델링 기법의 활용 가능성 검토", 비교문화연구, 서울대학교 비교문화연구소, 2016, 제22집 1호, pp.89~135.

31) Thomas L. Griffiths, Mark Steyvers, "Finding scientific topics", *PNAS*, 2004, Vol.101, suppl.1, pp.5228~5235.

32) 차윤정·이지혜·최지은·김희웅, 전게논문, p.74.

계열 분석에 사용하는 대표적인 확률적 시계열 모형으로[33], Box and Jenkins의 방법을 주로 사용한다.

$$W_t = \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad \text{식 (2)}$$

Box-Jenkins 방법은 1단계에서 상관도표와 편상관도표, 산점도를 사용하여 모형을 식별하고 적절한 p, d, q를 구한다. 시계열의 단위근 검정을 통하여 변수의 안정성을 판단 후 불안정 시계열이면 차분을 통하여 안정화과정을 거친다. 2단계는 시계열의 최적모형의 추정으로 모수 선택기준(criteria)은 가역성, 안정성, 모수절약등이며 추정된 모수가 통계적으로 유의하지 않으면 다시 1단계로 회귀하여 모형을 재식별 한다. 3단계는 모형의 진단으로 최종 적합모형을 선정하는 단계이다. 모형의 진단은 잔차가 자기상관의 존재 여부를확인하는 것으로, 이를 위하여 Ljung-Box Q-통계량과 p값 검정을 실시한다. 모든 유의도에서 p>0.05보다 클 경우 귀무가설이 채택되어 모형은 백색잡음으로 판단한다.[34] 4단계는 최종 선정된 모형을 활용하여 시계열 분석자료의 미래를 예측한다.

Ⅳ. 예측력 검증

1. 기본적 검정

1) 단위근 검정

단위근 검정은 시계열 변수의 안정성여부를 확인하기 위한 방법으로 시계열이 불안정할 경우 변수 내에 단위근이 존재하게 된다. 기본모형으로 사용하는 시계열 자료인 KB국민은행 2012년1월~2016년12월까지 전세가격지수와 비교모형의 분석자료로 사용하게 될 토픽비중을 X-13ARIMA-SEATS monthly seasonal adjustment method를 사용하여 계절 조정한 후, 전세가격지수와 월간 토픽 비중의 단위근을 검정하기 위하여 ADF검정법을 이용하였다. 기본모형으로 사용할 전세가격지수에 대한 단위근 검정결과 10% 유의수준에서 단위근을 갖는다는 귀무가설을 기각하지 못하고 있으므로, 전세가격지수는 단위근을 가지고 있는 것으로 판단된다. 시계열이 단위근을 갖고 있다는 것은 시계열이 추세를 내포하고 있는 것이므로 로그차분등을 실시하여 시계열의 안정성을 확보하여야 한다는 것을 의미하여, 전세가격지수는 로그 차분하여 안정된 시계열로 변경하였다.[35] 월간 토픽 비중은 수준변수에서 8개 변수 모두 1%와 5%의 유의수준에서 귀무가설을 기각하여 안정성을 갖는다고 판단할 수 있다. 결과적으로 전세가격지수는 1차로그차분변수를, 토픽 비중은 수준변수를 실증분석의 변수로 선정하였다.

〈표 3〉 단위근 검정결과

변수명	수준변수		로그 1차 차분변수	
	t 값	p-value	t 값	p-value
전세가격지수	-0.753	0.455	-2.345	0.023**
Topic1(전세)	-2.945	0.005***	※수준변수에서 단위근 없음이 판정됨	
Topic2(월세)	-2.193	0.033**		
Topic3(대출)	-3.023	0.004***		
Topic4(정책)	-9.551	0.000***		
Topic5(경기)	-2.815	0.007***		
Topic6(거래)	-6.957	0.000***		
Topic7(분양)	-2.899	0.005***		
Topic8(공급)	-6.073	0.000***		

주: *, **, ***는 유의수준 10%, 5%, 1%임

33) 우경·이성석, "지가변동률 예측을 위한 시계열 모형 분석", 부동산학보, 한국부동산학회, 2015, 60집, p.144.

34) 서정민, "시계열 모형을 활용한 사회서비스 수요·공급모형 구축", 한국콘텐츠학회논문지, 한국콘텐츠학회, 2015, Vol.15, No.6, pp.399~410.

35) 금기조·김병량, "KOSPI지수와 금융변수가 주택매매가격과 전세가격에 미치는 영향 분석", 부동산학보, 한국부동산학회, 2015, 60집, pp.188~189.

