2021-CR-05
작은연구 좋은서울 21-11

자연어 처리 모델을 활용한 인터넷 댓글과 부동산 가격의 관계 분석

서나래

서울연구원
The Seoul Institute

자연어 처리 모델을 활용한
인터넷 댓글과 부동산 가격의 관계 분석

연구책임

서나래 서울대학교 데이터사이언스대학원 석사과정

연구진

주종웅 서울대학교 건설환경공학부 박사수료

목차

표 목차

그림 목차

01. 서론

1. 연구 배경 및 목적

최근 주택 가격이 급격하게 상승하면서 부동산에 대한 관심이 높아지고 있다. 세금 인상, 대출 축소, 주택 공급 확대 등 부동산 관련 정책이 지속적으로 발표되고, 신문기사, 인터넷 카페 등을 통해 부동산에 대한 정보 공유 및 토론도 활발하게 이루어지고 있다. 우리나라의 가계 평균자산에서 부동산이 차지하는 비중이 70% 이상(전해정·양혜선, 2019)이라는 점에서 우리나라 사람에게 부동산은 중요하게 인식되는 주제라고 할 수 있다. 그동안 도시·부동산 분야에서 부동산시장 관련된 연구 대부분은 부동산 가격과 관련된 변수를 파악하기 위해 공공에서 제공하는 통계자료나 설문 조사, GIS 등을 통해 수집한 정형 데이터를 활용해 분석했다. 그러나, 정형 데이터는 부동산 가격과 관련이 깊은 소비자의 심리 태도를 충분히 반영할 수 없다는 점에서 한계가 있다(경정익·이국철, 2016; 박재수·이재수, 2019). 이에 최근에는 소비자의 심리 태도 분석을 위해 신문기사, 포털 검색 빈도, 트위터 자료 등 비정형 데이터를 수집하고, 이를 분석하는 연구가 진행되고 있다(김대원·유정석, 2016; 이종민 외, 2017a; 이종민 외, 2017b; 진창하·Gallimore, 2012).

하지만, 비정형 데이터를 활용한 선행연구 대부분은 형태소 및 단어를 단위로 분석을 진행하여 문장 내 긍·부정 단어가 혼재되어 있거나, 긍정 또는 부정의 의미는 갖는 부사가 반영되지 않는 등 한계가 있다. 또한, 최근 주택시장의 서울 강남 선호 현상에서 볼 수 있듯이, 소비자의 심리 태도는 지역에 따라 달라질 수 있지만, 기존의 연구들은 지역에 따른 차이를 고려하지 않았다.

따라서 이 연구에서는 부동산 카페 및 네이버 뉴스의 댓글 데이터를 GPT2 계열의 언어 모델에 학습시켜, 사전에 작성한 지역별 프롬프트(Prompt)[1]를 이용하여 문장을 생성하고, BERT 계열의 딥러닝 모델로 문장의 맥락을 고려한 감성 분석을 시행하고자 한다.

이를 통해 지역에 대한 심리 태도와 지역의 주택 가격 간 관계를 분석하고, 지역의 인식 개선을 위해 필요한 정책을 제안하고자 하였다.

1) 프롬프트(prompt)는 입력 예제로 사용하기 위한 텍스트로, 일부분만 작성되어있는 형태의 문장이다. 언어모델로 빈 부분을 예측(Masked language model)하기 위해 사용된다.(https://thegradient.pub/prompting/)

2_비정형 데이터를 활용한 도시·부동산 관련 연구

도시·부동산 분야에서는 AI의 발달로 인해 빅데이터를 분석에 활용할 수 있게 되면서 부동산시장의 분석에 비정형 데이터를 사용하기 시작했다. 사람들이 컴퓨터, 모바일 등을 활용하면서 이미지, 동영상, 음성, 텍스트 등과 관련된 대용량의 데이터가 형성되기 시작했고, 이에 따라 머신러닝, 딥러닝 등 데이터를 분석하기 위한 기술도 발전했다.
방법론의 발전 방향은 크게 두 가지로 구분할 수 있다. 첫 번째는 기존의 정형 데이터를 전통적 통계분석 방식이 아닌 SVM(support vector machine), RF(random forest), GRBT(gradient boosting regression tree), DNN(deep neural networks), RNN(recurrent neural network), LSTM(long short term memory), GRU(gated recurrent unit) 등 머신러닝 및 딥러닝에 기반한 분석방법을 적용한 것이다(배성완·유정석, 2018; 이태형·전명진, 2018; 전해정, 2020; 전해정·양혜선, 2019). 이들 연구는 기존의 전통적 통계분석방법을 머신러닝 및 딥러닝에 기반한 분석방법과 비교하여 부동산 가격의 예측력이 개선되는지 분석한다.
두 번째는 비정형 데이터를 부동산시장 분석에 적용하는 것이다. 이들 연구는 부동산시장에서 소비자의 심리 태도가 주택 가격과 관련이 있어, 소비심리지수 등 주택가치에 대한 전망이 부동산시장 변화를 예측할 수 있다고 보았다. 따라서 비정형 데이터를 활용해 소비자의 심리 태도를 분석하는 연구가 진행되었다(경정익·이국철, 2016; 김대원·유정석, 2016; 김진유, 2006; 박재수·이재수, 2019a; 박재수·이재수, 2019b; 박재수·이재수, 2021; 이종민 외, 2017a; 이종민 외, 2017b; 진창하·Gallimore, 2012). 신문기사 또는 트위터에서 텍스트 마이닝을 통해 부동산 관련 심리 태도에 대한 텍스트를 수집하고, 형태소 또는 단어에 대한 빈도 분석 또는 감성 분석(sentiment analysis)을 통해 중요도 또는 긍·부정 스코어를 산정한다. 그리고 산정된 심리 태도 관련 지표와 부동산 매매 및 전세가격과 비교하여 주 지표 간 관련성을 분석한다.
부동산에 대한 심리 태도는 지역에 따라 다름에도, 선행연구에서는 전국 또는 서울 지역 단위로 지표를 산출해 세부 지역에 대한 지표 결과는 알 수 없었다. 또한 부사 등 텍스트가 제외돼 감성 분석의 정확성이 낮다.
따라서 이 연구에서는 GPT2 언어 모델을 활용해 세부 지역별 부동산 심리 태도를 분석하고, 감성 분석의 정확도를 높이기 위해 BERT 계열의 딥러닝 모델을 적용한다.

3_언어 모델에 드러난 편견 관련 연구

과거의 기계학습은 고용량의 데이터와 고성능 모델, 지도학습을 통해 성능을 보여왔으나, 이러한 방식은 범용적으로 쓰이기보다는 과업별로 특화된(task-specific) 방식이었다. 이 때문에 매번 트레이닝 데이터셋을 라벨링할 필요가 없도록, 여러 과업(downstream task)에서 좋은 성능을 보이는 일반적인 시스템(general system)을 만들고자 하는 시도에서 등장한 것 중 하나가 GPT2이다(Alec Radford et al., 2018; Alec et al., 2019).

OpenAI에서 공개한 GPT2는 오토리그레시브 디코더(Autoregressive Decoder) 모델로, 범용성을 갖추기 위해 아주 크고 다양한 데이터셋을 가져와서 학습시켰다. 대용량 데이터를 학습시키기 위해, 웹 크롤링으로도 데이터를 수집하였으며, 고품질의 데이터만 선별하기 위해 Reddit에서 3 karma[2] 이상 받은 글만 가져왔다(Alec Radford et al., 2018; Alec et al., 2019). 이렇게 대용량으로 수집된 텍스트를 기반으로 만들어진 모델은, 문장 생성(generation)에서 특히 좋은 성능을 보인다. 하지만, 결국 학습 데이터에 나타난 인간의 편견마저 학습하게 되는 부작용이 있다.

