ISSN 1225-0740 Journal of the Korean Regional Science Association Vol.39, No.2(2023)
https: //doi.org/10.22669/krsa.2023.39.2.047 한국지역학회지 『지역연구』 제39권 제2호 pp.47-61

주거환경에 대한 거주민의 만족도와 영향요인 분석* – 직방 아파트 리뷰 빅데이터와 딥러닝 기반 BERT 모형을 활용하여 –

권준현** · 이수기***

Analysis of Resident's Satisfaction and Its Determining Factors on Residential Environment: Using Zigbang's Apartment Review Bigdata and Deeplearning-based BERT Model*

Kweon, Junhyeon** · Lee, Sugie***

국문요약 주거환경에 대한 만족도는 주거지 선택 및 이주 등에 영향을 미치는 주요인으로, 도시에서의 삶의 질과 직접적으로 연결된다. 최근 온라인 부동산 서비스의 증가로 주거환경에 대한 사람들의 만족도를 쉽게 확인할 수 있으며, 사람들이 평가하는 내용을 바탕으로 주거환경 만족 요인에 대한 분석이 가능하다. 이는 기존에 활용되던 설문조사 등의 방식보다 더 많은 양의 평가를 효율적으로 활용할 수 있음을 의미한다. 본 연구는 서울특별시를 대상으로 온라인 부동산 서비스인 '직방'에서 수집된 약 3만여 건의 아파트 리뷰를 분석에 활용하였다. 리뷰에 포함된 추천 평점을 토대로, 아파트 리뷰를 긍정적, 부정적으로 분류하고, 딥 러닝 기반 자연어 처리 모델인 BERT (Bidirectional Encoder Representations from Transformers)를 사용하여 리뷰를 자동으로 분류하는 모델을 개발하였다. 이후 SHAP(SHAPley Additive exPlanation)를 이용하여 분류에 중요한 역할을 하는 단어 토큰을 도출함으로 주거환경 만족도의 영향요인을 도출하였다. 더 나아가 Word2Vec을 이용하여 관련 키워드를 분석함으로써 주거환경에 대한 만족도 개선을 위한 우선 고려사항을 제시하였다. 본 연구는 거주자의 정성평가 자료인 아파트 리뷰 빅데이터와 딥러닝을 활용하여 주거환경에 대한 만족도를 긍정적, 부정적으로 자동 분류하는 모형을 제안하여 그 영향요인을 도출하는데 의의가 있다. 분석결과는 주거환경 만족도 향상을 위한 기초자료로 활용될 수 있으며 향후 아파트 단지 인근 주거환경 평가, 신규 단지 및 기반시설의 설계 및 평가 등에 활용될 수 있다.

주제어 주거환경, 주거환경 만족 영향요인, 아파트 리뷰, 빅데이터, 감정분석

Abstract: Satisfaction on the residential environment is a major factor influencing the choice of residence and migration, and is directly related to the quality of life in the city. As online services of real estate increases, peo-

* 이 논문은 한국연구재단의 지원(NRF-2018R1A5A7059549)을 받아 수행된 연구임.
** 한양대학교 도시공학과 석박통합과정(주저자: legojun98@hanyang.ac.kr)
*** 한양대학교 도시공학과 교수(교신저자: sugielee@hanyang.ac.kr)

ple's evaluation on the residential environment can be easily checked and it is possible to analyze their satisfaction and its determining factors based on their evaluation. This means that a larger amount of evaluation can be used more efficiently than previously used methods such as surveys. This study analyzed the residential environment reviews of about 30,000 apartment residents collected from 'Zigbang', an online real estate service in Seoul. The apartment review of Zigbang consists of an evaluation grade on a 5-point scale and the evaluation content directly described by the dweller. At first, this study labeled apartment reviews as positive and negative based on the scores of recommended reviews that include comprehensive evaluation about apartment. Next, to classify them automatically, developed a model by using Bidirectional Encoder Representations from Transformers(BERT), a deep learning-based natural language processing model. After that, by using SHapley Additive exPlanation(SHAP), extract word tokens that play an important role in the classification of reviews, to derive determining factors of the evaluation of the residential environment. Furthermore, by analyzing related keywords using Word2Vec, priority considerations for improving satisfaction on the residential environment were suggested. This study is meaningful that suggested a model that automatically classifies satisfaction on the residential environment into positive and negative by using apartment review big data and deep learning, which are qualitative evaluation data of residents, so that it's determining factors were derived. The result of analysis can be used as elementary data for improving the satisfaction on the residential environment, and can be used in the future evaluation of the residential environment near the apartment complex, and the design and evaluation of new complexes and infrastructure.

Key Words: Residential Environment, Determining Factors of Residential Environment Satisfaction, Apartment Reviews, Big Data, Sentiment Analysis

1. 연구의 배경 및 목적

주거환경(Residential Environment)은 주거와 생활 장소를 둘러싸고 있는 생활환경의 총체로, 주택 주변의 물리적인 환경뿐만이 아니라 사회적, 경제적, 문화적 환경을 모두 포함하는 개념이다(아사미 야스시, 2003). 즉, 주거환경의 질은 가족 돌봄, 건강뿐만이 아니라 경제활동, 사회적 상호작용 등 일반적인 삶의 전망에 영향을 미치며, 주거환경에 대한 만족도는 총체적 생활환경에 대한 사람들의 정서적 반응으로부터 치환된다(UN, 1989; 신은진·남진, 2012). 따라서, 주거환경 만족도는 도시의 지속가능성과 거주 가능성 등을 판단하는 주요한 지표로 작용하며, 생활환경 차이에서 기인하는 사회격차를 완화하고 삶의 질 향상을 위해 활용될 수 있다(안용진, 2016). 따라서 주거환경 만족도 및 그 영향요인을 분석함으로 주거환경 만족도 개선을 도모하여 주거환경의 질을 제고해야할 필요가 있다.

이러한 맥락에서, 본 연구의 목적은 서울특별시 내에 존재하는 아파트 리뷰를 활용하여 주거환경 만족도의 영향요인을 분석하는 것이다. 아파트는 주거생활의 총체적인 경험을 제공하며, 우리나라에서 가장 보편적인 주거형태이므로, 아파트 리뷰를 활용한 주거환경 만족도의 영향요인의 분석이 가능할 것으로 판단된다(이유재·문선희, 2016). 더 나아가 최근 온라인 플랫폼의 발달로 아파트를 포함한 다양한 부동산 유형에 대한 정보를 인터넷을 통해 쉽게 확인할 수 있으며, 웹 크롤링과 텍스트 마이닝을 통해 주거환경 만족도의 분석이나 부동산 임대료 예측 모형 개발이 가능하다(김보찬 외, 2018; 김선재·이수기, 2020). 따라서, 본 연구에서는 온라인 플랫폼 기반의 아파트 리뷰를 활용하여, 주거환경 만족도 분석을 위한 빅데이터를 구축하고, 새로운 방법론을 적용하여 기존의 설문조사나 수동적 영향요인 검출 방식 등의 한계를 극복하는 주거

환경 만족도 분석 과정을 구축한다. 이를 바탕으로 서울시 내 권역별로 나타나는 주거환경 만족도의 영향요인 및 연관어를 분석하고, 주거환경 만족도 개선을 위한 우선 고려사항을 도출하였다.

2. 관련 선행연구

1) 주거환경 만족도 및 영향요인

주거환경 만족도는 거주민이 체감하는 주거환경에 대한 만족도로, 이에 대한 영향요인을 분석하는 것은 주거환경의 환경적 차이에서 발생하는 사회적 격차를 완화하고 삶의 질 향상을 위해 필요한 과정이다(안용진, 2016). 특히, 인접해 있는 아파트들은 지역 내 마트, 지하철역, 학교 등에 대해 상호적 접근이 가능하며, 사회적 유대, 환경 위생, 범죄 등에 동시에 노출된다(Meng and Hall, 2006). 따라서 주거환경 만족도 영향요인을 분석하면 주거환경 모니터링 및 불편 사항 확인 등 주거환경 개선의 기초 자료로써 활용할 수 있을 것이다(김선재·이수기, 2020).

