Journal of Digital Convergence
Vol. 19. No. 8, pp. 33-42, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.8.033

외국인 관광객 리뷰데이터를 활용한 토픽모델링 기반의 공간분석: 대구광역시를 사례로

정지우[1], 김서윤[2], 김현유[3], 윤주혁[3], 장원준[4], 김건욱[5*]
[1]경북대학교 문헌정보학과 학부과정, [2]영남대학교 통계학과 학부과정, [3]경북대학교 컴퓨터공학과 학부과정,
[4]대구디지털산업진흥원 빅데이터활용센터 전임연구원, [5]대구디지털산업진흥원 빅데이터활용센터 센터장

Spatial analysis based on topic modeling using foreign tourist review data: Case of Daegu

Ji-Woo Jung[1], Seo-Yun Kim[2], Hyeon-Yu Kim[3], Ju-Hyeok Yoon[3],
Won-Jun Jang[4], Keun-Wook Kim[5*]
[1]Undergraduate, Department of Library and Information Science, Kyungpook University
[2]Undergraduate, Department of Statistics, Yeungnam University
[3]Undergraduate, Department of Computer Science, Kyungpook University
[4]Associate Research Engineer, Big Data Center, Daegu Digital Industry Promotion Agency
[5]Director, Big Data Center, Daegu Digital Industry Promotion Agency

요 약 스마트폰 기반의 관광 플랫폼들이 활성화되면서 리뷰 데이터를 활용한 정책 수립 및 서비스 고도화가 다양한 분야에서 이루어지고 있다. 관광 리뷰 데이터를 활용한 선행연구들의 경우 국내 관광객 중심의 연구가 대다수 수행되었으며, 외국인 관광객 연구의 경우 일부 언어로 수집된 데이터와 텍스트 마이닝 기법에 한정하여 연구가 수행되었다. 이에 본 연구에서는 온라인 리뷰 사이트를 통해 '대구 명소' 키워드를 지정하여 외국인들이 작성한 리뷰 데이터 3,515건을 수집하였다. 그리고 LDA 기반의 토픽모델링을 수행하여 관광 토픽을 도출하였으며, 각 토픽별 전역 및 국지적 공간분석을 수행한 점이 선행연구와 차별성이라 할 수 있다. 분석 결과 전역적 공간 자기상관이 존재하며, 외국인들이 주로 방문하는 관광지들이 국지적으로 결집되어 있음을 확인하였다. 또한 대다수 토픽에서 중구를 중심으로 핫스팟이 도출되었으며, 분석 결과를 바탕으로 지자체 외국인 관광정책 수립 및 토픽모델링 기반의 공간분석 연구의 기초연구로 활용되길 기대하며, 본 연구의 한계점 또한 제시하였다.

주제어 : 관광분석, 외국인관광객, 리뷰데이터, 토픽모델링, LDA, 공간자기상관

Abstract As smartphone-based tourism platforms have become active, policy establishment and service enhancement using review data are being made in various fields. In the case of the preceding studies using tourism review data, most of the studies centered on domestic tourists were conducted, and in the case of foreign tourist studies, studies were conducted only on data collected in some languages and text mining techniques. In this study, 3,515 review data written by foreigners were collected by designating the "Daegu attractions" keyword through the online review site. And LDA-based topic modeling was performed to derive tourism topics. The spatial approach through global and local spatial autocorrelation analysis for each topic can be said to be different from previous studies. As a result of the analysis, it was confirmed that there is a global spatial autocorrelation, and that tourist destinations mainly visited by foreigners are concentrated locally. In addition, hot spots have been drawn around Jung-gu in most of the topics. Based on the analysis results, it is expected to be used as a basic research for spatial analysis based on local government foreign tourism policy establishment and topic modeling. And The limitations of this study were also presented.

Key Words : Tourism Analysis, Foreign Tourists, Review Data, Topic Modeling, LDA, Spatial Autocorrelation

*Corresponding Author : Keun-Wook Kim(aut7767@dip.or.kr)
Received May 21, 2021 Revised June 16, 2021
Accepted August 20, 2021 Published August 28, 2021

1. 서론

최근 4차 산업혁명으로 선도 기업들을 중심으로 빅데이터를 활용한 신시장 창출 및 혁신적인 서비스들이 개발되어 사회 전 분야에 있어 디지털 혁신이 가속화되고 있다. 공공분야에서도 데이터 기반의 정책 수립과 공공혁신 사례들이 발굴되어 데이터 기반의 정책 수립은 필수 사항이라 할 수 있다.

관광 분야에서는 데이터 기반의 관광정책을 수립하기 위하여 전통적인 방법인 설문지 기반의 실태조사를 하고 있으며, IT 기술의 발달로 2000년대에는 대규모 관광지를 대상으로 센서 기반의 방문자 수집 체계를 구축하여 데이터 수집 및 활성화 정책을 수립하였다.

앞서 선행한 데이터 기반의 관광정책 수립은 객관적이고 관리가 용이하다는 장점이 있으나, 표본조사와 일부 대규모 관광지에 한정된 데이터 수집으로 광역단위 또는 지자체 관광정책을 수립하기에는 많은 한계점이 존재한다. 또한, 과거의 데이터를 기반으로 하기 때문에 빠르게 변화하는 관광 트렌드를 반영하기에는 어려움이 있다.

최근 스마트폰 기반의 플랫폼들이 증가하여 소셜미디어 등을 통한 트렌드 분석이 활성화되면서 관광 분야에서도 관광지를 방문한 관광객들의 경험과 의견을 공유할 수 있는 플랫폼들이 활성화되고 있으며, 이로 인해 수집되는 데이터 또한 지속적으로 증가하고 있다. 그중에서도 데이터 수집과 분석이 용이한 관광 리뷰 데이터의 경우 다양한 분야에서 주로 활용되며, 국내외에서 관광 트렌드 분석으로 관광정책 수립 및 마케팅 등 실제 관광산업에 적용하고자 하는 연구가 다양하게 시도되고 있다.