언어 모델에 학습된 편견에 대한 연구는 GPT2와 같이 사전학습된(Pre-trained) 대용량의 모델이 개발되기 이전부터 수행되어 왔다. Nikhil Garg et al(2017)에서는 단어 임베딩(embedding)에 반영된 시대별 성별 편견에 대해 다루었다. Nikhil Garg et al(2017)에서는 성별/민족별 단어들을 그룹핑하고, 워드투벡(Word2vec)으로 각 그룹의 벡터를 평균한 후 중립 단어들과의 거리를 비교하는 방식으로 임베딩(embedding)에 반영된 편견을 밝혔다.

이후에도 이런 시도들이 계속되어, Maarten et al(2020)에서는 GPT2 언어 모델을 통해 텍스트에는 명시적으로 드러나 있지는 않지만 내포되어 있는 사회적 편견을 연구하였다. Ben et al(2020)에서는 BERT(Bidirectional Encoder Representations from Transformers)에서 나타나는 장애에 대한 편견을 연구하였다. 이외에도 언어 모델에 나타난 종교 및 인종에 대한 연구 등이 진행되었다.

그간 언어 모델에 드러난 편견에 관한 연구는 대부분 성, 인종, 종교, 정치 성향 등에 집중되어 왔다.(Nikhil Garg et al., 2017; Maarten, Saadia and Lianhui, 2020;

---

2) Reddit 사이트에서 글이나 댓글에 사용하는 기능으로, 추천수(upvote)에서 비추천수(downvote)를 뺀 수치이다.

Ben et al., 2020; Ruibo Liu et al., 2021). 따라서 이 연구에서는 네이버 부동산 카페 및 네이버 뉴스 댓글 데이터로 학습시킨 언어 모델을 활용해 지역별 사람들의 인식을 분석하고, 지역별 주택 가격과 비교한다. 마지막으로 인식의 개선이 시급한 지역을 선별해, 해당 지역에 필요한 정책적 시사점을 제시할 것이다.

02. 연구 방법

1_데이터 수집 및 전처리

1) 지역 분석 단위

이 연구의 학습 데이터를 정의하기 앞서, 분석의 지역 단위에 대해 언급할 필요가 있다. 이 연구에서는 서울시의 행정동, 법정동, 자치구 등 행정 및 법정 구역의 명칭이 포함된 데이터를 학습 데이터로 만들었다. 키워드 집합을 만들 때, 행정동·법정동 등 명칭은 맨 뒤의 '동'을 제외하였고, 자치구는 맨 뒤의 '구'는 제외하였다. 예를 들어 '압구정동'은 '압구정'만을, '마포구'는 '마포'만을 키워드로 하였다. 다만 행정 및 법정 구역의 명칭이 한 글자인 경우에는, 맨 뒤의 '동' 혹은 '구'를 제외하지 않고 키워드에 포함하였다. 예를 들어 '명동', '중구'는 '동'이나 '구'를 제외하지 않고 그대로 키워드에 포함하였다. 단, 방배1동, 방배2동, 방배3동, 방배4동과 같이 행정 및 법정 구역의 명칭이 단순히 숫자로 구분되는 경우에는, '방배'만 키워드에 포함하였다. 최종적으로 분석의 지역 단위 키워드는 총 419개가 생성된다.

이를 통해 자치구와 비교해 세부 지역 단위인 행정동·법정동 단위의 편견을 파악할 수 있고, 같은 자치구 내에서도 인식이 크게 차이가 나는 행정동·법정동들에 대해서도 알 수 있도록 하였다.

2) 네이버 카페 및 뉴스 데이터

이 연구에서는 학습 데이터를 두 가지로 구성하였다. 첫 번째는 네이버 카페 '부동산 스터디' 댓글이다. 네이버 카페 부동산 스터디는 부동산 관련 국내 최대 규모의 카페로 2006년 설립되어 2021년 10월 26일 기준 회원 수가 173만여 명이다. 2021년 10월

26일 일요일 하루에만도 약 2,000여 개의 새 글이 작성되었다. 주로 부동산 투자에 관심이 있는 사람들이 이용하며, 투자 스터디를 위한 게시판, 매물을 홍보하는 게시판, 지역별 정보 나눔을 위한 게시판, 대출 및 세무, 법무 등을 문의하는 게시판 등으로 구성되어 있다. 이중 지역별 정보 나눔을 위한 게시판은 서울 지역의 자치구 단위로 분리되어 있다. 이 구별 게시판은 해당 구의 거주민들이 주로 글을 작성하는 것으로 추정되는 게시물들이 대다수다. 반면, 기타 다른 게시판들은 거주 지역과 상관없이 사람들이 글을 작성한다.

부동산 스터디 카페에서 2021년 3월 26일부터 8월 31일까지의 모든 글의 댓글을 수집하였다. 이때, 서울의 자치구별 게시판만을 크롤링 대상으로 하지 않고, 전체 게시물의 댓글을 크롤링하되, 서울의 행정 및 법정 구역 키워드가 포함된 댓글만을 선별하였다. 이와 같이 진행한 이유는, 자치구 등 지역 게시판은 해당 거주민들만 주로 이용하기 때문에 지역에 대한 거주민의 인식만을 수집하는 한계가 있기 때문이다. 따라서 지역에 대한 보편적인 인식을 파악하기 위하여, 전체 게시물의 댓글 중 서울 명칭이 포함된 댓글만을 먼저 선별하는 방식으로 크롤링하였다. 단, 단순 광고 및 홍보 글을 제외하기 위해 댓글이 5개 이상인 글만 크롤링하였다. 이렇게 선별된 댓글 중에서도, 특정 키워드('개발', '투자', '호재', '역세권', '부동산', '구역', '가치', '추진', '건축', '용적률', '임장', '안전진단', '거래', '재건축', 'GTX', 'gtx', '리모델링', '분담', '분담금', '조합')가 포함된 댓글은 제외하였다.

부동산 스터디 카페의 특성상 투자를 목적으로 해당 카페를 이용하는 경우가 많은데, '압구정 현대 용적률 300%로 높여줄 것입니다.', '북아현2구역 건축심의가 통과되었습니다.', '길음뉴타운 임장해 보세요.', '신길뉴타운 15억 거래 신고가'와 같은 댓글은 지역별 편견을 나타내지는 않는다. 위 키워드들이 포함된 문장 혹은 댓글을 데이터셋(dataset)에서 제거한 이유는 생성 모델(generation model)을 지역별 편견만을 생성(generation)하는 모델로 미세조정(finetuning)하기 위해서다.

두 번째는 네이버 뉴스 댓글 데이터를 수집하였다. 네이버 부동산 카페 데이터는 지역에 관련된 데이터만을 선별적으로 수집하는 데 유리하지만 대표성에 문제가 있을 수 있다. 부동산 카페 이용 회원은 부동산 투자에 관심이 있는 경향이 있으므로, 한국인을 대표한다고 보기는 다소 어렵기 때문이다. 따라서 더욱 범용적으로 이용되는 네이버 뉴스의 댓글을 추가로 크롤링하여 대표성의 문제를 보완하고자 하였다.

네이버 뉴스 댓글은 2019년 1월 1일부터 2020년 6월 11일까지의 댓글을 이용하였다.