기존의 연구에서는 설문조사를 활용하여 아파트 거주민들의 주거환경 만족도를 수집하고, 상관분석 혹은 신뢰성 검증을 통해 주거환경 만족도의 영향요인을 도출하였다(장한두, 2008; 오규만·이명훈, 2019). 거주자 특성 및 주거환경 특성과의 상관 분석을 통해 주거환경 만족도 영향요인을 도출한 연구에서는 거주하는 주택의 면적, 노후년수, 방 수 등만이 주거환경 만족도의 주요 결정 요인으로 나타났다(장한두, 2008). 반면, 아파트 인근의 교육요인, 환경요인, 접근성 요인 등 다양한 항목을 바탕으로 설문조사와 신뢰성 검증과 회귀분석을 진행한 연구에서는 브랜드 선망성, 자산 가치 상승, 노후생활 안정성, 조망 등이 주거만족도에 영향을 미치는 것으로 나타났다(오규만·이명훈, 2019). 이와 달리 서울시 내 아파트 단지에 대해 안전성, 편리성, 쾌적성 등의 객관적 지표를 설정하고, 설문조사를 통해 수집된 주거만족도와의 관계를 회귀분석을 통해 도출한 연구도 진행되었다(신은진·남진, 2012). 해당 연구에서는 계층적 군집분석을 활용하여 네 개의 주거환경 유형을 제시하였으며, 각 유형 별 주거환경 만족도 결정요인을 도출하였다. 전체적으로 가로등 설치밀도, 지하철 근접성, 도로폭, 남성가구주 비율, 주택매매가격 등이 주거만족도 영향요인으로 도출되었으며, 지역유형별 분석의 유효성을 입증하였다.

최근에는 빅데이터의 활용성이 높아짐에 따라 온라인 부동산 서비스인 직방의 아파트 리뷰를 활용한 연구가 진행되었다. 김선재·이수기(2020)는 수도권 2기 신도시 아파트들의 리뷰들을 수집하고, 리뷰에 나타난 부정적 요인을 수동으로 검토하여 활용하였다. 해당 연구는 15가지 세부 항목 별로 리뷰에 나타나는 키워드를 분류하고 도시별 항목의 중요도 순서를 도출하였다. 분석 결과, 신도시에서 주거환경 만족도 영향요인으로 대중교통, 상업시설, 교육환경 등이 주거환경 만족도의 상위 영향요인으로 도출되었다. 이는 신도시 개선 및 향후 개발을 위한 근거자료로 활용될 수 있다는 점에서 의의가 있다.

2) 주거환경 만족도와 감정 분석

인터넷의 진화로 온라인을 통해 사람들의 경험과 생각이 공유되기 시작하며 감정 분석의 활용이 주목 받기 시작했다(Vuong et al., 2019). 감정 분석은 다양한 문제에 대해 인터넷에 공유되는 사람들의 관점, 생각, 감정 등에 대해 사람들이 어떻게 생각하고 느끼는지 알기 위해 필요한 과정이며, 사람들의 행동을 이해하기 위한 필수적인 요소이다(Sailunaz and Alhaji, 2019; Mustafa et al, 2020). 최근에는 컴퓨터 기반의 자연어 처리 기술의 발달로 감정 분석의 활용 가능성이 증가하였으며, 기업의 마케팅뿐 아니라 문화, 스포츠, 정치 등 다양한 분야에서의 활용이 증가하고 있다(Araque et al., 2019; Luo et al,. 2019; Zhang et al., 2019).

최근에는 온라인 리뷰를 활용하여 주거환경에 대한 사람들의 인식을 분석하는 연구가 진행되었다. 미국 뉴욕시를 대상으로 한 연구에서는 주거지에 대한 텍스트 리뷰로부터 유의미한 주제를 도출하고, 온라

인 설문조사를 통해 각 리뷰와 이를 구성하는 주제에 대해 5점 척도의 점수를 수집하여 감정 분석에 활용하였다(Hu et al., 2019). 더 나아가 주제들과 관련된 사회경제적 속성 변수를 구축하고, 상관분석을 통해 리뷰로부터 수집된 사람들의 인식이 도시 공간의 객관적인 특성과 연관이 있음을 도출해냈다. 이와 비슷하게 온라인 주택 광고를 활용하여 감정 분석을 진행한 연구에서는 토픽 모델링(Topic Modeling) 기법을 활용하여 텍스트에 등장하는 주제를 도출하고 사전 훈련된 어휘집을 통해 감정을 분류하여 활용하였다(Wang et al., 2022). 이후 기계학습 방법론을 활용하여 광고에 나타나는 특성과 사회경제학적 변수 간의 연관성을 도출해 냈으며, 공간적 자기상관성 분석을 통해 온라인 주택 광고를 활용해 사회경제적 변수를 구축할 수 있음을 시사하였다. 이처럼 빅데이터를 활용해 주택의 속성을 정량화하고, 텍스트에 반영된 주관적 인식을 도출할 수 있다(Su et al., 2021).

3) 연구의 차별성

기존 연구들은 다양한 유형의 아파트에 대해 주거환경 만족도를 수집하고 그 영향요인을 분석하였으나 표본의 제한이 존재하거나, 설문조사를 통해 구축한 정량적인 자료만을 활용하였다. 또한 조사자가 정한 항목만을 고려하였다는 한계가 존재한다(장한두, 2008; 신은진·남진, 2012; 오규만·이명훈, 2019). 이와 달리 연구자의 개입 없이 사람들의 주관적 인식을 수집한 아파트 리뷰 자료를 활용한 연구에서는, 주거환경 만족도의 부정적 요인을 수동으로 검토하여 활용하였다는 한계가 존재한다(김선재·이수기, 2020). 이러한 맥락에서 본 연구의 차별성은 다음과 같다.

첫째, 서울특별시 전체를 대상으로 아파트 리뷰 데이터를 수집하여 활용하였다. 기존 설문조사 방식과 비교할 때 더욱 방대한 양의 데이터를 수집하여 활용하며, 사람들의 주관적인 감정이 고스란히 담겨져 있는 자료를 활용하였다는 점에서 차별성이 존재한다. 이를 통해 아파트의 생활 및 근린 환경에 대한 종합적인 주거환경 만족도 영향요인을 도출하였다.

둘째, 딥러닝 기반의 자연어 처리 모델을 활용하여 아파트 리뷰를 긍정 혹은 부정으로 자동 분류하고, 아파트 리뷰에 포함된 주거환경 만족도의 긍정적, 부정적 영향요인을 자동으로 도출하였다. 딥러닝 기반 모델은 사람들의 평가를 바탕으로 학습되며, 이를 통해 도출된 각 영향요인의 영향력은 요인 자체의 사전적 의미가 아닌, 해당 요인이 전체 데이터에서 등장하는 맥락을 바탕으로 계산된다. 따라서, 전체 데이터에서 해당 요인의 영향력을 보다 쉽게 판단할 수 있다.

셋째, 서울특별시를 다섯 개의 권역[1](서북권, 도심권, 동북권, 동남권, 서남권)으로 나누어 권역별 주거환경 만족도 영향요인을 도출하고 분석하였다. 생활권은 지역특성에 맞게 구성되며, 이를 바탕으로 주거환경 만족도 영향요인을 도출함으로 생활권 별 유의미한 특성 차이를 도출할 수 있을 것이다. 더 나아가 연관 키워드 도출을 통해 권역별 거주자들의 관심 특성 차이를 도출함으로 더욱 구체적인 주거환경 만족도 영향요인을 분석하였다.

