부동산학연구 제31집 제1호, 2025. 3, pp. 7~21 http://dx.doi.org/10.19172/KREAA.31.1.1

Journal of the Korea Real Estate Analysts Association
Vol.31, No.1 2025. 3, pp. 7~21

상업용 부동산 통계 사용자 감정분석 및 토픽모델링

Emotion Analysis and Topic Modeling of Commercial Real Estate Statistics Users

민 성 훈 (Min, Seonghun)*

< Abstract >

Interviews with experts are frequently employed as a research method to gain in-depth insights into non-quantifiable experiences and opinions. However, a major limitation of this approach is that researchers' subjectivity and preconceptions can significantly influence the interpretation of interview content and the derivation of implications. One way to overcome this limitation is to adopt analytical techniques that remain faithful to the text itself. This study applies Emotion Analysis and Topic Modeling, two representative text analysis techniques, to examine interviews conducted with users of commercial real estate statistics. The study investigates whether text analysis serves as a useful tool for comprehending interview content in an objective manner, independent of researchers' biases. First, after conducting the interviews, I initially perceived that users of commercial real estate statistics frequently expressed dissatisfaction. However, Emotion Analysis revealed that while some negative remarks were present, neutral statements constituted the largest proportion. This finding suggests that my own negative perception of commercial real estate statistics led me to focus more on users' critical remarks. Second, Topic Modeling extracted five key topics from the interviews: the role of the public sector, considerations for retail statistics, issues with commercial real estate statistics, the utility of these statistics, and commercial districts and sample selection. Notably, three key discrepancies emerged between my initial perceptions and the results of the text analysis. (1) Regarding the role of the public sector, I strongly perceived that users primarily demanded greater openness in public data. However, the analysis revealed that many participants also expressed support for expanding public statistics. (2) I initially believed that sample selection and commercial district were discussed as separate topics, but the results showed that they were consistently addressed together, indicating their deep interconnection in practice. (3) Retail-related statistical issues were the most frequently and prominently discussed topic throughout the interviews. While I had focused on other aspects, such as office statistics, interviewees continuously emphasized the unique characteristics of retail and their implications for statistical methodologies. Upon recognizing these discrepancies, I revisited the interview transcripts and confirmed that my initial interpretations had been influenced by personal biases and preconceptions, preventing me from fully grasping the content as presented in the text. This study thus demonstrates that text analysis can serve as a valuable supplementary tool for research methods involving interviews, enhancing objectivity in data interpretation.

Keyword : Commercial Real Estate, Statistics, Emotion Analysis, Topic Modeling, DMR

* 본 학회 정회원, 수원대학교 도시부동산학과 교수, smin@suwon.ac.kr

I. 서론

비즈니스, 교육연구, 정부정책 등 합리적 의사결정이 필요한 모든 분야에서 통계는 중요한 역할을 한다. 이는 부동산 분야에서도 마찬가지다. 하지만, 개별성이 크고, 거래가 빈번하지 않은 부동산의 특성상 시장의 변화를 잘 반영하는 통계를 작성하는 일이 부동산 시장에서는 쉽지 않다. 따라서 대부분 국가가 국민 생활에서 큰 비중을 차지하는 토지나 주택에 대해서는 정부가 나서서 공공통계를 작성하고 있다.

최근 국내에서도 기관투자자의 부동산투자가 활발해짐에 따라 오피스나 리테일과 같은 상업용 부동산에 대한 정보 수요가 주택 못지않게 커지고 있다. 하지만, 상업용 부동산은 주택보다 개별성과 거래의 희소성이 더 크기 때문에 통계의 공급 또한 더 부족하다. 한국부동산원이 제공하는 공공통계인 상업용부동산 임대동향조사와 민간 부동산회사가 제한된 지역을 대상으로 발간하는 마켓리포트 정도가 전부라고 할 수 있다.

신뢰성과 적시성을 가진 상업용 부동산 통계가 절실하다는 지적은 2000년대 이후 꾸준히 제기되었다. 하지만, 원론적 주장의 강도에 비해 이를 심도 있게 분석한 학술연구는 많지 않다. 원론적 주장에 대한 공감대가 형성되고 통계 작성에 대한 의지가 확립되어야 구체적인 문제와 해법에 관한 연구가 이루어질 텐데, 현실은 그렇지 못하기 때문이다. 이러한 환경에서도 몇몇 의미 있는 학술연구가 진행되었는데, 대부분 전문가의 식견으로 국내 현황과 해외 사례를 분석하여 시사점을 도출하는 방법을 취하고 있다(박원석 · 이성화, 2010; 이태리 외, 2017).

상업용 부동산 통계의 발전을 위해서는 전문가의 식견과 함께 사용자의 의견을 폭넓게 청취하는 것도 필요하다. 통계의 사용자야말로 문제와 해법을 실질적으로 파악하고 있기 때문이다. 하지만, 그간 이러한 연구가 충분히 이루어졌다고 보기는 어렵다. 여기에는 부동산 분야에서 의견과 같은 텍스트를 분석하는 기법이 확산하지 않은 것도 중요한 이유가 되고 있다.

부동산학 분야에서 의견을 조사하는 방법으로 가장 널리 사용되는 것은 설문이다. 설문은 수집된 의견의 정리와 분석에는 유용하지만, 사전에 작성한 질문으로 내용이 국한되고, 답변 또한 사전에 제시한 항목으로 제한되는 한계를 가진다. 이를 보완하기 위해 인터뷰(Interview)도 자주 활용된다. 인터뷰는 대상자를 잘 선정하고, 대화를 개방적으로 진행할 경우 유용한 통찰을 도출하지만, 의견의 해석 과정에서 연구자의 주관이 개입될 여지가 크다. 다양한 이야기 중 연구자가 관심 있는 것에만 주목하거나, 긍정 또는 부정적 평가를 연구자 선입관대로 이해하는 것이 대표적이다.

한편, 최근 인문학과 사회학 분야에서 문학작품, 인터뷰, 기사, 댓글과 같은 대량의 말뭉치(Corpus)를 분석하는 텍스트 분석(Text Analysis)이 발전하고 있다. 텍스트 분석은 질적 연구와 양적 연구의 중간적 또는 양면적 성격을 가지고 있는데, 일반적으로 말뭉치를 형태소(Morpheme) 단위로 분리하고, 형태소의 분포나 형태소 간의 관계를 분석하여 전체 말뭉치에 대한 이해를 높이는 방법을 사용한다. 이를 통해 말뭉치에 내재된 화자의 감정(Emotion)을 포착하거나, 여러 주제가 섞인 말뭉치에서 주된 토픽(Topic)을 추출하는 것이 가장 대표적인 사례다.

아직 부동산 통계와 관련해서 질적 연구가 시도된 사례를 찾아보기는 어렵다. 사실 계량 분석이 주를 이루는 국내 부동산학의 특성상 통계 뿐 아니라 대부분 주제에 대해 질적 연구는 희소한 편이다. Creswell (2015)이 분류한 바와 같이 전형적인 질적 연구는 네러티브 연구, 현상학적 연구, 문화기술지, 근거이론 연구, 사례 연구 등으로 나뉘는데, 이들은 모두 특정 개인이나 집단 즉 사람에 대한 총체적 이해를 목적으로 한다. 이를 위해 인터뷰, 관찰, 수집 등 여러 방법을 통해 다양한 자료를 수집하는데, 그중 텍스트 자료를 분석하는 방법을 객관적으로 발전시킨 것이 텍스트 분석이다. 부동산학의 접근방법을 질적 연구로 확장하기에 좋은 출발점이라고 할 수 있다.

본 연구는 대표적인 텍스트 분석기법인 감정분석(Emotion Analysis)과 토픽모델링(Topic Modeling)을 사용하여 상업용 부동산 통계의 사용자를 대상으로 진행한 인터뷰를 분석함으로써, 그들이 통계의 현황을 어떻게 평가하는지(긍정, 부정 등), 주요 문제들이 무엇이고 각 문제의 중요도가 어떠한지 등을 심층적으로 탐구하는 데 목적이 있다. 또한, 그 과정에서 텍스트 분석의 유용성도 함께 확인한다. 즉 인터뷰 내용을 정리하고 해석하는 과정에서 연구자의 관심과 선입관에 좌우되지 않고, 대상자의 의견에 충실하는 데 텍스트 분석이 도움이 되는지 살펴보는 것이다.