부동산 카페에서와 마찬가지로 네이버 뉴스 댓글도 행정 및 법정 구역 키워드가 포함된 댓글로 한정하였다. 다만 부동산 카페의 글과 댓글은 대부분이 부동산과 관련이 있기 때문에, 행정 및 법정 구역 키워드가 포함된 댓글로만 선별하여도 문제가 없다. 반면, 행정 및 법정 구역 키워드로 선별한 네이버 뉴스 댓글은 지역 명칭과 동음인 다른 단어가 포함된 문장이 많았다. 예를 들어 부암동은 자궁경부암이 포함된 기사의 댓글까지 포함되는 것이다. 이와 같은 문제를 해결하고자 미캡(Mecab)[3]을 이용하여 수집한 댓글을 형태소 단위로 쪼개고, 지역 명칭과 동일한 고유명사 형태소가 있는 경우에만, 선별하여 학습 데이터로 수집하였다. 그리고 네이버 카페와 마찬가지로 투자 관련 키워드가 포함된 댓글은 제외하였다.

3) 데이터 요약

이렇게 수집된 데이터는 총 124MB, 약 95만 개 댓글이며, 상세 지역 명칭별로 수집된 댓글 수는 아래와 같다. 특정 지역으로의 쏠림이 뚜렷한 형태로 강남, 여의도, 이태원, 종로 등 지역에 관한 댓글이 많았다. 반면 충현동, 가락본동, 동빙고, 필운동, 행촌동의 경우에는 해당 지역이 언급된 댓글이 5개에 불과하였다.

3) 오픈소스 형태소 분석기 중 하나이다.

[표 2-1] 댓글에서 언급이 가장 많이 된 상위 20개 지역 및 언급 빈도수

지역	빈도	전체 댓글 수 대비	지역	빈도	전체 댓글 수 대비
강남	211,448	22.2%	강북	7,590	0.8%
여의도	26,146	2.8%	대치	7,367	0.8%
이태원	20,994	2.2%	반포	7,108	0.7%
종로	18,564	2.0%	구로	6,986	0.7%
용산	13,745	1.4%	마포	6,461	0.7%
잠실	11,320	1.2%	동대문	6,152	0.6%
서초	11,187	1.2%	마곡	6,021	0.6%
목동	10,221	1.1%	대학	6,015	0.6%
송파	9,044	1.0%	내자	5,638	0.6%
광진	8,980	0.9%	압구정	5,027	0.5%

[표 2-2] 댓글에서 언급이 가장 많이 된 하위 20개 지역 및 언급 빈도수

지역	빈도	전체 댓글 수 대비	지역	빈도	전체 댓글 수 대비
용신	17	0.002%	냉천	11	0.001%
중림	16	0.002%	예관	9	0.001%
잠실본	15	0.002%	예장	8	0.001%
입정	15	0.002%	노고산	8	0.001%
은천	15	0.002%	왕십리도선	5	0.001%
남대문로	15	0.002%	행촌	5	0.001%
교북	12	0.001%	필운	5	0.001%
홍지	12	0.001%	가락본	5	0.001%
염곡	12	0.001%	동빙고	5	0.001%
연건	12	0.001%	충현	5	0.001%

또한, 미캡(Mecab) 형태소 분석기를 활용하여, 특정 지역과 가장 많이 포함된 키워드들이 무엇인지 살펴보았다. 다만 실질적인 의미가 포함된 형태소만을 포함하기 위하여 품사가 일반명사, 고유명사, 외국어인 형태소들만 가져왔다. 〈표 2〉는 각 지역별로 댓글에서 함께 많이 언급된 일반명사, 고유명사, 외국어인 형태소들과 각 형태소들이 등장한 빈도수를 보여주고 있다. 강남 지역은 사람, 구청장, 아파트, 집값 등의 키워드가 포함된 댓글이 많은 반면, 동대문은 교회, 옷, 중국, 시장 등의 키워드들이 포함된 댓글이 많았다.

[표 2-3] 지역별 함께 언급된 일반명사, 고유명사, 외국어

지역	빈도	키워드(빈도)
강남	211,448	강남(230,386), 사람(22,692), 구청장(22,560), 아파트(20,547), 집값(19,172), 집(17,609), 서울(17,507), 돈(12,023), 말(10,081), 국민(7,682)
여의도	26,146	여의도(27,374), 강남(2,860), 국민(2,016), 서울(1,923), 사람(1,873), 집회(1,683), 광화문(1,428), 용산(1,417), 교회(1,349), 선(1,340)
이태원	20,994	이태원(22,167), 클럽(7,257), 사람(2,814), 강남(2,134), 코로나(1,953), 확진(1,935), 홍대(1620), 게이(1,550), 발(1,200), 검사(1,197)
종로	18,564	종로(19,032), 이낙연(2,202), 강남(2,157), 황교안(1,948), 서울(1,936), 사람(1,754), 구민(1,692), 교회(1,188), 출마(1,046), 대표(973)
용산	13,745	용산(13,776), 강남(4,789), 서울(2,473), 아파트(2,347), 마포(1,604), 서초(1,499), 여의도(1,413), 미군(1,228), 송파(1,225), 사람(1,079)
잠실	11,320	잠실(12,290), 강남(2,962), 서울(1,233), 아파트(1,046), 사람(1,038), 운동장(904), 송파(842), 반포(782), 서초(772), 여의도(705)
서초	11,187	서초(8,778), 강남(7,915), 송파(2,816), 서울(1,710), 서초동(1,688), 용산(1,566), 아파트(1,497), 여의도(1,402), 서초구(1,348), 사람(1,166)
목동	10,221	목동(11,048), 강남(2,218), 학군(1,517), 학원(1,436), 아파트(1,324), 여의도(1,278), 서울(823), 선(805), 마포(773), 사람(719)
송파	9,044	송파(8,607), 강남(5,696), 서초(2,851), 서울(1,669), 잠실(1,207), 아파트(1,200), 위례(1,196), 용산(1,180), 송파구(1,119), 사람(877)
광진	8,980	광진(8,849), 구민(1,788), 고민정(1,753), 오세훈(1,263), 사람(942), 광진구(906), 서울(870), 수준(758), 주민(739), 강남(669)
강북	7,590	강북(7,915), 강남(5,564), 서울(1,524), 아파트(1,020), 집값(774), 선(736), 지역(708), 사람(700), 집(647), 곳(456)
대치	7,367	대치(4,674), 대치동(2,442), 강남(2,061), 학원(1,611), 학군(1,245), 잠실(1,006), 반포(920), 목동(782), 서울(678), 아파트(640)
반포	7,108	반포(6,398), 강남(1,430), 아파트(1,148), 압구정(947), 잠실(875), 서초(737), 대치(696), 서울(639), 반(639), 개포(583)
구로	6,986	구로(3,071), 강남(1,273), 서울(1,190), 삼성(870), 거리(733), 아파트(695), 대림(673), 선(650), 구로구(648), 신구(645)
마포	6,461	마포(6,091), 강남(2,005), 용산(1,461), 서울(1,070), 여의도(1,035), 아파트(962), 목동(828), 마포구(686), 성동(585), 동작(540)
동대문	6,152	동대문(6,211), 교회(947), 옷(876), 중국(726), 시장(617), 사람(602), 명동(598), 서울(575), 남대문(418), 때(410)
마곡	6,021	마곡(6,633), xa(1,640), 서울(967), 공항(723), 자구(638), 여의도(630), 선(615), 김포(576), 강서구(554), 강남(524)
대학	6,015	대학(4,495), 강남(2,114), 삼성(1,764), 서울(1,167), 서울대(765), 때(716), 국민(695), 병원(678), 애(657), 대학교(649)
내자	5,638	내자(4,777), 끝(975), 대한민국(834), 힘(827), 국민(607), 나라(606), 세금(583), 삼성(529), 일본(507), 때(368)
압구정	5,027	압구정(5,043), 강남(1,257), 교회(1,002), 아파트(981), 반포(917), 현대(876), 청담(630), 사람(470), 대치(456), 잠실(432)

