부동산학연구 제27집 제2호, 2021. 6, pp. 47~62 http://dx.doi.org/10.19172/KREAA.27.2.3

Journal of the Korea Real Estate Analysts Association
Vol.27, No.2 2021. 6, pp. 47~62

기계학습 기술을 이용한 부동산 감성지수 개발 모형 연구*

A Study on the Development of the Real Estate Sentiment Index Model Using the Machine Learning Techniques

박 재 수 (Park, Jaesoo)**
이 재 수 (Lee, Jae-Su)***

< Abstract >

The market sentiment analysis is a useful way to understand the participants' sentiment in the real estate market by extracting their opinion, attitude and tendency from the irregular text data. This study aims to develop the sentiment index model, using real estate-related online newspaper articles. The procedures are as follows: First, online news articles are web-crawled from major daily and economic news websites. Second, topics and words are extracted through the Latent Dirichlet Allocation (LDA) topic analysis. Third, the sentiment dictionary is established using the TextRank algorithm with sentences which contain the words extracted by the topic analysis in the second step. Finally, using the Term Frequency - Inverse Document Frequency (TF-IDF) and the Naive Bayes classification model, we assign polarity to the sentences and calculate the weights, producing the monthly real estate sentiment index. The result shows a good performance of the proposed method with an accuracy of 88%. The model is more advanced than existing methods which have been used in the real estate studies, proposing a novel analytical framework and model for the unstructured big data analysis. It also provides an index which reflects participants' sentiment immediately and flexibly, therefore helps to explain and predict changes in the real estate market.

Keyword : Housing Price, Sentiment Analysis, News Article, Naive Bayes, Machine Learning

I. 서론

1. 연구 배경 및 목적

부동산은 우리나라 가계 총자산 가운데 큰 비중을 차지하고 있고, 사회적 영향을 고려했을 때 중요한 경제 이슈 중 하나로 다루어진다. 부동산이 가계 자산에서 차지하는 비중은 우리나라가 다른 어떤 선진국보다 높은 편이다. 우리나라의 가계와 비영리단체 순자산 중 비금융자산 비율은 75.4%로 일본(43.3%)과 미국(34.8%)을 포함한 주요 선진국보다 상당히 높은 수준이다(황의영, 2018).

주택경기의 변화는 가계와 국가 모두에서 관심이지만, 미래의 변화와 흐름을 예측하는 것은 쉬운 일이 아니다. 주택 수요와 공급 등 기본적인 경제 요인을

* 이 논문은 "박재수, 2020, 주택시장 예측을 위한 부동산 감성지수 개발 연구"의 박사학위 논문 일부를 정리하였음.
** 강원대학교 부동산학 박사, goodman@hanil.com, 주저자
*** 본 학회 정회원, 강원대학교 부동산학과 교수, jslee25@kangwon.ac.kr, 교신저자

포함하여 정부의 주택정책, 시장금리와 경제 주체들의 심리 등 다양한 요인이 주택시장에서 가격 결정에 영향을 미치기 때문이다. 중장기를 떠나 단기적인 주택경기의 예측도 어려운 이유는 주택경기를 예측하기 위한 적정한 선행지표가 많지 않기 때문이다. 주택담보대출, 주택수급 동향지수와 금리 등이 일부 이용되지만, 선행시차가 적고 발표가 되는 시점도 서로 달라서 선행지표의 역할을 충분히 하지 못하고 있다.

부동산 부문에서 통계학적, 경제학적 또는 생애주기적 접근 등 전통적인 분석방법이 주로 이용되었으나, 최근에는 빅데이터를 활용한 다양한 분석방법의 개발과 적용 필요성이 높아지고 있다(김민희, 2014). 본 연구는 부동산 시장의 참가자인 매수인과 매도인의 심리에 직·간접적인 영향을 주는 언론매체에 주목하였다. 최근까지 언론매체의 보도자료는 비정형 데이터로 구성되어 정형 데이터인 주택매매 통계와 같이 조사·수집 및 처리를 통해 지수 등의 형태로 제공되기 어려웠다. 그러나 최근 뉴스 자료가 데이터베이스화 되어 자연어 처리와 의미연결망 분석 등 내용분석에 비교적 쉽게 활용할 수 있는 도구들이 제공되었다. 이에 따라 대량의 온라인 뉴스 비정형 텍스트 데이터로부터 감성지수를 산출할 수 있게 되어 다양한 계량적 분석이 가능해졌다. 그러나 감성지수와 같은 부동산 관련된 지수를 정량화하여 산출하기에는 여전히 학술적, 기술적 어려움이 잔존한다.

신문과 방송은 일반 대중이 새로운 정보를 얻는 대표적인 언론 매체이며, 신문은 가장 오랜 기간 사회의 중심적인 매체로 인식되어 왔다. 신문은 텔레비전 매체와 달리 시간의 제약이 없기 때문에 자세히 보기가 가능하고 정보의 권위나 신뢰성이 더 부여되는 경향이 있다(김재휘, 2009). 부동산 시장은 주식, 외환 및 채권 등과 같은 금융시장과 달리 매일 시장에서 형성되는 가격을 일반 대중이 쉽게 획득하기 어렵다. 이런 문제로 사람들은 신뢰성을 담보하는 신문이나 방송 매체에서 전달하는 정보에 의존하여 의사결정을 하는 경향이 크다. 따라서 부동산시장 참여자의 심리에 직접 영향을 미치는 뉴스에 초점을 맞추어 연구를 진행한다.

본 연구의 목적은 부동산시장의 변화를 설명 또는 예측하기 위해 온라인 부동산 뉴스 내용을 활용하여 부동산시장의 심리를 분석하는 감성지수를 개발하는 방법론을 정립하는 것이다. 감성지수는 기계학습 기법으로 신문기사에 나타난 부동산 시장의 분위기를 긍정과 부정으로 나누고, 이를 수치화 한 것이다.

최근 국토연구원이나 KB 국민은행 등 부동산 기관에서 부동산에 관계된 소비자의 심리지표를 설문 및 전화 면접조사를 통해 정기적으로 조사·발표하고 있다. 그러나 설문조사를 활용해 분석되는 지수는 조사항목에 관한 자료를 중심으로 반영되므로 부동산 시장의 특정 이슈 및 이벤트 발생이 시장참여자의 인식에 미친 영향을 분석하기 어렵다(송민채·신경식, 2017). 이와 같은 조사에 기반한 기존 심리지수의 단점을 보완하고, 빅데이터를 활용하여 부동산 감성지수 산출을 위한 새로운 방법론을 제시하고자 한다.

2. 연구 범위 및 방법

본 연구는 일간지와 경제전문지의 인터넷 사이트에서 제공하는 부동산 관련 텍스트 데이터를 분석대상으로 하여 감성지수를 개발한다. 온라인 뉴스기사 수집을 위한 시간적 범위는 2012년 1월부터 2018년 12월까지로 한다.

분석대상 일간지는 동아일보, 조선일보와 중앙일보, 경제전문지는 매일경제신문, 서울경제신문과 한국경제신문이다. 국내에서 비교적 영향력이 높은 일간지 및 경제지를 고려하여 주요 신문사를 선정하였다. 부동산 관련 텍스트 자료는 웹크롤링을 이용하여 해당 신문사의 인터넷 사이트에서 직접 수집하였다.

분석방법은 토픽모형을 활용하여 수집된 부동산 기사에 포함된 단어를 분석하였다. 그리고 텍스트랭크 및 TF-IDF를 이용하여 단어의 중요도를 파악한 후 나이브 베이즈 분류기로 분석 문장에 대한 긍정 또는 부정 극성을 부여하였다. 나이브 베이즈 분류기에서 도출한 점수를 활용하여 감성지수를 산출하였다.

II. 이론 및 선행연구

1. 빅데이터

빅데이터(Big Data)는 일반 데이터베이스 소프트웨어를 통해 저장, 관리 및 분석하는 크기를 넘어선 규모의 데이터이다(Manyika et al., 2011). 빅데이터의 정의는 큰 규모의 데이터 세트에서 출발하였지만, 이제

는 데이터의 분석을 통해 얻을 수 있는 가치로 의미가 확장되었다. 빅데이터의 개념은 급증하는 데이터의 양과 품질을 넘어 지능형 의사결정을 위한 분석에 중점을 둔다(Hilbert, 2016). 정보과학(IT) 기술의 발전으로 빅데이터의 분석 기술을 포괄하는 개념으로써 빅데이터 기술(Big Data Technologies)을 규정한다.