본 연구는 다음의 순서로 진행된다. 2장 이론적 고찰에서는 국내에서 진행된 상업용 부동산 통계에 관한

선행연구를 정리하고 본 연구의 차별성을 제시한다. 그리고, 본 연구에서 사용하는 두 가지 연구방법인 감정분석과 토픽모델링의 개념, 방법, 선행연구를 차례로 살펴본다. 3장 전문가 인터뷰에서는 15명의 전문가를 대상으로 총 5회에 걸쳐 진행한 인터뷰 개요를 소개하고, 인터뷰 내용을 필자의 전문적 식견으로 정리한다. 4장 감정분석에서는 인터뷰 텍스트의 전처리를 시행하고, 내재된 도덕감정의 분포를 분석한다. 5장 토픽모델링에서는 개방적으로 진행된 인터뷰 내용으로부터 핵심적인 토픽을 추출하고, 그것이 사용자 유형에 따라 어떻게 다른지 살펴본다. 그리고 6장에서 연구결과를 정리하고 시사점을 도출한다.

II. 이론적 고찰

1. 상업용 부동산 통계 선행연구

부동산 통계에 대한 종합적인 검토는 2000년대에 활성화된 정보체계 구축에 관한 연구에서 사례를 찾을 수 있다. 이재우(2006)는 감정평가 데이터베이스의 구조적 문제에 주목하여 데이터 수집 및 관리 과정의 비효율성을 분석하고, GIS(Geographic Information System)를 통한 공간 데이터 통합 방안을 제안하였다. 박원석・이성화(2010)는 부동산 통계의 개선을 위한 정책을 연구하는 과정에서 문헌 분석과 함께 전문가 인터뷰를 진행하였다. 그 결과 전담 조직의 설립, 통계 작성 프로세스의 표준화, 데이터 수집 체계의 일원화가 필요하다고 주장하였다. 경정익(2011) 또한 유사한 분석을 통해 부동산 정보화 정책의 성공적인 수행을 위해서는 리더십, 사용자 중심의 접근, 기술적 역량 및 정책 평가의 중요성을 강조하였다.

2010년대 후반부터는 상업용 부동산에 주목한 연구가 이루어졌다. 이태리 외(2017)는 국내 정보체계의 문제점을 개선하기 위해 미국 NCREIF(National Council of Real Estate Investment Fiduciaries)와 싱가포르 URA(Urban Redevelopment Authority) 사례를 분석하였다. 그 결과 정보의 신뢰성 및 효율성을 확보하기 위해 데이터 수집의 표준화, 체계적인 모니터링 시스템 구축, 정보 접근성 강화를 위한 단계별 확장 방안이 필요하다고 주장하였다.

방보람 외(2017)는 상업용 부동산 정보의 우선순위를 평가하기 위해 AHP(Analytic Hierarchy Process)를 수행하였다. 그 결과 거래 정보, 가격, 임대료, 접근성 등이 투자자와 정책 입안자에게 가장 중요한 정보라는 것을 발견하였다.

한편 양완진 외(2024)는 주택과 상업용 부동산의 중간 영역에 해당하는 준주택(오피스텔 등)에 대한 국가승인통계가 없는 문제를 지적하며, 개선 방안을 제시하였다. 이를 위해 건축물대장과 건축허가자료를 통합하여 시계열 데이터를 구축하고, 준주택 현황 및 주거용도로 활용되는 비율을 분석하였으며, 그 결과 다양한 데이터 원천의 통합과 시계열 기반의 체계적인 데이터 관리가 필요하다고 주장하였다.

지금까지 살펴본 선행연구가 공통되게 지적하는 문제는 첫째, 데이터 수집과 관리의 비효율성, 둘째, 부동산 통계 특히 상업용 부동산 통계를 전담하는 조직의 부재, 셋째, 정보 또는 통계에 대한 낮은 접근성, 넷째, 정보체계 표준화와 통합의 미비 네 가지다. 그리고 이러한 결론을 얻기 위해 사용한 연구방법으로는 문헌조사와 계량분석이 주를 이루고 있으며, 일부 전문가 인터뷰와 AHP가 부차적인 수단으로 활용되었다.

본 연구는 방법적인 면에서 인터뷰에 주목하되, 연구자의 주관과 전문성에 의존하여 내용을 해석하는 접근방법 대신 참가자의 발언에 충실하게 내용을 해석하는 텍스트 분석을 활용하는 점에서 선행연구와 차별화된다. 텍스트 분석은 일반적으로 알려진 상업용 부동산 통계의 문제가 실제로도 비중 있게 언급되는지, 그 간 중요하게 인식되지 않은 다른 문제는 없는지, 이들 문제의 상대적인 중요도는 어떻게 되는지 등에 대해 객관적인 시사점을 제공할 수 있을 것으로 기대된다.

2. 감정분석의 접근방법

감정분석은 텍스트 형태의 데이터를 분석하여 화자의 감정 상태를 이론에 근거한 여러 감정유형에 할당하는 자연어 처리(Natural Language Processing; NLP) 기법이다. 이는 텍스트를 긍정, 부정, 중립과 같이 비교적 간단한 상태로 식별하고 추출하는 감성분석(Sentiment Analysis)에 비해 발전된 분석기법이라고 할 수 있다. 감정분석은 주로 소셜 미디어, 고객의견, 뉴스 기사 등의 텍스트에서 사람들의 감정, 의견, 태도를 파악하기 위해 사용된다. 감정분석은 감정의 강도나 세부 감정 유형(분노, 슬픔, 기쁨 등)을 식별

하는 것까지 확장될 수 있다.

감정분석에는 크게 네 가지 접근방법이 사용된다. 첫째, 규칙 기반 접근방법(Rule-based Approach)은 사전에 정의된 규칙과 감정 어휘 사전을 사용하여 텍스트의 감정을 분석하는 방법이다. 둘째, 지도학습 접근방법은 사전에 라벨링(Labeling)된 데이터셋, 예를 들어 기쁨, 슬픔 등의 감정이 라벨링 된 문장들을 기반으로 학습된 모델을 사용하여 새로운 텍스트의 감정을 예측하는 방법이다. 주로 SVM(Support Vector Machine), 나이브베이즈(Naive Bayes), 로지스틱회귀(Logistic Regression)와 같은 전통적인 기계학습모델이나, 최근 발전하고 있는 딥러닝 기반 모델이 사용된다. 셋째, 비지도학습(Unsupervised Learning) 접근방법은 라벨링이 없는 데이터에 군집분석(Clustering)이나 토픽모델링(Topic Modeling)을 사용하여 텍스트의 감정 경향을 파악하는 방법이다. 넷째, 하이브리드(Hybrid Approach) 접근방법은 규칙 기반 접근방법과 기계학습을 결합하여 텍스트의 구조와 통계적 특성을 모두 고려하는 감정분석 방법이다.

Pang et al.(2002)은 지도학습 접근방법으로 영화 리뷰 데이터에서 감정을 예측한 초기 연구다. 이들은 나이브베이즈와 SVM을 사용하여 텍스트를 긍정과 부정으로 분류하였다. Hu and Liu(2004)는 고객 리뷰 데이터를 분석하여 특정 제품 속성(attribute)에 대한 감정을 예측하였다. 예를 들어, “배터리 성능은 좋지만, 화면 해상도는 아쉽다”와 같은 리뷰에서 배터리에 대한 긍정적 감정과 화면에 대한 부정적 감정을 분리하여 분석하였다. 이 연구는 감정분석이 텍스트 뿐 아니라, 세부 속성에도 적용될 수 있는 것을 보여주었다.

2010년대에 들어서는 딥러닝을 통해 문맥을 이해하고 세부 감정유형을 분석하는 것으로 연구가 확장되었다. Socher et al.(2013)은 RNN(Recursive Neural Network)을 사용하여 감정분석을 문맥의 이해로 확장하였다. 이들은 각 단어와 상하위 문장의 감정 관계를 학습하여, 단어 간 종속성과 문장 구조를 반영한 분석을 수행하였다. 이 연구를 통해 기존 긍정, 부정 분석에서 세밀한 감정유형 분석이 가능해졌다. Kim (2014)은 CNN(Convolutional Neural Network)을 사용하여 트윗과 같은 짧은 텍스트에서 세부 감정유형을 분류하였다. 그 결과 단어 간의 위치 정보를 반영하여 짧은 문장에서 핵심적인 감정 패턴을 추출하는 데 CNN이 효과적이라고 주장하였다. Devlin et al.(2018)은 BERT(Bidirectional Encoder Representations from Transformers)를 사용하여 문맥을 양방향으로 학습함으로써 보다 정교하게 감정적 의미를 분석하였다. 그 결과 감정의 강도나 문맥의 미묘한 차이를 파악하는 데 효과적이라고 보고하였다.