관광 리뷰 데이터를 활용한 선행 연구들은 대다수 국내 관광객들을 대상으로 수행되었으며, 외국인 대상의 연구인 경우 주로 일부 언어로 작성된 리뷰 데이터와 텍스트 마이닝 기법에 한정된 연구가 수행되었다. 또한 지자체 관광 빅데이터 분석 시 통신사 관광인구 데이터를 활용하여 공간분석을 수행하고 있으나, 외국인들의 로밍 특성상 일부 관광지에 있어 과소·과대 추정되는 문제가 존재하여 공간분석에 있어 한계점을 가지고 있다.

2019년을 기준으로 국내를 방문한 외국인 관광객 중 약 3.1%가 대구광역시를 방문하였으며, 2014년 1.2%로 낮았던 점을 고려하면 다소 증가하였다고 볼 수 있으나, 영남권 주요 광역도시임을 고려하면 낮은 방문율이라 판단되며 이로 인한 세부적인 관광정책과 산업 활성화 연구 또한 부족한 상황이다.

이러한 배경 하에, 공간적으로는 대구광역시를 설정하였으며, 시간적 범위는 코로나 19가 전 세계적으로 확산되기 이전인 2015년 1월부터 2019년 12월 31일까지 약 5년간 외국인들이 작성한 관광 리뷰 데이터를 활용하여 연구를 수행하였다.

본 연구에서는 영어, 중국어 이외에도 다양한 언어로 작성된 외국인들의 리뷰 데이터를 수집하고자 하며, LDA 기반의 토픽모델링 분석으로 대구를 방문하는 외국인들의 관광 토픽을 추출하여, 이를 공간적으로 접근하는 점이 선행 연구와 차별성이라 할 수 있다. 이러한 분석 결과는 대구광역시 외국인 관광정책에 기초자료로 활용될 수 있을 것으로 판단되며, 토픽모델링 기반의 공간분석 연구가 부족한 만큼 기초연구로 활용되길 기대한다.

2. 선행 연구 고찰

2.1 온라인 리뷰 관련 선행 연구

국내 관광학 분야에서의 온라인 리뷰 데이터를 활용한 연구들은 주로 영문으로 작성된 리뷰 분석이 주를 이루고 있으며, 분석 기법으로는 토픽모델링, 감성 분석 등의 기법을 적용하였다[1-6]. 국외 연구로는 토픽모델링 기반의 만족도 요인 분석, 관광 마케팅 등의 연구가 수행되어 왔다[7-10].

국내 연구들의 세부적인 내용을 살펴보면 이새미・유승의(2020)은 흰여울문화마을을 대상으로 2019년 10월 31일 기준 구글맵 지역 리뷰 데이터 1,162건을 수집하였으며, LDA 토픽모델링 분석을 통해 관광 명소의 만족·불만족 요인을 도출하였다[1].

하지영・이승현(2018)은 전라북도 임실군을 대상으로 온라인 사이트에서 방문 목적과 장소 정보를 수집하여 LDA 기반의 토픽모델링으로 문화 관광자원 토픽을 도출한 후 지역 경제 활성화를 위한 정책을 제언하였다[2].

임종훈・김영현(2020)은 네이버와 다음 포털에 작성된 익산지역 관광지 온라인 리뷰 데이터를 통해 네트워크 분석을 수행하여 익산 관광객들의 관광 형태와 관심사를 파악하고 실무적 함의를 제시하였다[3].

유윤희・이희찬(2019)은 지리정보시스템 기반의 ESDA기법을 적용하여 지역별 공간분포에 대한 군집분석으로 관광수요 증대 및 지역별 관광 발전 정도를 객관화하는 분석을 수행하였으며 관광객 수의 결정 요인을 추출하고 지역 간 관광격차를 해소하기 위한 방안을 제시했다[4].

이외에도 트립어드바이저, 트위터에 작성된 리뷰 데이

터를 수집하여 키워드 분석을 통해 최적의 여행경로를 제안하였으며[5], 트립어드바이저의 온라인 리뷰를 이용하여 국내 관광지를 방문한 외국인 관광객을 대상으로 관광지 만족도 요인분석 연구[6] 등이 진행되었다.

Table 1. Prior Study Summary

Author (Year)	Purpose of Study	Data	Method
S. M. Lee, S. E. Ryu (2020)	A Study on the Tourism Promotion Plan of Huinnyeoul Culture Village Using Online Review Big Data Analysis	Google Maps reviews (1,162 reviews)	LDA
J. Y. Ha, S. H. Lee (2018)	Exploring Potential Tourism Resources by Analyzing Online Reviews of Local Visitors	Daum, Naver, Facebook and Twitter reviews (2017~2018, 367 reviews)	Network analysis, LDA
Y. H. Yu, H. C. Lee (2019)	A Study on The Factors Influencing Regional Tourism Performance Considering Spatial Autocorrelation	Statistics city yearbook of Korea -Tourist DB	ESDA, Moran's I, LISA
J. H. Im, Y. H. Kim (2017)	An Analysis of Experience of Visit to Gyeongbokgung Palace Using Big Data	Tripadvisor reviews (2005~2016, 2,794 reviews)	ANOVA, t-test analysis
S. H. Cho, B. S. Kim, M. S. Park, G. C. Lee, P. S. Kang (2017)	Extraction of Satisfaction Factors and Evaluation of Tourist Attractions based on Travel Site Review Comments	Tripadvisor reviews (21,620 reviews)	LDA
S. T. Park, Y. K. Kim (2019)	A Study on Deriving an Optimal Route for Foreign Tourists through the Analysis of Big Data.	Tripadvisor, Twitter reviews (17,241,823 reviews)	Sentiment Analysis, LDA
A. P. Kirilenko, S. O. Stepchenkova (2020)	Automated Topic Modeling of Negative Tourist Reviews	Tripadvisor reviews (14,273 reviews)	LDA
I. R. Putri, R. Kusumaningrum (2017)	LDA for Sentiment Analysis Toward Tourism Review in Indonesia	Tripadvisor reviews (100 reviews)	Sentiment Analysis, LDA
B. Mathayomchan, K. Sripanidkulchai (2019)	Utilizing Google Translated Reviews from Google Maps in Sentiment Analysis for Phuket Tourist Attractions	Google map reviews (22,094 reviews)	Sentiment Analysis
Y. Sugiyama, J. Zheng, T. Matsuo, H. Iwamoto (2018)	Multilingual Review Analysis for Attracting Foreign Visitors to Local Cities	Tripadvisor, Expedia, Agoda, Booking, Ctrip (about 1,500 reviews)	Review Analysis

국외 연구로는 A. P. Kirilenko · S. O. Stepchenkova (2020)는 Museum of Qin Terracotta Warriors(China)를 방문한 외국인 관광객의 리뷰데이터를 LDA기법을 활용하여 관광객 만족도 요인을 분석하였다[7].