빅데이터는 구조화된 정형 데이터(Structured Data)와 함께 소셜 미디어 및 신문기사와 같은 비정형화된 다양한 데이터들을 포함한다. 빅데이터는 기존 데이터의 규모를 넘어서는 큰 양과 데이터의 생성과 변화가 매우 급속하게 이루어지는 특성이 있다.

정형화 정도에 따라 빅데이터는 정형, 반정형, 비정형 데이터로 구분된다. 정형 데이터는 일정한 규칙하에서 체계적으로 정리되고 고정된 필드에 저장된 데이터로 대표적으로 관계형 데이터베이스가 있다. 반정형 데이터는 고정 필드에 저장되지는 않지만, 메타데이터, 스키마 등을 포함한 데이터를 의미한다. 정형 데이터의 한 형태이지만, 정형 구조의 데이터 모델을 따르지 않는 데이터로 XML이 대표적인 예이다. 비정형 데이터는 고정된 필드에 저장되지 않는 데이터로 고정된 형태가 없으며, 연산이 가능하지 않은 데이터로 텍스트, 이미지, 음성 데이터가 대표적이다.

빅데이터 기술은 기존의 통계분석 기법에 새로운 데이터의 수집, 저장 및 처리 기술을 결합하여 대규모 데이터 활용의 효율을 높일 수 있다. 특히 비정형 데이터를 수집 및 처리하기 위한 기술, 그리고 분석 및 처리 속도를 높이기 위한 기술이 결합된 것이다(경정익, 2014). 보다 많은 데이터 양과 빈도의 측정이 가능하다는 사실 이외에도 정형 데이터와 함께 이미지, 영상, 텍스트 등의 비정형 데이터도 대상으로 하여 지수의 산출과 복잡한 변화를 예측하는데 사용할 수 있다.

2. 기계학습(Machine Learning)

기계학습(Machine Learning)은 컴퓨터 알고리즘과 프로그램을 이용하여 컴퓨터가 인간이 학습하는 것처럼 학습하고, 그 내용을 기반으로 정보를 산출하고 의사결정을 하는 인공지능의 한 분야이다(이요섭・문필주, 2017). 기계학습은 1950년대 인공신경망(Artificial Neural Networks: ANN)이 출현하면서 발전하기 시작하였다. 그러나 80년대 후반 이후 정체기를 겪었으나, 최근 인공지능 분야에서 딥러닝(Deep Learning) 기술의 발전으로 주목받고 있다(배성완・유정석, 2018). 최근에는 기계학습이 산업 전반에 걸친 문제를 해결하기 위한 도구로 활용도가 크게 증가하고 있다. 또한 다양하게 나타나는 문제와 이를 해결하기 위한 방안을 실시간으로 빨리 분석 및 처리함으로써 정확성과 신뢰성이 높은 해답을 제시하고 있다.

기계학습은 학습방법에 따라 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)의 형태로 구분할 수 있다. 지도학습은 입력 및 출력 값을 갖는 자료를 활용하여 학습하는 방법으로 분류(Classification), 회귀(Regression)분석에 응용된다. 대표적으로는 나이브 베이즈(Naïve Bayes), 서포트벡터머신(Support Vector Machine), 의사결정나무(Decision Tree), 인공신경망(ANN), 릿지 회귀(Ridge Regression)가 있다(배성완・유정석, 2018).

비지도 학습은 출력 값이 알려지지 않은 데이터를 컴퓨터가 스스로 학습하여 데이터 내부의 상호관계와 패턴을 찾는 학습 알고리즘이다. 대표적으로 텍스트랭크(TextRank), 주성분분석(Principal Component Analysis), 비음수 행렬 분해(Non-negative Matrix Factorization), k-평균 군집(k-means Clustering), DBSCAN (Density Based Spatial Clustering of Applications with Noise)이 있다(배성완・유정석, 2018).

지도학습이 비지도학습과 기본적으로 다른 점은 결과 값을 알 수 있는 데이터를 활용한 학습인지 아닌지이다. 지도학습과 비지도 학습 이외에 강화학습의 학습 유형도 존재한다. 강화학습은 컴퓨터 알고리즘이 스스로 답을 찾기 위해 환경과 상호작용하면서 보상이 강화되는 방향으로 진행되는 학습이다. 학습이 해답에서 멀어질수록 벌점이 부과되는 성질을 이용한다.

3. 선행연구 검토

비정형 빅데이터를 이용하여 현상의 설명 및 예측을 수행한 연구는 두 가지 방향에서 진행되었다. 우선, 언론사 및 소셜 미디어에서 추출한 텍스트 데이터를 가공 및 분석 과정을 통해 시장 변화의 예측에 활용하는 방법과 글로벌 포털사이트인 구글, 페이스북, 트위터, 네이버, 카카오 등에 저장된 접속 및 검색기록의 빈도 수를 계측하여 보이지 않는 정보를 창출 및 해석하는 연구이다. 본 연구는 언론사에서 제공하는 뉴스

텍스트를 이용한 선행연구를 중심으로 검토하였다. 김진유(2006)는 '투기'가 포함된 신문기사가 부동산가격에 미친 영향을 연구하였다. 분석 결과, 투기를 포함한 관련 신문기사 수가 전국 및 서울시의 아파트 가격과 인과관계를 형성하고 있음을 실증하였다. 투기와 관련되어 보도된 시기별 신문기사 건수는 부동산가격과 상호간에 영향을 미치는 양방향 그랜저인과관계가 있다. 그러나 이 연구는 투기라는 단어에만 한정되고, 기사의 내용과 논조보다는 단순히 보도된 기사건수만 대상으로 하는 한계가 있다.

우윤석 · 이은정(2011)은 부동산 시장의 참여자들이 갖는 가격상승 등의 기대심리에 크게 영향을 주는 요인으로 언론보도에 주목하였다. 언론보도에 따라 시장 참여자들이 갖는 기대가 정책의 효과에 유의미한 영향을 준다고 주장하였다. 이를 검증하기 위해 시기별로 집계된 언론 보도 수가 주택가격의 변동에 미친 영향을 분석하였다. 분석 결과, 강남 아파트 가격의 상승과 관계되는 언론기사가 기타 지역의 아파트 가격의 상승을 시차를 두고 이끈다고 주장하였다.

김대원 · 유정석(2016)은 트위터에서 만들어져 유통된 정보와 아파트의 매매 및 전세가 사이의 관계를 그랜저인과관계와 벡터자기회귀모형을 이용하여 분석하였다. 트위터 자료를 수집하여 월별 상승 및 하락 단어의 빈도수를 산정하고, 전국과 서울지역 아파트 가격과의 관계를 분석하였다. 연구 결과, 트위터 정보는 아파트 매매가격 변동에 영향을 미쳤다. 또한 상승이 하락보다 매매가 변동률에 더 큰 영향을 주는 것을 확인하였다.

진창하 · Gallimore(2012)는 부동산시장 참여자들이 객관적 정보와 함께 직관과 투자심리에도 영향을 받는다고 생각하였다. 이 가설 하에 미국 애틀랜타 CMSA를 사례로 기사 내용과 부동산가격 사이의 관계를 그랜저인과관계 검정과 오차수정모형을 이용하여 분석하였다. 이 연구는 긍정적 용어보다 부정적 용어를 사용한 신문기사의 내용이 주택시장의 가격 변화와 더 높은 연관성을 보인다는 것을 밝혔다.

부동산 시장 이외의 분야에서 뉴스기사를 이용한 선행연구를 살펴보면, 송치영(2002)은 뉴스가 주식 및 외환시장에 미치는 영향을 실증분석하고 주식시장에 더 큰 영향력이 있음을 확인하였다. 그리고 Li et al.(2016)은 로이터에서 수집한 오일 관련 뉴스를 감성분석과 빅데이터 분석도구를 사용하여 뉴스기사에서 도출한 지수가 실제 오일 가격과 방향성이 동일함을 확인하는 결과를 도출했다. 이와 같이 다양한 분야에서 뉴스기사를 활용한 연구가 활발히 진행되고 있다.

선행연구는 대부분 특정 단어를 포함한 기사의 수가 부동산 가격에 미치는 영향을 분석하는데 주목하였다. 그러나 신문기사와 SNS 등은 내용과 논조가 있어 이를 고려하지 않고 부동산시장과의 영향 관계를 명확히 밝히기는 어렵다. 최근에는 신문기사 등의 내용을 분석하기 위한 시도가 있으나, 부동산 분야에서는 이를 분석하기 위한 방법론을 정립한 연구는 거의 없다.