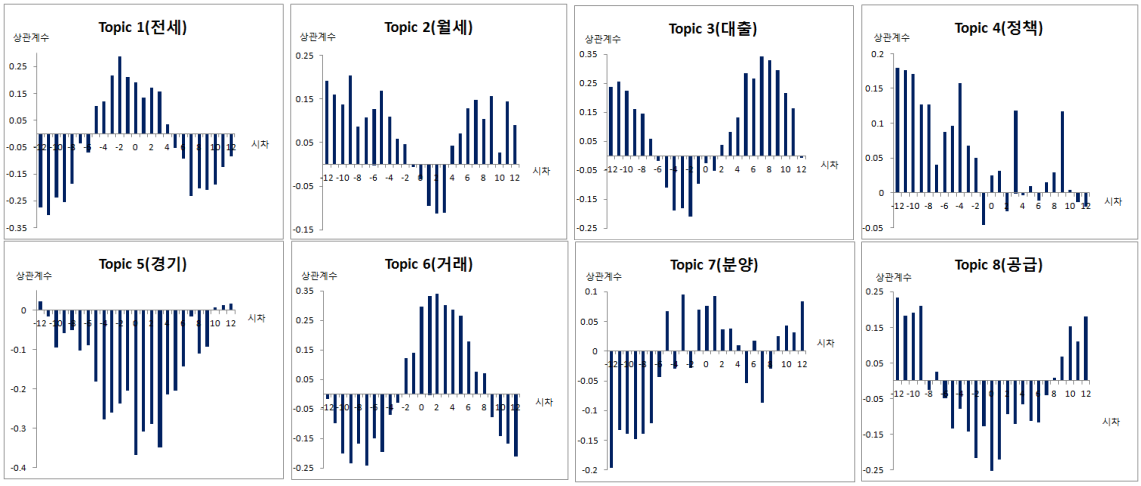
2) 교차상관 분석

토픽 비중과 전세가격변동률과의 시차별 상관관계를 분석하기 위하여 ±12개월의 시차로 교차상관분석을 실시하였다.[36] 첫째, 방향성 분석에서는 Topic1(전세), Topic4(정책), Topic6(거래), Topic7(분양)은 전세가격 변동률과 교차상관계수[37]가 양의 상관관계를 보여, 전세가격변동률과 같은 방향으로 변화는 것으로 확인되었으며, Topic2(월세), Topic3(대출), Topic5(경기), Topic8(공급)은 음의 교차상관관계를 보여 전세가격변동률과 반대방향으로 변화는 것으로 분석되었다. 선행성분석[38]에서는 Topic1(전세), Topic2 (월세), Topic4(정책), Topic7(분양)이 전세가격변동률에 선행하였다. 이것은 부동산정책과 신규주택분양이 일정기간이후에 전세가격의 변동에 영향이 미치고 있음이 실증된 것이며[39], Topic5(경기), Topic8(공급)은 전세가격변동률과 동행하는 것으로 확인되었으며, Topic3(대출)과 Topic6(거래)는 후행하는 결과를 보였다. 이것은 미디어의 특성인 현실성이 반영된 것으로 전세가격변동으로 대출 수요 및 거래량 변동 등이 시장에서 후행하여 반응하는 것과 일치하는 결과이다.

3) 그랜저 인과관계 검정

토픽 비중과 전세가격변동률간의 인과관계 검증을 위하여 그랜저 인과관계 검정을 실시하였다. 그랜저 인과관계 검정은 시차분포모형을 이용하여 원인변수와 결과변수의 구분을 명확하게 위하여 사용하는 방법이다. 토픽 비중과 전세가격변동률의 그랜저 인과관계 검정분석결과 Topic7(분양)는 시차3에서 5%, 시차4에서는 10% 유의수준에서 주택가격변동률에 영향을 미치는 것으로 확인되었으며, Topic8(공급)은 시차1에서 10% 유의수준에서 주택가격변동률과 상호 인과관계가 존재하였다. Topic7(분양)

〈그림 2〉교차상관 분석결과



36) 이강용·이종아·정준호, "주택시장과 주식시장의 동적 네트워크 구조 비교", 부동산학보, 한국부동산학회, 2015, 61집, p.198.

37) 교차상관계수가 양이면 같은 방향으로 변화, 음이면 반대 방향으로 변화, 0이면 서로 경기 중립적.