I R Putri · R Kusumaningrum(2017)은 인도네시아의 외국인 관광객 리뷰 데이터에 대한 감성분석을 위해 LDA기법을 활용하여 관광객 만족도 요인을 분석하여 관광 활성화 정책을 제언하였다[8].

또한 B. Mathayomchan · K. Sripanidkulchai(2019)는 푸켓의 관광명소에 대한 관광객들의 영어 리뷰와 자동번역된 비영어 리뷰를 이용하여 영어리뷰로 구축한 감성분석 모델이 번역된 리뷰에도 효과적이라는 사실을 증명하였으며[9], Y. Sugiyama · J. Zheng · T. Matsuo · H. Iwamoto(2018)은 외국인 관광객의 지방 유치를 위해 하마마쓰 시를 대상으로 NLPIR를 이용한 리뷰 분석으로 국적별 효과적인 서비스 방안을 제시하였다[10].

앞서 선행 연구에서는 온라인 리뷰 데이터를 활용하여 LDA 기반의 토픽을 추출하여 관광객 정책 수립, 시사점을 도출한 장점이 있으나, 대부분 일부 언어로 수집된 리뷰 데이터로 분석이 수행된 점과 토픽모델링 기반의 텍스트마이닝 기법에 한정되어 수행된 한계점을 가지고 있다. 이에 본 연구에서는 영어 이외에 중국어, 스페인어 등 다수의 언어를 대상으로 온라인상에 작성된 관광 리뷰 데이터를 수집하여 분석을 수행하였으며, 수집된 데이터로 토픽모델링 기반의 공간분석을 수행한 점이 선행 연구와의 차별성이라 할 수 있다.

2.2 이론적 고찰

2.2.1 LDA 토픽모델링

토픽모델링은 문서 집합을 대상으로 본문에 잠재된 의미구조나 주제를 찾는 텍스트 마이닝 기법이다[11]. 확률적 토픽모델링 기법의 하나인 LDA는 문서에 포함된 단어 수의 분포를 통해 주제를 예측하는 기법으로 각 토픽의 특성을 도출하는데 용이하다[12].

LDA모형은 설정한 토픽 수에 따라 상이한 분석 결과가 나타나므로 최적의 잠재 토픽 수를 결정하는 것이 중요하며, 토픽 수는 주로 Perplexity를 적용하나 결과 해석에 어려움이 다소 존재하기도 한다.

본 연구에서는 관광지 리뷰 데이터에 잠재된 주요 토픽을 파악하기 위해 LDA 기반의 토픽모델링 기법을 이용하여 분석을 수행하였으며, 유의미한 해석이 가능한 최

적의 군집 개수(k)를 도출하기 위해 선행 연구를 참고하여 군집의 모델 적합도를 정량적으로 표현하는[13] Perplexity와 Coherence 지표로 최적의 토픽 수를 선정하고, 전문가 3명의 의견을 수렴하였다.

2.2.2 공간 자기상관분석

공간 자기상관은 공간정보가 담긴 데이터 간의 상호의존적인 특성을 지니고 있으며, Moran's I를 이용하여 분석 공간 단위 내에서 변수의 공간적 분포와 상관성을 살펴보는 방법으로[14], 분석범위에 따라 전역적, 국지적 규모로 나누어진다.

전역적(Global) Moran's I 통계량은 전체 지역에 대한 공간 자기상관성을 파악하여 하나의 값을 나타내는 지수이며, 값의 범위는 −1부터 +1사이의 값으로 표현된다. 인접한 영역이 비슷한 속성의 값이면 정(+)의 값 1로 나타나며, 인접한 영역의 속성이 다르면 부(−)의 값 −1로 나타난다. 또한, 공간적으로 상관성이 없을 경우는 0으로 표현된다[15].

국지적 LISA 통계량은 어떤 특정 지역이 전체 지역의 공간 자기상관에 영향을 미치고 있는지 파악할 수 있는 기법으로 LISA 지표를 활용하여 공간 자기상관이 높게 나타나는 특정 지역의 군집 분포(Hot Spot)를 시각화할 수 있다[16].

본 연구에서는 전역적 Moran's I 분석으로 관광 토픽별 공간 자기상관성을 분석하였고, LISA 기법을 통해 국지적 규모에서의 공간적 상관관계를 파악하여 온열지점(Hot Spot) 맵으로 시각화하였다.

3. 데이터 수집 및 전처리

3.1 데이터 수집

데이터 수집은 공간적으로는 대구광역시로 설정하고, 외국인들이 주로 이용하는 관광 플랫폼인 트립어드바이저와 구글맵 사이트에 리뷰가 10개 이상인 관광지 136곳을 대상으로 데이터를 수집하였다. 한국어가 아닌 리뷰 데이터는 외국인들이 작성한 리뷰로 가정하고 외국인 관광지 리뷰 데이터를 수집하였으며, 세부 정보로는 관광지, 일자, 추천평점, 주소, 위경도 등이 포함되어 있다.

데이터의 시간적 범위는 코로나 19가 전 세계적으로 확산되기 이전인 2015년 1월부터 2019년 12월 31일까지의 약 5년간의 데이터 총 3,572건의 리뷰 데이터를 수집 하였으며, 데이터 수집 과정에서 이상치와 결측치가 일부 존재하여 이를 제거하여 최종적으로 3,515건의 데이터를 수집하였다.

플랫폼별 수집한 데이터 수는 트립어드바이저 1,424건, 구글맵 2,091건이며, Table 2와 같이 나타난다.