이 연구는 부동산시장 참여자들의 심리를 파악하는 국토연구원의 부동산 소비자 심리지수와의 차별성도 있다. 부동산 소비자 심리지수는 부동산 중개업소와 일반 가구를 대상으로 전화 설문조사를 통해 구득한 자료를 활용한다. 하지만 이 지수는 부동산시장의 주요 지표인 가격과 거래를 분리하거나 지수를 산정하면서 시장 상황을 고려한 가중치를 적용하기 어렵고, 월별 자료로써 분석의 즉시성에 한계가 있다. 본 연구에서 개발하는 새로운 감성지수 모델은 부동산시장에 대한 사람들의 심리 및 의견 형성의 토대가 되는 신문기사 문장에 대한 내용과 논조를 분석하여 지수를 도출한다. 이 방법은 비정형 텍스트 데이터를 이용하여 월간, 주간, 일간까지 분석할 수 있는 유연성도 있다.

III. 분석 절차 및 방법

1. 분석 절차 및 자료

1) 용어 정의

감성은 외부에서 들어오는 자극에 대해 사람이 오감으로 느끼는 감정의 변화를 의미한다. 이득환 외(2013)는 주식시장에서 주가정보에 포함되는 감성을 분노, 미움, 싫음, 사랑, 두려움, 수치심, 슬픔, 바람, 기쁨으로 분류하였다.

일반적으로 긍정과 부정을 나타내는 단어들은 '즐겁다', '아름답다', '재밌다', '멋있다', '지루하다', '졸리다', '망치다', '슬프다' 등과 같다. 하지만 부동산과 같은 경제를 다루는 신문기사나 뉴스에서는 주로 가격의 상승과 하락을 예측하거나 현재의 상황을 묘사하는 단어들이 많이 등장한다.

이런 경향을 반영하여 본 연구에서는 부동산 시장의 가격 상승 움직임을 나타내는 단어인 '오르다', '상승', '활황', '최고' 와 같은 단어들은 긍정단어로, '내리다', '하락', '위축', '최저' 와 같은 단어들은 부정 단어로 정의하였다. 즉 본 연구에서 의미하는 감성은 아파트 매수자의 관점에서 신문기사에 나타나는 부동산과 관련된 긍정적 문구와 부정적 문구에 대한 느낌으로 정의한다. 긍정적 문구는 아파트 매매가격이 상승한다는 뉘앙스를 포함한 문장을 말한다. 반대로 부정적 문구는 아파트 매매가격이 하락한다는 뉘앙스를 포함한 문장을 의미한다. 또한 기준의 모호성과 조작 가능성을 배제하기 위하여 연구자를 포함한 3인이 같은 문장에 대해 긍정과 부정의 극성을 부여하고 이를 교차 검증하는 절차를 수행하였다.

긍정적인 문장과 부정적인 문장을 토픽분석과 텍스트랭크, 나이브 베이즈 등 기계학습 기법을 활용하여 긍정적 문장에 대한 긍정 감성과 부정적 문장에 대한 부정 감성을 계량화하고, 이들 간의 차이를 감성지수로 정의한다.

2) 분석 절차

부동산시장 심리지수인 감성지수는 <그림 1>과 같이 신문기사 수집 및 전처리, 기계학습을 활용한 감성분석, 각 문장에 대한 감성지수 도출, 감성지수 도출의 순서로 이루어진다. 우선, 일간지와 경제지 웹사이트의 부동산 관련 텍스트 기사 중 '아파트'와 '매매'가 포함된 신문기사를 웹 크롤링(Web Crawling)한다.

2단계로 불용어와 특수문자 제거 같은 전처리 절차를 수행한 후 토픽분석을 통해 수집된 신문기사 빅데이터의 내용을 8개 토픽과 토픽별 30개 단어를 추출한다.

3단계로 토픽분석에서 추출한 240개의 각 단어가 포함된 총 9,600개 문장을 무작위로 추출하여 감성사전 구축을 위한 샘플 데이터를 만든다. 그리고 연구자를 포함한 3인이 상의하여 모든 샘플 문장의 긍정 또는 부정을 판별한다.

4단계로 텍스크랭크(TextRank)를 이용하여 샘플 문장에 포함된 단어들 사이의 관계를 파악하고, 이를 이용하여 감성사전을 구축한다. 5단계로 TF-IDF를 활용하여 분석하고자 하는 문장에 포함된 단어들의 점수를 산정하고, 이를 나이브 베이즈 분류 모델의 입력값으로 산입한다.

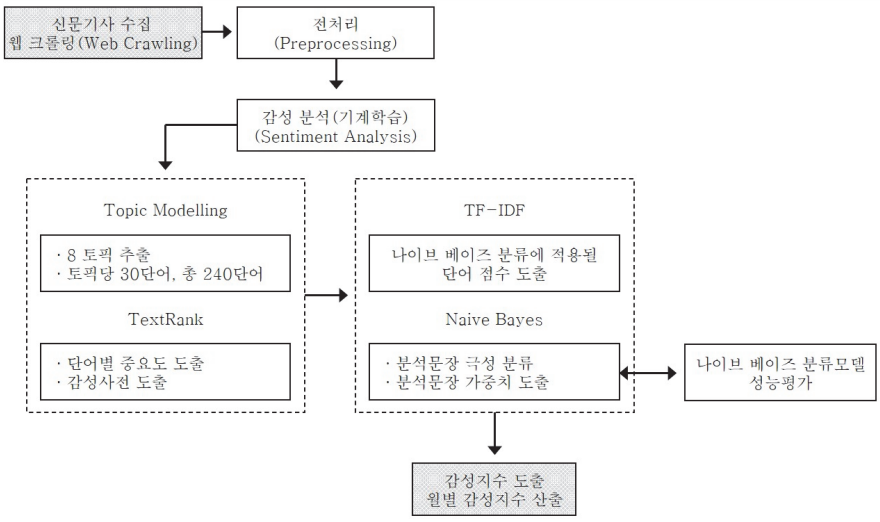
6단계로 나이브 베이즈 분류 모델을 적용하여 문장에 긍정과 부정의 극성을 부여하고, 문장의 긍정 또는 부정의 가중치를 산출한다.

7단계로 감성지수 산정 식에 따라 월단위로 신문기사 전체에 대한 최종 감성지수를 도출한다. 마지막으로 ROC와 정확도 값을 비교하여 분류 모델의 성능을 검증한다.

3) 분석 자료

이 연구는 2012년부터 2018년 말까지 84개월 간 영향력이 높은 국내 일간지와 경제지를 적절히 고려하

<그림 1> 분석 절차도



여 각각 주요 3개 일간지와 경제지를 선정하였다. 일간지는 조선일보, 동아일보와 중앙일보이고, 경제지는 매일경제신문, 한국경제신문, 서울경제신문이다. 한국ABC 협회의 2017년도 일간신문 발행자료에 의하면, 유료부수를 기준으로 조선일보, 동아일보, 중앙일보가 각각 1위, 2위, 3위를 차지하고, 매일경제는 4위 한국경제는 5위, 서울경제는 24위이다.

조선일보에서 기사 3,129건과 67,162건의 문장, 동아일보에서 기사 2,560건, 46,134건의 문장, 중앙일보에서 기사 2,563건과 78,868건의 문장을 추출하였다. 한국경제에서 기사 8,264건과 문장 147,363건, 매일경제에서 기사 7,847건과 문장 137,714건, 서울경제에서 기사 4,031건과 문장 75,096건을 추출하였다.

2. 분석 방법

1) 토픽 모델(Topic Model)

토픽 모델은 텍스트 문서에 숨겨진 주제(Topic)를 분석하기 위한 통계 모델이다. 텍스트 문서 내용에 내재한 의미구조를 추출하기 위해 이용되는 텍스트 마이닝 기법 중 하나이다(안정욱 외, 2015). 특정 주제와 관한 문서에서 해당 주제에 관련되는 단어가 다른 단어들보다 더 많이 나타난다. 문서 내에 특정한 주제가 포함되어 있고, 주제 간 비중이 어느 정도인지는 문서 집합 내의 단어 통계에 대한 분석을 통해 파악할 수 있다. 토픽 모델은 문서를 구성하는 키워드들을 활용하여 문서를 구성하는 주제를 찾아내기 위한 분석방법이다. 특히 대규모 문서 집합에 활용되는데, 다양한 종류의 데이터에도 활용할 수 있다(신규식 외, 2015).

