

텍스트 마이닝을 활용한 지역 특성 기반 도시재생 유형 추천 시스템 제안*

김익준

아주대학교 e-비즈니스학과
(kij2648@ajou.ac.kr)

김효민

아주대학교 e-비즈니스학과
(hyomin1028@ajou.ac.kr)

이준호

아주대학교 e-비즈니스학과
(dlwngshw@naver.com)

강주영

아주대학교 e-비즈니스학과
(jykang@ajou.ac.kr)

.....

현 정부의 주요 국정사업 중 하나인 도시재생 뉴딜사업은 매년 100 곳씩, 5년간 500곳을대상으로 50조를 투자하여 낙후된 지역을 개발하는 것으로 언론과 지자체의 높은 이목이 집중되고 있다. 그러나, 현재 이 사업모델은 면적 규모에 따라 “우리동네 살리기, 주거정비지원형, 일반근린형, 중심시가지형, 경계기반형” 등 다섯 가지로 나뉘어 추진되어 그 지역 본래의 특성을 반영하지 못하고 있다. 국내 도시재생 성공 키워드는 “주민 참여”, “지역특화” “부처협업”, “민관협력”이다. 성공 키워드에 따르면 지자체에서 정부에게 도시재생 사업을 제안할 때 지역주민, 민간기업의 도움과 함께 도시의 특성을 정확히 이해하고 도시의 특성에 어울리는 방향으로 사업을 추진하는 것이 가장 중요하다는 것을 알 수 있다. 또한 도시재생 사업 후 발생하는 부작용 중 하나인 젠트리피케이션 문제를 고려하면 그 지역 특성에 맞는 도시재생 유형을 선정하여 추진하는 것이 중요하다.

이에 본 연구는 ‘도시재생 뉴딜 사업’ 방법론의 한계점을 보완하기 위해, 기존 서울시가 지역 특성에 기반하여 추진하고 있는 “2025 서울시 도시재생 전략계획”의 도시재생 유형을 참고하여 도시재생 사업지에 맞는 도시재생 유형을 추천하는 시스템을 머신러닝 알고리즘을 활용하여 제안하고자 한다. 서울시 도시재생 유형은 “저이용저개발, 쇠퇴낙후, 노후주거, 역사문화자원 특화” 네 가지로 분류된다 (Shon and Park, 2017). 지역 특성을 파악하기 위해 총 4가지 도시재생 유형에 대해 사업이 진행된 22개의 지역에 대한 뉴스 미디어 10만여건의 텍스트 데이터를 수집하였다. 수집된 텍스트를 이용하여 도시재생 유형에 따른 지역별 주요 키워드를 도출하고 토픽모델링을 수행하여 유형별 차이가 있는 지 탐색해 보았다. 다음 단계로 주어진 텍스트를 기반으로 도시재생 유형을 추천하는 추천시스템 구축을 위해 텍스트 데이터를 벡터로 변환하여 머신러닝 분류모델을 개발하였고, 이를 검증한 결과 97% 정확도를 보였다. 따라서 본 연구에서 제안하는 추천시스템은 도시재생 사업을 진행하는 과정에서 신규 사업지의 지역 특성에 기반한 도시재생 유형을 추천할 수 있을 것으로 기대된다.

주제어 : 도시재생, 텍스트 마이닝, 토픽모델링, 머신러닝 분류 알고리즘, 도시재생 유형 분류

.....

논문접수일 : 2020년 4월 10일 논문수정일 : 2020년 9월 7일 게재확정일 : 2020년 9월 14일
원고유형 : 일반논문 교신저자 : 강주영

* “본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음” (IITP-2020-2018-0-01424)

1. 서론

도시재생 활성화 및 지원에 관한 특별법(2017)에 따르면 도시재생이란 인구의 감소, 산업구조의 변화, 도시의 무분별한 확장, 주거환경의 노후화 등으로 쇠퇴하는 도시를 지역 역량 강화, 새로운 기능의 도입과 창출 및 지역자원의 활용을 통해, 경제적, 사회적, 물리적, 환경적으로 활성화하는 것을 의미한다. 5년간 50조원을 투자하여 500개의 구도심과 노후 주거지를 재생시키겠다는 ‘도시재생 뉴딜사업’을 1호 공약으로 내세운 문재인 정부 출범 이후, 도시재생에 대한 정부 관심의 정도가 높다. 2018년 8월 31일 정부에 의해 의결된 ‘2018년도 도시재생 뉴딜사업 선정안’을 통해서 전국 99곳의 지역이 도시재생 뉴딜사업지로 선정되었다. 이는 2017년 시범 사업 68곳에 비해 대폭 확대된 뉴딜사업 선정지이며, 전국적인 인구감소 지역 증가와 고령화 가속화 등에 따른 도시 소멸 위기에 대응하기 위한 조치라고 할 수 있다. 특히 앞서 언급한 사업들은 지역 경제 활성화, 기초 생활 인프라 확충, 산업단지 환경개선 등 시민들의 삶의 질을 향상하고 생활 여건을 개선할 수 있도록 지역 밀착형 생활 SOC(사회간접자본)를 구축하는 데 중점을 두고 있다. 이는 과거 한국전쟁 이후로부터 계속되는 급속한 도시화가 2008년 저성장기의 영향을 받으면서 전반적인 패러다임이 ‘개발’과 ‘재개발’에서 ‘재생’으로 바뀐 것에서 영향을 받은 것이라고 분석할 수 있다 (Seoul, 2015).

국토교통부는 ‘도시재생 뉴딜사업’에서 다음과 같이 2가지 방법으로 도시재생 사업을 분류하였다. 첫 번째 방법은 ‘선정방식별’로 도시재생 사업을 분류하였고 두 번째 방법은 ‘사업유형별’로 도시재생 사업을 분류하였다 (Molit, 2017).

첫 번째 방법은 도시재생 사업을 선정한 주체에 대한 분류인데, 중앙 선정, 공공기관 제안, 광역 지자체 선정으로 나눌 수 있다. 해당 분류는 단순히 어떤 기관이 도시재생 사업을 주관했는지에 따라 유형을 정의한 것이므로 도시재생의 목표, 도시 속성과는 큰 관련이 없다.

두 번째 방법인 사업유형별 도시재생은 경제 기반형, 중심 시가지형, 일반 근린형, 주거지 지원형, 우리동네 살리기형으로 유형으로 분류가 되며, 이는 모두 도시재생의 사업 유형별 방법론을 제시한 것이므로 도시재생의 목표, 도시 속성과 관련이 있는 분류 방법이다.

국내 도시재생을 관리하는 국토교통부의 도시재생 종합정보체계에 따르면 도시재생 성공 키워드는 “주민 참여”, “지역특화” “부처협업”, “민관협력”이다. 성공 키워드에 따르면 지자체에서 정부에게 도시재생 사업을 제안할 때 지역주민, 민간기업의 도움과 함께 도시의 특성을 정확히 이해하고 도시의 특성에 어울리는 방향으로 사업을 추진하는 것이 가장 중요하다는 것을 알 수 있다. 도시재생 유형이 주민들과 민간기업들이 공감할 만하게 도시의 특성에 따라 잘 분류가 되어있고 도시재생을 진행할 수 있으면, 지속가능한 도시재생으로 성공 확률이 높다는 것이다.

한편, 현재 뉴딜 정책과는 달리 2015년도 12월에 발표된 ‘2025 서울시 도시재생 전략계획’에서는 도시재생 사업을 진행할 때 도시의 면적보다 특성에 따라 유형을 분류했다. ‘저이용·저개발 중심지역’, ‘쇠퇴·낙후 산업(상업)지역’, ‘역사문화자원 특화지역’, ‘노후 주거지역’이라는 4가지 분류로 서울형 도시재생을 정의했다. 먼저, 저이용·저개발 중심지역은 대중교통 접근성이 양호한 철도(지하철)역세권 일대나 가용토지 확보가 가능한 공공기관, 공장 이전 적지이지만 사람들

의 이용이 적은 지역을 도시재생 하는 것이다. 다음으로 쇠퇴·낙후 산업(상업)지역은 과거 산업(상업)활동이 활발하였으나 산업구조의 변화로 활력이 저하된 지역을 도시재생 하는 것이다. 다음으로 역사문화자원 특화지역은 역사·문화·자연 자원을 보유한 지역으로서 자원의 복원·활용을 통한 지역 활성화 가능성이 높은 지역, 개발로 인해 지역자원 훼손의 우려가 있어 선제적 대책이 필요한 지역을 도시재생 하는 것이다. 마지막으로 노후 주거지역은 주거환경이 노후 불량하고 지역 특성상 전면철거형 정비사업이 바람직하지 않거나 불가한 지역 또는 예방적 차원에서 주거지의 보존 및 관리가 필요한 지역을 도시재생 하는 것이다 (Seoul, 2015). 이와 같은 도시 특성을 고려한 유형 분류는 유형별로 도시의 특성이 구분할 수 있다는 점에서 도시재생 지역을 분류하는데 적합한 기준이라고 판단하였다.