Table 2. Summary of collected data

	Google maps	TripAdvisor
Keyword	대구 명소	
Period	2015/01/01~2019/12/31	
Components	Site name, Date, Rating, Site category, Address, Latitude, Longitude, Review text	
Reviews	2,091	1,424
	3,515	
Attractions	136	

3.2 데이터 전처리

외국인들이 온라인에 작성한 리뷰 데이터는 비정형 데이터 중 하나로 분석에 용이하게 정제하기 위해서는 별도의 전처리 작업이 필요하며, 수집된 데이터 중 텍스트를 포함하지 않았거나 중복으로 작성된 데이터는 제거하였다. 또한, 중국어, 일본어, 스페인어 등 다국어로 작성된 데이터는 구글 번역기 프로그램을 활용하여 국문 텍스트로 일괄 변환하였으며, 고유명사인 'jīnguāng shí(김광석)', 'yàolǐng(약령시)', 'míngdé(명덕)', 'xīmén(서문)' 등은 연구자가 개별적으로 변환 작업을 수행하였다.

또한, 언어의 특성상 자주 발생하고, 분석을 수행함에 있어 의미가 없는 '대구', '매우'와 같은 단어는 Table 3와 같이 불용어로 처리하였으며, 동일한 의미를 가지는 단어들은 유사어로 하나의 단어로 통합하였다. 대표적인 예로는 '김광석 거리' 또는 '김광석 길'은 '김광석거리'로 변환하였다.

Table 3. Stop-words

Stop-words	'당신','근처','대구','매우','우리','그것','가지','아주','때문','동안','곳','방문','장소','사람' 등 51개

Table 4. Representative words

Representative words	야간 ← 밤 관광명소 ← 관광 명소 김광석거리 ← 김광석 거리, 김광석 길, 김광석길 등 11개

4. 분석과정

4.1 네트워크 시각화

네트워크 분석은 객체 간의 관계를 단순화하여 직관적으로 시각화하는 기법 중 하나로, 외국인 관광 리뷰 데이터를 활용하여 단어의 출현 빈도와 연관성을 기반으로 네트워크 분석을 수행하였다.

전체 단어를 대상으로 네트워크 분석을 수행하기에는 자료의 복잡성과 시각적 표출 등의 제한으로 빈도수가 높은 상위 50개 단어를 대상으로 네트워크 분석을 수행하였으며, 대표적인 네트워크 분석 기법인 PFnet(Path Finder Network Scaling)[1]을 활용하여 분석하였다.

네트워크 분석 결과 총 5개의 중심점을 기반으로 Fig. 1과 같이 시각화를 도출하였으며, 주요 중심 단어별 세부적인 내용을 살펴보면 다음과 같다.

첫째, '박물관' 단어를 중심으로 '역사', '흥미', '전시회'가 연계되어 나타났으며, 박물관 투어와 연관된 키워드로 판단된다.

둘째, '시장', '음식' 단어를 중심으로 '전통', '야간', '경험' 단어가 연결되어 있는 것으로 나타났으며, 전통시장과 연계한 관광 리뷰로 판단된다.

셋째, '버스', '케이블카' 단어를 중심으로 '사원', '역', '풍경', '경치' 등의 단어가 높게 나타나며, 이는 케이블카와 연계한 경관 관광 리뷰인 것으로 판단된다.

넷째, '공원', '도시' 단어를 중심으로 '아이', '산책', '도시' 단어가 연결되어 있으며, 도시 내 공원과 연계한 관광 코스 리뷰인 것으로 보인다.

마지막으로 '카페', '사진' 단어와 연계하여 '레스토랑', '친구', '타워'의 연관성이 높게 나타나며, 대구 83타워와 관련된 관광 리뷰로 판단된다.

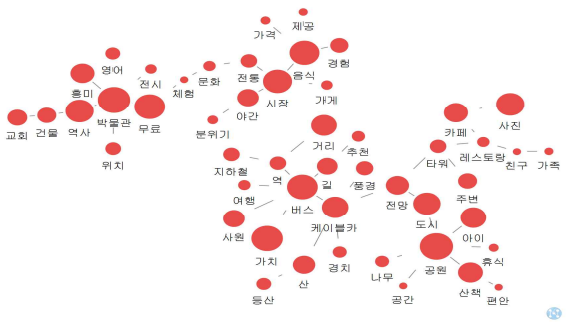


Fig. 1. Top 50 Word Network

4.2 LDA 토픽모델링

앞서 전처리된 리뷰 데이터를 활용하여 관광지별 특성을 파악하기 위해 LDA 기반의 토픽모델링 분석을 수행하였으며, 최적의 토픽 수를 찾기 위해 선행연구를 바탕으로 토픽 개수의 범위를 3개에서 20개로 지정하여 Perplexity 값과 Coherence 값을 통해 확인한 결과 Fig. 2, Fig. 3와 같이 나타났다.

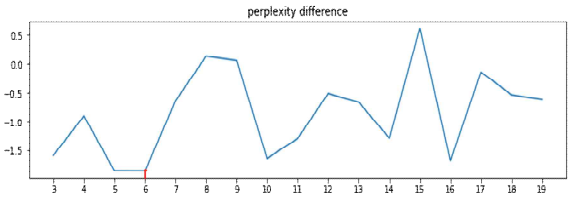


Fig. 2. Perplexity values on different topic numbers

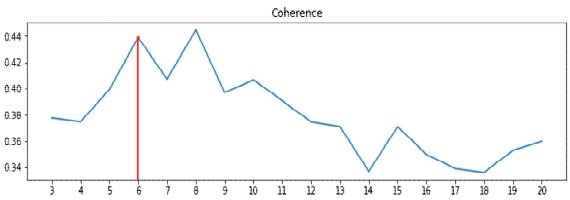


Fig. 3. Coherence values on different topic numbers

토픽의 수가 5개, 6개, 10개, 16개인 지점에서 Perplexity 값이 비교적 낮게 나타났으며, Coherence 값은 6개, 8개인 지점에서 높게 나타났다. 따라서 정량적인 수치로는 6개인 지점이 최적의 토픽 개수라 판단하였으며, 해당 분야 전문가 3명의 의견을 수렴하여 최종적으로 토픽의 수를 6개로 결정하였다.