토픽 모델의 대표적 방법론은 잠재디리클레할당(Latent Dirichlet Allocation: LDA)이다(차윤정 외, 2015). LDA 알고리즘은 생성모델로 문서 내의 숨어 드러나지 않는 주제들을 찾는 모델이다. 문서와 단어 등 관찰된 변수를 통해 문서 구조와 같은 관찰되지 않는 변수를 추론하는 것을 목적으로 한다. <그림 2>는 이 연구에서 적용한 토픽분석 절차이다.

1단계로 신문사 웹 사이트에서 부동산 관련 뉴스를 웹 크롤링하여 분석 자료를 수집한다. 수집된 분석 자료에서 분석 대상인 아파트와 매매 관련 기사만을 분류하기 위하여 '아파트'와 '매매'가 포함된 뉴스기사만 재분류하여 최종 분석 대상을 추출한다. 2단계로 파이썬(Python) 프로그램을 이용하여 수집된 뉴스기사에서 분석에 필요 없는 불용어와 특수문자, 영문과 숫자를 제거한다.

3단계는 불용어를 적절히 처리한 뉴스기사를 이용하여 형태소 분석을 실시하여 명사만 추출한다. 4단계는 LDA 방법을 이용하여 뉴스기사에 대한 토픽분석을 실시한다. 마지막 단계에서는 토픽분석 결과로 8개 토픽과 각 토픽별로 30개 단어를 추출한다.

2) 텍스트랭크(TextRank)

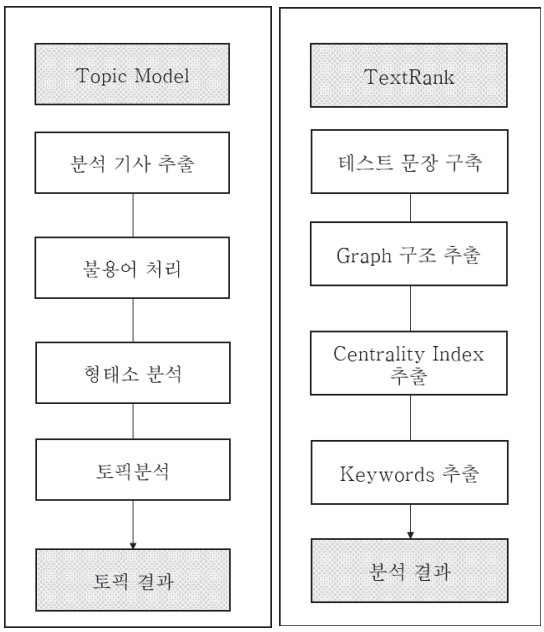
텍스트랭크는 텍스트를 기반으로 하는 그래프 기반 순위 모델이다. 문서의 요약과 단어의 추출에 활용되는 모델로 구글 검색 엔진인 페이지랭크(PageRank)에 가중치를 부가한 알고리즘이다(배영준 외, 2018). 최근 언어 특성에 제약받지 않는 비지도학습 알고리즘에 기반한 문서 요약 기법이 개발되는데, 이 중 대표적인 기법이 텍스트랭크이다(Brin and Page, 1998).

텍스트랭크 알고리즘은 A 텍스트에서 B 텍스트로 연결되는 링크를 A가 B에게 던진 하나의 표로 산정하여 특정 텍스트가 얻은 총 득표수를 준거로 중요도를 부여한다. 텍스트랭크의 값은 텍스트의 중요도를 감안하여 중요한 텍스트로부터 표를 받으면 링크된 텍스트에 더 큰 값을 부여한다. 식은 다음과 같다(이현우 외, 2009).

$$TR(p_i) = (1-d) + d \sum_{p_j \in m(\pi)} \frac{TR(p_j)}{L(p_j)} \qquad (1)$$

$TR(p_i)$와 $TR(p_j)$는 특정한 텍스트가 갖는 텍스트랭크의 값이다. 본 연구는 <그림 3>과 같이 텍스트랭크 알고리즘을 적용하여 단어들의 중요도 점수를 도출하고 감성사전을 구축하는데 활용한다. 1단계로 토픽분석에서 산출한 단어들이 포함된 9,600개 문장을 무작위로 추출한다. 2단계로 문장을 구성하는 단어 사이의 관계를 그래프 구조로 변환한다. 3단계는 그래프의 기하학적 구조를 이용하여 각 단어 노드의 중요도인 중심성 지수(Centrality Index)를 산정한다. 마지막으로 중심성 지수를 활용하여 상위의 n개 키워드, 즉 단어를 선택하여 감성사전을 도출한다.

<그림 2> 토픽분석 절차 <그림 3> TextRank 분석 절차



3) TF-IDF 분석

TF-IDF 분석은 텍스트 마이닝의 절차 중 중요 단어를 추출하고 이를 기반으로 단어사전을 만들 때 많이 사용된다. 이 분석은 여러 문서에서 출현한 단어별 중요성을 점수화한다. TF(Term Frequency)는 문서에 출현하는 단어의 빈도수를 나타낸다. IDF(Inverse Document Frequency)는 역문헌 빈도를 의미하고 특정 문서에만 많이 출현하는 정도를 나타낸다.

TF-IDF는 일반적으로 TF와 IDF를 곱한 값으로 정의한다. 단어의 중요도를 평가하는 척도는 중요한 단어일수록 그 빈도수가 많다고 판단한다. 따라서 단어 빈도(Term Frequency: TF)가 중요한 척도가 될 수 있다. 그러나 모든 문서에서 자주 출현하는 단어는 중요하지 않은 단어인 경우가 많고 빈도수는 문서의 수에 비례하여 증가하므로 좋은 척도가 아니다.

따라서 이 연구는 역문헌 빈도(IDF)를 도입한다. IDF는 전체 분석대상 문서 수를 특정 단어가 포함된 문서의 수로 나눈 값을 로그로 변환한 값이다. 이로 인해 특정한 문서에만 나타나는 단어의 점수가 높게 나타난다. 그러므로 TF와 IDF를 곱한 TF-IDF는 특정한 문서 내의 빈도수가 높을수록, 그리고 전체 분석대상 문서에서는 빈도수가 낮을수록 값이 높아지게 된다. 이 특성으로 인해 TF-IDF는 단어의 중요성을 평가하는 좋은 척도이다. <그림 4>는 이 연구에서 분석한 TF-IDF의 분석 절차도이다.

우선, 감성지수를 산출하기 위한 테스트 문장에서 불용어를 처리하고 형태소 분석을 수행하여 분석에 이용되는 단어를 선택한다. 이후 선정된 단어에 대한 TF-IDF 점수를 산정한다. 마지막 단계에서 각 문장에 적용된 단어별 TF-IDF 점수를 매트릭스로 구축한다.

4) 나이브 베이즈(Naïve Bayes) 분류 모델

정보사회에서는 기술 발전에 따라 노출되는 정보가 폭증하면서 많은 대상 중에서 각자 원하는 조건에 부합하는 정보를 찾는 일이 흔해지고 있다. 이에 따라 추천 시스템에 대한 사회적 관심과 연구가 최근에 증가하고 있다. 아마존(Amazon)이 제공하는 상품 추천과 넷플릭스(Netflix)가 제시하는 영화 추천 시스템이 구매자의 특성에 맞는 정보를 추천하는 대표적인 사례이다(유석종, 2019).

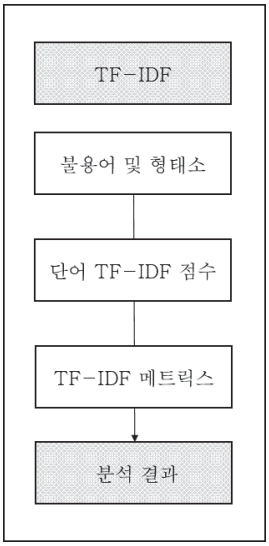
본 연구는 사람들의 많이 접하고 있는 주요 언론사의 온라인 뉴스를 이용하여 비정형 텍스트 데이터인 부동산 기사에 이를 적용하고자 한다. 이 목적을 달성하기 위하여 나이브 베이즈 분류(Naïve Bayes Classification) 모델을 적용한다. 이 분류 모델은 단어가 서로 독립적이라고 전제하면 특정한 단어들을 포함한 문서가 어느 주제에 해당하는지를 분류하는 모델이다. 특정 단어를 포함한 문서가 속하게 되는 확률이 가장 큰 주제로 분류하게 된다. 아래 식 (2)와 같이 베이즈 정리(Bayes' Theorem)를 활용한 나이브 베이즈 분류 모델은 자료량이 증가할수록 정확도가 높아지는 특성이 있다.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1)+\ldots+P(A_n)P(B|A_n)} \quad (2)$$

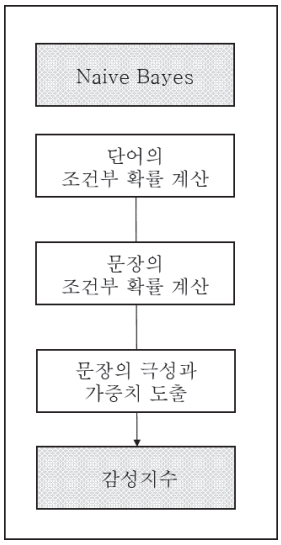
<그림 5>는 이 연구에서 적용한 나이브 베이즈 분류를 위한 절차도이다. 먼저 문장에 포함된 단어가 긍정 또는 부정에 속하게 될 확률을 계산한다. 다음 단계는 문장을 구성하는 전체 단어가 문장을 긍정 또는 부정 문장으로 판단할 조건부 확률을 계산한다. 마지막으로 분석 대상 문장의 긍정과 부정 가중치를 비교한 후 이

중 큰 값을 나타내는 쪽을 해당 문장의 극성 및 감성지수로 산출한다.