38) 선행성 측정: 시차상관계수의 값이 최대가 되는 시차가 양이면 해당변수는 기준지표 대비 후행지표, 음이면 선행지표, 시차상관계수가 최대가 되는 시차가 0이면 동행지표임.

39) 주택준공물량이 10~20% 증가시 전세가격은 0.3~06.%p 하락하고, 매매가격은 0.2~0.4%p 하락함. (국토연구원 부동산시장연구센터, "2017년 주택시장 영향요인 분석과 전망", 국토정책 Brief, 국토연구원, 2017, No.607, pp.1~8)

〈표 4〉 그랜저 인과관계 검정결과

Null Hypothesis	Lags 1	Lags 2	Lags 3	Lags 4	Lags 5	Lags 6
Topic1(전세) ↛ 전세가격변동률	1.1121	1.5608	1.1928	1.8912	1.4794	1.3647
Topic2(월세) ↛ 전세가격변동률	0.0712	0.4592	0.2119	0.2076	0.2993	0.2928
Topic3(대출) ↛ 전세가격변동률	1.5031	1.2810	1.0035	0.7054	0.7237	0.6794
Topic4(정책) ↛ 전세가격변동률	0.1565	0.5257	0.5162	1.1343	0.8671	0.6969
Topic5(경기) ↛ 전세가격변동률	0.2535	0.2355	0.3881	1.9881	1.5192	1.6259
Topic6(거래) ↛ 전세가격변동률	1.7372	1.1367	1.4820	0.9162	1.1920	0.8839
Topic7(분양) ↛ 전세가격변동률	0.0364	2.2946	2.8708**	2.2321*	1.7518	1.6741
Topic8(공급) ↛ 전세가격변동률	3.5142*	2.1347	2.0062	1.1700	1.0727	0.8791
전세가격변동률 ↛ Topic1(전세)	0.8330	0.6757	0.3651	1.2883	1.7404	1.5914
전세가격변동률 ↛ Topic2(월세)	0.5391	0.3423	0.2256	1.3910	1.3363	1.0736
전세가격변동률 ↛ Topic3(대출)	0.1485	0.8737	1.0091	0.7030	1.5899	1.3028
전세가격변동률 ↛ Topic4(정책)	0.0402	0.3009	1.9605	1.6448	1.3025	1.2631
전세가격변동률 ↛ Topic5(경기)	6.7219**	4.2380**	2.6714*	1.9687	1.3747	1.0345
전세가격변동률 ↛ Topic6(거래)	8.2855***	5.4217***	3.0277**	2.7815**	2.1058*	1.6726
전세가격변동률 ↛ Topic7(분양)	0.4208	0.5472	0.4597	0.2994	0.5102	0.5917
전세가격변동률 ↛ Topic8(공급)	2.9198*	2.3317	1.6311	1.2892	1.3782	1.1700

주: 1) *, **, ***은 F-Statistic이 10%, 5%, 1%의 유의수준에서 귀무가설을 기각하므로 변수간 인과관계가 있음을 의미함

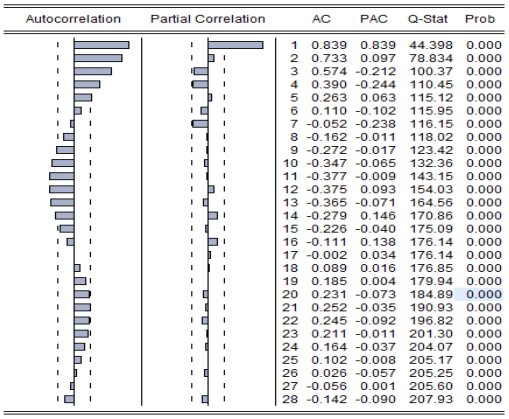
은 시차교차상관분석에서도 전세가격변동률에 선행성이 확인되었고, 주택의 신규 분양 공급량의 변화가 일정기간 이후 전세시장의 공급변화로 전세가격이 변동하는 시장의 순환구조와도 일치하는 결과이다. 전세가격변동률과 토픽 비중간의 인과관계 검정 결과는 Topic5(경기)에 대하여 시차1~2에서는 유의수준 5%, 시차3에서는 유의수준10%에서 영향을 미치는 미치는 것으로 확인되었으며, Topic6(거래)에 대하여는 시차1~2에서 유의수준1%의 높은 인과관계를, 시차3~5에서는 유의수준 5% 및 10%에서 영향을 미치는 것으로 확인되어, 전세가격변동이 일정기간 후 전세시장의 가격 및 거래 등의 뉴스에 반영된다는 것을 알 수 있다. 시차1에서는 전세가격변동률 → Topic6(거래) → Topic8(공급) → 전세가격변동률의 순환적 인과관계가 있음이 확인되었다.