3. 분석방법

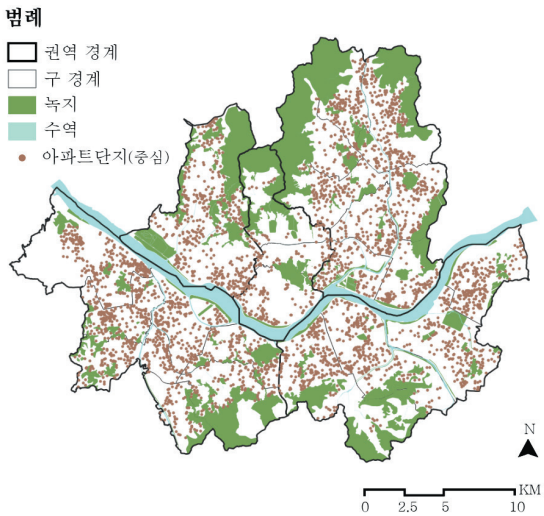
1) 분석자료

본 연구에서는 직방의 아파트리뷰 데이터를 활용하였다. 직방은 2012년 국내 최초로 모바일 애플리케이션 기반의 부동산 중개 서비스를 시작하였으며, 아파트를 포함한 다양한 부동산 정보를 제공한다. 또한 웹상에서도 쉽게 다양한 정보를 확인할 수 있다. 직방의 리뷰는 최근 5년 이내에 해당 아파트 거주 경험이 있는 거주자의 자발적인 참여로 작성된다. 따라서 작성된 리뷰를 활용하면 주거환경에 대한 사람들의 직접적이고 주관적인 인식의 분석이 가능할 것으로 판단된다. 각 리뷰는 〈그림 1〉과 같이 '추천 점수', '교통 여건', '주변 환경', '단지 관리', '거주 환경'의 다섯 가지 항목에 대한 5점 척도 점수와 50자 이상, 500자 이하의 텍스트로 구성된다. 본 연구에서는 다양한 항목에 대해 종합적인 의견이 서술된 추천리뷰의 내용과 점수

〈그림 1〉 아파트 리뷰 예시

를 분석에 활용하였다.

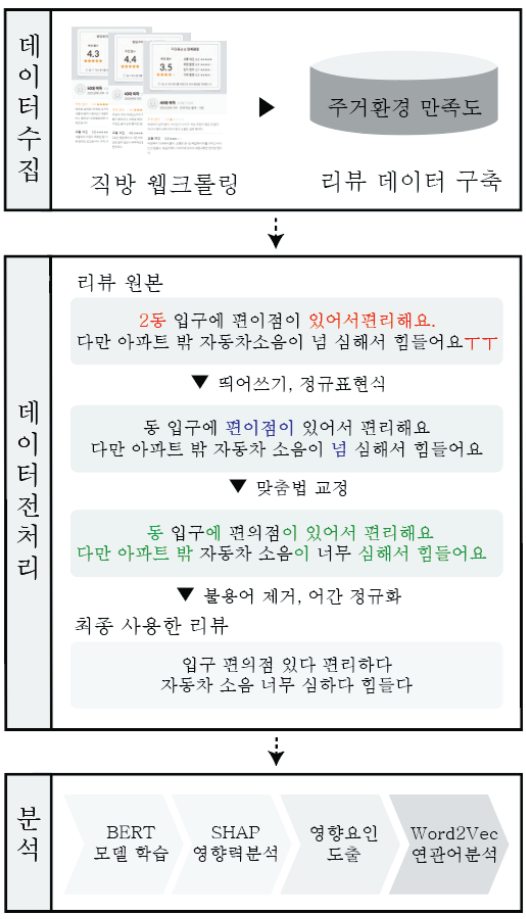
아파트 리뷰의 경우, 파이썬 기반의 웹 크롤링을 통해 서울특별시에서 조회되는 모든 아파트 단지와 작성된 리뷰들을 수집하였다. 또한, 시간적에 따른 주거환경의 차이를 고려하여 거주 시점 및 작성 시점이 5년을 경과한 리뷰는 제외하였다. 이후 주관적 척도 점수에 따라 4~5점의 경우 긍정적인 리뷰로, 1~2점의 경우 부정적 리뷰로 분류하여 분석에 활용하였다[2]. 총 3,602개 단지에서 2017~2022년 사이에 작성된 28,725개의 리뷰를 활용하였으며, 분석에 활용된 아파트 단지들의 분포는 〈그림 2〉와 같다.



〈그림 2〉 리뷰 수집된 아파트 단지 분포

2) 분석방법

본 연구의 전체적인 분석과정은 〈그림 3〉과 같다. 수집된 직방의 아파트 리뷰 데이터는 동일한 의미로 작성된 내용일지라도 작성자에 따라 활용된 문자 표기가 다르며, 모델 예측에 영향을 미칠만한 불용어들을 포함한다. 따라서 분석의 정확도를 높이기 위해 〈그림 3〉과 같이 리뷰 데이터를 전처리하는 과정을 진행하였다. 우선 띄어쓰기 없이 작성되어 두 개 이상의 단어가 잘못 분석되는 것을 방지하기 위해 PyKoSpacing을 활용하여 띄어쓰기를 진행하였다. PyKoSpacing는 대용량 말뭉치를 바탕으로 학습된 딥러닝 모델로, 한국어의 띄어쓰기를 위한 오픈소스 파이썬 패키지이다. 이후 정규표현식을 활용해 분석에 영향을 미치지 않을



〈그림 3〉 연구의 분석 과정

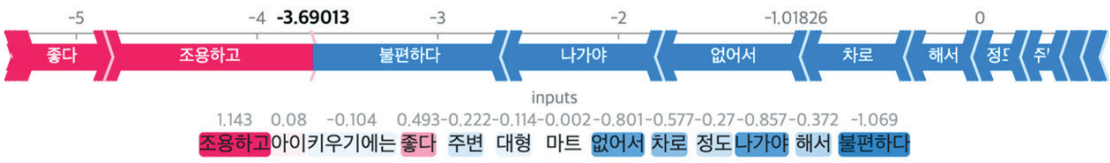
것으로 판단되는 숫자와 영문자, 특수문자 및 반복되는 한글 낱글자를 제거하였다. 맞춤법 교정의 경우, 네이버 맞춤법 검사기를 기반으로 한 파이썬용 한글 맞춤법 검사기인 HanSpell을 활용하였다.

불용어의 경우 한국어의 조사를 포함하여, 문장을 연결하는 부사나 단위에 대한 단어, '작성하였다' 등 리뷰 작성의 행위에 관한 단어들을 제거하였다. 추가로 '아파트'와 '단지'와 같이 제거하여도 문장 의미에 영향이 없는 단어들을 제거하였다. 마지막으로 파이썬 한국어 형태소 분석기인 KoNLPy를 활용하여 단어들의 어간을 정규화하였다. 이는 의미는 같지만 문장에서의 역할, 위치에 따라 형태가 바뀌는 한국어의 특징을 제어하기 위해 활용하였다. 예를 들어 '편리하고', '편리하며', '편리해서' 등은 모두 '편리하다'라는 의미를 담고 있지만 어미에 따라 형태가 달라 모델이 각각의 형태를 다른 단어로 받아들이게 되기 때문이다. 본 연구에서는 수집된 모든 리뷰에 대해 이처럼 전처리를 진행하여 사용하였다.

이후 주거환경 만족도를 긍정 혹은 부정으로 분류하고 영향요인을 도출하기 위해 딥러닝 기반의 자연어 처리 모델인 Bidirectional Encoder Representations from Transformers(BERT)와 기계학습 방법론인 SHapley Additive exPlanation (SHAP)을 활용하였다. 구글(Google)이 2018년에 개발한 BERT는 기존의 모델들과 달리 양방향으로 데이터의 맥락을 반영하여 높은 성능을 자랑한다(Devlin et al., 2018). 이를 활용해 텍스트 데이터의 긍정 및 부정의 분류를 자동화할 수 있으며, 한국어에 적용하여 뉴스 분류, 문서분석 등에도 활용이 가능하다(장은아 외, 2020; 황상흠·김도현, 2020). 본 연구에서는 이를 활용하여 아파트 리뷰를 긍정 혹은 부정으로 분류하는 모형을 학습하였다. 구체적으로 서울특별시 전체 리뷰와 각 권역 별 리뷰만을 포함한 총 6가지의 모델을 학습 후 활용하였다.