<그림 4> TF-IDF 분석 절차



<그림 5> 나이브베이즈 분석 절차



Ⅳ. 분석 결과

1. 전처리 및 토픽분석 결과

형태소는 언어의 형태론적 수준에서 최소단위를 말한다. 어절을 형태소 단위로 구분하고 개별 형태소에 맞는 범주 값을 부여하는 것이다(심광섭 · 양재형, 2004). 텍스트 자료에서 작성자의 감정 또는 의견을 추출하기 위해서 형태소로 구분하고, 형태소별 극성을 판단한 후 텍스트 전체의 긍정 또는 부정의 극성을 분류하는 방법을 이용한다(김동영 외, 2014).

이 연구는 파이썬 KoNLPy(Version 0.5.2)를 이용하여 형태소 분석을 시행하였다. 뉴스기사에서 불용어 등을 제거하고, 토픽분석에서 선정된 단어가 포함된 문장을 활용하여 감성사전을 구축하기 위한 테스트 데이터로 이용하였다.

감성분석을 수행하기 위해서 먼저 각 문장을 대상으로 형태소 분석을 실시한다. <표 1>과 같이 분석에 필요하지 않은 불용어 제거와 유사어를 통합하는 작업을 진행한다.

<표 1> 불용어 및 유사어 예시

구 분	단 어	
불용어	조사	은, 는, 이, 가
	숫자	0, 1, 2, 3, · · · · · · , 9
	특수문자	%, △, +, -, 등
	영문	a, b, c, · · ·, z, A, B, · · ·, Z
	불필요한 단어	개월, 건물, 공단, 대해, 리포터, 만원, 억원, 등
유사어	가격	가격, 값
	거래	거래, 거래량
	공시	공시지가, 공시가격
	금리	대출금리, 기준금리, 주택담보대출금리, 주담대
	급매물	급매물, 급매
	매도	매도, 매도세, 팔자
	회복	회복, 회복세

부동산 관련 신문기사의 내용을 주제별로 파악하기 위해서 토픽 모델링 기법의 대표적 방법인 LDA를 적용하였다. 먼저 <그림 6>과 같이 워드클라우드로 토픽분석을 통해 얻은 주요 단어들을 확인하였다.

<그림 6> 신문기사 워드클라우드

토픽분석 결과 중 4개 토픽만 선정하여 토픽별 주요 단어를 예시로 제시하면 <표 2>와 같다. 2012년부터 2018년 말까지 기사를 월 단위로 구분하지 않고 전체 신문기사를 이용하여 토픽분석을 실시하였다. 각 토픽을 구성하는 주요 단어들을 종합하여 개별 토픽의 주제로 제시하였다.

Topic 1은 아파트 가격 및 거래와 관련된 주제이다. Topic 2는 주택금융과 관련된 단어가 주를 이룬다.

Topic 3은 분양과 관련된 주제이고, Topic 4는 재건축과 관련된 주제로 분석된다, Topic 5는 부동산 투자와 관련된 단어, Topic 6은 상가매매와 관련된 단어가 주를 이룬다, Topic 7은 정부의 부동산 정책과 관련된 단어, 그리고 Topic 8은 외국인의 부동산 투자와 관련된 단어가 주를 이룬다.

<표 2> 토픽분석 결과 예시

토픽1	토픽2	토픽3	토픽4
가격	주택	분양	재건축
주택	대출	청약	강남
전세	금리	지구	가격
매매	담보	인근	상승
거래	소득	개발	하락
토픽5	토픽6	토픽7	토픽8
투자	면적	정부	외국인
오피스텔	상가	규제	제주
수익	매매	거래	호텔
가격	공인	대책	주택
대표	보증금	경매	직원

주: 토픽별로 상위 5개 단어만 제시함.

2. 텍스트랭크 및 감성사전 구축

1) 문장 샘플링과 극성 판별

토픽분석을 통해 부동산과 관련된 신문기사가 어떤 단어들로 구성되는지 파악하였다. 토픽분석에서 도출된 240개 단어 중 각 단어가 포함된 문장을 40개씩 무작위로 추출하여 총 9,600개 샘플 문장을 추출하였다. <표 3>과 같이 부정적 문장에는 부정 태그, 긍정적 문장에는 긍정 태그를 붙이는 작업을 실시하였다.

연구자의 주관적 판단에 따라 문장에 극성을 부여하는 주관성의 문제를 방지하기 위하여 연구자를 포함한 3인이 9,600개 테스트 문장 중에서 100개를 임의 추출하여 홀스티(Holsti) 공식을 이용한 신뢰도를 검증하였다. 홀스티 계수는 0.9 이상이면 신뢰도가 있다고 판단할 수 있다(유수정 외, 2016). 부동산 뉴스의 속성상 가격의 상승 또는 하락을 표현하는 단어들이 분명하기 때문에 다른 경우는 거의 없다. 그러나 '보합'이 들어간 문장이나 한국이나 국제 경제와 관련된 문장에 의견이 다른 경우가 있어 홀스티 계수가 <표 4>와 같이 0.87로 나타났으나, 이는 0.9에 근접한 것으로 판단하였다.

<표 3> 문장별 긍정 및 부정 판별 결과 예시

단어	문 장	극성
가격	한국감정원에 따르면 전월 대비 전국 주택 가격 상승률이 모두 상승세를 유지하고 있다	긍정
	보합세를 이어가던 아파트 가격이 8주 만에 하락세로 돌아섰다	부정
거래량	주택 거래량이 늘어나고 매매가격이 상승세로 전환하는 등 주택시장은 회복 기미를 보이는 것으로 분석됐다	긍정
	정부의 대출 규제로 거래시장이 위축되면서 주택 매매 거래량은 작년 94만 7천 건보다 13.4% 줄어든 82만 건으로 예상했습니다	부정
규제	정부의 매매전환 지원책과 부동산 규제 완화책도 약발이 먹혔다	긍정
	강력한 대출 규제와 다주택자들에 대한 종부세 강화 등 9.13 대책의 영향이 본격화하고 있다는 분석입니다	부정
급매	목동 공인 관계자는 급매물이 없어 호가가 쉽게 내려가지 않는다고 말했다	긍정
	금리까지 오르면 갭 투자자들이 아파트를 급매로 내놓을 수도 있다	부정
상승	지난 6월부터 상승세가 가팔라지고 있다	긍정
	재건축 아파트는 7월 1.24의 4분의 1 수준인 0.34%로 상승률이 급락했다	부정
하락	2011년 3월 이후 하락세를 이어온 분당 등 1기 신도시 아파트 매매가격도 0.02% 올라 25개월 만에 상승세로 돌아섰다	긍정
	서울 아파트값은 하락세가 계속되고 있습니다	부정

<표 4> 문장별 신뢰도 계수 분석 결과

온라인 뉴스의 문장 특성	홀스티 신뢰도 계수
긍정단어가 포함된 문장	1.00
부정단어가 포함된 문장	1.00
'보합'이 포함된 문장	0.87
한국경제 상황에 대한 문장	0.85
국제경제 상황에 대한 문장	0.80

2) 텍스트랭크(TextRank) 분석 결과

부동산 뉴스의 감성분류 전에 핵심 내용을 반영하는 부동산시장의 변화와 관련된 특정 단어(Feature)를 추출하기 위해 텍스트랭크를 이용하였다. 텍스트랭크는 단어 사이의 관계 네트워크에서 단어의 중요성을 산출하므로 더 많은 정보를 축약할 수 있다. 신문기사

에서 추출한 문장을 대상으로 한 형태소 분석, 불용어 처리와 유사어 적용 작업을 완료하고, 단어별 중요도 점수를 산출하면 <표 5>와 같다.