2_KoGPT2 미세조정

위에서 수집된 학습 데이터를 가지고, SKT에서 공개한 KoGPT2 version 2(이하 KoGPT2)를 미세 조정(finetuning)하였다. SKT에서 공개한 KoGPT2는 한국어 텍스트로 학습한 GPT2모델이다. OpenAI에서 공개한 GPT2와 같이 오토리그레시브 디코더(Autoregressive Decoder) 모델로, 토크나이저(tokenizer) 또한 OpenAI와 GPT2와 동일하게 BPE(Bytepair Encoding)[4]방식을 적용하였으며, 모델의 주요 구조는 아래 [표 2-3]와 같다. (Alec et al., 2019; SKT-AI, 2021)

[표 2-4] KoGPT2의 주요 구조

Model	# of params	Type	# of layers	# of heads	ffn_dim	Hidden_dims
KoGPT2	125M	Decoder	12	12	3072	768

KoGPT2는 한국어 위키 백과, 뉴스, 모두의 말뭉치 v1.0, 청와대 국민청원 등 40GB 이상의 다양한 데이터로 학습되었다고 한다. (SKT-AI, 2021) 이모지, 이모티콘 등이 단어(vocab) 리스트에 포함된 것으로 보아 학습 데이터에는 공식적이고 정제된 데이터뿐만 아니라, 댓글과 같이 비공식적이고 정제되지 않은 데이터가 포함되어 있을 것이라고 추정된다.

학습(Training)과 검증(validation) 데이터셋(dataset)은 7:3으로 구분하였으며, 주요 학습 아규먼트(training argument)들은 아래와 같이 수행하였다.

number of training epochs: 10
batch size for training: 32
batch size for evaluation: 64
Number of update steps between two evaluations: 400
number of warmup steps for learning rate scheduler: 500

4) Bytepair Encoding 서브 워드 분리 알고리즘으로, 연속적으로 가장 많이 등장한 글자의 쌍을 찾아서 하나의 글자로 병합하는 방식 수행한다. (Gage and Philip, 1994; Rico et al., 2016)

토크나이저(Tokenizer)는 일부 수정하였다. KoGPT2는 총 51,200개의 단어(Vocab) 리스트가 있는데, 이 중에는 〈unused0〉 ~ 〈unused99〉와 같이 총 100개의 미사용 토큰이 있다. 이 연구에서는 앞서 구 및 동의 명칭 리스트를 총 419개로 정의하였는데, 이중 이미 KoGPT2에 단어(Vocab)로 포함된 명칭은 54개[5]이다. 나머지 365개 중에서 앞선 탐색적 데이터 분석(EDA, Exploratory Data Analysis)를 통해, 댓글에서 많이 등장한 지역인 97개를 단어(Vocab) 리스트에 포함하기 위하여, 〈unused0〉 ~ 〈unused96〉을 97개의 지역명으로 대체하였다[6]. 따라서 앞으로의 연구는 총 151개의 지역에 대해 진행하였다.

KoGPT2 미세조정(finetuning)에는 Google Colab Pro가 사용되었으며, GPU는 Nvidia cuda version 11.2가 사용되었다. 학습시간은 약 18시간 내외로 소요되었다.

5) KoGPT2에 기포함된 명칭은 다음과 같다. ['주자', '당인', '구의', '구로', '인사', '왕십리', '연지', '암사', '상계', '미아', '원서', '일원', '개화', '대학', '신정', '돈의', '인의', '삼성', '신사', '우면', '천연', '망원', '신설', '방이', '송정', '인수', '천왕', '중앙', '삼각', '장사', '통의', '주교', '산림', '청구', '수송', '남산', '대신', '합동', '서원', '가락', '성동', '동자', '장교', '상도', '공항', '대조', '동작', '대방', '증산', '정동', '동화', '도화', '사당', '강남']

6) Unused 토큰 대신 사용한 지역 명칭은 다음과 같다. ['위례', '을지로', '잠실', '명륜', '신도림', '구기', '장지', '이촌', '원지', '잠원', '반포', '용산', '강동', '수색', '내수', '내자', '신천', '거여', '초동', '마곡', '수하', '도곡', '중계', '행당', '재동', '양재', '수서', '천호', '한남', '이태원', '오류', '중랑', '상일', '삼전', '이문', '염창', '둔촌', '강서', '광진', '성북', '청담', '성수', '송파', '계동', '신길', '관악', '광장', '종로', '양천', '시흥', '도림', '고덕', '대림', '노원', '개포', '신촌', '상암', '중구', '길음', '아현', '역삼', '가산', '서대문', '압구정', '인헌', '방배', '명일', '강북', '노량진', '평창', '흑석', '예지', '여의도', '문정', '창동', '조원', '대치', '공덕', '장안', '동대문', '동선', '금호', '명동', '신림', '우장산', '다산', '서초', '중학', '은평', '목동', '성산', '방화', '개봉', '마포', '당산', '청량리', '영등포']

3_프롬프트 및 문장생성

미세조정(Finetuning)한 KoGPT2를 가지고 지역별로 아래의 프롬프트(Prompt)들에서 뒷부분을 생성(generation)하였다.

프롬프트(Prompt)는 크게 두 가지로 나누어 작성하였다. 첫 번째는 주거만족도 및 근린만족도에 관한 선행연구에서 중요한 변수로 제시된 교통/교육/지역 환경 프롬프트(prompt)다. 이를 통해 전통적으로 이미 입증된 변수들에 대한 지역별 스코어를 분석하고자 하였다. 두 번째는 댓글 데이터에서 빈번하게 나타나는 토큰(token)들의 조합으로 작성한 프롬프트다. 학습 데이터에서 충분히 등장한 문장일수록 생성 모델의 성능이 우수할 것이며, 댓글에서 자주 나타난 지역별 편견을 파악할 것이기 때문이다.

트랜스포머(Transformers)의 파이프라인(pipeline)을 이용하여 지역별로 한 프롬프트(prompt) 당 총 5개의 문장을 생성(generation)하였다. 지역명은 앞서 정의한 419개의 키워드 중에서 미세조정(finetuning)한 단어(vocab) 리스트에 포함된 151개에만 한정하였다. 그래서 지역 151개 X 프롬프트(prompt) 종류 6개 X 한 프롬프트(prompt)에서 생성(generation)하는 문장 수 5개 = 총 4,530개의 문장을 생성하였다.

[표 2-5] KoGPT2로 생성할 프롬프트

구분		프롬프트
선행연구	○ 교육 환경	○ 〈지역명〉 아이들은 _____.
	○ 지역 환경	○ 〈지역명〉 주변에 _____.
	○ 교통 환경	○ 〈지역명〉은/는 교통이 _____.
빈도 분석		○ 〈지역명〉 사람들은 _____. ○ 〈지역명〉 민도는 _____. ○ 〈지역명〉 비싼 이유는 _____.

4_KoBERT를 이용한 감성 분석

앞에서 생성한 4,530개의 문장에 대해 감성 분석(sentiment analysis)를 수행하였다. 감성 모델(Sentiment model)은 두 가지의 후보군을 고려하였다. 첫 번째는 카카오 브레인(Kakao Brain)에서 공개한 Pororo[7]다. 카카오(Kakao)에서 2021년에 공개한 Pororo는 다양한 서브태스크(subtask)에서 활용할 수 있는 자연어 처리 모델이다[8]. Pororo가 수행 가능한 서브태스크(subtask) 중 하나는 네이버 영화 댓글 데이터(Naver sentiment movie corpus)로 학습된 감성 분석 모델이다. 이 모델은 해당 문장이 긍정인지 부정인지를 알려줄 뿐만 아니라 긍·부정의 확률(probability)도 함께 산출해 준다. 하지만, 아래의 테스트 문장으로 Pororo의 성능을 검증해 본 결과, 맨 마지막의 '사랑해요'를 제외한 나머지 모든 문장들을 부정적(Negative)로 예측하여 사용하지 않았다.