선정된 토픽 수를 기반으로 LDA 기반의 토픽모델링 분석을 수행하였으며, 토픽별 도출된 키워드를 기반으로 연구자가 상위 주제어를 네이밍 하였다. 토픽모델링 분석 결과는 Table 5와 같이 나타나며, 토픽별 세부내용은 다음과 같다.

토픽 1은 박물관, 무료, 영어, 역사, 전시 등의 키워드를 포함하고 있으며, 이는 대구 명소 중 '대구 박물관', '약령시 박물관'을 방문하는 외국인 관광객이 많다는 것으로 해석할 수 있다. 이는 역사와 관련된 문화체험을 선호하는 것으로 유추하여 '한국의 역사체험'으로 분류하였다.

토픽 2는 시장, 음식, 가게, 서문시장, 상점 등의 키워드를 포함하고 있으며 이는 '식도락 관광'으로 분류하였다. 전통시장 방문을 통해 지역의 고유문화를 직접 체험할 수 있는 특화된 콘텐츠에 대한 관심을 두는 것으로 해

1) 그래프 이론에 기반한 심리학적 스케일링 방법

Table 5. LDA Topics' keyword and Label

Topic Number	Topic Label	Top Keyword	Number of Places	Number of Reviews
Topic1	한국의 역사체험	박물관, 무료, 영어, 역사, 전시	18 (13.2%)	655 (18.6%)
Topic2	식도락 관광	시장, 음식, 가게, 거리, 상점	17 (12.5%)	482 (13.7%)
Topic3	근·현대가 어울린	교회, 건물, 역사, 거리, 사진	18 (13.2%)	471 (13.4%)
Topic4	아름다운 경치	타워, 카페, 야간, 사진, 가치	18 (13.2%)	382 (10.8%)
Topic5	산행하기 좋은 곳	케이블카, 산, 버스, 사원, 등산	22 (16.2%)	702 (20.0%)
Topic6	도심 속 힐링	공원, 도시, 산책, 휴식, 운동	43 (31.6%)	823 (23.4%)

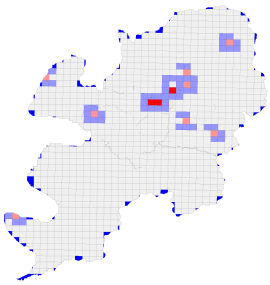
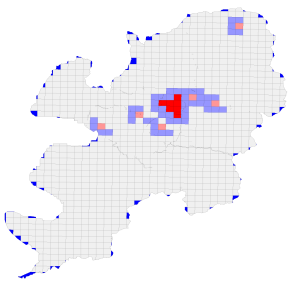
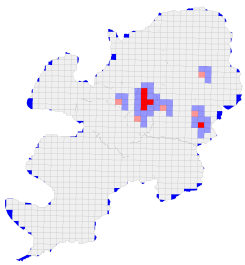
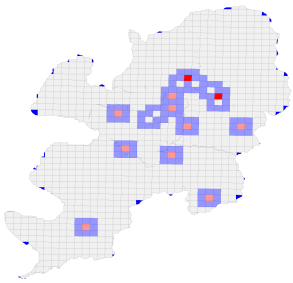
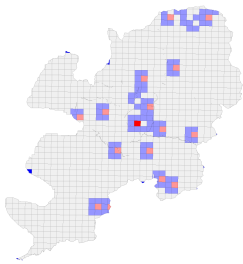
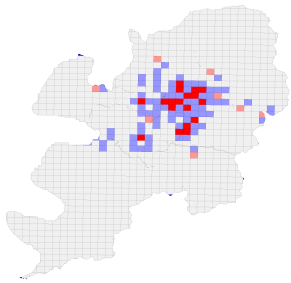
Method	Topic1(한국의 역사체험)	Topic2(식도락 관광)	Topic3(근·현대가 어울린)
LISA			
Moran's I	0.0587	0.2043	0.1813
p-value	0.005	0.001	0.001
	Topic4(아름다운 경치)	**Topic5(산행하기 좋은 곳)**	**Topic6(도심 속 힐링)**
LISA			
Moran's I	0.0585	0.0217	0.1544
p-value	0.002	0.096	0.001

Fig. 4. LISA Result
High-Low(HL)

High-High(HH), Low-Low(LL), Low-High(LH),

석할 수 있다.

토픽 3은 교회, 건물, 역사, 거리, 사진 등의 키워드를 포함하고 있으며 이는 '근·현대가 어울린 곳'으로 분류하였으며, 이는 토픽 1과 유사하게 역사·문화 및 유적 관광을 방문하는 관광객이 많다는 것으로 해석할 수 있다.

토픽 4는 타워, 카페, 야간, 사진, 가치 등의 키워드를 포함하고 있으며, '83타워'를 방문하는 외국인 관광객 수가 많은 것을 알 수 있다. 따라서 83타워를 중심으로 새로운 랜드마크를 형성하여 경쟁력을 갖춘 관광 상품 개발로 지역의 관광산업을 발전시킬 수 있을 것으로 해석할 수 있다.

토픽 5는 케이블카, 버스, 사원, 산, 등산 등의 키워드를 포함하고 있으며 외국인 관광객에게 인기 있는 관광 콘텐츠인 템플스테이를 체험하기 위해 사원과 산을 방문하는 관광객이 많다는 것으로 해석하였고, 이를 '산행하기 좋은 곳'으로 분류하였다.

토픽 6은 공원, 도시, 산책, 휴식, 운동 등의 키워드를 포함하고 있으며 여행 트렌드인 '힐링'을 목적으로 도심 속 휴식을 선호하는 관광객이 많다는 것으로 유추하여 이를 '도심 속 힐링'으로 분류하였다.

이상의 토픽모델링 분석을 통해 외국인 관광객 리뷰 데이터를 활용하여 잠재적 토픽을 추출하였으며, 앞서 수행한 네트워크 분석과 유사한 결과를 도출하였다.