한편, 한국어 텍스트에 대해 감정분석을 시행한 사례는 김재홍 외(2023)에서 찾을 수 있다. 이들이 개발한 KOME(Korean Online Moral Emotion) 모델은 BERT를 사용한 지도학습 접근방법으로서 한국어 데이터셋인 KOTE(Korean Online That-gul Emotions)을 기반으로 다양한 도덕 감정을 정교하게 분류하였다. 특히 정의, 배려, 순수성, 권위와 같은 도덕 감정을 심리학 이론에 기반하여 분류하고, 한국어 텍스트 내에서의 도덕 감정 표현을 추출하였다.

한국어로 진행된 인터뷰를 분석하는 본 연구는 KOME를 이용하여 감정분석을 시행한다. 세부 감정유형에 대해서는 4장에서 자세히 설명한다.

3. 토픽모델링의 접근방법

토픽모델링의 가장 고전적인 방법은 LDA(Latent Dirichlet Allocation)다. LDA는 하나의 텍스트가 여러 토픽의 혼합물이고, 각 토픽은 특정 단어의 분포로 표현된다고 가정한다. 그리고 단어-텍스트 행렬을 기반으로 토픽을 학습하여 각 텍스트가 어느 정도의 비율로 각 토픽을 포함하는지, 각 토픽이 어떤 단어들로 구성되는지 추론한다. 이 과정에서 디리클레(Dirichlet) 분포를 사용하여 텍스트별 토픽의 분포와 토픽별 단어의 분포를 확률적으로 추정한다.

DMR(Dirichlet Multinomial Regression)은 회귀모형의 형태를 가진다. 즉 저자, 연도, 범주와 같은 텍스트의 메타데이터를 함께 고려하여 각 텍스트의 토픽 분포를 예측한다. LDA가 각 텍스트의 토픽 분포가 일정한 디리클레 분포를 따른다고 가정하는 반면, DMR은 토픽 분포의 초매개변수(Hyperparameter)가 텍스트의 메타데이터에 의해 달라지도록 설정하여 더 정교한 예측을 한다. DMR은 Mimno and McCallum (2008)에 의해 처음 제시되었다.

DMR은 학술논문의 메타데이터를 이용하여 학술지, 저자의 소속이나 신분, 시간의 흐름 등에 따라 특정 분야의 연구주제가 어떻게 변화했는지 추적하는 데 많이 활용되었다. 국내에서는 김용환 · 김유신(2019)이

헬스케어 관련 연구의 트렌드를 분석하는데 DMR을 활용하였다. 또한, 보다 실용적인 분야에도 다양하게 적용되었는데, 국내에서는 박영욱・정규엽(2021)이 5성급 호텔의 외국인 이용객 리뷰를 분석하고, 고객 만족도와 리뷰 토픽 간의 상관관계를 파악하는데 DMR을 활용하였다. 이 연구는 감정분석과 토픽모델링을 결합하여 둘 간의 상관관계를 추적한 사례에 해당한다.

토픽모델링에서 유념할 점은 LDA와 DMR 모두 텍스트에 내재 된 토픽들을 구, 절 또는 문장의 형태로 제시하지 않는다는 사실이다. 토픽모델링은 확률적으로 하나의 토픽을 구성한다고 판단되는 단어의 조합들을 제시할 뿐이다. 토픽모델링에 의해 제시된 각 단어의 조합을 하나의 구, 절 또는 문장으로 표현하는 일은 연구자의 몫이다. 따라서 이 과정에서 연구자의 해석이 영향을 미치게 되는데, 이는 질적 연구가 가지는 공통된 특징이라고 할 수 있다. 연구자의 해석은 질적 연구의 객관성을 떨어트릴 수도 있고, 반대로 연구대상에 대한 이해를 풍부하게 할 수도 있다. 토픽을 하나의 구, 절 또는 문장으로 표현하는 데는 일정한 규칙이 없다. 연구대상에 대해 충분한 이해를 가진 연구자가 텍스트에 근거하여 해당 작업을 수행해야 한다.

본 연구는 DMR을 사용하여 토픽모델링을 시행한다. 특히 인터뷰의 메타데이터인 통계 사용자의 유형에 따라 어떤 차이가 있는지도 함께 분석한다.

III. 전문가 인터뷰

1. 인터뷰 개요

본 연구는 상업용 부동산 통계의 문제와 해법을 파악하기 위해 두 집단의 수요자에 대한 인터뷰를 진행하였다. 첫째, 상업용 부동산에 투자하는 기관투자자, 자산운용회사, 컨설팅회사 등 금융투자 종사자와 둘째, 상업용 부동산 및 통계에 대한 교육과 연구를 수행하는 교수, 연구원 등 교육연구 종사자가 그들이다. 상업용 부동산의 특성상 모든 대상자는 일반인이 아닌 전문가로 구성되었다.

인터뷰 대상자는 총 15명이며, 금융투자 종사자 9명, 교육연구 종사자 6명으로 구성하였다. 충분한 전문성과 현장감을 확보하기 위해 10~20년 경력을 보유한 전문가를 섭외하였으며, 그 결과 최단 9년, 최장 24년의 경력자가 인터뷰에 참여하였다. 금융투자 종사자의 평균 경력은 16.3년, 교육연구 종사자의 평균 경력은 18.9년이다.

각 전문가 집단의 구성을 좀 더 자세히 설명하면 다음과 같다. 금융투자 종사자는 부동산펀드나 리츠의 운용자, 부동산금융 관련 상품을 취급하는 판매자, 부동산 관련 정책자금을 운용하는 공공기관 실무자로 구성하였다. 이들은 자신의 업무를 위해 상업용 부동산 통계를 수시로 이용하는 대표적인 사용자다. 교육연구 종사자는 부동산학과 통계학 관련 교수와 연구자 중 상업용 부동산이나 부동산 관련 통계에 대한 연구 경험이 풍부한 학자로 구성하였다. 이들 역시 심층적인 시장분석을 위해 상업용 부동산 통계를 이용하는 대표적인 사용자라고 할 수 있다.

인터뷰는 2024년 7월 중 총 5회에 걸쳐 진행하였다. 한 번에 3인이 참석하는 회의 형식으로 진행하였으며, 모든 대상자가 단 한 번의 회의에만 참석하였다. 회의는 필자가 질문하고 대상자가 대답하는 형식을 취하였으나, 미리 준비한 질문에만 국한하지 않았다. 대상자가 원하는 발언은 무엇이든 할 수 있게 개방적으로 진행하였다. 각 질문마다 발언의 기회를 참석자 모두에게 제공하였으나, 발언의 횟수와 시간을 제약하지는 않았다. 어느 대상자가 제기한 주제가 있을 경우 이에 대해서 나머지 대상자들도 발언할 수 있도록 기회를 제공하였다. 이러한 형식을 통해 대상자가 가진 지식과 경험이 충분히 공유될 수 있도록 유도하였다.

각 회차마다 주되게 다룬 주제는 다음과 같다. 1회차: 상업용부동산 임대동향조사, 오피스 마켓리포트 등 현재 발표되고 있는 통계에 대한 전반적인 평가, 2회차: 상업용 부동산 권역설정(통계구)에 대한 의견, 3회차: 용도, 규모, 입지 등 표본선정의 기준에 대한 의견, 4회차: 향후 제공되기를 바라는 통계지표에 대한 의견, 5회차: 해외 우수 상업용 부동산 통계 사례 소개. 하지만, 각 주제는 해당 회차의 첫 번째 질문이었을 뿐 이에 대한 논의가 끝나면 곧바로 다른 주제를 다루었기 때문에 결과적으로 대부분 인터뷰 대상자로부터 다섯가지 주제에 대한 의견을 모두 들을 수 있었다. 전문가 집단에 따른 주제의 비중을 나누자면 둘째, 셋째 주제는 교육연구 종사자와, 넷째, 다섯째 주제는 금융투자 종사자와 더 많이 논의하였고, 첫째 주제는 모두와 고르게 논의하였다.

인터뷰 내용은 녹음하였으며, 이를 전사하여 학술연

구에 사용한다는 취지를 충분히 고지하였다. 특히 텍스트 분석을 시행하는 연구방법까지 간단하지만 구체적으로 설명하였다. 인터뷰 대상자 15인 모두 이러한 내용에 동의하였으며, 전자적인 방법으로 동의서를 제출하였다. 하지만 대상자들의 요청에 따라 구체적인 성명과 소속은 밝히지 않기로 하였다.