이후 학습된 모델에 SHAP 방법론을 적용하여 각 리뷰에 나타난 영향요인을 도출하였다. SHAP는 기계학습 모형에서 예측에 활용된 다양한 특성들이 모델에 얼마나 기여하는지를 도출하는 방법론이다(Lundberg and Lee, 2017). SHAP는 BERT와 같은 자연어 처리 모델에서도 활용이 가능하며, 문장의 구성 단위인 토큰(단어)의 유무가 모델 평가에 미치는 영향을 분석함으로 각 토큰의 상대적인 영향력을 도출해 낸다(Kokali et al., 2021; Catelli et al., 2022). 본 연구에 활용한 모델에 SHAP를 적용할 경우, 긍정적 분류 예측에 영향을 미친 토큰의 경우 양(+)의 값을, 부정적 예측 결과에 영향을 미친 토큰의 경우 음(−)의 값을 가진다. 〈그림 4〉의 경우 하나의 단일 리뷰에 포함된 각 토큰을 부호 별로 상대적 크기에 따라 나열한 예시로, 토큰 위의 숫자는 각 토큰이 해당 리뷰에서 미치는 영향력과 비례한다. 이후 모델의 예측 결과는 각 부호 별 모든 토큰 값을 합산하고, 합의 절댓값이 큰 쪽으로 결과가 예측된다. 해당 리뷰의 경우, 양(+)과 음(−) 토큰들 총합의 절댓값이 각각 1.716, 4.368로, BERT 모델은 이를 부정적인 리뷰로 예측한다.

마지막으로 Word2Vec 방법을 활용해 각 단어 토큰과 연관성이 있는 키워드를 도출해냈다. Word2Vec는 신경망을 기반으로 하여 방대한 말뭉치로부터 단어 간의 연관성을 학습시키는 방법이다(Mikolov et al., 2013). Word2Vec는 문장을 구성하는 단어들의 맥락을 보존한 채 벡터화를 진행하며, 이를 바탕으로 단어들 사이의 거리를 계량화 할 수 있다. 본 연구에서는 SHAP를 통해 주거환경 만족도 영향요인 중 영향력이 큰 것으로 도출된 단어들의 연관 키워드를 분석하였다.



〈그림 4〉 SHAP를 활용한 각 토큰의 상대적 영향 평가 예시

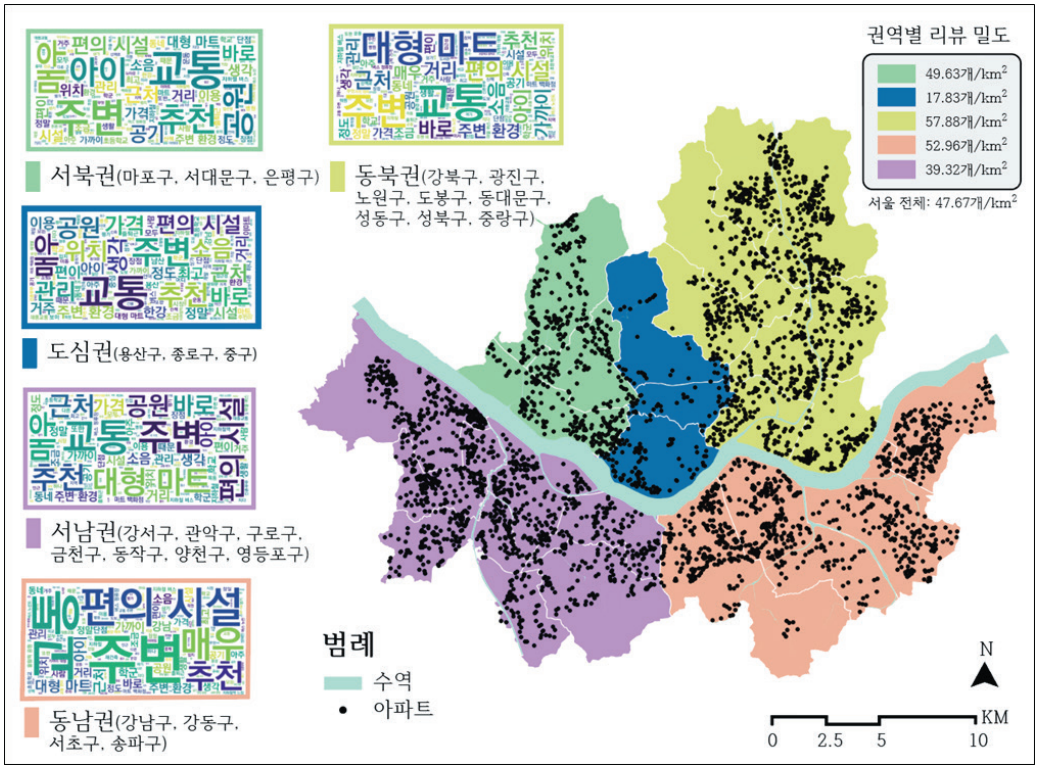
4. 분석 결과

1) 기초통계분석

전처리가 완료된 아파트 리뷰의 기초 통계 분석 결과는 〈표 1〉과 같다. 서울시 전체적으로 아파트당 평균 리뷰 개수는 7.97개로, 최소 5.62개의 리뷰가 작성된 것을 확인하였다. 전체적으로 긍정적 리뷰의 비율이 높았으며, 이는 평균 평점으로도 확인할 수 있다. 미세하나 도심권(6.74%)과 서남권(6.43%)에서 부정적 리뷰의 비율이 상대적으로 높게 나타났다. 토큰은 리뷰를 구성하는 각각의 단어를 의미하며, 토큰 개수가 많을수록 더 길게 작성된 리뷰를 의미한다. 전체적으로 등장하는 토큰의 종류는 1,312개로 다른 권역들의 리뷰 간에 동일하게 등장하는 토큰이 충분히 존재함을 확인할 수 있다. 토큰의 수가 4개인 리뷰의 경우

〈표 1〉 분석에 활용한 아파트 리뷰 기초 통계

권역	아파트 수 및 리뷰 평균 개수	평균 평점	리뷰 개수 (비율)	긍정 리뷰 개수 (비율)	부정 리뷰 개수 (비율)	등장 토큰 종류	리뷰 구성 토큰 개수 평균 (범위)
서북	427 / 8.27개	4.21	3,533(12.30%)	3,349(94.79%)	184(5.21%)	910	21.89(5~124)
도심	177 / 5.62개	4.40	994(3.46%)	927(93.26%)	67(6.74%)	760	23.15(4~106)
동북	1,009 / 9.80개	4.14	9,893(34.44%)	9,303(94.86%)	590(5.96%)	1,101	21.90(4~116)
동남	862 / 6.61개	4.23	5,700(19.84%)	5,407(94.86%)	293(5.14%)	1,011	21.61(4~110)
서남	1,120 / 7.68개	4.14	8,605(29.96%)	8,052(93.57%)	553(6.43%)	1,078	22.25(4~113)
합계	3,602 / 7.97개	4.17	28,725	27,038(94.13%)	1,687(5.87%)	1,312	21.99(4~124)



〈그림 5〉 서울특별시 권역별 리뷰 다빈출 토큰 워드클라우드 및 밀도

데이터 전처리 과정에서 불용어 등이 제거되어 전체적으로 리뷰를 구성하는 토큰의 수가 적어진 경우로, 확인해 본 결과 ['가장', '저렴하다', '신혼부부', '추천'] 혹은 ['오래되다', '치고', '좋다', '같다'] 등 토큰의 수는 적지만 충분한 의미를 포함하고 있는 것으로 확인되었다. 전체 리뷰에 대한 평균 토큰 수는 약 22개로 도출되었다.