2. 기본모형의 설정

토픽의 비중이 전세가격 예측력 향상에 유의미한 영향을 미치는지에 대하여 실증검증을 위하여 KB국민은행에서 제공하는 전세가격지수를 활용한 ARIMA(Auto Regressive Integrated Moving Average)기본모형과 기본모형에 BIGKINDS에서 수집한 전세뉴스의 토픽 비중를 추가한 ARIMA모형(이후 토픽포함모형이라고 한다)을 비교분석하였다.

1단계로 전세가격변동률의 계열상관도를 진단하여 기본모형의 AR 및 MA차수를 결정하였다. 자기상관함수(ACF)는 지수적으로 감소하여 0으로 수렴하고, 편자기상관 함수(PACF)는 시차1에서 스파이크가 신뢰 한계선 밖으로 벗어났지만 시차2 부터는 신뢰 한계선 내에서 0으로 수렴

〈그림 3〉 전세가격변동률 계열상관도



	AC	PAC	Q-Stat	Prob
1	0.839	0.839	44.398	0.000
2	0.733	0.097	78.834	0.000
3	0.574	-0.212	100.37	0.000
4	0.390	-0.244	110.45	0.000
5	0.263	0.063	115.12	0.000
6	0.110	-0.102	115.95	0.000
7	-0.052	-0.238	116.15	0.000
8	-0.162	-0.011	118.02	0.000
9	-0.272	-0.017	123.42	0.000
10	-0.347	-0.065	132.36	0.000
11	-0.377	-0.009	143.15	0.000
12	-0.375	0.093	154.03	0.000
13	-0.365	-0.071	164.56	0.000
14	-0.279	0.146	170.86	0.000
15	-0.226	-0.040	175.09	0.000
16	-0.111	0.138	176.14	0.000
17	-0.002	0.034	176.14	0.000
18	0.089	0.016	176.85	0.000
19	0.185	0.004	179.94	0.000
20	0.231	-0.073	184.89	0.000
21	0.252	-0.035	190.93	0.000
22	0.245	-0.092	196.82	0.000
23	0.211	-0.011	201.30	0.000
24	0.164	-0.037	204.07	0.000
25	0.102	-0.008	205.17	0.000
26	0.026	-0.057	205.25	0.000
27	-0.056	0.001	205.60	0.000
28	-0.142	-0.090	207.93	0.000

하여, 가능성 있는 모형[40]을 ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(1,1,2), ARIMA (2,1,0) 4가지로 선정하였다.

〈표 5〉 ARIMA 모형 식별통계량

구 분	ARIMA (1,1,0)	ARIMA (1,1,1)	ARIMA (1,1,2)	ARIMA (2,1,0)
adj R^2	0.7185	0.7166	0.7352	0.7185
AIC	-11.1680	-11.1454	-11.1929	-11.1516
SC	-11.0633	-11.0058	-11.0183	-11.0120
HQC	-11.1271	-11.0908	-11.1246	-11.0970

2단계로 전세가격변동률의 ARIMA 모형 선정은 adj R^2이 가장 높고, AIC, SC 및 HQC 값이 작은 모형이 적합도가 높은 것으로 추정되므로, 적합모형으로 예상되는 4개의 모형중 ARIMA(1,1,2)를 전세가격변동률의 기준모형으로 선정하였다.

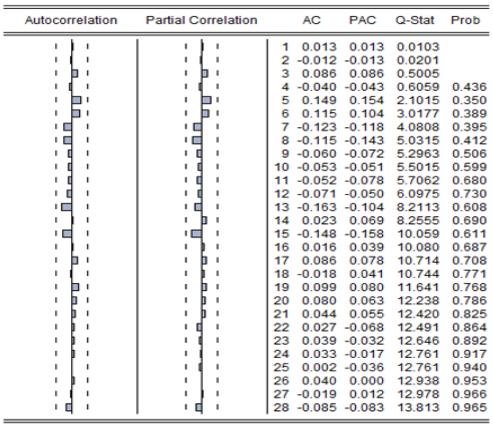
〈표 6〉 ARIMA(1,1,2) 모형 추정치

구 분	계수	t 값	adj R^2	F 값	DW
C	0.0032	4.1458***	0.7352	41.9495	1.9671
AR(1)	0.7935	7.2336***			
MA(1)	-0.0492	-0.2596			
MA(2)	0.3613	2.9690***			

주: *, **, ***는 유의수준 10%, 5%, 1%임

선정된 모형의 계수가 통계적으로 유의하여 다음 단계인 선정된 기준모형의 적합성을 진단하였다. 검정은 Ljung-Box Q-통계량값을 이용하였다. 잔차의 독립성 검정 결과 Q통계량이 이론적 임계값보다 작은 것으로 확인되어(p값〉0.05) 자기상관이 없다는 귀무가설을 기각하지 못하므로 선정된 기본모형을 최종 모형으로 선정하였다.