2. 인터뷰 내용 사전 정리

인터뷰를 진행하는 동안 필자는 대화 내용을 기록하였다. 상업용 부동산 통계는 필자에게 익숙한 분야이고, 회의 진행 또한 필자가 했기 때문에 주요 내용을 정확하고 상세하게 기록하는 데 어려움이 없었다. 감정분석과 토픽모델링을 시행하기 전에 필자의 전문가적 식견으로 사용자 의견을 요약하면 다음과 같다.

먼저 상업용 부동산 통계에 대한 사용자의 전반적인 감정은 긍정보다는 부정에 가까웠다. 다른 선진국에 비해 통계의 종류와 역사가 부족하고, 현재 발표되고 있는 정보도 시장을 정확하게 반영하지 못한다는 것이 그 이유였다. 공공통계의 경우 전국을 모집단으로 하여 소규모 오피스와 리테일까지 조사를 하고 있는데 이것이 실제 금융투자 종사자의 수요와 맞지 않고, 민간통계의 경우 표본선택에는 큰 문제가 없으나 실효임대료나 자본환원율과 같이 조사의 부담이 큰 항목이 정확하지 않다고 지적하였다.

주되게 논의된 토픽은 크게 네 가지였다. 첫째, 가장 많이 거론된 토픽은 상업용 부동산의 표본선택과 관련된 문제였다. 복합건물 내 오피스, 저층부에 리테일을 포함한 오피스 등 변모하는 건물의 형태를 고려하여 어떻게 오피스를 식별할 것인가, 연면적과 같은 규모가 얼마 이상인 것을 표본으로 선정할 것인가, 선정된 표본에 대해서 어떻게 등급을 구분할 것인가 등이 주된 내용이었다. 이는 리테일에 대해서도 마찬가지였는데, 오피스에 비해 논의가 더 복잡하고 통일되지 않았다. 둘째, 다음으로 많이 거론된 토픽은 가격지수, 수익률지수의 부재와 관련된 문제였다. 공공통계인 상업용부동산 임대동향조사가 수익률을 발표하고 있지만, 표본이 사용자의 수요와 맞지 않고 평가기반지수의 특성상 평활화 문제가 있다고 지적하였다. 하지만 이를 민간부문에서 제공하기에는 렌트프리(Rent Free)와 같은 요소를 고려한 순영업소득을 조사하는 데 소요되는 비용을 감당하기 어려울 것이라는 것이 공통된 의견이었다. 셋째, 가격지수, 수익률지수와 연관되어 논의된 토픽으로 공공부문의 데이터 개방 문제 역시 활발하게 논의되었다. 주로 한국부동산원과 국세청이 관리하는 데이터가 개방되면 민간통계가 활성화될 것이라는 의견이었다. 넷째, 통계작성의 공간적 단위인 권역설정에 관한 의견도 있었다. 오피스의 경우 주요 권역에 대한 공감대가 형성되어 있지만, 리테일의 경우 그렇지 못해서 상권을 어떻게 나누는 지가 통계의 품질에 큰 영향을 미친다는 의견이었다.

부동산학 분야에서 인터뷰를 활용한 연구는 대부분 이와 같이 전문가의 식견으로 대화 내용을 요약하는 방법을 사용하고 있다. 하지만 여기에는 필자가 평소에 가진 감정이나 중요하게 생각하는 토픽이 영향을 미쳤을 가능성이 크다. 이러한 한계를 극복하고 보다 객관적으로 인터뷰를 이해하는데 텍스트 분석이 유용한지 살펴보기 위해 지금부터 감정분석과 토픽모델링을 시행한다.

Ⅳ. 감정분석

1. 텍스트 전처리

수치 데이터와 마찬가지로 텍스트 데이터도 정확한 분석을 위해 전처리를 해야 한다. 본 연구에서 수행한 인터뷰는 대상자가 많은 만큼 텍스트 또한 방대한데, 인사, 안내, 담소 등 본 연구의 목적과 무관한 대화를 삭제하고도 총 2,025개 문장에 달한다.

참고로 텍스트 분석은 학술논문 서지정보나 대통령 연설문과 같이 규격화되고 정제된 텍스트를 대상으로 하는 경우가 많다. 텍스트 데이터는 수치 데이터와 비교할 수 없을 정도로 비정형적이어서 전처리에 많은 시간과 노력을 기울여야 하기 때문이다. 본 연구가 대상으로 하는 인터뷰는 비록 전문가의 발언이기는 하지만 자유로운 토론 형식을 취하고 있어서 지난한 전처리 과정을 거쳤다. 하지만, 여전히 불완전한 문장을 포함하고 있을 것이라는 점을 미리 밝혀둔다.

감정분석과 토픽모델링은 분석하는 텍스트의 형태가 다르다. 전자는 전체 문장을 대상으로 하고, 후자는 문장 내에서 의미 있는 형태소를 추출한 토큰(Token)을 대상으로 한다. 특히 토큰을 분석할 때는 한 문장 내에 토큰의 개수가 일정 수 이상인 문장만 대상으로

하는 것이 일반적이다. 토큰의 개수가 과소할 경우 토큰 간 관계를 분석하는 것이 불가능하기 때문이다. 본 연구는 감정분석과 토픽모델링의 데이터 일관성을 확보하기 위해 감정분석도 토큰의 개수가 일정 수 이상인 것만 추출하여 시행한다.

전처리의 첫 번째 절차는 인터뷰에서 사용된 전문용어를 사용자 사전에 등록하는 것이다. 일반적인 국어 사전으로 형태소 분석을 할 경우 복잡한 전문용어가 해체될 수 있기 때문이다. 본 연구가 사용자 사전에 등록한 형태소는 '상업용부동산', '오피스', '리테일', '인더스트리얼', '호스피탈리티', '데이터센터', '재생에너지', '지식산업센터', '레지덴셜', '데이터', '상업용부동산통계', '오피스통계', '리테일통계' 13개다.

두 번째 절차는 2,025개의 문장에 대해 각각 형태소 분석을 하는 것이다. 형태소란 단어보다 작은 개념으로서 하나의 의미를 가지는 최소 언어 단위를 말한다. 예를 들어 '회의를 했다'라는 문장에는 '회의(일반명사)', '를(조사)', '하(동사)', '았(과거 시제 선어말 어미)', '다(종결 어미)'와 같이 다섯 개의 형태소가 포함되어 있다. 교착어 즉 어근과 접사의 구조를 가지는 한글의 특성상 형태소 분석에는 많은 시간과 노력이 필요하다. 특히 동사와 형용사는 형태소 분석이 제대로 되지 않는 경우도 많다. 따라서 많은 연구가 여러 품사 중 일반명사와 고유명사만으로 형태소 분석을 하고 있는데, 이는 명사만으로도 감정이나 토픽을 추출하는 것이 어느 정도 가능할 뿐 아니라 잘못된 동사와 형용사에 의한 오류도 방지할 수 있기 때문이다. 본 연구도 한글 연구에서 가장 널리 사용되는 키위(Kiwi) 형태소 분석기를 사용하여 명사만을 추출하였다.

세 번째 절차는 형태소 중 연구에 사용하지 않을 불용어를 제거하는 것이다. 불용어는 대화에서 자주 등장하지만 연구내용과 무관한 형태소를 말한다. 본 연구는 빈도수가 가장 높은 형태소 30개 내에 불용어가 하나도 없을 때까지 제거를 반복적으로 시행하였다. 그 결과 '거', '것', '수', '때', '말씀', '생각', '때문', '개', '경우', '사실', '정도', '필요', '데', '다음', '입장', '번', '얘기', '지금' 18개 형태소가 제거되었다.

네 번째 절차는 불용어와 마찬가지로 한 음절 형태소를 제거하는 것이다. 일반적으로 한 음절 형태소는 불용어에서 제거된 '것'과 같이 문장의 맥락에서 중요하지 않은 경우가 많기 때문이다. 하지만 연구주제와 관련된 형태소는 제거하지 않아야 한다. 본 연구는 '층', '평', '시', '도', '군', '구', '읍', '면', '동', '리' 10개만 남기고 나머지 한 음절 형태소를 제거하였다.