'대치는 학군이 좋아요'
'대치는 학군이 안 좋아요'
'대치는 학원 가기가 편리해요'
'대치는 학원 가기가 불편해요'
'사랑해요'

두 번째로 고려한 모델은 KoBERT를 네이버 영화 댓글 데이터(Naver sentiment movie corpus, NSMC)로 미세조정(finetuning)한 모델이다. 트랜스포머(Transformers)에 공

---

7) Pororo는 Platform Of neuRal mOdels for natuRal language prOcessing의 준말이다. (Heo et al., 2021)

8) Available tasks are ['mrc', 'rc', 'qa', 'question_answering', 'machine_reading_comprehension', 'reading_comprehension', 'sentiment', 'sentiment_analysis', 'nli', 'natural_language_inference', 'inference', 'fill', 'fill_in_blank', 'fib', 'para', 'pi', 'cse', 'contextual_subword_embedding', 'similarity', 'sts', 'semantic_textual_similarity', 'sentence_similarity', 'sentvec', 'sentence_embedding', 'sentence_vector', 'se', 'inflection', 'morphological_inflection', 'g2p', 'grapheme_to_phoneme', 'grapheme_to_phoneme_conversion', 'w2v', 'wordvec', 'word2vec', 'word_vector', 'word_embedding', 'tokenize', 'tokenise', 'tokenization', 'tokenisation', 'tok', 'segmentation', 'seg', 'mt', 'machine_translation', 'translation', 'pos', 'tag', 'pos_tagging', 'tagging', 'const', 'constituency', 'constituency_parsing', 'cp', 'pg', 'collocation', 'collocate', 'col', 'word_translation', 'wt', 'summarization', 'summarisation', 'text_summarization', 'text_summarisation', 'summary', 'gec', 'review', 'review_scoring', 'lemmatization', 'lemmatisation', 'lemma', 'ner', 'named_entity_recognition', 'entity_recognition', 'zero-topic', 'dp', 'dep_parse', 'caption', 'captioning', 'asr', 'speech_recognition', 'st', 'speech_translation', 'ocr', 'srl', 'semantic_role_labeling', 'p2g', 'aes', 'essay', 'qg', 'question_generation', 'age_suitability'] (Heo et al, 2021)

개된 NSMC 데이터로 미세조정(finetuning)한 모델들 가운데 위의 테스트 문장의 라벨을 모두 맞춘 모델은 "sackoh/bert-base-multilingual-cased-nsmc"가 유일하였다[9]. 따라서 이 연구에서는 이 모델을 이용하여 감성 스코어(sentiment score)를 산출하였다.

이 모델을 가지고 앞서 추출한 4천 5백여 개의 문장의 스코어를 산출하였다. 각 문장별 최종 스코어는, 긍정 라벨일 때에는 스코어(score)를 그대로 가져오고, 부정 라벨일 때에는 스코어(score)×(-1)을 하여 산출하였다. 그래서 같은 긍정(positive)이라도 더욱 명확한 긍정의 표현은 높은 점수를, 상대적으로 불명확한 긍정(positive)은 낮은 점수를 받도록 하였다. 또한 긍정이 부정보다 높은 점수를 받도록 하여, 긍정에 가까울수록 1점을, 부정에 가까울수록 -1점이 되도록 하였다. 한 지역의 한 프롬프트(prompt)마다 총 5개의 문장을 생성(generation)하였기 때문에, 한 지역의 한 프롬프트(prompt)별로 스코어를 구하기 위하여 다섯 개의 문장의 점수를 단순 평균하여, 최종적으로는 지역별로 프롬프트(prompt)별로 점수가 산출되도록 하였다.

9) https://huggingface.co/sackoh/bert-base-multilingual-cased-nsmc

5_감성 스코어 집계 및 주택 가격과 비교 분석

위에서 산출한 지역별 스코어는 크게 세 가지 방식으로 분석을 진행하였다. 첫 번째는, 구별로 스코어를 집계하고 이를 최근 아파트 실거래가와 비교하는 방식이다. 학습 데이터인 댓글은 익명성을 기반으로 작성되어, 특정 지역에 대한 사람들의 인식이 드러나 있다. 가격에는 사람들의 인식이 반영되므로, 구별 스코어와 아파트 실거래가를 비교 분석하였다. 이때 실거래가와 스코어가 차이가 큰 지역의 경우에는 프롬프트별 스코어와 생성된 문장을 면밀히 분석해 보았다.

두 번째는, 동일 구 내에서도 동별로 스코어 차이를 비교하였다. 같은 구 내에서도 동·프롬프트별 스코어의 차이가 있는지 살펴봄으로써, 같은 구 내에서도 동별로 사람들의 인식이 다르게 나타나는지 파악하고자 하였다.

세 번째는, 지역별로 특징적인 문장 사례를 선별하여 살펴보는 방식이다. 지역·프롬프트별 스코어가 모두 집계되어 있기 때문에, 스코어가 낮은 지역 및 프롬프트 조합은, 생성된 문장 사례를 살펴보아, 해당 지역에 어떠한 문제점이 있는지 파악하고자 하였다.

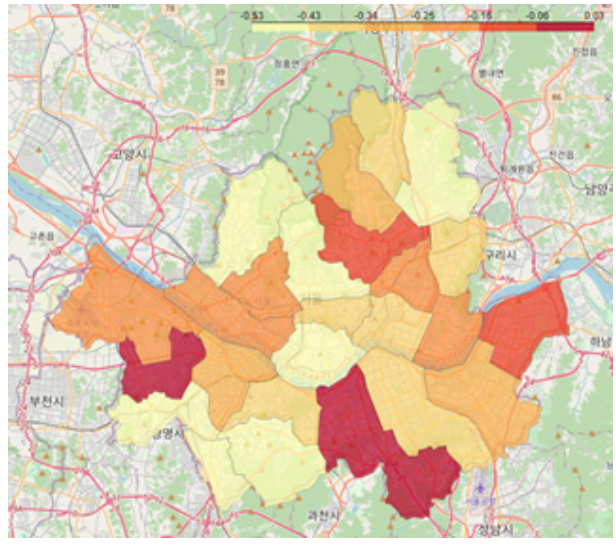
03. 감성 스코어와 실거래 평단가 관계 분석

1_구별 감성 스코어 및 실거래 평단가

구별로 프롬프트별 감성 스코어(이하 스코어)를 모두 단순 평균하여 구별 스코어를 집계하였다. 앞서 언급한 바와 같이 총 151개의 지역 키워드 집합을 연구의 범위로 한정하였는데, 151개의 키워드를 구별로 구분하여 평균 스코어를 구하였다. 예를 들어 강남구는 신사, 강남, 개포, 대치, 도곡, 삼성, 수서, 압구정, 역삼, 일원, 청담이라는 지역 키워드에 대한 프롬프트별 스코어를 평균하였다. 강남구의 경우에는 강남이라는 구 명칭과 신사, 개포 등과 같은 동 명칭이 모두 포함되어 스코어가 산출되었다. 이때, 스코어는 생성된 문장별로 -1~+1사이의 값을 가질 수 있는데, 구별로 집계하였을 때 최대값은 서초구의 약 0.0, 최소값은 금천구 약 -0.5였다.