〈그림 4〉 ARIMA(1,1,2) 잔차 독립성 검정



3. 유용성 분석

전세가격변동률과 토픽비중이 교차상관분석을 통하여 일정한 시차를 두고 시차상관계수가 높아지고, 그랜져 인과관계 분석을 통하여 토픽 비중이 전세가격변동률에 영향을 미치는 것으로 분석되었다. 그러므로 토픽 비중이 전세가격변동률의 예측력 향상에 유용하게 이용될 수 있다고 예상된다. 전세가격변동률 기본모형과 토픽포함모형과의 예측력 비교를 위하여 실제값과 예측치값의 제곱근오차인 RMSE[41](Root mean squared error)값과 예측치 차이 절대값의 평균인 MAE[42](Mean absolute error)를 사용하여 모형간의 예측력의 정확도를 검증하였다.[43]

토픽포함모형의 예측력 향상 정도를 분석하

40) ARIMA적정모형을 결정하기 위한 차수는 최대 시차6으로 지정.

41) RMSE = $\sqrt{\frac{1}{n-k}\sum_t (Y_t - \widehat{Y_t})^2}$, k는 모형의 자유도이며 Y_t는 실제값, $\widehat{Y_t}$는 예측값으로 값이 낮을수록 예측력이 좋다

42) MAE = $\frac{1}{n}|Y_t - \widehat{Y_t}|$, Y_t는 실제값, $\widehat{Y_t}$는 예측값으로 값이 낮을수록 예측력이 좋다

43) 고용호·한승우·이상엽, "주택매매지수 예측을 위한 빅데이터 기반 웹데이터 속성 분석", 부동산학보, 한국부동산학회, 2016, 65집, p.16.

기 위해 기본모형(Model(0))에 8개의 토픽 비중을 포함하여 비교분석 하고자 한다.

〈표 7〉 기본모형과 토픽포함모형

Model명	토픽 비중	토픽포함모형
Model(0)	$\Delta \log(y_t) = \beta_0 + \beta_1 \Delta \log(y_{t-1}) + e_t - \theta_0 - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$	
Model(1)	w_t^1(Topic1(전세))	Model(0) + $\beta_{21}(w_t^1)$
Model(2)	w_t^2(Topic2(월세))	Model(0) + $\beta_{22}(w_t^2)$
Model(3)	w_t^3(Topic3(대출))	Model(0) + $\beta_{23}(w_t^3)$
Model(4)	w_t^4(Topic4(정책))	Model(0) + $\beta_{24}(w_t^4)$
Model(5)	w_t^5(Topic5(경기))	Model(0) + $\beta_{25}(w_t^5)$
Model(6)	w_t^6(Topic6(거래))	Model(0) + $\beta_{26}(w_t^6)$
Model(7)	w_t^7(Topic7(분양))	Model(0) + $\beta_{27}(w_t^7)$
Model(8)	w_t^8(Topic8(공급))	Model(0) + $\beta_{28}(w_t^8)$