다섯 번째 절차는 동의어 처리를 하는 것이다. 이는 실제로 같은 의미를 가지지만 발음이 다른 형태소를

<표 1> 동의어 처리

기준단어	동의어
상업용부동산	상업용 부동산, 커머셜프라퍼티, 커머셜 프라퍼티
오피스	사무실, 사무용부동산, 사무용 부동산, 업무용부동산, 업무용 부동산
리테일	상가, 점포, 매장, 업장, 가게, 판매점, 상가용부동산, 상가용 부동산, 점포용부동산, 점포용 부동산, 매장용부동산, 매장용 부동산, 판매용부동산, 판매용 부동산
인더스트리얼	물류, 창고, 공장, 물류센터, 물류 센터, 물류창고, 물류 창고, 물류부동산, 물류 부동산, 산업용부동산, 산업용 부동산
호스피탈리티	호텔, 숙박시설, 숙박용부동산, 숙박용 부동산, 리조트
레지덴셜	주택, 아파트, 공동주택, 주거, 주거용부동산, 주거용 부동산
지식산업센터	지식산업 센터, 지식 산업센터, 지식 산업 센터, 지산, 아파트형공장, 아파트형 공장
데이터	자료, 정보, 데이타
임대동향조사	임대 동향 조사, 임대동향 조사, 임대 동향조사, 상업용부동산 임대동향조사, 상업용부동산 임대 동향 조사, 상업용부동산 임대동향 조사, 상업용부동산 임대 동향조사
가격지수	가격 지수, 매매가격지수, 매매가격 지수, 매매 가격지수, 거래가격지수, 거래가격 지수, 거래 가격지수, 매매가지수, 매매가 지수
임대료지수	임대료 지수, 임대가 지수, 임대가지수, 월세 지수, 월세지수
수익률지수	수익률 지수
상업용부동산 통계	상업용부동산 통계, 부동산시장통계, 부동산시장 통계, 부동산 시장통계, 부동산 시장 통계, 부동산관련통계, 부동산관련 통계, 부동산 관련통계, 부동산 관련 통계
오피스통계	오피스 통계, 오피스시장통계, 오피스시장 통계, 오피스 시장통계, 오피스 시장 통계, 오피스관련통계, 오피스관련 통계, 오피스 관련통계, 오피스 관련 통계
리테일통계	리테일 통계, 리테일시장통계, 리테일시장 통계, 리테일 시장통계, 리테일 시장 통계, 리테일관련통계, 리테일관련 통계, 리테일 관련통계, 리테일 관련 통계

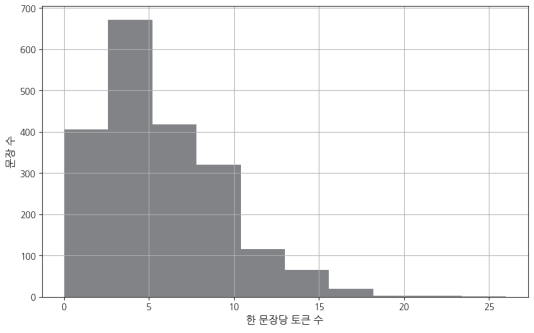
같은 것으로 인식하기 위한 것이다. 여기에는 띄어쓰기에 의한 차이를 제거하는 것도 포함된다. 본 연구가 동의어로 등록한 형태소는 <표 1>과 같다.

이상의 절차를 수행한 후 남은 형태소가 바로 본 연구에서 사용할 토큰이다. 하지만, 전술한 대로 한 문장 내에 토큰 수가 과소하면 분석에 방해가 되므로 이를 제거해야 한다. 한 문장당 형태소의 개수를 센 결과는 <표 2> 및 <그림 1>과 같다. 평균 5.67개의 토큰이 한 문장을 구성하고 있으며, 토큰이 하나도 없거나 3개 이하인 문장이 전체의 25%를 차지하고 있다. 일반적으로 허용하는 문장당 최소 토큰 수는 3개다. 본 연구도 이를 따라 토큰 수가 2개 이하인 문장을 분석대상에서 제거하였는데, 남은 문장 수는 총 1,619개다.

<표 2> 한 문장당 토큰수

문장수	평균	표준편차	최소	25%	50%	75%	최대
2025	5.67	3.61	0	3	5	8	26

<그림 1> 한 문장당 토큰수



참고로 가장 빈번히 등장한 토큰을 추출해 보면 <표 3>과 같다. 토큰의 빈도수는 단순히 개수를 센 TF (Term Frequency)와 전체 문장에서 해당 토큰이 출현한 문장의 비율의 역수인 IDF(Inverse Document Frequency)를 구해 이를 TF에 곱한 TF-IDF 두 가지로 산출하였다. TF-IDF는 너무 많은 문장에서 출현한 토큰은 오히려 중요도가 낮을 수 있다는 관점에서 TF에 패널티를 부여한 지표다. <표 3>에서 보는 바와 같이 두 지표는 토큰의 종류와 순위 면에서 큰 차이를 보이지 않았다. <그림 2>는 TF 기준으로 더 많은 토큰으로 워드클라우드(Word Cloud)를 그린 것인데, 대체로 예상한 단어들이 출현하고 있다. 하지만, 이러한 단어의 나열만으로는 큰 시사점을 얻기 어렵다.

<표 3> 자주 언급된 토큰

구분	TF 토큰	TF 빈도수	TF-IDF 토큰	TF-IDF 빈도수
1	데이터	305	데이터	133
2	리테일	208	조사	83
3	조사	190	상권	78
4	상권	179	리테일	78
5	통계	178	통계	74
6	표본	130	표본	61
7	부분	125	부분	59
8	오피스	121	오피스	53
9	지역	112	지역	51
10	임대료	106	임대료	51

<그림 2> 자주 언급된 토큰

2. 감정분포

도덕감정(Moral Emotions)이란 비난과 죄책감처럼 도덕적 판단과 평가에 관련된 감정을 말한다(김재홍 외, 2023). KOME는 도덕감정의 배경이론으로 Haidt (2003)의 분류를 따르고 있는데, 그는 도덕감정을 타인에 대한 비난(Other-Condemning), 칭찬(Other-Praising), 고통 공감(Other-Suffering), 그리고 자신의 판단과 행동에 대한 의식(Self-Conscious)으로 분류하였다. KOME는 KOTE 데이터셋에 통계적 방법론을 적용하여 세부(fine-grained) 도덕감정 데이터셋을 구축한 것이다(김재홍 외, 2023).

KOTE는 12개의 온라인 플랫폼으로부터 50,000개의 댓글을 수집하여 44개의 감정(43개의 감정과 1개의

감정없음)을 라벨링한 멀티라벨(multi-label) 데이터셋이다. KOME는 이 데이터셋으로 딥러닝을 진행하여 <표 4>와 같이 6가지 도덕감정과 각 도덕감정에 속하는 세부 도덕감정을 추출하였다. 특정한 텍스트에 KOME 모형을 적용하면 각 도덕감정의 확률분포를 구할 수 있다.

<표 4> 최종 도덕감정 분류표

구분	도덕감정	세부감정
도덕 감정	Other-Condemning	anger, contempt, disgust
	Other-Praising	admiration, gratitude
	Other-Suffering	compassion
	Self-Conscious	shame, guilt, embarrassment
비도덕 감정	Non-Moral	care, comport, pride, anxiety, boredom, exhaustion, fear, gessepany, despair, laziness, reluctant, sorrow, fed up
	Neutral	arrogance, resolute, no-emotion, realization, surprise

출처: 김재홍 외(2023)

전처리를 마친 1,619개의 문장을 대상으로 KOME를 시행한 결과는 <표 5> 및 <그림 3>과 같다. KOME는 각 문장이 6가지 도덕감정을 표현할 가능성을 0에서 1 사이의 값으로 계산한다. 그리고 <표 5>의 각 셀은 1,619개 문장에 대해 계산된 값의 기초통계량을 보여준다.

전체 인터뷰에서 가장 강하게 드러난 유형은 특정한 감정이 없는 Neutral이었다. Neutral은 1,619개 문장에서 평균 0.581의 가능성으로 출현하였으며, 3분위수 0.796, 최대값 0.957로 애매하지 않고 분명하게 표현된 경우도 매우 많았다. 또한, 감정은 있으나 도덕감정에 해당하지 않는 Non-Moral-Emotion도 평균 0.293의 가능성으로 출현하였다.