유형별 도시재생 사업이 지역 활성화에 미친 영향에 대해 연구(Choi and Lee, 2016)의 경우 도시재생의 유형은 상당히 다양하며, 도시재생 유형에 따라 지역활성화의 정도에 차이가 있기 때문에 도시재생의 유형별 특성을 고려하여 앞으로의 도시재생을 계획한다면, 양적 성장의 측면에만 치중한 현재의 도시재생 관련 사업의 문제를 해결할 수 있는 가능성을 보여주었다.

이에 본 연구에서는 ‘2025 서울시 도시재생 전략계획’에서 정의된 도시재생 유형이 지역의 특성이 반영된 도시재생을 실시하였는지 확인하기 위해, 빅카인즈(Bigkinds)의 뉴스데이터를 전처리한 후 토픽모델링 분석 기법을 사용하여, 유형별 유의미한 특성을 발견하고자 한다. 이러한 유형별 특성을 토대로 다양한 머신러닝 분류모델을 사용하여, 신규 지역에 대한 도시재생 유형을 추천하는 시스템을 제안하였다. 총 4가지 도

시재생 유형에 대해, 10만건의 텍스트 데이터를 Word Count Vector와 TF-IDF 벡터로 변환 후 분류 모델을 적용하고 비교 분석하여 최적의 분류 모델을 선정하여 도시 특성에 맞는 도시재생 유형 추천시스템을 제안하고자 한다.

본 연구의 구성은 다음과 같다. 우선 2장에서는 본 연구에 관한 선행 연구에 대해 소개한다. 3장은 본 연구의 연구 방법에 대해 서술한다. 4장은 도시재생 유형별 속성을 분석한 것에 대해 설명한다. 5장에서는 머신러닝으로 분석한 연구 모델을 평가할 것이며, 마지막으로 6장에서는 연구로부터 도출된 결론과 시사점 그리고 더 나아가 한계점에 대해 논하고자 한다.

2. 선행연구 분석

2.1. 도시재생 유형 분류

<Table 1>을 보면 2018년도 ‘도시재생 뉴딜사업’의 5가지 사업유형별 기준으로 경제 기반형 3곳, 중심 시가지형 17곳, 일반 근린형 34곳, 주거지 지원형 28곳, 우리 동네 살리기형 17곳의 대상지가 선정되었다. 하지만 이러한 국토교통부의 도시재생 유형은 한계점이 존재하는데, 뉴딜정책의 도시재생 유형은 유형별 권장면적이 있다. <Table 1>에서 권장 면적에 따른 분류를 살펴보면 경제 기반형은 50만 m^2 내외, 중심 시가지형은 20만 m^2 내외, 일반 근린형은 10~15만 m^2 내외, 주거지 지원형은 5~10만 m^2 내외, 우리동네 살리기형은 5만 m^2 내외이다 (Molit, 2017). 뉴딜정책의 도시재생 유형은 특성에 따라 나누었지만, 이 면적에 따라 도시재생이 진행되는 경우가 대부분이다. 2018 도시재생 뉴딜 사업 선정지 99곳

〈Table 1〉 New Deal Urban Regeneration Type and List

New Deal Urban Regeneration Type	Recommended Area Size	Regeneration Area List
Economic Based	Local Industry and Economy (approx. 500,000m ²)	Jungheung-dong, Jangseong-dong, Songdo-dong (3 areas)
Center Urban District	Center Urban District (approx. 200,000m ²)	Spaldong, Sanyeok-dong, Pojeong-dong, Seoknam-dong, Jungheung2-dong, Daechon-dong, Suseong-dong, Yochong-dong, Geumnam-dong, Gwangyeong-dong, Hwango-dong, Wonpyeong-dong, Namhae-eup, Daechung-dong, Saman-an-dong (17 areas)
General Neighborhood	Commercial-Residence, Commercial Alley (approx. 100,000 ~150,000m ²)	Muk 2-dong, Natural-dong, Baegwa 2-dong, Sinpyeong 1-dong, Dongsan-dong and Jukjeon-dong. Ojeong-dong, Seobu-dong, Eonyang-eup, Chochiwon-eup, Jeonmyeon, Ilsan-dong, Shincheon-dong, Paengseong-eup, Wolpidong, Gyeongang-dong, Hakseong-dong, Seongnae-dong, Munhwa-dong, Onyang 1-dong, Hwaji-dong, Hwaji-dong, Hwashun-eup, Gangjin, Gangjin, Winsang-eup, Seosang-eup, Seosang-eup, Seosang-eup, Seosang-eup, Seosang-eup (34 areas)
Residential Supportive	Residence (approx. 50,000 ~100,000 m ²)	Suyu 1-dong, Bulkwang 2-dong, Nankok-dong, Geumsadong, Yeongju-dong, Geoje-dong, Bisan-dong, Shinheung-dong, Hyosung 1-dong, Ganghwa-eup, Dongmyeong-dong, Nongseong 2-dong, Sajik-dong, Samsong-dong, Cheorwon, Yeongwol, Naedeok 1-dong, Eumseong-eup-eup, Gosu-dong, Tong, Gosu-dong, Tong, Gosu-dong, Gosu-dong, Tong, Tae-dong, Tae-dong, Gosu-dong, Tong-dong, Tong-dong (28 areas)
Town Rehabilitation	Small Residence (Under 50,000m ²)	Jegi-dong, Doksan 1-dong, Dongdaesin 2-dong, Bokhyeon 1-dong, Ichon-dong, Baekryeong-myeon, Daedong, Okdong, Byeongyeong 2-dong, Daeyodong, Sabuk-eup, Volcano-dong, Hongseong-eup, Bulgyo-eup, Sinheung-dong, Sancheong (17 areas)

을 보면, 대부분의 사업 면적 유형별 권장 면적에 해당한다. 이는 각 지자체에서 도시재생 사업 계획을 할 때 도시의 특성을 고려하기보다는 도시재생을 할 수 있는 면적에 따라 유형을 끼워 맞추는 형식으로 변질될 가능성이 존재한다는 것을 의미한다.

먼저 경제 기반형은 국가·도시 차원의 경제 쇠퇴가 심각한 지역을 대상으로 복합 앵커시설 구축 등 新(신) 경제거점을 형성하고 일자리를 창출하는 사업이다. 다음으로 중심 시가지형은 역사·문화·관광과의 연계를 통해 상권 경쟁력을 확보하고 청년·소상공인 창업 인큐베이팅, 사회

적 경제 주체 육성 등을 지원하는 사업이다. 세 번째 유형인 일반 근린형은 주거지와 골목상권이 혼재된 지역을 대상으로 골목상권과 무인택배 등 주민 체감형 시설 개보수를 지원하여 공동체를 활성화하고 영세 상권을 보호하는 사업이다. 네 번째 유형인 주거지 지원형은 원활한 주택 개량을 위해 골목길 정비 등 소규모 주택정비의 기반을 마련하고, 소규모 주택 정비 사업 및 생활 편의시설 공급 등으로 주거지 전반의 여건을 개선하는 사업이다. 마지막으로 우리동네 살리기 유형은 생활권 내에 도로 등 기초 기반 시설은 갖추고 있으나 인구 유출, 주거지 노후화로

활력을 상실한 지역에 대해 소규모 주택 정비 사업 및 생활 편의시설 공급 등으로 마을 공동체를 회복시키는 사업이다 (Molit, 2017).

도시재생을 도시의 특성에 맞게 도시재생을 진행하는 이유는 도시재생은 무분별한 재건축, 재개발이 아니라 마을 주민들, 도시를 방문하는 방문객들 모두 만족하는 도시를 재생시킬 수 있도록 지속 가능한 도시재생을 목표로 하기 때문이다. 지속가능한 도시재생은 도시의 기능, 성격을 고려하고 정체성, 역사·문화자원을 보존하고 활용하며 주민의 참여가 포함된 형태의 도시재생이다 (Jeon, 2011). 또한 도시재생은 사회 및 경제적 측면을 활성화하는 사회적 요구에 부응하여 종합적이고 프로그램 중심으로 진행되어야 한다고 전제한 후 모든 계획은 다양한 사회적 요구 및 해당 지역의 공간적 특수성을 고려하여 지역의 원주민이 지속적인 생활을 영위할 수 있도록 삶의 질도 개선되는 개발이 되어야 한다 (Sang, 2010). 따라서 해당 지역의 특색과 지역 원주민이 고려되지 않는 무분별한 랜드마크 건설이나 먹자골목, 로데오 거리를 만드는 도시재생은 비효율적인 도시재생사업이 될 수 있다. 따라서 해당 도시와 주민들을 깊게 이해하고 도시와 어울리는 도시재생을 제안해야 지속 가능한 도시재생으로 인정받을 수 있을 것이다.