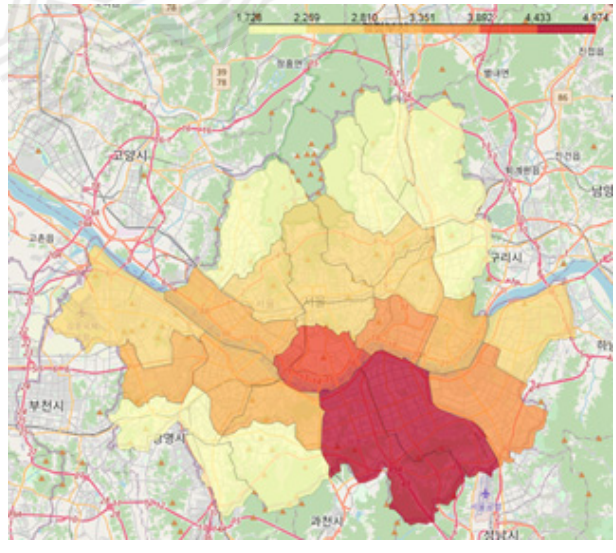
[표 3-6] 구별 평단가 및 스코어

구	평단가(만 원)	스코어	구	평단가(만 원)	스코어
강남구	4,974	-0.4	서대문구	2,602	-0.2
서초구	4,687	0.0	성북구	2,437	-0.1
용산구	4,002	-0.4	동대문구	2,384	-0.2
성동구	3,839	-0.3	강서구	2,332	-0.2
송파구	3,545	-0.3	노원구	2,154	-0.5
마포구	3,290	-0.2	관악구	2,137	-0.4
광진구	3,248	-0.2	은평구	2,036	-0.4
동작구	3,033	-0.4	구로구	1,953	-0.4
영등포구	2,969	-0.4	강북구	1,952	-0.3
양천구	2,897	0.0	중랑구	1,938	-0.4
중구	2,738	-0.4	금천구	1,871	-0.5
강동구	2,738	-0.1	도봉구	1,728	-0.4
종로구	2,733	-0.4			



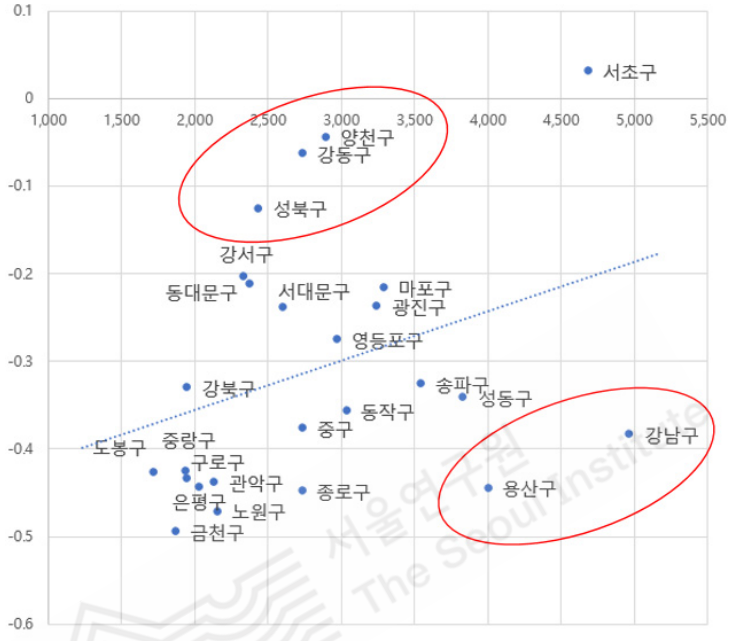
[그림 3-1] 구별 스코어

구별로 살펴보면, 서초구, 양천구가 높은 스코어를 보이며, 노원구, 은평구, 구로구, 금천구, 용산구, 종로구는 상대적으로 낮은 스코어를 보이고 있다.
이 스코어와 비교를 하기 위하여 구별로 아파트 실거래 평단가(2020년 7월 기준)를 함께 살펴보았다.



[그림 3-2] 구별 아파트 실거래 평단가(만 원)

강남구, 서초구가 각각 4,974만 원, 4,687만 원으로 높은 평단가를 보이며, 노원구, 도봉구, 은평구, 금천구, 구로구가 상대적으로 낮은 평단가를 보이고 있다.
스코어와 아파트 실거래가를 비교하기 위하여, x축을 아파트 실거래 평단가로, y축을 스코어로 하여 산점도를 그리면 아래와 같다



[그림 3-3] 구별 스코어-평단가 산점도

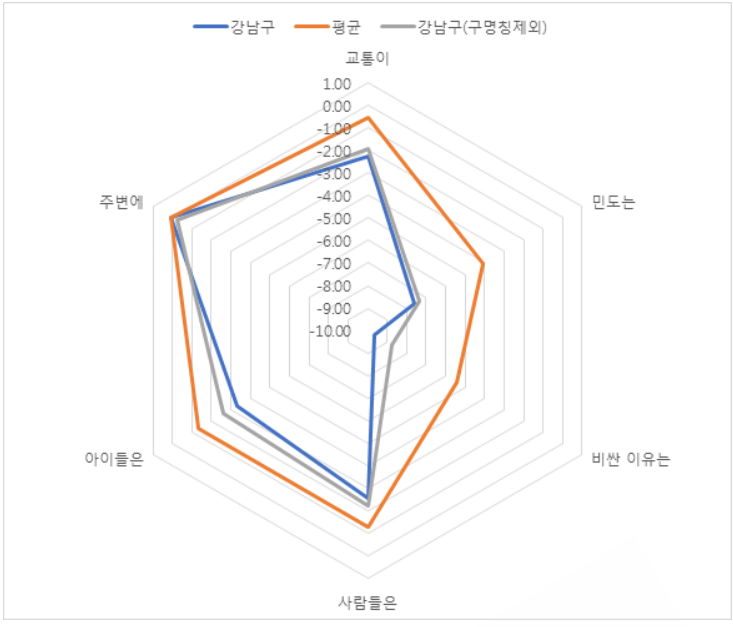
이때, 회귀선과 상대적으로 거리가 먼 지역이 일부 관측되었다. 평단가 대비 스코어가 높은 지역으로는 성북구, 양천구, 강동구가 있었다. 각 구별로 스코어가 평단가 대비 상대적으로 높은 이유를 분석해 보면, 세 구가 공통적으로 구 평균 대비 평단가가 높은 동만 분석 대상이 포함되었기 때문이었다. 성북구의 경우, 구 전체의 평균 실거래 평단가 2,437만 원 대비 실거래 평단가가 높은 길음동(2,991만 원)만 스코어 산출에 포함되었다. 양천구의 경우도, 구 전체 평균 실거래 평단가 2,897만 원 대비 평단가가 높은 목동(3,687만 원)만 스코어 산출에 포함되었다. 강동구의 경우, 구 전체 평균 실거래 평단가 2,738만 원 내에서 동별 실거래가 평단가가 가장 낮은 길동(2,173만 원), 성내동(2,229만 원)만 스코어 산출에 미포함되었다. 댓글 내에서 해당 구 내에서 평단가가 높은 특정 동만 언급된 영향이므로, 사람들의 인식 내에서 성북구, 양천구, 강동구는 구

내에서 평단가가 높은 일부 동에만 관심이 집중되었음을 추정할 수 있다.