텍스트랭크에서 단어에 부여된 중요도가 높은 단어들을 이용하여 <표 6>과 같이 감성사전을 구축한다. 감성사전을 구축하는데 있어 연구자의 주관성을 제거함으로써 객관성을 극대화하기 위함이다.

<표 5> 텍스트랭크 결과 예시

Word	Value	Rank
아파트	7.7654	9
가격	6.1853	20
주택	5.1490	28
부동산	4.4967	31
집값	4.7715	41
거래	3.9812	43
상승	3.3070	46
규제	3.5770	58
대출	3.0300	64
하락	2.9481	65
재건축	2.4224	93

<표 6> 긍정 및 부정 감성사전 구축 예시

긍정 감성사전	부정 감성사전
거래, 최고치	가격, 하락
관심, 커지다	거래, 위축
규제, 완화	거래, 급감
높은, 청약, 경쟁률	거래, 가격, 약세
대출, 금리, 인하	거품, 우려
매매, 가격, 오르다	공급, 과잉, 먹구름
매매, 부동산, 규제, 완화	규제, 강화
매물, 거두다	금리, 인상, 부담
매물, 품귀, 가격, 오르다	금리, 인상, 집값, 하락
신규, 분양, 활기	아파트, 매매, 가격, 하락
아파트, 품귀	재건축, 아파트, 호가, 하락
취득세, 면제	주택, 불확실, 커지다
취득세, 인하	집값, 거품, 미분양, 증가
투자, 문의, 증가	집값, 하락, 가능성
호가, 상승	호가, 내려, 매물, 내놓다

3. TF-IDF 와 나이브 베이즈 분류 분석

토픽분석에서 단어를 추출하고 텍스트랭크 분석을 통해 감성사전을 구축하였다. 이를 바탕으로 문장에 대해 TF-IDF 알고리즘과 나이브 베이즈 분류 분석을 실시하여 긍정 및 부정점수를 도출한다.

<표 7>은 나이브 베이즈 분류기를 이용해 신문기사 문장에 대한 분석 결과 예시이다. 이 연구에서 나이브 베이즈 분류의 목표값(Target)은 예측할 부동산 뉴스의 긍정 또는 부정 극성이다. 또한 특성 값(Feature)은 나이브 베이즈 분류 모델에 투입되는 특정 단어이다.

나이브 베이즈 분류 분석의 가중치는 특정 단어로 구성된 문장이 목표값의 긍정 또는 부정에 미치는 영향을 의미한다. 이 단어로 구성된 문장이 긍정 또는 부정으로 판단될 확률에 영향을 미치는 정도를 나타낸다. <표 7>에서 '분양시장의 열기는 계속되고 있다'는 문장은 '계속'에 0.63, '분양'에 0.54, '열기'에 0.55를 부여하고, 나이브 베이즈 모델에 투입되어 긍정 극성과 이에 따른 가중치 0.8604가 도출된다. 이렇게 도출된 가중치를 긍정과 부정의 감성지수를 산출하는 근거자료로 활용한다.

<표 7> 나이브 베이즈 분류 분석 결과

	문장	가중치
긍정	KB국민은행에 따르면 지난달 말 기준 고가 아파트 시가총액 기준 50개 단지의 매매가격은 1년 전보다 1.25% 올랐다	0.5637
	집값이 상승세를 타면서 9월 말 현재 매매값이 전고점의 98.3% 수준으로 바짝 다가섰다	0.6938
	분양시장의 열기는 계속되고 있다	0.8604
	송파구는 잠실 지역 재건축단지와 새 아파트가 모두 올랐고 서초구는 주상복합 단지에 투자자들이 몰리면서 상승했다	0.8250
	은평구는 수색역세권 개발과 카톨릭병원 개원 예정 등 개발 호재로 상승했다	0.9186
부정	강남구 개포동 개포주공 기준 아파트는 11억8000만원으로 12억원에서 대책 발표 전보다 2,000만원 하락했다	0.5825
	1.27 부동산대책에 따른 기대심리로 값이 일시적으로 올랐던 강남지역 재건축 아파트를 중심으로 가격이 빠졌다	0.6297
	부동산114에 따르면 이번 주 서울 재건축 아파트 가격은 0.25% 하락했다	0.7046
	다주택자에 대한 양도소득세 중과 시행 여파로 서울 아파트 거래량이 감소세를 지속하면서 거래절벽이 심화되는 모습이다	0.8915
	서울 아파트 매매시장은 다주택자 양도세 강화 여파로 관망세로 돌아서 거래량이 절반 이상 급감해 반토막이 난 바 있다	0.9194

4. 감성지수 산출과 분류 성능 검증

1) 감성지수 산출

이 연구는 감성분석(Sentiment Analysis)을 이용하여 부동산 감성지수를 산출하는 모형을 제시한다. 감성분석은 긍정과 부정의 극성을 판별 및 추출하고 범주화 및 분류하여 정량화하는 작업이다(남길임 · 보은경, 2017). 감성사전의 구축 방법은 연구자 직관 중심, 통계적 방식 또는 기계학습 기반의 감성사전 구축으로 분류할 수 있다(서덕성 외, 2017).

이 연구는 통계적 방식과 기계학습 모델 방법을 결합한 감성사전의 구축을 시도하였다. 감성지수를 산출하기 위하여 우선, 1개의 문장에 대한 긍정과 부정 가중치를 계산한다. 1단계에서는 나이브 베이즈 분류기법을 활용하고 문장을 감성사전과 비교하여 개별 문장에 대한 일별 가중치 합을 식 (3)과 같이 계산한다.

$$(\text{긍정})D.NPSI_p = \sum_{i=1}^{n} sentence_p(i) \quad (3)$$
$$(\text{부정})D.NPSI_n = \sum_{i=1}^{n} sentence_n(i)$$

2단계는 1단계에서 일별로 산출한 문장의 긍정 또는 부정 가중치의 합을 해당 월까지 확장한다. 해당 월에 속하는 모든 문장의 긍정 또는 부정 가중치 합하여 해당 월의 날짜로 나누어 월별 가중치의 평균을 식 (4)와 같이 산정한다.

$$(\text{긍정})M.NPSI_p = \frac{\sum_{i=1}^{n} D.NPSI_p(i)}{n} \quad (4)$$
$$(\text{부정})M.NPSI_n = \frac{\sum_{i=1}^{n} D.NPSI_n(i)}{n}$$

3단계는 2단계에서 산출한 월별 신문기사의 긍정 가중치 평균과 부정 가중치 평균의 차이를 계산한다. 평균값의 차이가 양(+)의 값이면 해당 월은 부동산 시장에 대해 전반적으로 긍정적인 감성이 나타난 것으로 해석한다. 이와 반대로 음(-)의 값이면 해당 월은 부동산 시장에 대해 전반적으로 부정적인 감성을 나타낸 것으로 판단한다.

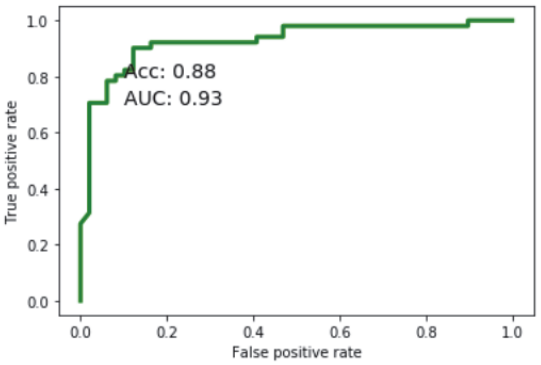
$$\text{신문감성지수}(NPSI_i) = \sum_{i=1}^{n} M.NPSI_P(i) - \sum_{i=1}^{n} M.NPSI_n(i) \quad (5)$$

<표 8>은 신문기사 문장에 대한 월별 감성지수를 산출한 결과 중 일부를 제시한 것이다.

2) 분류 성능 비교

ROC 곡선은 판별의 정확도를 도표로 검증하기 위하여 사용된다. 곡선 아래의 면적을 나타내는 AUC (Area Under Curve)는 분류에 활용되는 모형 또는 변수의 성능을 검증하기 위해 이용된다(김지현 외, 2018). ROC 곡선이 대각선에 일치할수록 AUC는 0.5에 근접하고, 1.0에 근접할수록 모형의 성능, 즉 판별의 정확도가 높음을 의미한다(Simundic, 2009).