일본 무코지마 지역은 도쿄 스카이트리가 위치한 곳이며, 과거 스카이트리의 공사가 진행됨에 따라 지가가 상승하면서 고층 아파트가 들어서고 빈 공장이 늘어나기 시작했다. 따라서 기존 원주민들이 이탈하는 ‘젠트리피케이션’ 현상이 발생하게 되고, 도시 전체가 침체되어가는 중이었다. 하지만 무코지마 지역의 주민들은 지자체의 도움을 구해서 빈 공장을 개조하여 예술행사를 개최하고, 무코지마 지역에서 오래전부터 있

었던 예술가들을 배려하면서 도시가 다시 재 활성화된 사례도 있다. 이처럼 도시재생 시킬 도시와 주민들을 깊게 이해한 후 그 도시와 어울리는 도시재생을 진행하는 것은 지속 가능한 도시재생을 하는 것의 전제 조건이라고 할 수 있다.

2.2. 도시재생 연구

국내에서 도시재생 유형에 관한 연구는 주로 국토교통부에서 발표한 도시재생 유형에 기반하여 수행되었다. 해당 연구들은 유형별 특성 분석, 유형별 성과 관리, 도시 특성에 따른 유형 계획 등으로 구분되어 실시되었다. 현재 국토교통부가 발표한 뉴딜정책의 도시재생의 유형은 도시의 특성에 기반한 유형 분류가 아닌 도시의 물리적인 조건에 따른 유형을 분류한 것이 특징이다. 이에 대해 도시재생 유형에 대한 연구 논문들은 물리적 조건의 유형 분류의 한계점을 극복하려는 논문들이 다수 존재한다.

도시재생 유형에 따른 대처나 특성을 도시재생 유형별 성과관리에 대해 연구를 수행한 Jin and Lee (2018)에 따르면 현재 도시재생 성과관리에 있어 유형별 평가가 이루어지지 않고 있으며, 도시재생의 올바른 성과관리를 위해서는 사업유형별로 성과지표의 중요도를 달리 두어 평가를 진행해야 한다고 밝히고 있다. 유사한 연구로, 대구 광역시를 중심으로 도시재생 유형별 성과관리 연구를 진행한 Shin and Shin (2017)은 뉴딜정책의 유형인 경제기반형과 근린재생형의 성과측정지표 간 중요도가 다르게 측정된다는 것을 AHP 분석을 통해 보여주었다. 또한 Lee et al. (2017)는 독일의 도시재생 사례를 심층 분석하여 자연 친화적이고 에너지 순환형 도시재생을 통한 도시의 지속 가능성을 높일 수 있는 생태적

도시재생의 유형과 특성을 도출하였다.

도시 정체성 확립을 위한 도시재생의 유형 및 계획 특성에 대한 분석한 연구에서는 해외의 사례를 기반으로 도시재생 전과 후의 정체성을 계승한 도시재생 계획 요소를 분석하고자 했다 (Lee et al., 2009). 도시재생 전의 정체성을 계승한 정도에 따라 계승 및 강화형, 변화 및 창조형으로 나누어 도시재생 활성화에 있어 중요한 요소에 대해 연구하였다.

도시재생 활성화 계획(안)을 중심으로 사업유형 및 재원조달 특성을 분석한 연구에서는 뉴딜사업의 유형인 경제기반형과 근린재생형의 재원 특성면에서는 다르지만, 사업내용으로는 큰 차이가 없다는 것을 보여주었다 (Jung et al., 2016). 또한 도시재생 선도지역의 사업 내용 및 기대효과 유형 분류 및 분석에 대한 연구를 진행한 Bae and Park (2018)은 근린재생형 도시재생 선도지역의 사례를 통해, 사업 내용의 특성에 따라 14가지 새로운 기준을 제시하고, 이러한 기준에 기반한 도시재생 사업의 특성을 분석하고자 하였다. 문화적 도시재생 유형의 사례 분석에 대한 연구의 경우 창원시 마산 원도심 재생 사업이 문화적 재생 유형에 해당된다는 것을 명시하고 이에 대한 전략과 특징을 규명하였다 (Seo, 2016).

IPA분석을 활용하여 원도심 쇠퇴 상권의 도시재생적 활성화 방안에 관한 연구를 진행한 Ryu and Kim (2018)은 상권의 12가지 선택 속성에 대한 중요도와 만족도 조사를 하였으며, 쇠퇴 상권의 활성화를 위한 필요한 요소를 분석하였다. 도시재생에 대한 국제 연구 동향과 국내 정책결정자 담화를 분석한 연구는 의미 연결망 분석과 토픽 모델링 분석 기술을 활용하여, 국내 정책적 맥락에서 정책 결정자의 담화문을 통해 정책적

지향점을 알아보았다 (Park et al., 2018). 이를 통해 주거환경 개선 유형에 있어, 박근혜 정부와 문재인 정부 간에 시각차가 있다는 것을 보여주었다.

2.3. 텍스트 벡터 모형 및 머신러닝 분류 모형

본 연구에서는 도시재생 유형을 추천하기 위해 뉴스 데이터를 텍스트 벡터로 변환한 후 머신러닝 분류 모형을 활용하고자 한다. 텍스트를 벡터로 표현하는 방법은 단순히 빈도수만을 고려한 방법부터 복잡한 순서까지 포함하는 임베딩 방법 등 다양한 방법들이 연구되어 오고 있다. 본 연구에서는 텍스트에서 사용되는 단어의 빈도수에 따라 문서의 의미를 분석하는 기본 방식인 Bag-of-Words 기법(Zhang et al., 2010)을 도입하고 단순 빈도 분석(Count Vector)과, 단어-역문서 빈도 분석((Term Frequency-Inverse Document Frequency, TF-IDF)을 통하여 문서를 벡터로 변환하고자 한다 (Aizawa, 2003).

분류를 위한 머신러닝 모형은 매우 다양하며 활용하는 데이터에 따라 각 알고리즘 별로 성능에 차이를 보이고 있다 (Caruana and Niculescu-Mizil, 2006; Williams et al., 2006). 본 연구에서는 기본적으로 텍스트를 벡터로 변환하여 분류모형에 적용하고자 한다. 따라서, 본 연구에서 활용하는 머신러닝 분류모형은 많은 연구에서 가장 빈도수나 활용도가 높은 SVM, Decision Tree, Random Forest, Logistic Regression, Gradient Boost와 같이 5가지를 활용하고자 한다 (Ikonomakis et al., 2005; Khan et al., 2010; Patra and Singh, 2013). 이들 분류 모형 구축을 위해 파이썬 싸이킷런 (Python Scikit-learn) 패키지를 활용하였다.

〈Table 2〉 Urban Regeneration Types of Seoul

Urban Regeneration Types of Seoul and Area List (2015)	
Regeneration Type	Regeneration Area List
Under Use Under Used and Developed	Seoul Station Station, Changdong Sanggye, Kwangwoon University Station, Sangam Search, Yeongdong MICE (4 areas)
Declining Industry	Sewoon Sangga, Jangnapeong, Southwest G-Valley (2 areas)
Historical and Cultural Resources	Sejongdaero, Mapo Petroleum Non-Foreign Base, Nodeul Island, Namsan Yejangjak, Dangin-ri Power Plant, Nakwon Sangga-Donhwa Munho, Donuimun Historical and Cultural Park (6 areas)
Decrepit Residence	Seonggyo Village, Baeksa Village, Haebangchon, Seongsu-dong, Sinchon-dong, Amsa-dong, Jangwi, Sangdo 4-dong, Bukhansan Mountain, Seochon, Changshin Approved Lead Area, Garibong-dong (10 areas)

2.4. 선행논문과의 차별성

본 연구는 <Table 2>와 같이 서울시에서 2015년에 발표한 도시재생 유형에 기반한 연구를 진행하고자 한다. 도시재생 유형에 관한 선행논문을 살펴보았을 때, 도시재생 특성에 기반한 도시재생 유형 분석을 하는 연구는 있었으나, 새롭게 분류하는 연구는 없었다. 국토교통부에서 발표한 도시재생 유형은 지역의 물리적 조건에 기반한 유형 분류로서 도시의 차별적인 특성을 반영하지는 못하고 있다. 반면에 서울시의 도시재생 유형은 역사문화자원지역, 저이용·저개발지역, 쇠퇴·낙후지역, 노후주거지역으로 지역의 특성을 반영한 유형 분류를 실시하였다. 따라서 본 논문은 서울시 유형 분류의 효과성을 검증하고, 서울시 외의 지역에 도시 특성을 반영하는 분류기준을 적용할 수 있는 모델을 구축하고자 한다.