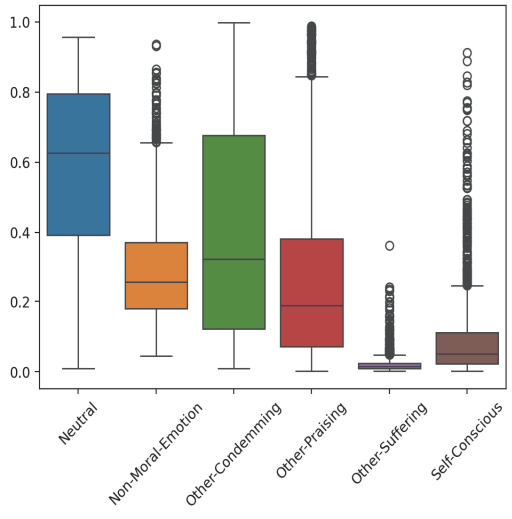
이는 필자가 체감한 것과는 매우 다른 결과다. 필자는 인터뷰 중 상업용 부동산 통계에 대해 부정적이거나 비관적인 발언을 한 경우가 대부분을 차지한다고 느꼈는데, 감정분석 결과는 이것이 필자의 선입관에 의해 왜곡된 인식일 수 있다는 것을 시사하고 있다.

도덕감정 중에서는 Other-Condemming의 가능성이 평균 0.407로 Other-Praising의 가능성 평균 0.263보다 월등히 높고, 3분위수와 최대값도 Other-Condemming이 Other-Praising보다 커서 의사표시 또한 더 분명한 것을 알 수 있다. 하지만 여기서도 Other-Praising의 가능성을 보면 부정적인 문장으로만 대화가 이어지지 않았다는 것을 알 수 있다. 이에 비해 Other_Suffering과 Self_Conscious의 가능성은 매우 낮게 나타났다.

<표 5> KOME 분석결과

구분	Neutral	Non-Moral	Other-Cond.	Other-Prais.	Other-Suff.	Self-Cons.
문장수	1619	1619	1619	1619	1619	1619
평균	0.581	0.293	0.407	0.263	0.022	0.096
표준편차	0.258	0.152	0.317	0.246	0.027	0.126
최소	0.009	0.046	0.009	0.002	0.002	0.002
25%	0.393	0.180	0.122	0.071	0.009	0.022
50%	0.627	0.256	0.321	0.189	0.015	0.051
75%	0.796	0.371	0.677	0.381	0.025	0.111
최대	0.957	0.937	0.998	0.988	0.362	0.913

<그림 3> KOME 분석결과



V. 토픽모델링

1. 토픽 추출

토픽모델링은 감정분석에서 사용한 1,619개 문장을 대상으로 하되, 문장 자체가 아닌 문장으로부터 추출한 토큰을 사용하여 시행한다. 텍스트 전처리에서 설명한 바와 같이 분석에 투입된 토큰은 한 문장당 최소 3개 이상 추출된 것들이다. 본 연구는 한글 텍스트에 대한 DMR에 가장 널리 사용되는 Tomotopy 라이브러리를 사용하여 토픽모델링을 시행한다.

DMR을 위해서는 먼저 토픽의 개수를 정해야 한다. 이를 위한 계량적인 방법도 존재하지만, 텍스트 데이터의 특성상 계량적 지표에 의지하기보다는 토픽의 개수를 증가시키면서 중복되거나 탈락되지 않는 적당한 개수를 연구자가 선택하는 방법을 많이 사용한다. 본 연구는 토픽의 개수를 3개부터 1개씩 증가시킨 결과 5개를 적용하기로 하였다. 토픽의 개수와 함께 토큰이 출현한 문장의 최소 개수도 정해야 한다. 지나치게 적은 수의 문장에 출현한 토큰은 큰 의미를 가지지 않을 수 있기 때문이다. 본 연구는 다수의 토큰을 분석에 포함 시키기 위해 최소 문장 수를 3개로 설정하였다.

토픽모델링의 가장 중요한 과정은 토큰의 중요도를 계산하는 것이다. 이를 위해 앞에서 설명한 TF-IDF를 사용하기도 하지만, 본 연구는 더 종합적인 방법인 PMI(Pointwise Mutual Information)를 사용한다. PMI는 특정 토큰과 토픽이 함께 나타날 확률과 독립적으로 나타날 확률을 비교하여 계산한다. 특정 토큰과 토픽이 자주 함께 나타나면 PMI가 커지는데, 값이 클수록 그 토큰이 해당 토픽을 잘 대표한다고 해석한다.

DMR과 같은 토픽모델링 모형은 요구한 개수만큼 토픽을 추출하고, 해당 토픽에 자주 나타나는 토큰을 나열해 준다. 이때 각 토큰이 해당 토픽에 나타날 확률도 함께 계산해 주는데, 이 확률이 높을수록 해당 토픽에서 중요한 토큰이라고 해석할 수 있다. 본 연구는 기계학습을 20회씩 나누어 총 500회 실시하였으며, 그 결과는 <표 6>과 같다. 여기서 토픽의 순서는 무작위로 할당된 것이어서 의미를 가지지 않는다.

DMR은 토픽을 텍스트 형태로 기술하지 않기 때문에 연구자가 높은 확률을 가지는 토큰을 통해 제목 또는 내용을 유추해야 한다. 토픽 1부터 하나씩 살펴보면 다음과 같다.

토픽 1의 경우 상업용 부동산 통계의 작성과 관계된 [공공부문의 역할]에 대한 담론을 나타낸다. 10개의 키워드는 '통계의 작성에는 방대한 조사가 필요한 만큼 국가가 적극적으로 역할을 할 필요가 있다.', '국가승인통계인 상업용부동산 임대동향조사의 범위와 내용을 확대할 필요가 있다.', '상업용 부동산 통계는 (수익률과 같은) 투자 관점의 정보를 포함해야 하는데, 이러한 조사는 민간 기관이 수행하기 쉽지 않다.' 등의 의견에서 함께 자주 거론된 것이다.

토픽 2의 경우 상업용 부동산 중에서 [리테일 통계의 고려요소]에 대한 담론을 나타낸다. 10개의 키워드는 '리테일은 지역이나 상권에 따라 특성에 차이가 커서 통계구 설정에 유의해야 한다.', '리테일의 임대료는 한 건물 내에서도 층이나 계약방식에 따라 차이가 커서 통계를 작성하기가 어렵다.', '서울이 아닌 지방의 경우 조사가 더 어렵다.' 등의 의견에서 함께 자주 거론된 것이다. 앞에서 살펴본 '자주 언급된 토큰'에서 토

<표 6> DMR 분석결과

토픽 1		토픽 2		토픽 3		토픽 4		토픽 5	
통계	0.0209	상권	0.0236	임대	0.0173	활용	0.0237	리테일	0.0271
국가	0.0184	데이터	0.0150	데이터	0.0173	가격	0.0188	평	0.0151
투자	0.0156	지역	0.0138	부동산	0.0153	설명	0.0187	데이터	0.0124
승인	0.0150	임대료	0.0135	조사	0.0147	데이터	0.0178	상권	0.0120
조사	0.0147	리테일	0.0125	문제	0.0138	지수	0.0167	표본	0.0120
분기	0.0140	층	0.0116	단위	0.0135	상황	0.0166	업무	0.0112
민간	0.0131	년	0.0115	상업	0.0120	반영	0.0165	층	0.0108
기관	0.0124	계약	0.0111	가격	0.0120	가치	0.0160	오피스	0.0105
관점	0.0122	지방	0.0111	지수	0.0118	측면	0.0157	전체	0.0099
년	0.0119	부분	0.0100	통계	0.0114	부분	0.0151	지역	0.0098

픽 2와 관련된 것의 순위가 높았던 것을 고려할 때 전체 인터뷰에서 이 토픽이 가장 중요하게 다루어진 것으로 판단된다.

토픽 3의 경우 [상업용 부동산 통계의 문제]를 전반적으로 다룬 담론을 나타낸다. 10개의 키워드는 상업용 부동산의 임대료, 매매가(가격), 이와 관련된 지수, 데이터 조사 등과 관련된 문제를 두루 포괄하고 있다. 다른 토픽들이 특정한 주제에 초점을 두고 있는 점을 고려할 때 토픽 3은 기타 담론이라고도 볼 수 있다. 이러한 토픽의 존재는 5개라는 토픽 수가 포화성을 가진다는 해석을 가능하게 한다.

토픽 4의 경우 [통계의 활용성]에 대한 담론을 나타낸다. 10개의 키워드는 '상업용 부동산 통계의 경우 가격이나 가치를 나타내는 지수가 제공되어야 쓸모가 있다.', '지역의 상황이 잘 반영된 상권이 설정되지 않으면 사용하기 어렵다.', '데이터나 통계에 대한 설명이 충분히 제공되어야 활용성이 높아진다.' 등의 의견에서 함께 자주 거론된 것이다.