각 권역별 리뷰의 밀도 및 빈출 빈도가 높은 토큰에 대한 워드클라우드는 〈그림 5〉와 같다. 권역별 리뷰 밀도의 경우 동북권이 가장 높게 나타났으며, 동남권이 뒤를 이었다. 도심권과 서남권의 경우 서울특별시 전체의 리뷰 밀도보다 낮게 나타났으며, 이로부터 아파트 리뷰 분포가 특정 권역에 밀집되어있음을 알 수 있다. 워드클라우드는 토큰의 빈도수에 기반한 시각화 기법으로, 토큰의 등장 빈도에 비례하여 크기가 시각화되어 표시된다. 모든 권역에서 '교통'과 '주변'의 빈도가 높게 나타났으며, 서북권의 경우 '아이'와 '공원'이, 도심권과 서남권, 동남권에서는 '편의 시설'이, 동북권에서는 '대형 마트'의 빈도가 상대적으로 높게 나타났다.

2) BERT 모델 학습 결과

BERT 모델 학습을 위해 긍정적인 리뷰는 1로, 부정적인 리뷰는 0으로 라벨을 추가하였다. 각 모델은 이를 바탕으로 텍스트 리뷰[3]를 긍정 혹은 부정으로 예측하기 위해 학습된다. 모델 학습을 위한 설정[4]으로, 리뷰의 최대 입력의 길이는 128개로, 최적화 함수는 AdamW을, 학습률은 0.0002로, 훈련 횟수는 5회로 하여 학습을 진행하였다. 훈련과 검증을 위한 데이터 비율은 8: 2로 구성하여 활용하였다. 이와 같은 과정을 거쳐 학습 완료된 BERT 모델 학습 결과는 정확도(Accuracy)[5]로 계산되며, 그 결과는 〈표 2〉와 같다. 본 연구에서 학습한 BERT 모델은 전체적으로 우수한 성능을 보였으며, 이는 아파트 리뷰를 활용한 주거환경 만족도 감정 분석의 가능성을 보여준다.

〈표 2〉 BERT 모델 학습 결과

정확도(Accuracy)	훈련(Training)	검증(Test)
전체 모델	0.97	0.88
서북권	0.96	0.93
도심권	0.95	0.96
동북권	0.96	0.85
동남권	0.96	0.94
서남권	0.93	0.95
평균	0.95	0.92

3) 권역별 상위 영향요인 분석

SHAP 방법론을 통해 리뷰에 등장하는 모든 토큰에 대해 각 토큰의 영향력을 도출하였다. 이를 위해 학습이 완료 된 6가지 BERT 모델에 대해 각 모델의 권역에 해당하는 리뷰를 적용[6]하였으며, 이 과정에서 각 토큰의 영향력은 단일 리뷰에 대해 해당 토큰이 해당 모델의 예측 과정에 기여하는 정도에 대해 상대적인 수치로 계산된다. 본 연구에서는 각 토큰의 평균 SHAP 값을 해당 요인의 영향력으로 상정하였다. 예를 들어, '좋다'라는 토큰이 서북권 리뷰에 10번 등장하면 SHAP 값은 총 10번 도출되며, 서북권에서 '좋다'의 영향력은 평균 값으로 상정하였다. BERT 모델의 결과를 바탕으로 주거환경 만족도의 감정 평가에 대한 권역별 상위 영향요인을 도출한 결과는 〈표 3〉과 같다.

우선, 서울시 전체 및 각 권역별 리뷰 평가에 영향을 미친 상위 단어 토큰을 구체적으로 분석하면 다음과 같다. 우선 서울시 전체적으로 '리모델링(0.436)', '향후(0.432)'의 긍정적 영향이 크게 나타났다. 이에 대해 리뷰를 검토한 결과, '향후 리모델링 가능성 농후하다', '재건축 통합 리모델링 추진', '리모델링 소식 들려오다' 등 미래의 개발에 관한 기대감이 나타났다. 이는 이와 같은 요인이 긍정적 영향요인의 도출에 기여하는 것으로 판단된다. 상위 5개 영향력에 포함되지는 않았지만, 이와 관련된 '개발(0.378)', '건축(0.296)' 및 '재개발(0.130)' 등의 영향력도 크게 나타났다. 이는 서울시 전체적으로 미래의 개발과 관련한 요인이 아파트 리뷰에 나타나며, 심리적인 측면에서 주거환경 만족에 영향을 미치는 것으로 판단된다. 그 다음으로는 '유치원

(0.410)', '산책(0.406)', '기준(0.395)' 등의 요인이 뒤를 이었다. 서울시 전체 모델에서 부정적 영향요인의 경우 '불편하다(−0.538)'가 가장 높게 도출되었으며, 구체적인 원인 파악을 위해 연관 키워드 분석을 활용할 수 있다. 다음으로 높게 도출된 '소음(−0.409)의 경우 모든 권역에서 부정적 영향이 높게 도출되었으며, 일반적으로 주거환경 만족에 끼치는 부정적 영향이 크다고 판단할 수 있다. 따라서 구체적인 소음의 원인을 분석하여 그에 따른 해결책을 마련해야 할 필요가 있다. 그 다음으로는 '혼자(−0.386)', '차(−0.240)', '낡다(−0.213)' 등이 뒤따랐다. '혼자'의 경우 '아이 혼자 다니다 걱정되다', '혼자 조금 무섭다' 등의 부정적 리뷰에 나타나는 것으로 확인하였으며, '혼자 살다 편하다'와 같이 1인 거주자로 추천하는 리뷰에서 나타나는 것으로 확인되었다. 전자의 경우 모델 예측에 미치는 영향이 더 큰 것으로, 부정적 영향력이 크게 도출된 것이라 판단된다. 차의 경우 '차 엄청 막히다', '차 지나 다니다 아이 통학 안전 염려 되다', '차 없다 경우 다소 불편하다' 경우 등 차의 존재로 인한 안전, 생활편의 등의 불편성이 주거환경 만족도에 부정적 영향을 미치는 것으로 도출되었다.

서북권의 경우 '초등학교(0.206)'의 긍정적 영향력이 가장 크게 나타났으며, 부정적 영향력 상위 요인으로 '키우다(−0.059)'와 '자녀(−0.002)'가 도출되었다. 이는 서울시 전체 권역 중 상대적으로 서북권에서 육아 및 교육에 대한 영향력이 상대적으로 크다는 것을 시사한다. '키우다'의 경우 '아이 키우다 좋다'와 같은 긍정적 리뷰에서도 나타났지만, '아이 키우다 곤란하다', '아이 데리고 이동 불편하다' 등의 부정적 리뷰에서 나타나는 빈도가 높아(부정적 리뷰중 6.6%가 이를

〈표 3〉 서울시 권역별 긍정적 및 부정적 상위 요인 5개(상대적 영향력)

		1순위	2순위	3순위	4순위	5순위
서울시	긍정	리모델링 (0.436)	향후 (0.432)	유치원 (0.410)	산책 (0.406)	기준 (0.395)
	부정	불편하다 (−0.538)	소음 (−0.409)	혼자 (−0.386)	차 (−0.240)	낡다 (−0.213)
서북권	긍정	초등학교 (0.206)	지하철 (0.197)	주변 (0.188)	시내 (0.178)	공원 (0.162)
	부정	키우다 (−0.059)	차 (−0.030)	젊다 (−0.025)	소음 (−0.007)	자녀 (−0.002)
도심권	긍정	조용하다 (0.680)	시내 (0.652)	한강 (0.558)	용산 (0.515)	교통 (0.511)
	부정	비싸다 (−0.406)	소음 (−0.338)	혼자 (−0.310)	오래되다 (−0.048)	어리다 (−0.006)
동북권	긍정	뉴타운 (0.336)	버스정류장 (0.332)	평지 (0.328)	대중교통 (0.278)	왕십리 (0.272)
	부정	소음 (−0.306)	오래되다 (−0.063)	넘다 (−0.042)	창동 (−0.027)	은행 (−0.019)
동남권	긍정	깨끗하다 (0.768)	대중교통 (0.495)	초등학교 (0.425)	백화점 (0.422)	교통 (0.416)
	부정	주차 (−0.311)	낡다 (−0.276)	소음 (−0.258)	어리다 (−0.241)	비싸다 (−0.080)
서남권	긍정	강서구 (0.551)	가족 (0.441)	공원 (0.413)	편리하다 (0.410)	주차장 (0.405)
	부정	오래되다 (−0.246)	주차 (−0.179)	낡다 (−0.138)	집값 (−0.133)	소음 (−0.128)