4.3 공간 자기상관분석

대구를 방문하는 외국인 관광객들의 활동을 공간적인 측면에서 접근하기 위해 온라인 리뷰 데이터 내 포함되어 있는 관광지명을 활용하여 위도, 경도 정보를 추출하였으며, 토픽별 공간 자기상관분석으로 외국인이 방문하는 관광지를 공간적으로 분석하였다. 공간 자기상관분석은 일반적인 통계분석으로는 종속변수 간의 자기상관 등의 문제로 분석이 어렵기 때문에 본 연구에서는 특정한 지역의 외국인 방문 관광지와 인접한 지역의 관광지 간의 관계를 분석하기 위해 전역적 접근으로는 Moran's I 기법을 활용하였고, 국지적으로는 LISA기법을 적용하였다.

분석 결과는 Fig. 4와 같이 나타나며, 전역적 공간 상관관계의 지표인 Moran's I는 토픽 5를 제외한 모든 토픽에서 높은 정(+)의 상관관계가 나타났다. 이는 공간 자기상관이 존재하지 않는다는 귀무가설을 기각하는 것으로 공간 자기상관이 존재함을 알 수 있다.

국지적 공간 상관관계 지표인 LISA를 통해 살펴보면 High-High 군집에 해당하는 적색 지점은 토픽별 특정 지역의 외국인 방문 관광지 수가 높게 나타나며, 인접 지역 또한 관광지 수가 높은 지역을 의미한다.

Low-Low 군집의 청색 지점은 토픽별 특정 지역의 외국인 방문 관광지 수가 낮은 지역을 의미하고, 인접지역 또한 관광지 수가 낮은 지역을 의미한다. 그리고 통계적으로 유의하지 않은 지역은 회색으로 나타났다.

이를 통해 토픽별 외국인들이 주로 방문하는 관광지들의 영역을 살펴볼 수 있고, 공간 간의 자기상관분석으로 외국인 방문 관광지들의 결집 지역을 도출할 수 있다.

5. 결론

5.1 분석 결과 요약

최근 4차 산업혁명으로 빅데이터, 인공지능이 각광받고 있으며, 선도 기업들을 중심으로 빅데이터를 활용한 신시장 창출 및 혁신적인 서비스들이 개발되어 사회 전 분야에서 디지털 혁신이 이루어지고 있다.

관광 분야에서는 내·외국인 관광객 유치를 위한 웹사이트·소셜 미디어 등을 활용한 빅데이터 분석이 증가하고 있으며, 그중에서도 데이터 수집이 용이한 온라인 리뷰데이터를 활용한 트렌드 분석이 관광정책 수립에 유용하게 활용되고 있다.

하지만 대다수의 선행연구들이 국내 관광객들을 대상으로 수행되었으며, 외국인 관광객들을 대상으로 하는 연구는 주로 일부 언어와 텍스트마이닝 기법에 한정되어 연구가 진행되었다. 또한, 지자체별로 관광 빅데이터 분석을 위해 통신사 관광인구를 활용하고 있으나, 외국인들의 로밍 특성상 일부 관광지에 있어 과소·과대 추정되는 문제가 발생하는 한계점을 가지고 있다.

이에 본 연구에서는 외국인들이 실제 작성한 온라인 리뷰 데이터를 수집하여 분석을 수행하였으며, 한국어가 아닌 모든 외국어 리뷰 데이터 3,515건을 수집하였다.

수집된 데이터를 LDA 기반의 토픽모델링으로 토픽별 키워드를 추출하였고, 토픽별 공간 자기상관분석 기법으로 공간분석을 수행하여 외국인 관광지 결집지역 및 정책적 시사점을 도출한 점이 선행연구와의 차별점이라 할 수 있다.

분석된 결과를 요약하면 다음과 같다.

첫째, 수집된 리뷰 데이터를 활용하여 네트워크 분석을 수행한 결과 박물관, 시장·음식, 버스·케이블카, 공원·도시, 카페·사진 키워드를 중심점으로 리뷰 데이터 내 키워드들이 연계되어 있음을 알 수 있다.

둘째, 토픽모델링 분석 결과 전체 토픽 6개 중 토픽 1(한국의 역사체험)과 토픽 3(근·현대가 어울린)은 한국의 역사 및 문화와 관련된 테마라는 점에서 유사한 성격을 가지고 있으며, 외국인 관광객들을 대상으로 한국의 전통문화를 체험할 수 있는 연계 상품을 제공한다면 효과적일 것으로 판단된다.

셋째, 토픽 4(아름다운 경치)와 토픽 5(산행하기 좋은 곳)은 대구시 경관과 관련된 테마라는 점에서 유사한 성격을 가지고 있으며, 토픽 6(도심 속 힐링) 또한 자연경관과 연관된 장소로 이루어져 있다.

넷째, 토픽별 Global Moran's I 기법으로 전체 관광지에 대한 공간 자기상관을 분석한 결과 토픽 5(산행하기 좋은 곳)을 제외한 모든 토픽에서 공간 자기상관이 통계적으로 유의한 것으로 나타나 공간 자기상관이 존재하는 것으로 분석되었다. Moran's I 계수는 토픽 2(식도락 관광), 토픽 3(근·현대가 어울린), 토픽 6(도심 속 힐링)

순으로 높게 나타나 공간적 결집 정도를 계량적으로 분석하였다.

마지막으로 LISA 지표를 통해 국지적 공간 자기상관 분석을 수행한 결과, 토픽 4(아름다운 경치), 토픽 5(산행하기 좋은 곳)를 제외한 나머지 토픽들은 중구를 중심으로 Hot-Spot이 도출되었으며, 대구광역시 중심지인 중구를 기반으로 외국인 관광객들이 주로 선호하는 관광지들이 분포하고 있음을 알 수 있다. 토픽 4(아름다운 경치)의 경우 벽화마을, 망우당 공원, 공항교, 83타워 등 야간 경관이 우수한 관광지의 외국인 방문이 높게 나타났다. 또한 토픽 5(산행하기 좋은 곳)는 팔공산, 동화사, 달성 한일우호관, 불로동고분군, 건들바위, 벽화마을 등으로 공간적으로 분산되어 나타났으며, 토픽 6(도심 속 힐링)은 하중도, 강정보, 화랑공원 등에 집중되어있는 것으로 나타났다.