전세가격변동률 기본모형에 토픽포함모형의 예측력을 비교한 결과 시차0에서는 토픽 포함모형 Model(1)~(3)의 RMSE향상률이 -1.08~-0.06%, Model(1),(3),(4),(7)은 MAE가 -2.26~-0.47%로 일부모형에서는 음의 향상률을 보였지만, 나머지 Model에서는 최대 Model 5가 3.67%의 예측력 향상을 보이는 등 전세가격기준모형에 토픽 비중을 추가할 경우 예측력이 향상될 수 있다는 가능성을 보여 주었다. 최대 시차12 이하의 토픽 포함모형의 예측력을 비교한 결과 시차 10~11에서 각 Model의 RMSE 및 MAE가 최대의 향상률을 보였다. 최소의 향상을 보인 Model(7)은 시차10에서 RMSE가 15.22%, MAE는 12.14%의 예측력 향상 결과를 나타냈고, Model(3)은 시차10에서 RMSE가 17.39%, 시차11에서 MAE는 15.72%로 Model중 최대의 예측력 향상을 보였다. 시차10과 11의 그랜저인과관계 분석결과 유의수준 10% 범위내에서 각 토픽이 전세가격변동률에 직간접적으로 영향을 미치는 것으로 확인되었다. 시차10의 그랜저인과관계 분석에서는 Topic7 → Topic6 → Topic1 → Topic2, 3 → Topic4, 8 → Topic5 → 전세가격변동률, 시차11의 그랜저인과관계 분석에서는 Topic6 → Topic1 → Topic2 → Topic4, 8, 3 → Topic5 → 전세가격변동률로 인과관계가 있음이 확인되었다. 이것은 토픽포함모형 Model의 예측력 향상이 인과관계분석으로도 효과가 있음이 실증분

〈표 8〉 기본모형과 토픽포함모형 예측력 비교

Model	시차(lag)	RMSE	RMSE 향상률	시차(lag)	MAE	MAE 향상률
Model(0)	0	0.001662	-	0	0.001285	-
Model(1)	0	0.001680	-1.0830%	0	0.001301	-1.2451%
	10	0.001399	15.8243%	10	0.001116	13.1518%
Model(2)	0	0.001663	-0.0602%	0	0.001280	0.3891%
	11	0.001389	16.4260%	11	0.001107	13.8521%
Model(3)	0	0.001675	-0.7822%	0	0.001314	-2.2568%
	10	0.001373	17.3887%	11	0.001083	15.7198%
Model(4)	0	0.001661	0.0602%	0	0.001298	-1.0117%
	11	0.001408	15.2828%	11	0.001123	12.6070%
Model(5)	0	0.001601	3.6703%	0	0.001252	2.5681%
	11	0.001408	15.2828%	10	0.001124	12.5292%
Model(6)	0	0.001652	0.6017%	0	0.001278	0.5447%
	10	0.001397	15.9446%	10	0.001129	12.1401%
Model(7)	0	0.001656	0.3610%	0	0.001291	-0.4669%
	10	0.001409	15.2226%	10	0.001129	12.1401%
Model(8)	0	0.001621	2.4669%	0	0.001265	1.5564%
	10	0.001384	16.7268%	10	0.001118	12.9961%

주: 1) 시차(lag)10~11은 토픽 비중의 선행 최대 시차 12개월내 RMSE 및 MAE가 최소가 되는 시차임.
2) 향상률은 기본모형(Model(0))대비 토픽포함모형(Model1~8)의 RMSE 및 MAE의 향상 정도.

석되었다.

안성원·조성배(2010)[44]는 주가의 예측 뉴스 텍스트 마이닝과 주가의 시계열 데이터 분석 기법을 추가할 경우 주가 예측이 평균 5.5%의 예측 성능향상을 증명하였으며, Dick Stevens(2014)[45]도 부동산 가격 예측에 텍스트 마이닝을 이용할 경우 18~38%의 예측력이 향상됨을 확인하였다.

본 연구도 전세가격변동률 기본모형대비 토픽포함모형의 예측력이 12.14~17.39% 향상됨을 보여 선행연구와 유사한 결과를 보였다고 할 수 있다.

Ⅴ. 결론 및 시사점

본 연구는 토픽모형과 ARIMA모형을 이용하여, 2012년 1월부터 2016년 12월까지 한국언론진흥재단 BIGKINDS에서 수집한 '전세' 관련 뉴스 토픽 비중의 월별 변화가 전세가격 예측에 효과가 있는지 여부와 효과가 있다면 어떠한 토픽이 가장 유용성이 높은 결과를 보이는지 실증분석 하였다. 결과적으로 토픽 비중이 경제예측의 보조지표로서 대체성이 있는지의 여부를 검증한 선행연구와 유사하게 전세시장 예측에 '전세'뉴스 토픽 비중을 활용할 경우 전세가격변동률 예측의 정확도 향상에 유의미한 결과가 있음이 검증 되었다.