포함) 부정적 영향요인으로 도출된 것으로 판단된다. '자녀(−0.002)'의 경우도 이와 비슷한 맥락으로 부정적 영향요인으로 도출되었다. 또한 일부 '매매 비추', '관리비 비싸다' 등 다른 부정적 항목과 하나의 리뷰에 같이 작성되어 부정적 리뷰로 분류되었는데, 이러한 리뷰에 대한 검토를 진행하여 보다 상세적인 영향력 분석이 진행될 필요가 있다. 또 다른 긍정적 영향요인으로 '지하철(0.197)', '주변(0.188)', '시내(0.178)', '공원(0.162)'이 도출되었으며, 이러한 요인들은 다른 권역에서도 긍정적인 요인으로 도출되었으며 주거환경 만족도에 긍정적으로 작용함을 시사한다. 부정적 영향요인 중 '젊다(−0.025)'의 경우 '젊다 사람 살기 너무 비싸다'와 같은 부정적 리뷰에 나타났다.

도심권의 경우 '조용하다(0.680)'와 '소음(−0.338)'이 각각 가장 높은 긍정적, 부정적 요인으로 도출되었다. 이는 소리에 관한 요인이 도심권에서 상대적으로 중요하게 작용한다는 것을 시사한다. 더불어 '비싸다(−0.406)'가 가장 큰 부정적 요인으로 도출되어 도심권에서 아파트의 가격이 주거환경 만족에 부정적인 영향을 미치는 주요인임을 알 수 있다. 또한 긍정적 요인에 '한강(0.558)', '용산(0.515)' 등의 요인이 상위 요인으로 도출된 것은 도심권의 주요 입지가 주거환경 만족에 큰 영향을 끼친다는 것을 시사한다.

다음으로 동북권에서는 다른 권역들과 달리 '뉴타운(0.336)'의 긍정적 영향력이 가장 크게 나타났다. 동북권의 경우, 준공된 곳을 제외한 재정비촉진지구 39곳이 존재하며, 이는 서울시 전체 117곳 중 33%로, 가장 큰 비율을 차지한다(서울 열린데이터광장, 2022). 따라서 재정비와 관련한 개발현황이 주거환경 만족도의 요인으로도 작용하며, 동북권의 경우 그 영향이 가장 크게 나타난 것으로 판단된다. 또 다른 긍정적 요인으로 '버스정류장(0.332)', '대중교통(0.278)' 등 교통 관련한 요인이 나타났다. 또한 '왕십리(0.272)'의 경우 '왕십리역 다양하다 노선', '왕십리 와도 가깝다' 등의 리뷰에 나타난 경우로, 이는 교통 관련 요인과 더불어 교통 시설 및 주요 지점에 관한 접근성과 이용이 동북권에서 긍정적 요인으로 중요하게 작용함을 알 수 있다. 다른 권역과 달리 '평지(0.328)' 또한 상위의 긍정적 영향요인으로 도출되었다. 동북권에서만 도출된 부정적 영향요인의 경우 '넘다(−0.042)', '창동(−0.027)', '은행(−0.019)'로 도출되었다. 이중 '넘다'의 경우 '복도 넘다(넘어) 시끄럽다', '전세가 넘다(넘는다)', '청계천 넘다(넘어) 걸리다' 등 다양한 의미로 부정적으로 사용되는 것을 확인했으며, '창동'의 경우 등장 횟수가 3회 밖에 되지 않아 영향력이 과다 추정된 것으로 판단된다. '은행'의 경우 '은행 차로 나가야 한다', '은행 없다'와 같이 부정적 리뷰에 직접적으로 나타나 은행의 부재는 유의미한 부정적 영향을 미친다는 것을 확인할 수 있다.

동남권의 경우 '깨끗하다(0.768)'의 긍정적인 영향력이 가장 큰 것으로 나타났다. 이는 실제로 '주변 환경 깨끗하다'와 같은 리뷰에서 확인할 수 있다. 또한 다른 권역과 달리 '백화점(0.422)'의 긍정적 영향력이 큰 것으로 나타났다. 부정적 영향요인의 경우 '주차(−0.311)'가 가장 큰 요인으로 나타났는데, 이는 '주차 대수 모자라다', '주차 때문 스트레스' 등의 부정적 리뷰로부터 확인할 수 있다. 또 다른 부정적 요인으로는 다른 권역에서도 나타나는 요인들이 도출되었다.

마지막으로 서남권의 경우 긍정적 영향요인으로 '강서구(0.551)'가 가장 높게 도출되었는데, 서남권에서 수집된 아파트 리뷰들 중 23%가 강서구에서 수집된 것으로, 해당 리뷰들을 확인해본 결과, '강서구 학군 좋다', '강서구 인근 교통 가장 좋다', '강서구 최고 랜드마크' 등 지역명이 들어가는 경우가 많이 나타났다. 이는 서남권의 아파트 리뷰 중 강서구에서 작성된 리뷰의 수가 가장 많기 때문으로 판단되며, 향후 미시적인 지역 구분을 통해 지명에 관한 요인의 영향력을 통제하여 살펴 봄으로 더욱 구체적인 분석이 가능할 것이다. '가족(0.441)'의 경우 '가족 함께 정말 좋다', '가족 단위 외식 좋다', '가족 바람 쐬다(쐬러) 가다(가기에) 좋다' 등 긍정적인 리뷰에 나타나는 요인으로 유의미한 긍정적 영향요인으로 도출됐다. 이는 이후 순위로 도출된 '공원(0.413)' 및 '편리하다(0.405)' 와 더불어 다른 권역에서 상위 요인으로 도출되지 않는 요인으로, 서남권이 다른 권역에 비해 가족 단위의 생활 및 주변 시설 인프라 이용으로 주거환경에 이점이 있음을

〈그림 6〉 서울특별시 권역별 주거환경 만족도 영향요인

시사한다. 반면 '주차장(0.405)'과 '주차(−0.179)'의 경우 각각 긍정적 요인과 부정적 요인으로 나타나 서남권 내에서 지역 간 주차에 대한 주거환경 만족도가 상이한 것으로 판단된다. 이는 주차와 관련된 부정적 리뷰를 모니터링함으로, 해당 리뷰가 수집된 지역을 확인하고 주차 관련 불편사항을 우선적으로 개선해야 할 필요가 있음을 시사한다.

권역별로 도출된 주요 영향요인[7]과 권역별 주거환경 만족도 영향요인의 중심 주제를 도출한 결과는 〈그림 6〉과 같다.

4) 권역별 상위 연관 키워드 분석

〈표 3〉에서 도출된 서울시 전체와 권역별 모델의 주거환경 만족도 상위 영향요인에 대해 Word2Vec를 활용하여 연관 키워드를 분석한 결과는 〈표 4〉와 같다. 권역이 다르더라도, 동일하거나 비슷한 영향요인에 대해 연관 키워드의 구성이나 순서는 대체로 비슷하게 도출되었다. 특히 '젊다', '자녀', '어리다'나, '소음', '오래되다', '주차' 등의 영향요인은 권역이 다르더라도 유사한 연관 키워드들이 도출됐다.

또한 서울시 전체 모델에서 긍정적 영향요인으로 도출된 '리모델링'과 '향후'와 연관된 키워드의 경우, 재건축, 투자 등의 키워드가 나타났다. 또한, '기준'과 관련된 키워드로 '절반', '평당', '시세' 등 전체적으로 가격과 관련된 내용이 나타났다. 더 나아가 도심권의 부정적 영향요인인 '비싸다', 동북권의 긍정적 영향요인인 '뉴타운' 및 서남권의 부정적 영향요인인 '집값'의 연관 키워드들과도 비슷한 연관성을 보였으며, 이는 아파트의 경제적 가치가 서울시 전체적으로 주거환경 만족도에 영향을 미치는 것으로 판단된다.