<그림 7> ROC 그래프



<표 8> 감성분석 결과 예시

매체	구분	12년 1월	12년 2월	12년 3월
조선일보	긍정	0.7406	0.5161	0.5818
	부정	0.6107	0.6101	0.6411
중앙일보	긍정	0.7325	0.5107	0.5537
	부정	0.5787	0.5769	0.5815
동아일보	긍정	0.6298	0.5051	0.5611
	부정	0.5956	0.5892	0.5915
매일경제	긍정	0.7466	0.5107	0.5635
	부정	0.5943	0.5864	0.5880
한국경제	긍정	0.7316	0.5085	0.5516
	부정	0.5965	0.6146	0.6016
서울경제	긍정	0.7325	0.5164	0.5627
	부정	0.6274	0.6370	0.6354
평균	긍정	0.7189	0.5112	0.5624
	부정	0.6005	0.6023	0.6065
감성지수		0.1184	-0.0911	-0.0441

이 연구에서 수행한 나이브 베이즈의 성능을 ROC 곡선과 오차행렬(Confusion Matrix)을 통해 확인하였다. <그림 7>은 테스트 데이터 100개를 이용한 ROC 곡선으로 나이브 베이즈 분류기의 성능을 나타낸다. ROC 곡선에서 정확도는 88%이고, AUC는 93%로 모형의 성능이 양호한 것으로 나타났다.

분석 결과, <표 9>와 같이 전체 문장 중에서 긍정 또는 부정 문장을 정확하게 분류한 정확도(Accuracy)는 88%로 나타났다. 나이브 베이즈 모델이 긍정 또는 부정 문장으로 분류한 문장 중 실제로 긍정 또는 부정으로 판별된 정밀도(Precision)는 각각 86%, 89%로 분석되었다. 마지막으로 실제 값이 긍정 또는 부정 문장인 문장들 중 모델이 긍정 또는 부정으로 판별한 재현도(Recall)는 각각 90%, 86%로 나타났다. 유석종(2019)은 부동산 매물의 속성정보를 이용하여 상승전망을 나이브 베이즈 분류 모델로 구축하고 정확도를 계산한 결과, 평균 85%로 나타났다. 이 연구는 88%로 이보다 높아 모델의 성능이 양호한 것으로 판단된다.

<표 9> 나이브 베이즈의 오차행렬

구분		예측		
		긍정	부정	합계
실제	긍정	46	5	51
	부정	7	42	49
	합계	53	47	100
분석 결과	정확도	88%		
	긍정정밀도	87%	부정정밀도	89%
	긍정재현도	90%	부정재현도	86%

Ⅴ. 결론

이 연구는 비정형 빅데이터인 신문기사를 이용하여 빅데이터 분석방법인 토픽분석과 TF-IDF, 기계학습 방법인 텍스트랭크와 나이브 베이즈를 활용하여 부동산 감성지수 산출을 위한 모형을 개발하였다. 개발한 부동산 감성지수의 성능을 오차행렬 값을 이용하여 검증하였다.

이 연구에서 산출한 감성지수는 긍정 또는 부정을 나타내는 특정 단어의 빈도를 이용한 단순 지수를 활용하지 않고 문장이 나타내는 긍정적인 극성 또는 부정적인 극성을 분석하고, 이를 지수로 활용하기 위한 빅데이터 및 기계학습 방법 모델을 제시하였다. 선정된 특정 단어의 단순한 출현 빈도를 이용한 분석 방법보다 문서를 구성하는 문장의 긍정 또는 부정 극성을 분석하는 방법을 적용한 점에서 부동산 분야에서 이용된 기존 방법들보다 진일보한 분석 모형이라고 할 수 있다. 문장의 극성을 판별한 나이브 베이즈 분류 모델은 정확도 88%의 양호한 성능을 나타냈다.

부동산가격지수는 정책 의사 결정자에게 제공하는 정보의 적절성과 정확성의 관점에서 중요하다. 주택가격 등 부동산시장의 변동과 참여자들의 심리 변화를 파악하는 선행지표 또한 중요하다. 이에 따라 국토연구원은 주택시장을 포함한 부동산시장 수요자들의 소비심리를 조사하고 분석하여 2011년부터 부동산시장 소비심리지수를 제공하고 있다.

그러나 부동산시장 소비심리지수는 시기별로 일반가구와 부동산 중개업소를 대상으로 월별로 조사한 자료를 활용하여 산출된다. 이 지수는 월별로 산출되어 주택가격 등 부동산시장의 변화를 즉각적으로 반영하는데 한계가 크다. 또한 부동산시장 소비심리지수가 주택시장 등 부동산시장의 변화를 파악하고 분석하는데 다양하게 활용되지 못하고 있다. 심리적 요인이 부동산 시장 분석에서 중요성과 유용성이 있음에도 불구하고 가격이나 거래 등 시장 참여자의 소비심리가 부동산시장의 변화와 어떤 관계를 가지고 있는지에 대한 연구도 많지 않다.

이 연구는 비정형 빅데이터를 분석할 수 있는 기계학습 기술을 활용하여 빅데이터에 내재하는 집단의 감성을 수치화하여 정형 데이터 형태로 변환하는 새로운 방법론을 체계화하였다. 아직까지 모형 설계와 주관적 감성사전으로 시장을 분석하고 있는 현실에서 본 연구는 부동산 관련 뉴스기사 등 비정형 빅데이터 연구에 새로운 분석 체계를 제시하였다. 이를 통해 그동안 어려웠던 부동산시장 참여자들의 심리를 즉각적이고 유연하게 지수화하고, 이를 이용하여 아파트가격 등 부동산시장의 변동을 설명하거나 예측하는 토대를 마련한 점에서 연구의 의의가 있다.

대부분의 부동산 대책은 부동산 가격이 크게 상승 또는 하락하면서 부동산 시장의 불안정성이 커지고, 일반 대중의 부동산 시장에 대한 불만이 분출한 뒤에 나오는 것이 일반적이다. 하지만 매일 생산되는 부동산 관련 뉴스기사의 분석을 통한 감성지수 도출과 이를 활용한 부동산 시장 심리 파악이 이루어진다면 부

동산시장의 변화를 사전에 예측하고 부동산시장의 안정을 도모하기 위한 다양한 정책을 시기적절하게 펼치는데 도움이 될 것이다.

본 연구에 사용된 신문은 보수적이라 평가되는 주요 일간지와 경제지에 한정되어 있다. 따라서 상대적으로 진보적 성향을 나타내는 다른 신문사의 부동산 관련 데이터를 반영하지 못하였다. 또한, 자연어의 처리를 위해 필요한 한글 말뭉치가 부족하여 체계화된 연구에도 어려움이 있다. 영어와 비교하여 우리 한글은 복합적인 의미를 갖거나 난해한 표현 형태가 많은 편이다. 이에 따라 한글 형태소 분석기에 따라 다른 분석 결과가 나타날 수 있는 점도 연구의 한계이다. 본 연구에서 산출한 감성지수가 주택가격 등 부동산시장의 설명 또는 예측에 유의미한지 여부에 대한 검증이 필요하다. 국토연구원의 부동산심리지수 등 다른 지표와 부동산 시장 변화에 대한 설명력 또는 예측력을 비교·분석하는 후속 연구도 필요하다.