토픽 5의 경우 [상권설정과 표본선정]에 대한 담론을 나타낸다. 10개의 키워드는 '오피스와 리테일은 지역적 특성이 달라서 상권을 달리 설정해야 한다.', '전체 지역보다는 밀집 상권을 중심으로 표본을 선정하는 것이 바람직하다.', '표본을 배분함에 있어서 층, 면적(평) 등 건물 이하 단위의 특성도 고려해야 한다.' 등의 의견에서 함께 자주 거론된 것이다.

토픽모델링의 결과는 필자가 인터뷰 내용을 사전적으로 정리한 것과 유사하지만 동시에 몇 가지 다른 모습을 보이고 있다.

첫째, 가격지수의 필요성에 대한 논의는 필자가 인식한 만큼 실제로 활발하게 이루어졌다.

둘째, 공공부문의 역할에 대해서 필자는 공공데이터 개방에 대한 요구를 강하게 받아들였는데, 이에 못지 않게 국가승인통계의 확대에 대한 논의도 있었다.

셋째, 표본선정과 권역설정에 대해 필자는 두 토픽이 별도로 논의되었다고 인식했는데, 실제로는 함께 또는 동시에 논의되었다. 이는 비록 필자가 두 질문을 별도로 했지만, 표본선정에 대한 질문에서도 권역설정이 활발히 논의되고, 권역설정에 대한 질문에서도 표본선정이 활발히 논의되었다는 것을 의미한다. 결국 필자의 선입관과 달리 두 문제는 함께 고려해야 하는 요소인 것을 알 수 있다.

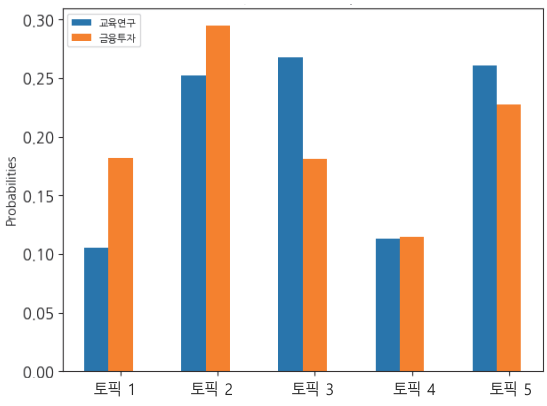
넷째, 상업용 부동산 통계의 문제가 무엇이고 그 활용성을 높이기 위해 어떤 노력을 해야 하는지에 대한 논의가 각각 별도의 토픽으로 추출되었다. 필자는 표본선정, 권역설정과 같은 실무적 프레임에서 대화에 임했지만, 인터뷰 대상자는 문제와 해법이라는 정책적 프레임도 함께 견지한 것이다.

다섯째, 필자의 인식과 가장 크게 차이나는 점은 리테일 통계와 관련된 이슈가 전체 인터뷰에서 가장 중요하고 빈번하게 논의되었다는 사실이다. 필자가 상업용 부동산의 세부적인 용도보다는 통계의 작성과 사용이라는 절차적 측면에 집중하는 동안 인터뷰 대상자들은 상업용 부동산 내에서 리테일이라는 용도가 가지는 특수성과 그것이 통계의 작성과 사용에 어떤 영향을 미치는지 끊임없이 알려주고 있던 것이다.

2. 사용자 간 비교

DMR은 LDA와 달리 토픽모델링의 결과가 특정한 명목변수에 따라 어떻게 차이 나는 지도 보여준다. 본 연구는 통계의 사용자 즉 금융투자 종사자와 교육연구 종사가 간 차이를 살펴보았는데, 그 결과는 <그림 4>와 같다. 참고로 총 1,619 문장 중 금융투자 종사자의 발언은 919개, 교육연구 종사자의 발언은 700개였다.

<그림 4> 사용자에 따른 토픽 비율



<그림 4>는 금융투자 종사자와 교육연구 종사자의 발언에서 각 토픽이 나타난 확률을 나타내고 있다. 전술한 바와 같이 토픽 2 [리테일 통계의 고려요소]가 전체적으로 가장 큰 비중을 차지하고 있다.

두 집단에서 각 토픽의 확률은 비슷한 모습을 보이고 있다. 다만, 토픽 1 [공공부문의 역할]은 금융투자

종사자에서, 토픽 3 [상업용 부동산 통계의 문제]는 교육연구 종사자에서 다소 큰 차이로 높게 나타났다. 이는 업계의 수요자는 공공부문의 역할을 강조했고, 학계의 수요자는 상업용 부동산 통계가 가지는 본질적 한계를 강조했다는 것을 의미한다. 단, DMR은 이러한 차이가 통계적으로 유의한 지는 검정하지 못한다.

VI. 결론

본 연구는 감정분석과 토픽모델링을 사용하여 상업용 부동산 통계의 사용자를 대상으로 진행한 인터뷰를 분석하였다. 인터뷰는 금융투자 종사자와 교육연구 종사자 15인을 대상으로 하였으며, 최소 9년 이상의 경력을 가진 전문가로 구성하였다. 인터뷰는 2024년 7월 중 총 5회에 걸쳐 진행하였다.

인터뷰 내용을 사전 정리하는 과정에서 필자는 상업용 부동산 통계에 대한 사용자의 전반적인 감정이 긍정보다는 부정에 가깝다고 느꼈다. 또한 주되게 논의된 토픽을 첫째, 상업용 부동산의 표본선택과 관련된 문제, 둘째, 가격지수, 수익률지수의 부재와 관련된 문제, 셋째, 공공부문의 데이터 개방 문제, 넷째, 통계작성의 공간적 단위인 권역설정에 관한 의견 네 가지로 정리하였다. 하지만, 이러한 인터뷰 요약에는 필자의 주관과 선입관이 반영되었을 수 있다. 이를 보완하기 위해 텍스트 분석을 실시한 결과는 다음과 같다.

먼저 감정분석을 시행한 결과 전체 인터뷰에서 가장 강하게 드러난 유형은 특정한 감정이 없는 Neutral이었다. 또한 감정은 있으나 도덕감정에 해당하지 않는 Non-Moral-Emotion도 낮은 가능성이지만 존재하였다. 이는 필자가 인지한 것과는 매우 다른 결과다. 필자는 인터뷰 중 상업용 부동산 통계에 대해 부정적이거나 비판적인 발언을 한 경우가 대부분을 차지한다고 느꼈는데, 텍스트는 그렇지 않았던 것이다. 실제로 사용자는 상업용부동산 임대동향조사나 오피스 마켓리포트를 업무에 활용하고 있으며, 이들 통계가 가지는 문제와 원인을 냉정하게 이해하고 있었다.

토픽모델링에서는 공공부문의 역할, 리테일 통계의 고려요소, 상업용 부동산 통계의 문제, 통계의 활용성, 상권설정과 표본선정 다섯 가지 토픽이 중요하게 논의된 것으로 나타났다. 이는 필자가 인터뷰 내용을 사전적으로 정리한 것과 유사하지만 동시에 몇 가지 다른 모습을 보이고 있다.

첫째, 가격지수의 필요성에 대한 논의는 필자가 인식한 만큼 실제로 활발하게 이루어졌다. 둘째, 공공부문의 역할에 대해서 필자는 공공데이터 개방에 대한 요구를 강하게 받아들였는데, 국가승인통계의 확대에 대한 논의도 있었다. 셋째, 표본선정과 권역설정에 대해 필자는 두 토픽이 별도로 논의되었다고 인식했는데, 실제로는 함께 또는 동시에 논의되었다. 필자의 선입관과 달리 두 문제는 함께 고려해야 하는 요소인 것을 알 수 있었다. 넷째, 상업용 부동산 통계의 문제가 무엇이고 그 활용성을 높이기 위해 어떤 노력을 해야 하는지에 대한 논의가 별도의 토픽으로 추출되었다. 다섯째, 필자의 인식과 달리 리테일 통계와 관련된 이슈가 전체 인터뷰에서 가장 중요하고 빈번하게 논의되었다. 필자가 통계의 작성과 사용이라는 절차적 측면에 집중하는 동안 인터뷰 대상자들은 리테일이 가지는 특수성과 그것이 통계의 작성과 사용에 어떤 영향을 미치는지 끊임없이 알려주고 있었던 것이다.