5.2 정책적 시사점 및 연구 한계점

본 연구의 분석 결과를 토대로 정책적 시사점을 정리하면 다음과 같다.

첫째, 전체 리뷰 데이터 분석 결과 '박물관', '시장' 등의 키워드 빈도가 높게 나타났으며, 대구를 방문하는 외국인 관광객들의 경우 전통문화 체험과 소통과 관련된 콘텐츠에 관심이 높은 것으로 나타났다. 지자체에서 진행하는 문화 축제와 관광 자원을 개선한다면 외국인들의 관광지 방문이 증가할 것으로 판단된다.

둘째, 토픽별 키워드 분석 결과, 토픽 1(한국의 역사체험), 토픽 2(식도락 관광), 토픽 3(근·현대가 어울린)은 공간적으로 중구를 기반으로 외국인들이 주로 방문하는 관광지가 결집하고 있으며, 인접한 북구 일부 지역 또한 핫스팟으로 도출되었다. 중구의 골목투어와 스마트 관광 등의 관광 상품과 역사·문화 관광지, 쇼핑·오락·음식 등의 관광 자원과의 연계시 시너지가 높을 것으로 예상된다.

셋째, 토픽 6(도심 속 힐링)의 경우 특정 관광지인 하중도, 강정보, 화랑공원을 기반으로 외국인 관광지가 결집되어 있고, 공간적으로 근거리에 위치하고 있는 것으로 분석되었다. 도심 속 힐링이라는 관광 콘텐츠로 연계 노선을 개발한다면 이동과 편의적인 측면에서 효과적일 것으로 보인다.

넷째, 자연 경관과 관련된 토픽 4(아름다운 경치)의 경우 공항교, 망우당 공원 등 야간 경관이 우수한 관광지들이 도출되었으며, 최근 지자체 관광 통계자료에서 동구를 방문하는 관광객들이 지속적으로 증가하고 있음을 감안할 때 생태 환경과 야간경관을 연계한 관광 상품 개발 시 동구의 관광객 방문이 활성화될 것으로 판단된다.

다섯째, 토픽 5(산행하기 좋은 곳)는 공간적으로 외국인들이 주로 방문하는 관광지가 도시 외곽으로 분산되어 있으며, 사원, 등산로, 수목원 등과 연계하여 주목받지 못한 잠재 관광지들을 개발한다면 팔공산 이외에도 외국인 관광객들이 자주 방문하는 관광지로 활성화될 수 있을 것으로 판단된다.

마지막으로 전통적인 지자체별 설문조사, 센서 데이터 기반의 관광정책 수립과 본 연구는 수집 데이터, 분석 기법 등에서 차이점을 보이며, 기존의 다년간 수행되어 온 관광자원과 연계하여 관광정책을 수립한다면 의미 있는 외국인 관광정책 수립이 이루어질 것으로 기대한다.

본 연구에서는 다음과 같은 한계점을 가지고 있다.

첫째, 수집된 데이터는 코로나 19 발생 이전의 데이터로 코로나 19로 인한 외국인 관광객들의 통행행태 변화와 포스트 코로나에 따른 관광 트렌드 변화에는 다소 차이가 있을 것으로 판단되며, 추후 코로나 19 이후의 외국인 관광객 분석과 비교한다면 의미 있는 연구결과가 이루어질 것으로 판단된다.

둘째, 온라인 리뷰 데이터의 경우 데이터의 수집 기간, 제공 플랫폼, 불용어, 유사어 지정 등의 다양한 조건에 따라 분석 결과가 상이하게 발생할 수 있으므로, 데이터 신뢰성 문제가 중요할 것으로 보인다.

셋째, 대구를 방문하는 중국인 관광객 비중이 높으나, 중국인들의 경우 C trip, 马蜂窝(마펑워)와 같은 별도의 여행 리뷰 플랫폼을 주로 활용하기 때문에 본 연구에서 수집된 데이터가 중국인 관광객들의 리뷰 특성을 충분히 반영하지 못했을 가능성이 존재한다.

넷째, 트립어드바이저와 구글맵에 작성된 외국어를 기준으로 외국인 관광객으로 정의하여 분석을 수행하였는데, 국내에 거주하는 외국인과 미군부대 관계자들이 상당수 존재하는 바 이에 따른 데이터 편향이 일부 존재할 것으로 예상된다. 향후 언어별 세분화된 분석을 수행할 경우 심층적인 연구가 수행될 것으로 판단된다.

다섯째, 트립어드바이저와 달리 구글맵에서 제공하는 리뷰 데이터에서는 특정기간이 경과하면 시계열 데이터를 제공하지 않아 시계열 분석을 수행하는데 한계가 존재한다. 추후 다양한 플랫폼들의 데이터가 수집되어 데이터의 품질과 규모가 보완된다면 시계열 분석도 가능할 것으로 판단된다.

마지막으로 외국인 관광객들이 작성한 리뷰 데이터를 수집하였으나, 빅데이터 분석으로 수행하기에는 다소 데이터의 수가 부족한 것으로 판단되며, 추후 지자체 수집

데이터, 통신사 관광 데이터 등 다양한 유형의 데이터와 결합으로 내국인과 외국인 관광객 비교 분석을 수행한다면 고도화된 연구가 가능할 것으로 판단된다.