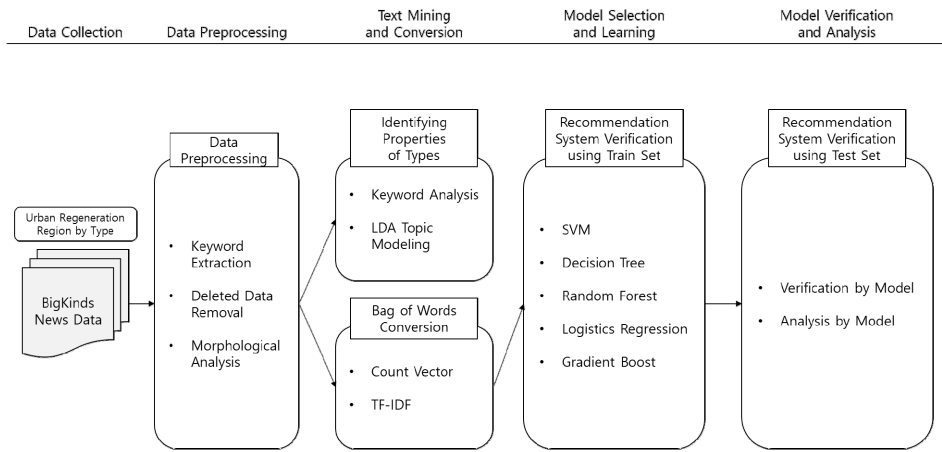
3. 연구방법

3.1. 연구 절차

본 연구는 <Figure 1>과 같이 빅카인즈로부터 수집된 도시재생 유형별 뉴스데이터를 바탕으로 토픽모델링을 통한 유형별 차이를 탐색하고, 머신러닝 분류 알고리즘을 활용하여 주어진 특성에 가장 맞는 도시재생 유형 추천 시스템을 제안하고자 한다.

3.2. 데이터 수집

본 연구에서는 뉴스데이터를 수집하기 위해 빅카인즈 웹 서비스를 활용했다. 빅카인즈 서비스는 한국언론진흥재단에서 제공하는 서비스로서 총 41개의 매체에 대해 1990년도 뉴스부터 현재까지 뉴스를 저장 및 제공하고 있다. 또한 이 서비스는 정형, 반정형, 비정형 뉴스를 모두 키



〈Figure 1〉 Overall Research Process

☐	[매일시론] 잔나비와 생활SOC, 그리고 스마트시티 인디밴드 '잔나비'의 음원차트 통권 현상 음악상가공연 등 경제적 파급효과 야기 '생활SOC'로 지역균형발전 기대한다면 지역 특성 등 고려한 실질... 출산매일 미분류 2019/05/06 이영규 출산정보산업협회 회장․아이티공간 CEO
☐	'악기 나눔'으로 꿈과 희망을...유명인사 기부 줄이어 [앵커] 아이들이 악기를 배워 꿈과 희망을 펼칠 수 있도록 하기 위한 유명 인사들의 악기 기증이 이어지고 있습니다. 악기 나눔 사업을 통해 학교 밖 ... YTN 사회>사회일반 문화>음악 문화>예술_문화재 2019/05/05
☐	'악기 나눔'으로 꿈과 희망을...유명인사 기부 줄이어 중복 [앵커] 아이들이 악기를 배워 꿈과 희망을 펼칠 수 있도록 하기 위한 유명 인사들의 악기 기증이 이어지고 있습니다. 악기 나눔 사업을 통해 학교 밖 ... YTN 사회>사회일반 문화>음악 문화>예술_문화재 2019/05/05
☐	300개 매장마다 색다른 음악 이야기... '악기 낙원' 속으로 서울 종로구 인사동과 낙원동을 가로지르는 삼일대로 위에 낙원악기상가가 자리잡고 있다. 1969년에 완공돼 올해로 50년을 맞은 건물. 한국을 대표... 동아일보 문화>요리_여행 문화>출판 문화>예술_문화재 2019/05/04 유재영
☐	결과 속 다른 세계 최대 규모의 '악기 아지트', 낙원상가의 매력은? 종로에서 놀 거리를 찾는 사람들은 종로2가 사거리에서 고민을 시작한다. 대체로 청계천 방면으로 걸어서 종로 '젊음의 거리'로 가든지, 아니면 안국... 동아일보 문화>요리_여행 문화>출판 문화>예술_문화재 2019/05/03 유재영
☐	"보컬만 NO, 악기도 보세요"...JTBC '슈퍼밴드' 악기 매력 재조명

〈Figure 2 – Results of a search for “Nakwon Sangga” in BigKinds〉

워드 단위로 세분하여 csv 형태로 다운받을 수 있도록 빅데이터 형태로 활용하기 쉽게 제공하고 있다. 따라서, 시기적으로나 양적으로 많은 뉴스에 대해 간편하게 분석할 수 있다. (Lee et al., 2017).

Yoon et al. (2019)은 도시 브랜드 특성을 파악하기 위해 빅카인즈 뉴스데이터를 활용하였는데, 본 연구에서도 빅카인즈 뉴스 데이터를 활용하여 도시의 특성을 분석하고자 한다. <Figure 2>는 빅카인즈에 서울시 도시재생 유형 중 하나인

〈Table 3〉 District and Data Count of Urban Regeneration Types (2016.1.1~2019.5.18)

Urban Regeneration Type	District	Number of Data
Decrepit Residence	Garibong-dong, Baxa-dong, Sangdo 4-dong, Seonggyo-dong, Haebangchon, Seongsu-dong, Shinchon-dong, Amsa-dong, Jangwi and Seochon	17,919
Declining Industry	Zhang Anpyeong, Seun Sangga	2,434
Historical and Cultural Resources	Nakwon Sangga, Namsan Yejangjak, Nodeul Island, Dangin-ri Power Plant, Mapo Petroleum Non-Factory Base, Sejong-daero	8,673
Under Used and Developed	Sangam Search, Seoul Station Yeokse-kwon, Young-dong MICE, Chang-dong OFFICE	78,728

“낙원상가”를 검색했을 때의 결과 데이터 중에 하나이다. 낙원상가의 경우 악기 상점이 많은 지역인데, 이러한 정보를 포함하고 있는 것을 알 수 있다.

분석에 활용한 뉴스데이터의 기간은 서울시의 ‘2025 서울시 도시재생 전략계획’ 발표 날짜인 2015년 12월 이후인 2016년 1월부터 2019년 5월까지로 총 3년 5개월로 분석 기간을 설정하였다. 발표 시점 직후부터 기간을 설정한 이유는 도시재생이 진행할 것임을 고시한 이후부터 미디어가 해당 지역에 관심을 갖고 뉴스 데이터를 생산할 것이며, 도시재생과 관련되고 해당 지역과 관련된 뉴스 데이터들이 많을 것으로 판단했다. <Table 3>은 유형별 수집 지역과 수집 건수를 보여주고 있는데 총 22개 지역에 대해 총 107,754건의 데이터를 수집하였다.

앞서 언급했듯이 서울시 도시재생 유형으로는 4가지 종류의 유형이 존재하며, 분석 대상 지역의 경우 ‘2025 서울시 도시재생 전략계획’ 보고서를 참조하여 지역 전체에 대해 전수조사를 실시했다. 노후 주거지역 유형으로는 “가리봉동, 백사마을, 상도4동, 서촌, 성곽마을, 성수동, 신촌동, 암사동, 장위, 해방촌”을 선정했으며 쇠퇴·낙후 산업(상업) 지역으로는 “세운상가, 장안평”이

있다. 또한 역사문화자원 특화지역으로는 “낙원상가, 남산예장자락, 노들섬, 당인리발전소, 마포 석유비축기지, 세종대로”이 있으며, 저이용·저개발 중심지역으로는 “상암수색, 서울역 역세권, 영동MICE, 창동 상계”을 분석하였다. 각 지역을 뜻하는 다른 용어들도 있을 것이나, 일단 주어진 지역명만을 토대로 데이터를 전수 수집하였다. 수집된 약 10만건의 데이터는 도시 특성을 설명하는 단어들이 얼마나 도시재생 유형을 잘 분류하는지를 분석하고 있기에 충분할 것으로 판단되어 지역 기반 유사어 사전을 구축하지 않고 연구를 진행하였다.

3.3. 텍스트 데이터 준비

빅카인즈에서 추출한 데이터 셋은 불필요한 열이 존재하기 때문에 토픽 모델링과 머신러닝 기법에 활용되기 전에, 크롤링한 데이터를 전처리할 필요가 있다. 각 지역별 크롤링한 뉴스기사 데이터의 키워드 부분만 남기고 제거하였으며, 도시재생 지역명과 유형을 새로운 열로 추가하여 데이터를 채워 넣었다. 이후 각 유형별로 도시재생 지역의 데이터 셋을 연결하여 하나의 데이터프레임을 만들었다. 또한 빅카인즈 데이터 셋의 키워드 열에서 발견된 NA값을 포함한 열은

분석에서 활용할 수 없기 때문에 제거하였다.