그러나, 전반적으로 많이 나타난 영향요인인 '소음', '주차', '오래되다' 등에 대한 키워드 분석 결과, 보편적인 키워드들만이 도출되었으며, 그 원인을 구체적으로 분석하기 어려운 것으로 나타났다. 이는 사람들이 하나의 리뷰에 여러 가지 요인에 대해 서술하기 때문으로 판단되며, 따라서 더욱 구체적인 데이터 전처리 과정이 필요할 것으로 판단된다.

반면, '산책', '공원', '버스정류장' 등, 주변의 인프라와 관련된 영향요인의 연관 키워드를 분석할 경우, '우이천', '선유도', '수유' 등 해당 권역 내에 해당 인프라의 실제 존재 위치가 도출되었다. 이는 연관 키워드 분

〈표 4〉 모델별 주거만족 상위 영향요인 연관 키워드

서울시 전체			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
리모델링	재건축, 추진, 수리, 새집, 기대하다	불편하다	지장, 무리, 뚜벅, 외지다, 혼잡하다
향후	효과, 투자, 방어, 잠재, 대폭	소음	방음, 공해, 외풍, 위약, 층간
유치원	초등학교, 공립, 사립, 고등학교, 혁신	혼자	신혼, 시끌벅적하다, 은퇴, 낳다, 아이
산책	등산, 조깅, 우이천, 흐르다, 배수지	차	완만, 경사지다, 드나들다, 움직이다, 매번
기준	절반, 평당, 시세, 평균, 메이저	낡다	부실하다, 허름, 수도관, 외풍, 녹물
서북권			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
초등학교	고등학교, 중학교, 유치원, 사립, 자녀	키우다	아이, 자녀, 어리다, 유해, 학군
지하철	노선, 이용, 버스, 대중교통, 다양하다	차	서비스, 고장, 분리 다르다, 유일하다,
주변	여건, 교통, 환경 우수하다, 상당하다	젊다	부부, 분위기, 아기, 최적, 최상
시내	중심지, 요지, 지점, 어디, 업무	소음	약간, 적다, 들다, 주차, 문제
공원	한강, 하늘, 홍제천, 불광천, 월드컵	자녀	어리다, 부부, 학군, 적합, 최적
도심권			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
조용하다	낡다, 살다, 대부분, 느낌, 비추	비싸다	이유, 오래되다, 정도, 관리, 리모델링
시내	중요, 설계, 상대, 평수, 남향	소음	불편하다, 방음, 없다, 낡다, 문제
한강	용산역, 가족, 운동, 박물관, 재래시장	혼자	문화생활, 크다, 덥다, 경우, 제외
용산	업무, 국제, 부지, 향후, 여의도	오래되다	리모델링, 넓다, 낡다, 집값, 건축
교통	타다, 편리, 중심부, 용이하다, 산책로	어리다	유치원, 찾다, 완벽, 나가다, 안전하다
동북권			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
뉴타운	길음, 착공, 예정, 향후, 신설	소음	낡다, 먼지, 오래되다, 심하다, 방음
버스정류장	수유, 마들역, 잇다, 월곡역, 여분	오래되다	낡다, 연식, 제외, 층간, 외관
평지	갑, 아프다, 성북구, 다섯, 자양동	넘다	참고, 벗어나다, 익숙해지다, 적응, 돌리다
대중교통	자가용, 이동, 수단, 약속, 수월하다	창동	청량리, 상계, GTX, 미아, 상봉
왕십리	성수, 청량리역, 상봉, 동대문, 망우역	은행	사거리, 가다, 학원, 병원, 중계동
동남권			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
깨끗하다	친절하다, 넓다, 치안, 조용하다, 조경	주차	지하, 문제, 엘리베이터, 넉넉하다, 부족하다
대중교통	자가용, 용이, 이동, 출퇴근, 분당선	낡다	노후, 부분, 방음, 좁다, 녹물
초등학교	중학교, 고등학교, 유치원, 건너다, 공립	소음	아쉽다, 심하다, 문제, 불편, 좁다
백화점	쇼핑몰, 할인, 전통, 종합병원, 영화관	어리다	부부, 아기, 자녀, 부모님, 가족
교통	굉장하다, 수단, 만족, 자연환경, 번화가	비싸다	저렴, 편입, 비교, 부담, 월세
서남권			
긍정적 영향요인 연관 키워드		부정적 영향요인 연관 키워드	
강서구	관악구, 경쟁력, 여겨지다, 구로구, 갑	오래되다	연식, 낡다, 인테리어, 아쉬움, 층간
가족	노인, 적합하다, 친구, 나이, 엄마	주차	넉넉하다, 엘리베이터, 외부, 협소, 부족
공원	우장산, 선유도, 현충원, 도림천, 둘레길	낡다	튼튼하다, 인테리어, 노후, 수리, 사무소
편리하다	수단, 약속, 구비, 중형, 다양	집값	시세, 아직, 오르다, 매매, 전세
주차장	엘리베이터, 넉넉하다, 공간, 지상, 여유	소음	방음, 오르막, 불편, 벌레, 심하다

석을 통해 권역별 주거환경 만족에 영향을 미치는 인프라 요소의 도출과 모니터링이 가능함을 시사한다.

5. 결론

본 연구는 서울시 아파트 거주민들이 직방에 작성한 주관적 리뷰를 가지고 주거환경 만족도 영향요인을 분석하였다. 웹 크롤링을 통해 자료를 구축하고, 딥러닝 기반의 자연어 처리 모델을 활용하여 아파트 리뷰를 자동으로 긍정적 리뷰 혹은 부정적 리뷰로 분류하는 모델을 구축하였다. 이후 각 영향요인의 상대적 영향력과 상위 영향요인별 연관 키워드를 분석하여 서울시 전체와 다섯 개의 권역별 차이를 도출해냈다. 본 연구는 빅데이터를 바탕으로 주거환경에 대한 포괄적이고 주관적인 만족 영향요인을 자동으로 추출하는 방법을 제안하였으며 본 연구가 제안한 방법론과 결과의 의의는 다음과 같다.

첫째, 서울시 전체적으로 주거지 인근의 재건축, 재개발로 인한 경제적 가치 상승에 대한 기대감이 주거환경 만족도에 영향을 미치는 것으로 나타났다. 특히, 아파트는 위치, 구조, 근린 환경 요소의 인식 정도에 따라 경제성을 지니며, 이는 거주자를 끌어들이는 요소로 작용한다(Hu et al., 2022). 따라서 향후 주거환경 만족도 개선을 위한 방안 마련 시 경제성에 영향을 미치는 사업 유무에 따른 지역 구분이 필요하며, 각 주거환경의 상황에 맞는 방안 마련이 필요할 것이다.

둘째, 권역별로 주요 영향요인의 순서가 다르게 도출되었다. 이는 아파트 리뷰를 활용한 모니터링 및 도시환경 계획의 수립이 가능함을 시사한다. 본 연구의 방법론을 활용하면 권역별, 더 나아가 아파트가 위치한 행정동별 리뷰에 나타난 주거환경 만족도 영향요인을 도출할 수 있으며, 영향요인의 유무를 분석해 주거환경 만족도 개선 방안 마련이 가능할 것이다.

셋째, 아파트 리뷰 빅데이터를 활용한 주거환경의 분석 및 모니터링이 가능하다. 딥러닝 기반의 모델을 사용하여 사람들의 주거환경에 부정적 영향을 미치는 요인들 자동으로 도출할 수 있으며, 연관어 분석을 통해 관련 인프라 및 주거환경 개선을 위한 우선 고려사항을 도출할 수 있다. 이는 지속적으로 활용 가능한 방법론으로, 새로운 아파트 리뷰 데이터가 축적될 때마다 자동화하여 결과를 갱신할 수 있다. 더 나아가 시간에 따라 작성된 리뷰를 구분하여 주거환경의 변화에 따른 주거환경 만족도에 대한 시계열적 비교 분석이 가능하다. 특히 아파트 리뷰가 아닌 다른 텍스트 데이터에도 적용하여 활용할 수 있다는 점에서 유용할 것으로 판단된다.