반면에 평단가가 스코어 대비 낮은 지역으로는 강남구, 용산구가 있었다. 용산구의 경우에는 이태원 코로나19 집단감염 관련 댓글의 영향으로 스코어가 낮게 나왔다. 학습 데이터를 구축할 때에 사람들의 인식을 반영하는 모든 댓글을 수집하는 것이 목적이었다. 이 때문에 수집된 데이터 중 '임장해 보세요.', '신고가 경신'과 같이 단순한 투자 관련 댓글만 제거하였고, 이태원 코로나19 집단감염으로 인한 부정적 인식이 포함된 댓글이 포함되어 해당 시기의 특정 사건(Event)에 대한 효과가 반영되었다. 이태원동 제외 시의 스코어는 성동구 수준으로 스코어가 올라간다.

강남구의 경우에는 특정 동만 포함되거나 특정 동이 미포함되어 스코어가 낮게 나온 것으로는 스코어가 낮은 현상이 설명되지 않았다. 강남구는 아래에서 좀 더 살펴보겠다. 모든 동의 데이터가 없거나, 특정 이슈가 있는 동이 포함되어 특이치가 나타난 성북, 양천, 강동, 용산구를 제외할 경우, 스코어와 아파트 실거래가 평단가의 상관계수는 0.4969 수준이다.

특정 동의 영향으로 설명되지 않는 강남구의 경우에는 프롬프트별로 스코어를 전체 평균과 비교하였다. '주변에'를 제외하고는, 모든 부분에서 강남이 전체 평균 대비 스코어가 낮아, 대부분의 영역에서 강남구에 대한 부정적 인식을 확인할 수 있었다. 다만 '강남'이라는 키워드는 '강남구' 외에 한강 이남 지역으로 해석될 수 있어, '강남'이라는 단어의 중의성 영향 여부 확인이 필요하다. 따라서 '강남'이라는 키워드는 제외하고 신사, 개포, 대치, 도곡, 삼성, 수서, 압구정, 역삼, 일원, 청담과 같이 '동'의 명칭만 포함하여 강남의 스코어를 재산출하였다. 그러나 '주변에'를 제외하고 모든 프롬프트에서 스코어가 소폭 상승한 수준에 불과하여, '강남'이라는 단어의 중의성으로 인한 스코어 하락 효과는 미미했다.



[그림 3-4] 강남구 프롬프트별 스코어

프롬프트별 강남의 스코어는, 특히 '민도는', '비싼 이유는'과 같이 부동산 카페에서 빈번하게 사용되어 프롬프트로 선정된 부분에서 특히 더 낮았는데, 이는 강남 내 타지역과 비교하는 댓글 영향으로 추정된다. 생성된 문장의 사례를 보면 다음과 같다.

(생성된 문장 사례)
신사 민도는 괜찮지만, 강남보다 불편함이 더 많음
대치 민도는 강남서초보다 못하고 압구정 청담 같은 데 가서 더 난리 치기는
청담이 비싼 이유는 뭘까???? 님이 그 가격이면 대치동, 청담동 살지

강남의 경우에는 타지역 대비 강남 내 타지역과 비교하는 댓글이 많이 생성되었으며, 이로 인해 스코어가 낮아지는 경향을 보였다. 즉, 강남구는 내의 동에 대한 평가는 강남구 내의 타지역과 상대평가를 하는 경향을 보였다.

2_동일 구 내에서의 감성 스코어 비교

동일 구 내에서도 동별로 스코어 차이가 나타나는지 분석하였다. 분석의 대상은, 상대적으로 많은 동들이 분석에 포함된 '강남구'와 이태원동 코로나19 집단감염 여파로 평균 스코어가 낮아진 '용산구'로 하였다. 우선 강남구의 경우에는, 동별로는 평균 스코어와 아파트 실거래가는 상관관계가 없었다. 앞서 언급된 것처럼 강남구의 경우에는 동별 스코어를 산출할 때, 강남구 내의 타지역과의 비교 댓글 영향으로 보인다. 또한 같은 구 내에서도 동별 프롬프트별 편차가 크게 나타났다. 예를 들어, 압구정동의 경우에는 '아이들은'을 제외하고 전반적으로 높은 스코어를 보였다. 반면에 일원동의 경우에는 '아이들은', '주변에'와 같이 아이들과 주변 환경에 대한 평가는 우수하였으나, 그 외 평가는 상대적으로 나쁘게 나왔다.

[표 3-2] 강남구 프롬프트별 스코어 및 평단가

강남구	스코어					평단가(만 원)
	아이들은	사람들은	주변에	교통이	평균	
신사	0.57	0.53	-0.86	-0.29	-0.01	3,576
강남	-0.69	-0.29	0.29	-0.35	-0.26	4,974
개포	-0.75	-0.31	-0.74	-0.27	-0.52	6,556
대치	-0.20	0.64	0.08	-0.24	0.07	4,238
도곡	0.11	-0.54	-0.29	0.20	-0.13	4,938
삼성	-0.79	-0.77	-0.32	-0.76	-0.66	4,952
수서	-0.25	-0.22	0.60	-0.22	-0.02	5,366
압구정	-0.28	0.58	0.46	0.68	0.36	6,335
역삼	-0.67	-0.66	-0.23	-0.13	-0.42	4,239
일원	0.31	-0.75	0.84	-0.71	-0.08	5,504
청담	-0.69	-0.75	0.21	-0.18	-0.35	4,893

용산구의 경우에는, 강남구와 달리 동별 평균 스코어와 아파트 실거래가 평단가와의 상관계수가 0.9230으로 매우 높게 나왔다. 댓글에서 나타나는 인식 및 편견과 평단가 트렌드가 매우 유사함을 알 수 있다. 또한 강남구와 마찬가지로 같은 구 내에서도 동별 프롬프트별 편차가 크게 나타났다. 예를 들어 이태원동은 전반적으로 모든 평가가 나쁘게 나왔다. 이촌동은 거주민에 대한 평가(프롬프트 '아이들은', '사람들은')는 우수하였으나, 주변 환경에 대한 평가(프롬프트 '주변에', '교통이')는 나쁘게 나왔다. 반면 한남동은 거주민에 대한 평가는 나빴지만, 주변 환경에 대한 평가는 우수하였다.

[표 3-3] 용산구 프롬프트별 스코어 및 평단가

용산구	스코어					평단가(만 원)
	아이들은	사람들은	주변에	교통이	평균	
동자	-0.30	-0.08	-0.23	-0.63	-0.31	3,657
용산	-0.10	-0.74	0.27	-0.34	-0.23	4,002
이촌	0.66	0.14	-0.69	-0.61	-0.13	4,447
이태원	-0.79	-0.72	-0.87	-0.63	-0.75	3,372
한남	-0.62	-0.24	0.85	0.25	0.06	4,532

3_특징적 문장 도출

마지막으로 지역·프롬프트별 조합 중 스코어가 낮은 조합을 중심으로 생성된 문장의 사례를 살펴보고, 사람들이 해당 지역에 어떤 문제가 있다고 인식하는지 추정하였다.
몇몇 사례를 살펴보면, 동대문은 주변 환경에 대한 프롬프트에서 노점상 문제를 언급하고, 은평은 지하철과 같은 대중교통 문제를 지적한다. 대림동에 관한 생성 문장을 통해서는, 조선족 밀집 거주 지역으로 인식하기 때문에 나타나는 대림동에 대한 부정적 인식이 나타났다. 신길동 또한 대림동과 유사한 편견이 존재함을 알 수 있다. 아현동은 상권의 쇠퇴 문제가 언급되었고, 우면동은 중소기업이 부족한 부분이 나타났다. 그 외 미아동은 교통이 불편한 부분이 언급되었고, 정동의 경우 교통이 나쁘지는 않지만, 버스가 정체되는 문제가 나타났다.
이러한 방법을 통해 해당 지역에 대한 댓글을 모두 읽지 않아도, 특징적인 이슈가 무엇인지 파악할 수 있고, 이를 통해 스코어가 낮은 지역의 문제를 파악할 수 있다.