논문접수일 : 2021년 2월 24일
논문심사일 : 2021년 3월 2일
게재확정일 : 2021년 3월 18일

참고문헌

1. 강소랑 · 최은영, “베이붐 세대와 이전 및 이후 세대 간 비교분석”, 「춘계학술대회발표논문집」, 한국정책분석평가학회, 2016, pp. 157-184
2. 경정익, “부동산분야의 빅데이터 할용 방안과 정책적 제언”, 「부동산경영」 제10집, 한국부동산경영학회, 2014, pp. 65-97
3. 김구회 · 김기홍 · 이주형, “아파트 규모별 하위시장과 소비심리지수의 선행성 및 인과성에 관한 연구”, 「한국산학기술학회논문지」 제17집 제4호, 한국산학기술학회, 2016, pp. 682-691
4. 김구회 · 김기홍 · 이주형, “주택시장 소비심리지수의 주택하위시장 및 경매시장과의 영향관계에 관한 실증연구-서울 및 수도권 아파트 시장을 대상으로”, 「GRI연구논총」 제18집 제1호, 경기연구원, 2016, pp. 147-167
5. 김대원 · 유정석, “트위터 정보와 아파트 매매 및 전세 가격 간 동적 관계 분석”, 「도시행정학보」 제29집 제1호, 한국도시행정학회, 2016, pp. 1-33
6. 김동영 · 박제원 · 최재현, “SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구”, 「한국IT서비스학회지」 제13집 제3호, 한국IT서비스학회, 2014, pp 221-233
7. 김민희, “검색데이터 주택시장의 단기예측에 유용하다”, LGERI 리포트, 2014
8. 김지현 · 신선화 · 강현철, “판별모형의평가에서 ROC 곡선과 AUC의 활용에 대한 사례 연구”, 「Journal of The Korean Data Analysis Society」 제20집 제2호, 2018, pp. 609-619
9. 김진유, “신문기사가 부동산가격변동에 미치는 영향 -‘투기’가 포함된 신문기사와 주택가격간의 그랜저인과관계분석을 중심으로”, 「주택연구」 제14집 제2호, 한국주택학회, 2006, pp. 39-63
10. 김재휘, 「광고심리학」, 커뮤니케이션스북스, 2009
11. 남길임 · 보은경, 「한국어 텍스트 감성 분석」, 커뮤니케이션스북스, 2017
12. 박재수 · 이재수, “아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석-온라인 뉴스기사의 비정형 빅데이터를 활용한 감성분석 기법의 적용”, 「국토계획」 제54집 제1호, 대한국토 · 도시계획학회, 2019, pp. 131-147
13. 배성완 · 유정석, “기계 학습을 이용한 공동주택 가격 추정: 서울 강남구를 사례로”, 「부동산학연구」 제24집 제1호, 한국부동산분석학회, 2018, pp. 69-85
14. 배영준 · 장호택 · 홍태원 · 이해연, “향상된 Text- Rank 알고리즘을 이용한 자동 회의록 생성 시스템”, 「한국정보전자통신기술학회논문지」 제11집 제5호, 한국정보전자통신기술학회, 2018, pp. 467-474
15. 서덕성 · 모경형 · 박재선 · 이기창 · 강필성, “워드임베딩과 그래프 기반 준지도학습을 통한 한국어 어휘 감성 점수 산출”, 「대한산업공학회지」 제43집 제5호, 대한산업공학회, 2017, pp. 330-340
16. 송민채 · 신경식, “뉴스기사를 이용한 소비자의 경기심리지수”, 「지능정보연구」 제23집 제3호, 한국지능정보시스템학회, 2017, pp. 1-27
17. 송치영, “뉴스가 금융시장에 미치는 영향에 대한 연구”, 「국제경제연구」 제8집 제3호, 한국국제경제학회, 2002, pp. 1-34
18. 신규식 · 최회련 · 이홍철, “신재생에너지 동향 파악을 위한 토픽 모형 분석”, 「한국산학기술학회논문지」 제16집 제9호, 한국산학기술학회, 2015, pp. 6411-6418
19. 심광섭 · 양재형, “인접 조건 검사에 의한 초고속 한국어 형태소 분석”, 「정보과학회논문지:소프트웨어 및 응용」 제31집 제1호, 한국정보과학회, 2004, pp. 89-99
20. 안정욱 · 이규현 · 김희웅, 2015, “정보시스템 연구 트렌드 변화 분석: 토픽모델링 과 네트워크 분석”, 「한국경영정보학회학술대회논문집」 2015년 11월, 한국경영정보학회, pp. 561-570
21. 우윤석 · 이은정, “언론보도와 시계열 주택가격 간의 관계에 관한 연구”, 「주택연구」 제19집 제4호, 한국주택학회, 2011, pp. 111-134
22. 유석종, “나이브 베이즈 분류를 활용한 부동산 추천기법 연구”, 「한국정보기술학회논문지」 제17집 제10호, 한국정보기술학회, 2019, pp. 115-120
23. 유수정 · 이석호 · 김균수, “O2O 관련 언론보도 내용분석을 통해 살펴본 국내 ‘ICT 저널리즘’의 현황”, 「정보통신정책연구」 제23집 제4호, 정보통신정책학회, 2016, pp. 117-149
24. 이득환 · 강형구 · 김수형 · 이창민, “빅데이터에 나타난 감성분석”, 「금융공학연구」 제12집 제2호, 한국금융공학회, 2013, pp. 79-96
25. 이상기 · 이병섭 · 박병용 · 황혜경, “나이브 베이즈 분류모델과 협업필터링 기반 지능형 학술논문 추천시스템 연구”, 「정보관리연구」 제41집 제4호, 한국과학기술정보연구원, 2010, pp. 227-249
26. 이요섭 · 문필주, “딥 러닝 프레임워크의 비교 및 분석”, 「한국전자통신학회 논문지」 제12집 제1호, 한국전자통신학회, 2017, pp. 115-122
27. 이현우 · 한요섭 · 김래현 · 차정원, “Text Rank를 이용한 키워드 정련: TextRank를 이용한 집단 지성에서 생성된 콘텐츠의 키워드 정련”, 「한국HCI학회 학술대회」 2009년 2월, 한국HCI학회, pp. 285-289
28. 정한조 · 박병화, “사전과 말뭉치를 이용한 한국어단어 중의성 해소”, 「지능정보연구」 제21집 제1호, 한국지능정보시스템학회, 2015, pp. 1-13
29. 진창하 · Paul Gallimore, “신문기사 내용과 주택가격: 인식, 사유, 그리고 투자심리”, 「부동산학연구」 제18집 제2호, 한국부동산분석학회, 2012, pp. 49-69
30. 차윤정 · 이지혜 · 최지은 · 김희웅, “소셜미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원”, 「지식경영연구」 제16집 제4호, 한국지식경영학회, 2015, pp. 69-87
31. 황윤태, “아파트 가격지수 산출에 관한 연구; 머신러닝 알고리즘을 중심으로”, 「금융연구」 제33집 제3호, 한국금융학회, 2019, pp. 51-82
32. 황의영, “주식 · 부동산값 뛰면서 지난해 가계 자산 7% 늘

어", 중앙일보, 2018. 06. 19
33. Brin, Sergey and Lawrence Page, "Anatomy of a Large-scale Hypertextual Web Search Engine", Computer Networks, Vol. 30, 1998, pp. 107-117
34. Hilbert, M. "Big Data for Development: A Review of Promises and Challenges", Development Policy Review, Vol. 34 No. 1, 2016, pp. 135-174
35. Li, Jian, Zhenjiang Xu, Lean Yu and Ling Tang, "Forecasting oil prices trends with sentiment of online new articles", Procedia Computer Science, Vol. 91, 2016, pp. 1081-1087
36. Manyika, J., Michael Chui., Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers, "Big data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute, 2011
37. Simundic, Ana-Maria, "Diagnostic Accuracy - Part1: Basic Concepts: Sensitivity and Specificity, ROC Analysis, STARD Statement", Point of Care, Vol. 11 No. 1, 2009, pp. 6-8

<국문요약>

기계학습 기술을 이용한 부동산 감성지수 개발 모형 연구

박 재 수 (Park, Jaesoo)
이 재 수 (Lee, Jae-Su)

감성분석은 비정형 텍스트 데이터에서 사람의 의견, 태도나 성향 등과 같은 정보를 추출하는 기법으로 부동산 시장에 참여자의 심리를 파악하는데 유용하다. 이 연구의 목적은 온라인 신문기사 중 부동산 관련 뉴스기사를 이용하여 부동산 시장의 변화를 설명 또는 예측할 수 있는 감성지수 모형을 개발하는 것이다. 주요 일간지와 경제지 웹사이트에서 부동산 관련 기사를 웹 크롤링하여 수집하고, 전처리 절차와 토픽분석을 통해 8개 토픽과 단어를 추출하였다. 토픽분석에서 추출한 단어가 포함된 문장을 선정하고 텍스트랭크를 이용하여 감성사전을 만든다. 이후 TF-IDF와 나이브 베이즈 분류 모델을 이용하여 문장에 극성을 부여하고 가중치 값을 산출하고, 월별 부동산 감성지수를 산출한다. 분석 결과, 나이브 베이즈는 정확도 88%의 양호한 성능을 나타냈다. 이 모형은 부동산 부문에서 이용된 기존 방법들보다 진일보한 감성지수 개발 모형이며, 비정형 빅데이터 분석 연구에 새로운 분석틀과 체계를 제시하였다. 부동산시장 참여자들의 심리를 즉각적이고 유연하게 지수화하고, 이를 이용하여 아파트가격 등 부동산시장의 변동을 설명하거나 예측하는 토대를 마련한 점에서 의의가 있다.

주 제 어 : 주택가격, 감성분석, 신문기사, 나이브 베이즈, 기계학습