(You et al., 2015).

3.3.1. 토픽모델링 분석 수행을 위한 전처리

키워드 데이터를 단어 수준으로 토큰나이징하여 문자열을 나누었다. 이후 데이터들을 tm 패키지의 Corpus를 사용하여 데이터를 Term Document Matrix 형식으로 바꿔주었으며, 추가적으로 topicmodels 패키지의 LDA 분석 툴의 입력 값으로 활용하기 위해 Document Term Matrix 데이터의 형태로 변환하여 토픽 모델링을 진행하였다.

3.3.2. 유형 추천 머신러닝 모델 학습을 위한 입력변수 설정

전체 뉴스키워드 데이터 중 80%는 머신러닝 모델을 만드는데 필요한 학습 데이터로 사용하였고, 20%는 머신러닝 모델을 검증하는데 필요한 검증 데이터로 추출하였다. 머신러닝 모델에 입력될 벡터 데이터로는 Count Vector, TF-IDF Vector로 크게 2가지 방식으로 나누었다. Count Vector는 여러 개의 말뭉치들이 포함된 문서들 속에서 고유의 토큰을 찾아 개수를 세는 벡터를 만드는 방법이다. 이러한 방법은 각각의 문서와 단어의 수를 모두 센다는 특징이 있다. TF-IDF Vector는 특정 단어가 문서 내에서 등장하는 빈도를 나타내는 TF(단어 빈도)와 특정 단어가 문서 전체 집합에서 등장하는 빈도를 나타내는 IDF(역문서 빈도)를 곱한 값으로 정의할 수 있다. 특정 단어가 모든 문서에서 빈번하게 나타나면 IDF를 높이기 때문에 단어의 중요도를 낮춘다. 즉, TF-IDF는 특정 단어가 문서에 자주 등장하지만 다른 주제의 문서에서는 출현하는 빈도가 비교적 낮을 때 단어의 중요도를 높게 평가하는 식으로 단어의 중요도를 평가하는 방법이다

3.4. 토픽모델링 분석

본 연구에서는 도시재생 유형별로 주요 키워드 분포나 구성하고 있는 토픽들에 차이가 있는지를 탐색적으로 살펴보기 위해 토픽모델링 분석을 실시하였다. 토픽모델링은 문서에 포함되어 있는 텍스트 데이터의 키워드를 추출하여 이를 바탕으로 특정 문서의 주제와 속성을 파악하는 기법이다. 가장 대표적인 방법은 LDA(Latent Dirichlet Allocation) 방법이며 이는 문서를 잠재적으로 주제에 대한 혼합된 텍스트 데이터로 이루어져 있는 것으로 여기며 문서에 포함되어 있는 텍스트 데이터의 분포를 통해서 문서의 주제를 파악하는 방법이다 (Blei et al., 2003). 이는 기존의 키워드 중심 텍스트 분류 방법에서 벗어나 더욱 효과적인 문서 분류를 가능하게 하는 방법이라 할 수 있다 (Kim et al., 2016a).

본 연구에서는 수많은 텍스트 데이터로 이루어져 있는 뉴스 문서에서 키워드를 추출한 후 도시재생 유형별로 LDA 토픽모델링을 실시하여 텍스트 데이터 안의 잠재적 주제를 추출하였다. 분석을 위해서는 토픽의 개수가 지정되어야 하는데 이는 많은 양의 데이터를 포함하는 문서를 분석할 때 토픽의 개수를 잘못 지정하면 과적합(overfitting) 등의 문제로 인해서 데이터의 객관성이 부족해질 수 있기 때문이다 (Kim et al., 2016b). 따라서 본 연구에서는 토픽개수를 5개, 10개, 15개로 LDA 토픽 모델링을 실시했으며 가장 도시의 특성을 명확하게 나뉘는 경우가 토픽 개수 10개였다. 이중 상위 5개 토픽의 특징이 잘 나타나 있었기 때문에, 최종적으로는 유형별 5개의 토픽을 사용하였고 이에 대한 라벨링을 실시하였다.

3.5. 머신러닝을 통한 유형 추천 모델 학습

머신러닝은 컴퓨터가 대량의 훈련 데이터를 학습할 수 있도록 알고리즘을 만들고 테스트 데이터로 모델을 검증하여 이를 통해 결과를 예측하는 것을 목표로 한다. 본 연구에서는 4가지 도시재생 유형을 분류 결과값으로 뉴스미디어 텍스트 데이터를 입력값으로 세팅하여 주어진 입력값에 대해 가장 분류를 잘 해주는 머신러닝 모델을 찾아 이를 기반으로 지역의 특성에 기반한 추천시스템을 제안하고자 한다.

본 연구는 텍스트 데이터를 머신러닝에 적용하기 위해 Bag-of-Words 방식으로 벡터로 변환하였다. 이를 위해 파이썬 언어의 싸이킷런 패키지를 활용하여 Count Vector와 TF-IDF Vector 데이터 형태로 변환하였다. 이렇게 변환된 벡터는 싸이킷런 패키지에서 제공하는 대표적인 분류 알고리즘인 SVM(Support Vector Machine), Decision Tree, Random Forest, Logistic Regression, Gradient Boosting를 적용하여 머신러닝 모델을 구축하고 성능을 평가하였다. 본 연구에서는 머신러닝 모델 구현을 위해 2가지 벡터 형태에 대해 5가지 알고리즘을 적용하였으므로 총 10가지 머신러닝 모델을 구축하여 성능을 평가 및 비교하였다.

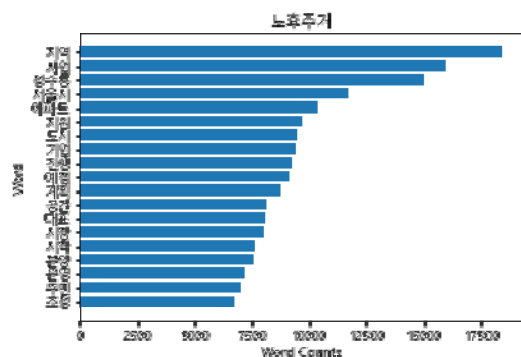
3.6. 탐색적 분석 결과

본 연구에서 제안하고 있는 도시재생 유형별로 실제 주요 키워드 분포나 구성하고 있는 토픽에 차이가 있는지 탐색적으로 살펴보기 위해 키워드 분석과 토픽모델링 분석을 수행하였고 도시재생 유형별로 아래 섹션에서 결과를 보여주고 있다. 분석 결과를 살펴보면 도시재생 유형별로 서로 주요 키워드의 차이를 보여주고 있으며, 토픽 구성도 확연한 차이를 보여주고 있다. 즉

지역별 특성이 반영되어 있는 뉴스 미디어 텍스트를 도시재생 유형별 차이를 보여주고 있음을 탐색적 분석결과 확인할 수 있었다. 따라서 이러한 결과를 바탕으로 다음단계에서 뉴스미디어 텍스트 데이터를 입력값으로, 도시재생 유형을 결과값으로 하여 머신러닝 분류 알고리즘을 적용하여 정확도가 가장 높은 모델을 기반으로 지역 특성을 나타내고 있는 텍스트를 활용한 도시재생유형 추천시스템을 제안하고자 한다.

3.6.1. 노후주거 지역

노후 주거 지역의 경우 키워드 분석결과인 <Figure 3>을 살펴보면, 노후 지역명, 사업, 진행 등 부동산 개발과 관련된 키워드들이 상위에 분포되어 있다. 또한, <Table 4>와 같이 토픽모델링을 통해서 도시재생(1회), 부동산(1회), 경제(2회), 문화(1회)에 대한 토픽을 확인할 수 있다. 서울에서 분류한대로 노후 주거지역의 부동산(재건축)에 대한 토픽도 있지만, 경제에 관련된 토픽도 있다는 것을 확인할 수 있다. 이는 거주민들이 요구하는 시설인 할인매장과 금융(은행)에 대한 지표가 드러나는 것이라 할 수 있다.