REFERENCES

[1] S. M. Lee & S. E. Ryu. (2018). A Study on the Tourism Promotion Plan of Huinnyeoul Culture Village Using Online Review Big Data Analysis. *Journal of Hotel & Resort, 19, 115-130.*

[2] J. Y. Ha & S. H. Lee. (2018). Discovering potential cultural tourism resources through online review of local visitors : Focused on Imsil county, Jeonbuk Province. *Journal of Region & Culture, 5 4), 67-85.* DOI:10.26654/iagc.2018.5.4.067

[3] J. H. Im & Y. H. Kim. (2020). Examining Public Opinion on Iksan Tourism Using Social Media Analytics. Tourism Research, 45(3), 427-441. DOI:10.32780/ktidoi.2020.45.3.427

[4] Y. H. Yu & H. C. Lee. (2019). A study on the factors influencing regional tourism performance considering spatial autocorrelation. International Journal of Tourism and Hospitality Research, 33(2), 35-46. DOI : 10.21298/IJTHR.2019.2.33.2.35

[5] S. T. Park & Y. K. Kim. (2019). A Study on Deriving an Optimal Route for Foreign Tourists through the Analysis of Big Data. *Journal of Convergence for Information Technology, 9(10), 56-63.* DOI:10.22156/CS4SMB.2019.9.10.056

[6] S. H. Cho, B. S. Kim, M. S. Park, G. C. Lee & P. S. Kang. (2017). Extraction of Satisfaction Factors and Evaluation of Tourist Attractions based on Travel Site Review Comments. *Journal of the Korean Institute of Industrial Engineers , 43(1), 62-71.* DOI:10.7232/JKIIE.2017.43.1.062

[7] A. P. Kirilenko & S. O. Stepchenckova. (2019). Automated Topic Modeling of Negative Tourist Reviews. *e-Review of Tourism Research, 17(4), 532-545*

[8] I. R. Putri & R. Kusumaningrum. (2017). Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia. *Journal of Physics: Conference Series, 801(1), 012073* DOI:10.1088/1742-6596/801/1/012073

[9] B. Mathayomchan. & K. Sripanidkulcha.. (2019, July). Utilizing Google Translated Reviews from Google Maps in Sentiment Analysis for Phuket Tourist Attractions. *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE).* pp(260-265). Chonburi : IEEE DOI : 10.1109/JCSSE.2019.8864150

[10] Y. Sugiyama, J. Zheng, T. Matsuo, & H. Iwamoto (2018, July). Multilingual Review Analysis for Attracting Foreign Visitors to Local Cities -About Sightseeing in Hamamatsu city-. *2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI) IIAI-AAI Advanced Applied Informatics (IIAI-AAI).* pp(741-746). Yonago : IEEE DOI:10.1109/IIAI-AAI.2018.00153

[11] M. Steyvers & T. Griffiths. (2007). Probabilistic topic models. Handbook of latent semantic analysis. *Psychology Press.* DOI : 10.4324/9780203936399.ch21

[12] W. S. Kim & D. S. K. (2020). Deriving the Determinants of Hotel Service Quality According to Hotel Class in Korea by Using LDA Topic Modeling. *Journal of International Trade & Commerce, 16, 779-791.* DOI:10.16980/jitc.16.5.202010.779

[13] S. G. Nam, J. H. Kim & Y. J. Yu. (2020) Understanding the Shopping Tourist Perceptions based on Text Analytics. Journal of New Industry and Business, 38(2), 3-21 DOI:10.30753/emr.2020.38.2.001

[14] H. J. Yun & M. H. Park. (2014). Multivariate Spatial Autocorrelation on the Number of Tourists and Natural Amenities. *Journal of Tourism Management Research, 18(58), 135-150.*

[15] J. M. Yun & C. K. Seo. (2010). Deriving the Declining Areas and Analysing Their Spatial Characteristics Using the Spatial Autocorrelation Measure. *Journal of the Korean Association of Geographic Information Studies, 13(3), 64-73.* DOI : 10.11108/kagis.2010.13.3.064

[16] D. H. Lee, S. B. Yoon & J. S. Kim (2015). Analysis of the Crime Pattern and Influencing Factors by the Spatial Autocorrelation in Busan. *Journal of the Korean Regional Development Association, 27(2), 259-276.*

정 지 우(Jiwoo Jung) [학생회원]

· 2020년 2월 : 대구대학교 호텔관광학과
· 2020년 3월 ~ 현재 : 경북대학교 문헌정보학과
· 2020년 12월 ~ 2021년 2월 : 대구디지털산업진흥원 현장실습
· 관심분야 : 관광학, 빅데이터, 텍스트마이닝
· E-Mail : brisbanejiwoo@gmail.com

김 서 윤(Seoyun Kim) [학생회원]

· 2020년 2월 : 한남대학교 비즈니스통계학과
· 2020년 3월 ~ 현재 : 영남대학교 통계학과
· 2021년 1월 : 대구디지털산업진흥원 현장실습
· 관심분야 : 통계학, 빅데이터
· E-Mail : sparklingyouth@naver.com

김 현 유(Hyeonyu Kim) [학생회원]

· 2017년 3월 ~ 현재 : 경북대학교 컴퓨터학과
· 2020년 12월 ~ 2021년 2월 : 대구디지털산업진흥원 현장실습
· 관심분야 : 인공지능, 빅데이터, 데이터마이닝
· E-Mail : jjpinklady98@gmail.com

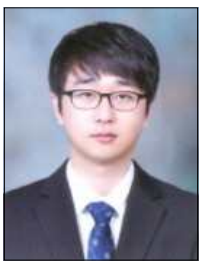
윤 주 혁(Juhyeok Yoon) [학생회원]

· 2019년 3월 ~ 현재 : 경북대학교 컴퓨터학부
· 2020년 12월 ~ 2021년 2월 : 대구디지털산업진흥원 현장실습
· 관심분야 : 인공지능
· E-Mail : juhyeok0123@gmail.com

장 원 준(Wonjun Chang) [정회원]

· 2021년 2월 : 영남대학교 통계학과
· 2020년 10월 ~ 현재 : 대구디지털산업진흥원 빅데이터활용센터 전임연구원
· 관심분야 : 통계학, 공간분석, 인공지능, 텍스트마이닝
· E-Mail : bd1jun@dip.or.kr

김 건 욱(Keunwook Kim) [정회원]



· 2009년 2월 : 영남대학교 도시공학(공학사)
· 2011년 8월 : 아주대학교 교통공학(공학석사-교통모델링)
· 2019년 7월 ~ 현재 : 대구디지털산업진흥원 빅데이터활용센터 센터장
· 관심분야 : 도시데이터분석, 빅데이터, 인공지능, 텍스트마이닝
· E-Mail : aut7767@dip.or.kr