토픽모형을 활용한 전세가격변동률 예측 실증분석결과는 다음과 같다. 첫째, 전세가격변동률과 토픽 비중의 최대±12개월의 교차상관분석 결과 Topic1(전세), Topic4(정책), Topic7(분양)은 전세가격변동률에 선행성이 있으며, 양의 상관관계로 전세가격변동률과 같은 방향으로 변동하였다. 최윤정·권상희(2014)[46]의 미디어의 이슈 해석방향은 일반대중의 이해와 해석에 영향을 미칠 수 있다는 연구결과와 마찬가지로 미디어의 전세와 부동산정책, 주택분양의 뉴스 기사화는 전세시장 참여자들의 장·단기 향후 의사결정에 영향을 준다고 해석할 수 있다. Topic2(월세)는 전세가격변동률에 선행성이 있으면서 음의 상관관계로 전세가격변동률과 반대방향으로 변동하고, Topic5(경기), Topic8(공급)은 전세가격변동률과 중립적인 관계에서 전세가격변동률과 반대 방향으로 변동하는 것으로 분석되었다. Topic3(대출)은 전세가격변동률에 후행하며, 반대 방향으로 변동하고, Topic6(거래)는 전세가격변동률에 후행하며, 같은 방향으로 변동하는 것으로 확인되었다. 이것은 미디어의 특성인 현실성이 반영된 것으로 미디어가 일정한 시차를 두고 전세가격변동에 후행하는 전세시장의 거래동향을 뉴스기사화하는 것이 반영된 결과이다. 둘째, 그랜저 인과관계 분석에서는 Topic7(분양)이 시차 3~4에서 전세가격변동률에 영향을 미치는 것으로 분석되었다. 이것은 교차상관분석에서도 선행성이 있는 것으로 확인된 것과 일치하는 결과이며, 전세가격변동률은 시차1~3에서 Topic5(경기)에, 시차1~5에서는 Topic6 (거래)에 영향을 미치는 것으로 확인되었다. 교차상관분석에서도 Topic6(거래)는 전세가격변동률에 후행하는 것으로 분석되었다. 셋째, 전세가격변동률 기본모형인 ARIMA모형에 토픽포함모형의 예측 정확도를 비교하기 위하여 RMSE, MAE를 비교 분석한 결과, 토픽포함모형 Model(7)은 시차10에서 RMSE가 15.22%, MAE는 12.14%의 예측력 향상 결과를 나타냈으며, Model(3)은 시차10에서 RMSE가 17.39%, 시차11에서 MAE는 15.72%로 Model중 최대의 예측력 향상 결과를 보이는등 전 비교 Model에서 예측력이 12.14~17.39% 향상되는 것으로 분석되었다.

44) 안성원·조성배, 전게논문, pp.364~369.

45) Dick Stevens, *op.cit,* p.37.

46) 최윤정·권상희, "'빅데이터'관련 신문기사의 의미연결망 분석", 사이버커뮤니케이션 학보, 사이버커뮤니케이션학회, 2014, 제31권 1호, pp.241~286.

본 연구는 토픽모형을 활용한다면 전세가격의 변동에 관심이 많은 주택 임대차 시장 참여자와 무주택 서민의 주거안정을 위한 정책입안자에게 보다 정교한 전세가격 예측 결과를 제공할 수 있는 근거를 마련하였다는데 의의가 있다.

끝으로 본 연구에 사용된 토픽 비중은 전국을 대상으로 기사를 제공하는 언론사의 기사를 수집하여 전국전세가격의 예측 유용성을 분석하였다. 그러나 실제 전세가격은 시군구별로 상이하게 변동하고 있으므로. 보다 더 정교한 예측을 위해서는 각 지역의 상황만을 정확하고 객관적으로 보도하는 뉴스를 수집하여 토픽 비중을 분석할 필요가 있을 것으로 판단된다. 이에 대한 연구는 향후 과제로 남긴다.

參考文獻

고숙자·정영호, "국민건강 미래예측 시스템 구축방안: 빅데이터를 활용한 건강위험 예측방안 모색을 중심으로", 보건복지포럼, 한국보건사회연구원, 2012, 통권193호.

고용호·한승우·이상엽, "주택매매지수 예측을 위한 빅데이터 기반 웹데이터 속성 분석", 부동산학보, 한국부동산학회, 2016, 65집.

공성배·장민석·주성관, "빅데이터 분석을 이용한 단기전력 수요예측", 대한전기학회 전력기술부문회 전력계통연구회 춘계학술대회 논문집, 대한전기학회, 2015.

국토연구원 부동산시장연구센터, "2017년 주택시장 영향요인 분석과 전망", 국토정책 Brief, 국토연구원, 2017, No.607.

금기조·김병량, "KOSPI지수와 금융변수가 주택매매가격과 전세가격에 미치는 영향 분석", 부동산학보, 한국부동산학회, 2015, 60집.