<Figure 3> Result of Keyword Distribution- Decrepit Residence

〈Table 6〉 Results of Topic Modeling - Historical and Cultural Resources

Theme	Topic
Politics; Government and Finance	Seoul, government, Sejong Boulevard, minister, support, chairman, Seoul City Hall, business, photography, day, government office, policy, finance, chairman, meeting, economy, Shinhan Bank, attendance, society, mayor
Politics; Assembly	Cheong Wa Dae, Park Geun-hye, Sejong-daero, the same day, assembly, police, the police, Seoul, Seoul, Gwanghwamun, citizens, candles, hosting candlelight vigils, candles, candles, impeachment, players, demonstrations, participants, demonstration
Culture; Performances and Events	Seoul, Sejong-daero, events, festivals, controls, sections, Seoul City, streets, performances, May, citizens, operations, Gwanghwamun, vehicles, progress, downtown, April, reenactment, participation
Culture; People and Story	people, thoughts, Korea, beginnings, movies, people, stories, music, Seoul, writers, works, love, oneself, heart, name, world, picture, world, stage
Urban Regeneration; Culture and Space	Seoul City, business, planning, creation, region, Nodeul Island, city, Seoul, space, regeneration, one-on-one, drive, history, connection, urban regeneration, culture, paradise shopping, next year, citizens, management

성격을 띠고 있기 때문에 문화와 관련된 주제도 드러났음을 알 수 있다.

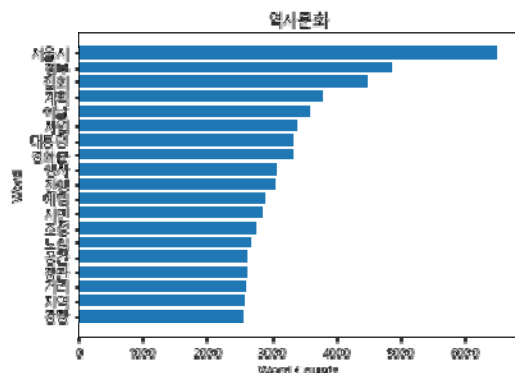
3.6.3. 역사문화자원 지역

역사문화자원 특화지역의 경우 키워드 분석 결과 역사와 관련된 지역명, 시민, 예정, 거리, 행사, 광장 등 역사나 문화 행사와 관련된 단어들로 구성되어 있음을 알 수 있다 토픽모델링 분석

결과 정치(2회), 문화(2회), 도시재생(1회)에 대한 토픽을 확인할 수 있다. 세종대로의 영향으로 정치와 관련된 주제도 확인되지만 대다수의 주제가 서울시에서 분류한대로 문화에 대한 토픽이 존재하며 도시재생을 주제로 두고 있는 키워드들 또한 문화적 성격이 강하다는 것을 확인할 수 있다.

3.6.4. 저이용·저개발 지역

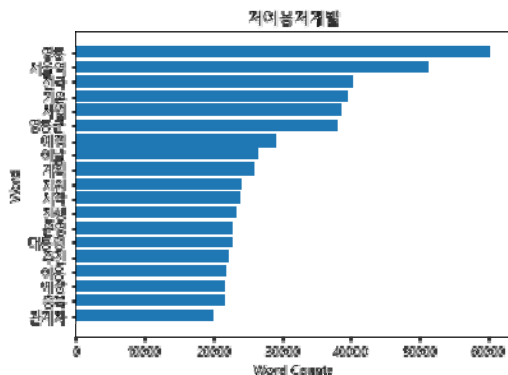
저이용·저개발 중심지역의 경우 키워드 분석 결과 해당 지역명이 특히 많이 나타나고 있으며 지원, 추진, 예정 등 부동산 개발과 관련된 단어들이 많이 보이고 있다. 또한, 토픽모델링 분석 결과 부동산(2회), 도시재생(1회), 경제(2회)에 대한 토픽을 통해 부동산에 관한 관심도가 높다고 볼 수 있다. 역세권과 관련된 주제가 확인되며 대중교통에 대한 접근성이 높은 지역이기 때문에 이를 매일 같이 이용하는 직장인들과 연관된 일자리 관련 주제도 드러난을 확인할 수 있다.



〈Figure 5〉 Result of Keyword Distribution-Historical and Cultural Resources

〈Table 7〉 Results of Topic Modeling – Under Used and Developed

Theme	Topic
Real Estate; Station	location, office-tel, lease, shopping mall, building, KTX, possible, Cheongna, scale, Seoul, Gwangmyeong Station, nearby, ground, first floor, utilization, basement, rent, office, reverse tax, hotel
Urban regeneration; Station	business, Seoul, the region, Seoul, Planning, Promotion, City, Seoul Station, Subway Station, Subway, Market, Greater, Urban Regeneration, Regeneration, Site, Housing, Government, Selection, Progress, Korail
Economy; Industry and Jobs	Changdong, Seoul, Seoul City, Sanggye, Planning, Creation, Culture, Jobs, Regional, Startups, Cities, Support, Economy, Northeast Asia, Industry, Growth, Space, Citizens, Centers, Facilities
Economy; Policies and Jobs	Seoul, Seoul City, Jobs, Citizens, Growth, Markets, Think, District Chief, Economy, Policy, People, Future, Welfare, Lawmakers, Park Won-soon, North Korea, Innovation, South Korea, Industry
Real Estate; New Town	schedule, complex, apartment, sale, utilization, availability, Cheongna, formation, furniture, Seoul, scale, opening, composition, dedicated area, location, Gimpo, design, Cheongna International City, Central, Stationary Area



〈Figure 6〉 Result of Keyword Distribution- Under Used and Developed

이와 같이 LDA 토픽 모델링을 통해서 서울시에서 제안한 도시재생 유형의 속성을 파악해봤을 때 각 유형별로 명확하게 주제가 나뉘는 것을 확인할 수 있다. LDA 토픽 모델링이 효과적인 키워드 추출과 분류를 가능하게 한다는 점과 주제가 명확하게 분류되었다는 점을 미루어 봤을 때 서울시의 도시재생 유형 분류가 적절하게 이뤄졌다고 평가해도 무관하다 (Yang et al., 2018).

3.7. 머신러닝 분석 결과

머신러닝 분석을 위해 파이썬 기반의 머신러닝 패키지 중 가장 잘 알려진 Scikit-learn 패키지를 사용하였다. 주어진 텍스트에 대해 Count Vector와 TF-IDF Vector를 추출하기 위해 각각 CountVectorizer와 TfidfVectorizer를 사용하였으며, 매개변수 최적화를 통해 양쪽 모두 추출 단어에 대한 최대 문서 빈도는 0.5로, 최소 문서 빈도는 0.01로 설정하였다. 랜덤 포레스트와

도시재생 유형을 예측하기 위한 머신러닝 모형은 총 5가지를 비교하였으며, 로지스틱 회귀분석의 경우, 수렴을 위해 최대반복수를 기본 100에서 1,000으로 늘리고 최적화알고리즘으로 'lbfgs'를 사용한 것 외에 나머지는 기본값을 사용하였다. SVM은 커널을 'linear'로 설정하였으며 그 외 알고리즘은 모두 Scikit-learn이 제공하는 기본값을 사용하였다.

<Table 8>에서와 같이 10가지의 머신러닝 모형을 비교해본 결과, TF-IDF 입력값에 대해 Gradient Boost가 97%로 가장 높은 성능을 보였

〈Table 8〉 Accuracy Comparison for Test data set

	Classifier				
	SVM	Decision Tree	Random Forest	Logistics Regression	Gradient Boost
Count Word Vector	0.924	0.911	0.930	0.934	0.945
TF-IDF	0.966	0.964	0.959	0.962	0.970

다. TF-IDF 입력값을 사용한 경우, 모든 모형이 95%가 넘는 정확도를 보였다. 각각의 모형이 고유의 매개변수가 있기 때문에, 그리드 서치 등의 방법을 이용하여 적절한 매개변수를 찾으면 현재의 성능보다 더 나은 성능을 보일 수도 있을 것으로 기대되나 이미 충분한 정확도를 보이고 있어 본 논문에서는 더 이상의 실험은 진행하지 않았다. 향후 연구에서는 이 부분이 더 개선될 수 있을 것으로 예상된다. Count Vector 보다 TF-IDF의 정확도가 높은 이유는 도시재생 유형 뉴스미디어 문서에 대부분 도시재생, 서울시 같은 단어들이 반복적으로 들어가므로 이러한 단어들을 제외할 수 있는 TF-IDF가 좀 더 정확한 분류 성능을 보였을 것으로 판단된다.

4. 결론 및 한계점

4.1. 결론 및 토의

본 연구는 서울시 도시재생 유형에 근거하여 유형별 특성을 분석하고, 서울특별시 외의 지역들의 뉴스데이터를 입력했을 때 지역의 특성에 근거한 유형을 추천하는 모델을 구현하였다. 현 정부가 들어서고, 도시재생에 대한 관심과 지원이 증가하고 있으며, 매년 새로운 도시재생 지역이 선정되고 이에 따른 사업이 진행되고 있다. 국토교통부 자료에 따르면, 도시재생의 유형을

다양한 기준에 따라 나누어 놓았지만, 도시의 특성에 기반하기 보다 주로 물리적인 조건에 기반한 유형 분류인 경향이 크다. 반면 서울시 도시재생 전략계획의 경우, 지역의 특성에 기반한 4가지 유형으로 분류해 놓았다. 이러한 유형 분류가 뉴딜사업의 유형분류보다 지역 특성에 기반한 도시재생 사업 테마를 잡는데 유의미하다고 판단하였고, 토픽 모델링을 통해 이러한 도시 특성 분류가 유의미했는지 살펴보고자 했다.