그러나, 본 연구는 다음과 같은 한계점을 가진다.

첫째, 한글 언어의 특성 상 모든 단어의 영향력을 상세하게 도출하지 못했다. 특히 고유명사나 인터넷상에서 활용되는 용어, 사전에 학습되지 않은 오탈자 등의 경우 분석에서 제외하지 못했으며, 단어의 등장 빈도에 따른 영향력의 과다추정 혹은 과소추정을 고려하지 못했다. 향후 지속적인 아파트 리뷰 데이터의 사용과 모니터링 활용을 위해 보다 구체적인 전처리를 진행하고, 각 리뷰를 정제하는 과정이 필요할 것이다.

둘째, 리뷰 작성자에 대한 개인특성 제어와 신뢰성 검증을 하지 못했다. 직방의 리뷰는 리뷰 작성자의 성별, 나이, 주거유형 등을 포함한다. 이는 주관적 평가에 크게 영향을 미칠 수 있는 요인으로 작용하며, 특히 주택 소유자는 일반적으로 더 큰 자존감과 복지를 가질 수 있어 고려가 필요하다(Chen et al., 2019). 특히 직방 리뷰의 경우, 온라인 플랫폼의 특성 상 긍정적 리뷰의 비율이 높으며, 허위 여부나 거주민 대표성을 판단하기 어렵다는 한계가 존재한다. 향후 연구에서는 개인특성을 제어한 추가 분석과 거주 경험에 대한 신뢰성을 확보한 리뷰의 활용이 필요하다.

마지막으로, 리뷰 작성자들의 주관성을 배제하지 못했다. 리뷰에 활용된 5점 척도는 작성자의 주관을 바탕으로 매겨지는 점수로, 절대적인 기준이 존재하지 않는다. 따라서 동일한 내용이더라도 점수가 다르게 측정될 수 있다. 또한, 하나의 리뷰에 긍정적 내용과 부정적 내용이 모두 포함될 수 있다. 향후 연구에서는 각 단어의 영향력이 크게 도출된 리뷰를 검토하여 실제 어떤 맥락에서 해당 요인들이 도출되었는지 상세히 검토하는 과정이 필요할 것으로 판단된다.

참고문헌

김보찬·김유현·김민정·이종석, 2018, 데이터마이닝을 활용한 서울 주요 대학가 주거용 부동산 임대료 모형 수립에 관한 연구, 「대한산업공학회지」, 44(4): 259–271.

김선재·이수기, 2020, 수도권 2 기 신도시 주거환경만족도 요인 분석: 웹크롤링과 텍스트 마이닝을 활용하여. 「국토계획」, 55(7), 5–20.

서울특별시, 2014, 2030 서울도시기본계획.

신은진·남진, 2012, 서울시 아파트 단지의 주거환경 유형별 주거만족도 결정요인에 관한 연구, 「국토계획」, 47(5): 139–154.

안용진, 2016, 거주지 교통사고 공간적 집중이 주거환경 만족도에 미친 영향: 다수준 분석을 활용한 서울시 25개 자치구 실증연구, 「도시설계」, 17(2): 5–18.

오규만·이명훈, 2019, 대도시 고가아파트 주거만족과 추천의도에 대한 영향 분석–서울과 부산을 중심으로, 「주거환경」, 17(4): 219–234.

이유재·문선희, 2016, 아파트 주거만족도 평가 모형의 개발과 적용: KARSI (Korea Apartment Resident Satisfaction Index) 모형을 중심으로, 「소비자학연구」, 27(2): 25–55.

장은아·최회련·이홍철, 2020, BERT를 활용한 뉴스 감성분석과 거시경제지표 조합을 이용한 주가지수 예측, 「한국컴퓨터정보학회논문지」, 25(5): 47–56.

장한두, 2008, "주거만족 영향요인과 주거환경평가: 서울시 중소규모 아파트의 거주자 특성별 분석을 중심으로, 「대한건축학회」, 24(5): 11–21.

아사미 야스시(淺見泰司), 2003, 『住居環境 : 평가방법과 이론』, 강부성·강인호·박인석·이규인· 최정민 역, 시공문화사.

황상흠·김도현, 2020, 한국어 기술문서 분석을 위한 BERT 기반의 분류모델, 「한국전자거래학회지」, 25(1): 203–214.

Araque, O., Zhu, G., and Iglesias, C. A., 2019, A semantic similarity-based perspective of affect lexicons for sentiment analysis, Knowledge-Based Systems, 165: 346-359.

Catelli, R., Pelosi, S., and Esposito, M., 2022, Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian, Electronics, 11(3): 374.

Chen, W., Wu, X., and Miao, J., 2019. Housing and subjective class identification in urban China, Chinese Sociological Review, 51(3): 221-250.

Devlin, J., Chang, M. W., and Lee, K. T., 2018, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv: 1810.04805.

Hu, Y., Deng, C., and Zhou, Z., 2019, A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments, Annals of the American Association of Geographers, 109(4): 1052-1073.

Hu, L., He, S., and Su, S, 2022, A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets, Computers, Environment and Urban Systems, 94: 101775.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., and Robnik-Šikonja, M., 2021, BERT meets shapley: Extending SHAP explanations to transformer-based classifiers, In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, 16-21.

Lundberg, S. M., and Lee, S. I., 2017, A unified approach to interpreting model predictions, Advances in neural information processing systems, 30.

Luo, Z., Huang, S., and Zhu, K. Q., 2019, Knowledge empowered prominent aspect extraction from product reviews, Information Processing and Management, 56(3): 408-423.

Meng, G., and Hall, G. B., 2006, Assessing housing quality in metropolitan Lima, Peru, Journal of Housing and the Built Environment, 21: 413-439.

Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013, Efficient estimation of word representations in vector space, arXiv preprint, arXiv: 1301-3781.

Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., and Gelbukh, A., 2020, A multiclass depression detection in social media based on sentiment analysis, In 17th International

Conference on Information Technology-New Generations, 659-662.

Sailunaz, K., and Alhajj, R., 2019, Emotion and sentiment analysis from Twitter text, Journal of Computational Science, 36: 101003.

Su, S., He, S., Sun, C., Zhang, H., Hu, L., and Kang, M., 2021, Do landscape amenities impact private housing rental prices? A hierarchical hedonic modeling approach based on semantic and sentimental analysis of online housing advertisements across five Chinese megacities, Urban Forestry and Urban Greening, 58: 126-968.

United Nations, 1989, Handbook of social indicators, studies in methods series F no. 49, New York: United Nations.

Vuong, T., Saastamoinen, M., Jacucci, G., and Ruotsalo, T., 2019, Understanding user behavior in naturalistic information search tasks, Journal of the Association for Information Science and Technology, 70(11): 1248-1261.

Wang, L., He, S., Su, S., Li, Y., Hu, L., and Li, G., 2022, Urban neighborhood socioeconomic status (SES) inference: A machine learning approach based on semantic and sentimental analysis of online housing advertisements, Habitat International, 124: 102572

Zhang, B., Xu, X., Li, X., Chen, X., Ye, Y., and Wang, Z., 2019, “Sentiment analysis through critic learning for optimizing convolutional neural networks with rules”, Neurocomputing, 356: 21-30.

서울 열린데이터 광장, “서울시 재정비촉진지구 개발 추진 현황 자료”, 2022.09.20 읽음. http: //data.seoul.go.kr/dataList/OA-2747/S/1/datasetView.do

게재신청 2023.03.21
심사일자 2023.06.16
게재확정 2023.06.16
주저자: 권준현, 교신저자: 이수기