김건·윤혜정, "토픽모델링을 활용한 서울지역 호텔서비스에 대한 고객인식의 변화 분석", 서비스경영학회지, 서비스경영학회, 2016, 제17권 제3호.

김대원·유정석, "뉴스기사가 아파트 매매·전세 가격 국면 지속기간에 미치는 영향 분석", 서울도시연구, 서울연구원, 2016, 제17권 제2호.

김유신·김남규·정승렬, "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형", 지능정보연구, 한국지능정보시스템학회, 2012, 제18권 제2호.

김진유, "신문기사가 부동산가격변동에 미치는 영향", 주택연구, 한국주택학회, 2006, Vol.14, No.2.

나상태·김자희·정민호·안주연, "토픽 모델링을 이용한 시뮬레이션 연구 동향 분석", 한국시뮬레이션학회논문지, 한국시뮬레이션학회, 2016, Vol.25, No.3.

남춘호, "일기자료 연구에서 토픽모델링 기법의 활용가능성 검토", 비교문화연구, 서울대학교 비교문화연구소, 2016, 제22집 1호.

박민주, "빅데이터를 활용한 교통상황 예측기술개발", 월간교통, 한국교통연구원, 2015, Vol.212.

박자현·송민, "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석", 정보관리학회지, 정보관리학회, 2013, 제30권 제1호.

변태우·이성우·김재광·이지형, "텍스트마이닝과 환율데이터를 활용한 주가의 위험성 예측에 관한 연구", *Proceedings of KIIS Fall Conference,* KISS, 2011, Vol.21, No.2.

서정민, "시계열 모형을 활용한 사회서비스 수요·공급모형 구축", 한국콘텐츠학회논문지, 한국콘텐츠학회, 2015, Vol.15, No.6.

송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구", 국제경제연구, 한국국제경제학회, 2012, 제8권 제3호.

안성원·조성배, "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측", 한국컴퓨터종합학술대회 논문집, 한국정보과학회, 2010, Vol.37, No.1.

우경·이성석, "지가변동률 예측을 위한 시계열 모형 분석", 부동산학보, 한국부동산학회, 2015, 60집.

우윤석·이은정, "언론보도와 시계열 주택가격간의 관계에 관한 연구", 주택연구, 한국주택학회, 2011, 제19권 제4호.

이강용·이종아·정준호, "주택시장과 주식시장의 동적 네트워크 구조 비교", 부동산학보, 한국부동산학회, 2015, 61집.

이윤홍, "전국 아파트 전세변동률 기준 시장 유형화와 유형별 전세가격 변동률 영향요인 분석", 부동산학보, 한국부동산학회, 2016, 64집.

차윤정·이지혜·최지은·김희웅, "소셜미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원", 지식경영연구, 한국지식경영학회, 2015, 제16권 4호.

최윤정·권상희, "빅데이터관련 신문기사의 의미연결망 분석", 사이버커뮤니케이션 학보, 사이버커뮤니케이션학회, 2014, 제31권 1호.

현윤진·김남규·조윤호, "토픽분석을 활용한 관심 기반 고객 세분화 방법론", *Journal of information Technology Applications&Management*, 2015, 제22권 제1호.

David M. Blei, "Probabilistic Topic Models", *Communication of the ACM*, 2012, Vol55(4).

David M. Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation", *the Journal of machine Learning research*, 2003, 3.

Dick Stevens, "Predicting real estate price using Text Mining", *HAIT Master Thesis,* Tilburg University School of Humanities, 2014.

Hannes Mueller, Christopher Rauh, "Reading between the lines: Prediction of political violence using newspaper text", *Cambridge working paper economics,* Cambridge-INET Institute, 2016, 1630.

Ko Ichinose, Kazutaka Shimada, "Stock market prediction from news on the Web and a new evaluation approach in trading", *2016 5th IIAI International Congress on Advanced Applied Informatics(IIAI-AAI)*.

Min Jeong Park, Doo Sup Choi, Il Chul Moon, "Does news affect the housing market?", *Preprint submitted to Journal of Housing Economics*, 2015.

Thomas L. Griffiths, Mark Steyvers, "Finding scientific topics", *PNAS,* 2004, Vol.101 suppl.1.

Yoosin Kim, Seung Ryul Jeong, Imran Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction", *Int. J. Advance. Soft Comput. Appl.,* 2014, Vol.6 No.1.

한국언론진흥재단 : www.bigkinds.or.kr

KB국민은행 : www.kbstar.com