데이터로는 빅카인즈의 뉴스데이터를 크롤링하여 분석에 활용하였으며, 서울시 도시재생 지역의 전체에 대한 데이터를 연구에 사용하였다. 이후 데이터를 전처리하여 토픽 모델링을 실시하였고 상위 토픽 5개에 대한 분석을 실시하였다. 결과로는 노후주거 지역의 경우 부동산과 경제와 관련된 토픽들이 다수 등장하였으며, 쇠퇴·낙후 지역의 경우 과거 산업활동이 활발했던 지역의 성격을 반영하고 있는 토픽이 등장함을 확인할 수 있었다. 역사문화자원 지역의 경우는 과거의 역사의 흔적이 담긴 지역인만큼, 오래된 지역이므로 정부와 관련된 키워드가 다수 등장하였다. 따라서 정치 관련 토픽과 각종 행사로 인한 문화 토픽을 확인할 수 있었다. 마지막으로 저이용·저개발 지역의 경우 부동산 및 접근성에 관한 토픽이 다수 등장하고 있어 접근성은 좋으며, 주로 개발이 예정된 지역이거나 개발이 될 가능성이 높은 지역의 특성을 가지고 있었다.

더 나아가 서울시 외의 지역에 대해 지역 특성

에 맞춘 도시재생 유형을 제안하는 모델을 구현하였다. 이 모델을 구현하기 위해 머신러닝 기술을 사용하였으며, 학습데이터와 검증 데이터를 8:2 비율로 랜덤 추출하여 활용하였다. 다양한 모델간 성능 비교를 위해서, 입력 변수를 Count Vector, TF-IDF Vector 두개 방식으로 데이터를 세팅하였으며, Classifier로는 SVM(Support Vector Machine), Decision Tree, Random Forest, Logistic Regression, Gradient Boosting로 5가지를 적용하여 총 10개 모델에 대한 성능 비교를 실시하였다. 가장 높은 성능이 나온 모델은 TF-IDF Vector 입력 데이터를 활용한 Gradient Boosting 방식이었으며, 정확도는 97%가 나왔다.

4.2. 연구의 의의 및 한계점

본 연구의 의의는 다음과 같다. 첫째, 도시재생 유형 제안 모델을 구현함으로써, 서울시를 제외한 지역들이 도시재생 사업의 테마를 결정할 때 의사결정에 도움을 줄 수 있는 도구를 제공했다. 기존 연구의 경우 도시 특성에 맞춘 유형 제안 모델을 한 연구는 없었다. 또한 현재 서울시를 제외한 지역의 경우 국가가 아닌 지자체에서 직접 도시재생의 테마를 결정하고 있기 때문에, 데이터 기반의 의사결정을 하기 어렵다. 하지만 해당 모델을 통해, 도시의 특성에 기반한 도시재생 테마를 결정하는데 도움을 줄 수 있게 되었다.

둘째, 도시재생의 유형을 최신 기술에 기반하여 분석하는 데에 의의가 있다. 기존 연구의 경우, AHP분석이나 사례분석, 자원 지표 분석을 통해 도시재생 유형을 분석한 것이 다수였다. 본 연구는 이러한 분석 기법에서 벗어나, 머신러닝, 텍스트 마이닝 기술을 통해 도시재생에 대해 연

구를 진행하였다. 탐험적인 해당 연구를 통해 추후 다른 연구과제에서 이를 참고할 수 있도록 하였다.

본 연구의 한계점은 다음과 같다. 첫째는 뉴스 데이터만을 활용했다는 점으로, 뉴스 데이터는 지역 주민이나 방문자의 의견을 온전하게 포함하지 못한다. 따라서 본 연구에서 제시한 4가지 분류 체계에 근거한 도시 재생 유형을 좀 더 신뢰성 높게 추천하기 위해서는 구글 리뷰 또는 네이버 플레이스 댓글을 포함한 분석이 요구된다. 추후 연구에서는 이들을 반영하는 경우 다른 분석 결과나 시사점이 나올 수 있을 것으로 기대된다. 그 다음은 뉴딜 도시재생 사업지에 대한 검증이 부족하다는 점이다. 본 연구는 2015년에 서울시가 발표한 도시재생 관련 보고서를 기반으로 진행됐기 때문에 최근 선정된 뉴딜 도시재생 사업지는 포함시키지 못했다. 따라서 추후 연구에는 2017년부터 매년 선정되고 있는 뉴딜 도시재생 사업지에 대한 분석을 포함하여 연구하는 것이 필요하다. 마지막으로 데이터 수집 기간이 도시재생 이전과 이후에 대한 인과관계를 담지 못한다는 점이다. 본 연구의 데이터 수집 기간은 근간이 되는 보고서인 '2025 서울시 도시재생 전략계획'의 발표로부터 한 달 후부터 현재까지로 실질적인 도시재생이 이뤄진 시점을 담지 못한다. 이는 도시재생 전후로 나타나는 특정 지역의 변화를 고려하지 못한 것이며 추후 연구에서는 이러한 변화까지 고려하여 보다 섬세한 모델을 구축하는 것이 요구된다.

참고문헌(References)

Aizawa, A., "An Information-Theoretic Perspective

- of Tf-Idf Measures", *Information Processing & Management*, Vol. 39, No. 1, 2003, 45-65.
- Bae, M.-K. and Park, S.-H., "Classification and Analysis of the Project Contents and Expected Effects of Urban Regeneration Pilot Projects", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, Vol. 18, No. 10, 2018, 527-544.
- Blei, D. M., Ng, A. Y. and Jordan, M. I., "Latent Dirichlet Allocation", *Journal of machine Learning research*, Vol. 3, No. Jan, 2003, 993-1022.
- Caruana, R. and Niculescu-Mizil, A., "An Empirical Comparison of Supervised Learning Algorithms", *Proceedings of Proceedings of the 23rd international conference on Machine learning*, 2006.
- Choi, J.-W. and Lee, J.-H., "An Impact Analysis of Idle Space Regeneration Types on Regional Revitalization", *Journal of the Korea Academia-Industrial cooperation Society*, Vol. 17, No. 5, 2016, 478-489.
- Ikonomakis, M., Kotsiantis, S. and Tampakas, V., "Text Classification Using Machine Learning Techniques", *WSEAS transactions on computers*, Vol. 4, No. 8, 2005, 966-974.
- Jeon, K. S., "Urban Regeneration and Sustainable Growth Strategies in Gwangju Metropolitan City", *Journal of the Korean Urban Geographical Society*, Vol. 14, No. 3, 2011, 1-17.
- Jin, E.-A. and Lee, W.-J., "Criticality Analysis of Performance Indicators by Urban Regeneration Project Type - with a Focus on Central City Type, General Neighborhood Type, Urban Residential District Regeneration and Town Regeneration Projects", *KIEAE Journal*, Vol. 18, No. 6, 2018, 29-41.
- Jung, K., Lee, J. and Lee, S., "An Empirical Study on Project Type and Funds Supply Characteristics of Urban Regeneration Priority Regions", *Journal of The Korean Regional Development Association*, Vol. 28, No. 2, 2016, 19-37.
- Khan, A., Baharudin, B., Lee, L. H. and Khan, K., "A Review of Machine Learning Algorithms for Text-Documents Classification", *Journal of advances in information technology*, Vol. 1, No. 1, 2010, 4-20.
- Kim, D., Kang, J. and Lim, J. I., "Comparative Analysis of Job Satisfaction Factors, Using Lda Topic Modeling by Industries : The Case Study of Job Planet Reviews", *Journal of Korea Society of IT Services*, Vol. 15, No., 2016a.
- Kim, T., Choi, H. and Lee, H., "A Study on the Research Trends in Fintech Using Topic Modeling", *Journal of the Korea Academia-Industrial Cooperation Society*, Vol. 17, No. 11, 2016b, 670-681.
- Lee, J.-H., Seo, E.-K. and Lim, J.-H., "Analyses on Type and Characteristics of Urban Regeneration for Establishment of Urban Identity", *KIEAE Journal*, Vol. 9, No. 5, 2009, 85-96.
- Lee, S.-H., Oh, D.-S. and Kim, J.-K., "Types and Characteristics of Ecological Urban Regeneration - Focused on Urban Regeneration Cases in Germany", *KIEAE Journal*, Vol. 17, No. 6, 2017, 213-220.
- Molit. (2017) "Special Act on Activation and Support of Urban Regeneration".
- Park, S., Han, E.-J., Lee, J. and Kim, D.-G., "A Recommendation System for Assisting Devices in Long-Term Care Insurance", *The Korean Journal of Applied Statistics*, Vol. 31, No. 6,