이러한 결과는 상업용 부동산 통계의 작성에 대해 다음과 같은 시사점을 알려준다. 첫째, 조사대상과 관련해서 리테일 섹터에 대한 보완이 절실하다. 여기에는 기관투자자를 위한 대형 리테일 뿐 아니라 소상공인과 같은 개인을 위한 소형 리테일도 포함된다. 둘째, 통계지표와 관련해서 가격지수, 수익률지수와 같은 투자지표의 개발이 절실하다. 이를 위해서는 임대정보, 거래정보 등에 대한 폭넓은 조사가 요구되며, 무엇보다 적절한 통계구(상권) 설정이 선행되어야 한다. 셋째, 앞의 두 과제는 민간부문에서 해결하기 어렵다. 정보와 예산 면에서 큰 역량을 가진 공공부문에서 역할을 해줄 필요가 있다.

본 연구는 인문학과 사회학 분야에서 대량의 말뭉치를 분석하는 데 사용되고 있는 텍스트 분석을 이용하여 상업용 부동산 통계와 같이 전문적인 분야의 인터뷰를 분석하였다. 그 결과 연구자가 자신의 관심과 선입관으로부터 거리를 두고 텍스트 자체에 충실하게 인터뷰 내용을 이해하는데 감정분석과 토픽모델링이 도움 된다는 것을 확인하였다.

하지만, 본 연구는 다음과 같은 한계를 가진다. 첫째, 지도학습 모형을 사용하는 텍스트 분석에는 감정 라벨링이 된 텍스트 데이터셋이 필요하다. 본 연구는 지도학습을 위해 온라인 포털의 댓글에 라벨링을 한 KOTE 데이터셋을 사용하였다. 하지만, 온라인 댓글과

전문가 인터뷰의 문장에는 차이가 있어서 기계학습의 결과를 신뢰하기 어려울 수도 있다. 사실 이러한 현실은 본 연구뿐 아니라 한국어 텍스트 분석이 겪고 있는 공통된 문제다. KOTE 데이터셋 만큼 방대하고 상세하게 라벨링 된 것이 없기 때문이다. 향후 학술연구와 같이 정제된 텍스트에 대해서도 데이터셋 작업이 이루어져야 할 것이다. 둘째, 토픽모델링에서 통계 사용자 간 차이를 살펴보기는 했지만, 본 연구는 인터뷰 문장의 메타데이터가 결과에 미치는 영향을 치밀하게 분석하지 못했다. 향후 인터뷰 대상자 특성을 다양하게 추출하고, 그들 간의 차이에 대한 유의성을 검정하는 방향으로 연구가 확장되어야 할 것이다.

논문접수일 : 2024년 10월 17일
논문심사일 : 2025년 2월 4일
게재확정일 : 2025년 3월 6일

참고문헌

1. 경정익, “부동산 정보화정책의 효율성을 위한 개선방안”, 「부동산학연구」 제17권 제1호, 2011, pp. 95-117
2. 김용환 · 김유신, “토픽모델링을 이용한 국내 헬스케어 학술연구 트렌드 분석”, 「한국웰니스학회지」 제14권 제1호, 2019, pp. 253-262
3. 김재홍 · 정채윤 · 차미영 · 이원재, “KOME(Korean Online Moral Emotion): 세부 분류를 위한 한국어 도덕감정 데이터셋”, 「한국정보과학회 학술발표논문집」, 2023
4. 박영욱 · 정규엽, “DMR(Dirichlet Multinomial Regression) 토픽모델링을 이용한 온라인 리뷰 빅데이터 기반 고객감성 분석에 관한 연구: 국내 5성급 호텔의 외국인 이용객 리뷰를 중심으로”, 「호텔경영학연구」 제30권 제2호, 2021, pp. 1-20
5. 박원석 · 이성화, “국내 부동산통계의 개선 및 정책 활용제고 방안”, 「한국지적학회지」 제26권 제2호, 2010, pp. 65-78
6. 방보람 · 이태리 · 조정희, “상업용 부동산 정보의 중요도 평가 연구”, 「부동산학연구」 제23권 제3호, 2017, pp. 29-39
7. 양완진 · 박과영 · 김동원, “준주택 현황과 통계 개선방안에 관한 연구”, 「부동산학연구」 제30집 제1호, 2024, pp. 37-51
8. 이재우, “부동산 감정평가정보체계 DB구축 개선방안에 관한 연구”, 「한국콘텐츠학회 논문지」 제6권 제8호, 2006, pp. 94-104
9. 이재우, “부동산 정보체계의 효율성 개선 방안에 관한 연구”, 「대한건축학회 논문집」 제36권 제5호, 2024, pp. 85-98
10. 이태리 · 조정희 · 최진 · 권건우, “미국, 싱가포르 사례를 통한 한국의 상업용 부동산 정보체계 구축 방안 연구”, 「정보화정책」 제24권 제4호, 2017, pp. 44-67
11. Creswell, J. W., 2015, “질적 연구방법론”, 조흥식 · 정선욱 · 김진숙 · 권지성 공역, 2013. 「학지사」, 원본출판
12. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2018, pp. 4171-4186
13. Haidt, J., “The moral emotions,” 2003
14. Hu, M. and Liu, B., “Mining and Summarizing Customer Reviews,” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2004, pp. 168-177
15. Kim, Y., “Convolutional Neural Networks for Sentence Classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751
16. Mimno, D. and McCallum, A., “Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression,” Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI), 2008, pp. 411-418
17. Pang, B., Lee, L., and Vaithyanathan, S., “Thumbs up? Sentiment Classification using Machine Learning Techniques,” Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79-86
18. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C., “Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013, pp. 1631-1642

<국문요약>

상업용 부동산 통계 사용자 감정분석 및 토픽모델링

민 성 훈 (Min, Seonghun)

전문가를 대상으로 한 인터뷰는 비계량적 경험이나 의견을 심층적으로 파악하기 위해 자주 사용되는 연구방법이다. 그러나 인터뷰 내용을 해석하고 시사점을 도출하는 과정에서 연구자의 주관이나 선입관이 크게 작용하는 단점을 가진다. 이러한 단점을 극복하기 위한 방법 중 하나는 인터뷰 내용 그 자체 즉 텍스트에 충실한 분석기법을 도입하는 것이다. 본 연구는 대표적인 텍스트 분석 기법인 감정분석과 토픽모델링을 사용하여 상업용 부동산 통계의 사용자를 대상으로 진행한 인터뷰를 분석하였다. 그리고 연구자의 주관이나 선입관에 좌우되지 않고 텍스트에 충실하게 내용을 이해하는데 텍스트분석이 유용한지 살펴보았다. 첫째, 인터뷰 이후 필자는 통계의 사용자들이 부정적 의견 즉 불만을 많이 토로한 것으로 인식하였다. 그러나 감정분석 결과 부정적 발언도 있었지만, 중립적 발언이 가장 큰 비중을 차지한 것을 알 수 있었다. 이는 상업용 부동산 통계에 대한 필자의 부정적 감정이 사용자의 부정적 발언에 더 주목하게 했다는 것을 암시한다. 둘째, 토픽모델링에서는 공공부문의 역할, 리테일 통계의 고려요소, 상업용 부동산 통계의 문제, 통계의 활용성, 상권설정과 표본선정 다섯 가지 토픽이 추출되었는데, 여기서도 필자의 인식과 다른 세 가지 차이가 발견되었다. 1) 공공부문의 역할에 대해 필자는 공공데이터 개방에 대한 요구를 강하게 인식했는데, 실제로는 공공통계의 확대에 대한 의견도 적지 않았다. 2) 표본선정과 권역설정에 대해 필자는 두 토픽이 별도로 논의되었다고 인식했는데, 실제로는 매번 함께 논의되었다. 이는 두 문제가 실무적으로 깊게 연관되어 있음을 나타낸다. 3) 리테일 통계와 관련된 이슈가 전체 인터뷰에서 가장 중요하고 빈번하게 논의되었다. 필자가 오피스와 같이 다른 내용에 집중하는 동안에도 인터뷰 대상자들은 리테일의 특수성과 그것이 통계에 미치는 영향을 끊임없이 알려주고 있었던 것이다. 이러한 차이를 인지한 후 다시 인터뷰 내용을 읽었을 때 필자의 초기 해석에 주관과 선입관이 영향을 미쳐 텍스트를 있는 그대로 이해하지 못했다는 것을 확인할 수 있었다. 이를 통해 텍스트분석이 인터뷰와 같은 연구방법을 사용하는데 유용한 보조수단이 된다는 것을 알 수 있다.

주 제 어 : 상업용 부동산, 통계, 감정분석, 토픽모델링, DMR