04. 결론

이 연구는 비정형 데이터인 인터넷 댓글 데이터를 분석하기 위해 KoGPT2 모델을 활용해 지역별로 대표 문장을 생성하고, 이를 KoBERT 모델로 감성 스코어를 산출해 아파트 실거래가와 비교했다.

서울시 내 25개 행정구의 아파트 실거래가와 온라인 댓글 데이터를 통해 미세조정된 GPT2 모델에서 생성된 프롬프트의 감성 스코어는 상관계수가 0.3557로 나타났다. 실거래가와 감성 스코어 사이의 차이가 큰 지역은 성북구, 양천구, 강동구, 용산구, 강남구로 나타났다. 성북구와 양천구는 구 내에서 평균 실거래가가 높은 길음동과 양천구 등 일부 동에 대한 댓글 데이터만 수집되었고, 강동구는 실거래가가 낮은 길동과 성내동에 대한 댓글 데이터가 없어 실거래가 대비 감성 스코어가 높게 분석되었다. 용산구는 자료 수집 기간 중의 코로나19 집단감염으로 인해 실거래가 대비 감성 스코어가 낮게 나타났다. 수집된 댓글 데이터가 부족하여 감성 스코어의 신뢰도가 부족한 성북구, 양천구, 강동구, 용산구를 제외할 경우 상관계수는 0.4969로 올라갔다.

강남구의 감성 스코어가 실거래가에 비해 낮게 산정된 이유는 강남의 경우 특정지역을 평가할 때 비교의 대상이 되면서, 강남에 대해서도 감성 스코어가 낮게 산정되기 때문이다.

이 연구는 데이터의 양이 많아 기존의 방법론으로는 분석이 어려운 인터넷 댓글을 분석할 수 있는 방법론을 제시했다. 특히, 공간적 범위를 고려하기 어려웠던 기존 방법론에 비해 동 단위의 지역별 분석이 가능하며, 프롬프트 설계에 따라 특정 주제에 대한 분석이 가능하다는 점에서 의의가 있다.

이 연구의 결과는 지역에 대한 시민의 인식을 실시간으로 파악하는 양적 지표 중 하나로 활용될 수 있을 것이다. GPT2 모델을 통해 생성된 문장은 지역에 대한 특성을 드러내기 때문에 시민이 특정 지역에 대해 어떠한 인식을 가지고 있는지 실시간으로 파악할 수 있다. 서울 전역에 대해 정기적으로 모니터링하여 스코어가 큰 폭으로 변화하는 지역이 나타나는 경우, 원인을 분석하고 이에 대한 정책적 대안을 수립할 수 있는 기초자료로

활용될 수 있을 것이다. 또한 타지역에 비해 스코어가 낮다는 것은 지역에 대한 부정적 인식 또는 편견이 있다는 것을 뜻하기 때문에, 특정 지역의 인식개선을 위한 정책을 우선적으로 시행하기 위한 기준이 될 수 있다.

하지만 인터넷 댓글은 조작 또는 선동의 목적으로 작성될 수 있으나, 이를 구분하기 어려우며, 특정 시기에 대두되는 이슈에 따라 지역에 대한 평가가 크게 달라질 수 있다. 또한 관심이 적어 데이터의 양이 부족한 지역은 정확도가 낮을 수 있다는 한계가 있어 이를 개선하기 위한 추가적인 연구가 진행될 필요가 있다.

참고문헌

경정익 · 이국철, 2016, “Textmining에 의한 부동산 빅데이터 감성분석 모형 개발”, 「주택연구」, 24(4), pp.115~136.

김대원 · 유정석, 2016, “트위터 정보와 아파트 매매 및 전세 가격 간 동적 관계 분석”, 「도시행정학보」, 29(1), pp.1~33.

김진유, 2006, “신문기사가 부동산가격변동에 미치는 영향: ‘투기’가 포함된 신문기사와 주택가격간의 그랜저인과관계분석을 중심으로”, 「주택연구」, 14(2), pp.39~63.

박재수 · 이재수, 2019a, “벡터자기회귀모형을 이용한 온라인 뉴스기사와 아파트 매매가격의 동태적 관계 연구: 비정형 빅데이터를 활용한 감성분석 기법의 적용”, 「감정평가학 논집」, 18(2), pp.83~113.

박재수 · 이재수, 2019b, “아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석: 온라인 뉴스 기사의 비정형 빅데이터를 활용한 감성분석 기법의 적용”, 「국토계획」, 54(1), pp.131~147.

박재수 · 이재수, 2021, “부동산 감성지수의 주택가격 예측 유용성: 뉴스기사와 방송뉴스 빅데이터 활용 사례”, 「국토계획」, 56(4), pp.99~111.

배성완 · 유정석, 2018, “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, 「주택연구」, 26(1), pp.107~133.

이종민 · 이종아 · 정준호, 2017a, “뉴스 빅데이터를 이용한 전세 가격 예측: 토픽모형 분석을 중심으로”, 「부동산학보」, pp.43~57.

이종민 · 이종아 · 정준호, 2017b, “포털 검색 지수를 활용한 전세 가격 예측: 네이버 · 구글을 중심으로”, 「부동산학보」, pp.134~148.

이태형 · 전명진, 2018, “딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구: 다변량 시계열 자료를 중심으로”, 「주택도시연구」, 8(2), pp.39~56.

전해정 · 양혜선, 2019, “딥 러닝을 이용한 주택가격 예측에 관한 연구”, 「주거환경」, 17(2), pp.37~49.

진창하 · Paul Gallimore, 2012, “신문기사 내용과 주택가격: 인식, 사유, 그리고 투자심리”, 「부동산학

연구」, 18(2), pp.125~142.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, 2018, Language Models are Unsupervised Multitask Learners.

Alec, Jeffery, Rewon, David, Dario and Ilya, 2019. Language Models are Unsupervised Multitask Learners.

Ben, Vinodkumar, Emily, Kelle, Yu and Stephen, 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities.

Gage and Philip, 1994. A New Algorithm for Data Compression.

Heo, Ko, Kim, Han, Park and Park, 2021. PORORO: Platform Of neuRal mOdels for natuRal language prOcessing, https://github.com/kakaobrain/pororo/blob/master/README.md.

Maarten, Saadia and Lianhui 2020. SOCIAL BIAS FRAMES: Reasoning about and Power Implications of Language.

Nikhil Garg, Londa, Dan and James, 2017. Word Embeddings Quantify 100 Years of Gender and Ethic Stereotypes.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, Soroush Vosoughi, 2021, Mitigating Political Bias in Language Models Through Reinforced Calibration

https://github.com/SKT-AI/KoGPT2(SKT-AI, 2021).

https://huggingface.co/sackoh/bert-base-multilingual-cased-nsmc.

https://www.kaggle.com/junbumlee/kcbert-pretraining-corpus-korean-news-comments(Beomi, 2020).

https://www.seoul.go.kr/seoul/autonomy_sub.do(서울특별시, 자치구별 동 현황(한글, 한자)).

https://thegradient.pub/prompting/(Tianyu Gao, 2021.07.03., "Prompting: Better Wasys of Using Language Models for NLP Tasks").

작은연구 좋은서울 21-11

자연어 처리 모델을 활용한
인터넷 댓글과 부동산 가격의 관계 분석

발행인 유기영
발행일 2021년 11월 9일
발행처 서울연구원
비매품

06756 서울특별시 서초구 남부순환로 340길 57