- 2018, 693-706.
- Patra, A. and Singh, D., "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms", *International Journal of Computer Applications*, Vol. 75, No. 7, 2013.
- Ryu, T.-C. and Kim, W.-H., "A Study on the Method to Activate of Urban Regeneration in Business Districts in Old Towns through Ipa Analysis - Focusing on Andong-Si Food Street", *RESIDENTIAL ENVIRONMENT : JOURNAL OF THE RESIDENTIAL ENVIRONMENT INSTITUTE OF KOREA*, Vol. 16, No. 2, 2018, 159-169.
- Sang, N. (2010) *Toward a Collaborative Implementation Model for Urban Regeneration in Seoul : With Special Reference to Se-Woon District*. Seoul City University.
- Seo, I., "An Emperical Analysis on the Type of Cultural Urban Regeneration : Focused on the Case of Masan Old Downtown in Changwon City", *RESIDENTIAL ENVIRONMENT : JOURNAL OF THE RESIDENTIAL ENVIRONMENT INSTITUTE OF KOREA*, Vol. 14, No. 4, 2016, 363-382.
- Seoul, C. O. (2015) "2025 Seoul Urban Regeneration Strategies".
- Shin, W.-H. and Shin, W.-J., "A Study on the Priority among Achievement Indexes Depending on the Types of Urban Regeneration Projects: Focused on the Cases in Daegu", *Journal of The Korean Regional Development Association*, Vol. 29, No. 5, 2017, 39-52.
- Shon, Y.-M. and Park, T.-W., "Priority Comparative Analysis for Urban Revitalization Factor of Decline Commercial Area - Focused on Shinchon in Seoul", *Journal of the Korean Urban Management Association*, Vol. 30, No. 4, 2017, 1-18.
- Williams, N., Zander, S. and Armitage, G., "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical Ip Traffic Flow Classification", *ACM SIGCOMM Computer Communication Review*, Vol. 36, No. 5, 2006, 5-16.
- Yang, N. Y., Kim, S. G. and Kang, J. Y., "Researcher and Research Area Recommendation System for Promoting Convergence Research Using Text Mining and Messenger Ui", *The Journal of Information Systems*, Vol. 27, No. 4, 2018, 71-96.
- Yoon, S., Shin, M. and J, K., "A Study on City Brand Evaluation Method Using Text Mining : Focused on News Media", *Journal of Korea IT Service*, Vol. 18, No. 1, 2019, 153-171.
- You, E.-S., Choi, G.-H. and Kim, S.-H., "Study on Extraction of Keywords Using Tf-Idf and Text Structure of Novels", *Journal of the Korea Society of Computer and Information*, Vol. 20, No. 2, 2015, 121-129.
- Zhang, Y., Jin, R. and Zhou, Z.-H., "Understanding Bag-of-Words Model: A Statistical Framework", *International Journal of Machine Learning and Cybernetics*, Vol. 1, No. 1-4, 2010, 43-52.

Abstract

Suggestion of Urban Regeneration Type Recommendation System Based on Local Characteristics Using Text Mining

Ikjun Kim* · Junho Lee** · Hyomin Kim** · Juyoung Kang***

"The Urban Renewal New Deal project", one of the government's major national projects, is about developing underdeveloped areas by investing 50 trillion won in 100 locations on the first year and 500 over the next four years. This project is drawing keen attention from the media and local governments. However, the project model which fails to reflect the original characteristics of the area as it divides project area into five categories: "Our Neighborhood Restoration, Housing Maintenance Support Type, General Neighborhood Type, Central Urban Type, and Economic Base Type," According to keywords for successful urban regeneration in Korea, "resident participation," "regional specialization," "ministerial cooperation" and "public-private cooperation", when local governments propose urban regeneration projects to the government, they can see that it is most important to accurately understand the characteristics of the city and push ahead with the projects in a way that suits the characteristics of the city with the help of local residents and private companies. In addition, considering the gentrification problem, which is one of the side effects of urban regeneration projects, it is important to select and implement urban regeneration types suitable for the characteristics of the area.

In order to supplement the limitations of the 'Urban Regeneration New Deal Project' methodology, this study aims to propose a system that recommends urban regeneration types suitable for urban regeneration sites by utilizing various machine learning algorithms, referring to the urban regeneration types of the '2025 Seoul Metropolitan Government Urban Regeneration Strategy Plan' promoted based on regional characteristics. There are four types of urban regeneration in Seoul: "Low-use Low-Level Development, Abandonment, Deteriorated Housing, and Specialization of Historical and Cultural Resources" (Shon and

* Department of e-business, Ajou University

** Department of e-business, Ajou University

*** Department of e-business, Ajou University

**** Corresponding author: Juyoung Kang

Department of e-business, Ajou University

206 Worldcup-ro Suwon 16499 Korea

Tel: 82-31-219-2910, Fax: +82-31-219-1616, E-mail: jykang@ajou.ac.kr

Park, 2017).

In order to identify regional characteristics, approximately 100,000 text data were collected for 22 regions where the project was carried out for a total of four types of urban regeneration. Using the collected data, we drew key keywords for each region according to the type of urban regeneration and conducted topic modeling to explore whether there were differences between types. As a result, it was confirmed that a number of topics related to real estate and economy appeared in old residential areas, and in the case of declining and underdeveloped areas, topics reflecting the characteristics of areas where industrial activities were active in the past appeared. In the case of the historical and cultural resource area, since it is an area that contains traces of the past, many keywords related to the government appeared. Therefore, it was possible to confirm political topics and cultural topics resulting from various events. Finally, in the case of low-use and under-developed areas, many topics on real estate and accessibility are emerging, so accessibility is good. It mainly had the characteristics of a region where development is planned or is likely to be developed.

Furthermore, a model was implemented that proposes urban regeneration types tailored to regional characteristics for regions other than Seoul. Machine learning technology was used to implement the model, and training data and test data were randomly extracted at an 8:2 ratio and used. In order to compare the performance between various models, the input variables are set in two ways: Count Vector and TF-IDF Vector, and as Classifier, there are 5 types of SVM (Support Vector Machine), Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting. By applying it, performance comparison for a total of 10 models was conducted. The model with the highest performance was the Gradient Boosting method using TF-IDF Vector input data, and the accuracy was 97%. Therefore, the recommendation system proposed in this study is expected to recommend urban regeneration types based on the regional characteristics of new business sites in the process of carrying out urban regeneration projects."

Key Words : Urban Regeneration, Text Mining, Topic Modeling, Classification, Machine Learning

Received : April 10, 2020 Revised : September 7, 2020 Accepted : September 14, 2020

Corresponding Author : Juyoung Kang

저 자 소 개



이 준 호

현재 아주대학교 e-비즈니스학과 학부생으로 재학 중이며, 일본지역연구를 복수전공으로 하고 있다. 텍스트 마이닝과 도시재생사업, 산업 분석 및 전략 수립, 위성정보 서비스 기획 등을 관심분야로 두고 있다. 켄트리피케이션과 도시재생 유형별 특성을 주제로 2건의 학술대회논문을 투고하였다.



김 익 준

아주대학교 e-비즈니스, ICT융합을 복수전공하여 졸업했다. 텍스트 마이닝과 도시재생사업, 산업 분석 및 전략 수립, 위성정보 서비스 기획 등을 관심분야로 두고 있다. 켄트리피케이션과 도시재생 유형별 특성을 주제로 2건의 학술대회논문을 투고하였다.



김 효 민

아주대학교 e-비즈니스, ICT융합을 복수전공하여 졸업했다. 텍스트 마이닝과 도시재생사업, 산업 분석 및 전략 수립, 위성정보 서비스 기획 등을 관심분야로 두고 있다. 켄트리피케이션과 도시재생 유형별 특성을 주제로 2건의 학술대회논문을 투고하였다.



강 주 영

현재 아주대학교 경영대학 e비즈니스학과 교수로 재직중이며, 포항공과대학교 컴퓨터공학과에서 학사, 서울대학교 컴퓨터공학과에서 석사, 한국과학기술원 경영공학전공에서 공학박사학위를 취득하였다. 주요 관심분야는 빅데이터, 텍스트마이닝, 시맨틱 웹, 지능형 전자상거래, 클라우드 컴퓨팅, ERP 등이다. 관련 분야에서 몇편의 저서를 기술하고, 국내외 학회 및 해외 저명 학술지 등에 50여건 이상의 논문을 발표 및 게재하였다. 수상 경력으로는 한국지능정보시스템학회 우수논문상 수상, 한국경영정보학회 최우수 사례상 수상, 한국 IT 서비스학회 최우수 논문상 수상 등이 있다.