관광학연구 제42권 제8호(통권 164호) pp.165~189
DOI http://dx.doi.org/10.17086/JTS.2018.42.8.165.189 한국관광학회 2018. 9.

온라인 리뷰 빅데이터 기반의 Word2Vec 기법을 활용한 관광지 개성과 여행객 평점 간 구조적 관계 분석*

An Analysis on the Structural Relationship Between Destination Personality and Traveler Rating with Word2Vec Technique based on Online Review Big Data

심 영 석** · 김 홍 범***
Sim, Young-Seok · Kim, Hong-Bum

요 약 : 온라인 리뷰를 바탕으로 형성된 관광지에 대한 긍정적인 이미지는 여행객의 의사결정에 중요한 원천이 되며, 이러한 이미지는 관광지 개성을 통해 형성된다. 현재까지 관광지 개성에 대한 연구는 제품범주의 브랜드 개성 측정척도(BPS)를 활용하여 척도개발, 관광지 이미지 및 행동의도 간 영향관계 규명 등 다양하게 이루어져 왔지만, 관광지가 가지는 본질적 가치를 반영하지 못한다는 한계점이 존재하였다. 이에 본 연구는 온라인 리뷰의 질적·양적 정보를 활용하여 의미론적 차원에서 관광지 개성을 확장하기 위해 신경망 언어 모델인 Word2Vec를 활용하였으며, 여행객 평점과의 영향관계를 추정하였다. 분석 결과, 관광지 개성은 기존 브랜드 개성의 구성개념과 달리 관광지 경험 후의 정서적인 감정을 표현하는 유사단어들이 도출되었다. 또한, 관광지 개성의 세련됨이 여행객 평점에 유의한 양(+)의 영향력이 가장 큰 요인으로 나타났으며, 강인함이 여행객 평점에 유의한 음(-)의 영향력을 가지는 설명요인인 것으로 나타났다. 마지막으로 텍스트 데이터를 수치화하여 인과관계를 추정할 경우 OLS모형보다 WLS모형이 보다 설명력이 높은 것으로 나타났다.

핵심용어 : 온라인 리뷰, Word2Vec, 단어 임베딩, 신경망언어모델, 관광지 개성, 가중최소제곱법

ABSTRACT : The current study examines destination personality in a contextual semantic approach by utilizing word embedding with deep learning via Word2Vec, a neural network language model using collected online review data from Tripadvisor.com. This study tested the relationship between destination personality and traveler rating. The results show that the traits of destination personality reflect expressed affective emotion after travelers' experience, unlike existing brand personality scales that measure tangible product. Furthermore, sophistication, which is one of the dimensions of destination personality, had the most significant positive impact on the traveler rating, but ruggedness, another dimension of destination personality, appeared to have a negative effect on the traveler rating. Finally, comparing the result of the WLS regression to the OLS regression, R-squared for WLS was found to be substantially superior to that of OLS when estimating relationship by quantifying textual data.

Key words : Online review, Word2Vec, Word embedding, Neural network language model, Destination personality, Weighted least square

* 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016S1A5A2A01022896).
** 세종대학교 대학원 호텔관광경영학과 박사과정. e-mail: iamssys@gmail.com
*** 세종대학교 호텔관광대학 교수(교신저자). e-mail: kimhb@sejong.ac.kr

Ⅰ. 서 론

관광산업은 인터넷 상의 정보기술(IT) 확대로 성장세가 가장 빠른 산업 중 하나로 발돋움하게 되었으며, 이러한 시대적 패러다임의 변화에는 소비자들이 관광활동 의사결정과정에서 온라인 리뷰에 내재되어 있는 여행경험 관련정보의 적극적인 수용(Wang & Fesenmaier, 2004)과 e-티켓 기반의 관광 상품의 유통활성화를 통한 편의성 도모 등이 중요한 요인으로 자리매김하였기 때문이다(전효재, 2018). 특히 온라인상에서 실시간으로 생산, 공유되고 있는 소비자 리뷰는 관광객의 욕구나 행태, 지역 및 관광지의 매력, 관광지와 기반시설에 대한 서비스 평가에 이르기까지 최신정보를 내포하여 온라인구전(eWOM) 효과를 통한 잠재 관광수요를 창출하는 순기능을 담당함으로서 관광산업 성장 동력의 원천으로 안착하였다. 실제로, Ludwig *et al.*(2013)의 연구에서 온라인 소비자의 92%가 리뷰를 내용 그대로 받아들이고 이를 토대로 구매의사결정에 활용하는 것으로 나타났으며, 이는 궁극적으로 소비자들이 영리목적 없이 제공되는 사용자 제작 콘텐츠(UGC)를 더 신뢰할만한 정보로 인식하고 있다는 것을 의미한다(Kang & Schuett, 2013). 더욱이 무형의 상품과 서비스를 제공하는 관광산업의 경우, 구매가 소비보다 선결된다는 점에서 발생하는 위험 및 불확실성에 대한 인식이 온라인 리뷰로 인해 감소한다는 연구결과(Morosan & DeFranco, 2016)와 온라인 리뷰를 통한 정보 교류는 소비자의 힘(power)을 증대시키는 역할을 한다는 연구결과(노미진・이경탁, 2012) 등을 통해 온라인 리뷰가 가지는 파급력은 더욱 커질 것이라는 것을 예상해 볼 수 있다.

이러한 배경에서 학계에서는 온라인 리뷰 상에서 텍스트로 표현된 정성적인 평가와 수치로 표현된 정량적 평가(평점)를 다각적으로 활용함으로써 데이터에 내재되어 있는 유의미한 정보를 추출 및 탐색, 분석할 수 있는 여러 방법들이 논의되고 있다(O'Connor, 2010). 예컨대 온라인 여행사(OTA)의 평가항목에 대하여 소비자가 남긴 양적 데이터인 평점을 활용하여 숙박시장의 가격결정요인을 도출한 연구나, 언어권에 따른 호텔등급별 선호도 및 평점의 차이 등 인과관계를 규명한 연구들이 이루어지고 있다(Schuckert, Liu & Law, 2015; Zhang, Ye & Law, 2011). 더불어 웹상의 블로그 게시물, 온라인 여행 커뮤니티의 여행객 리뷰 등의 질적 데이터를 활용하여 사회연결망 분석을 통한 여행객의 관광인식 및 이미지 조사(오익근・이태숙・전채남, 2015; 한지연・김홍범, 2017), 검색엔진 키워드를 바탕으로 한 수요예측 등의 연구가 이루어지고 있으며(박수지・신진옥・송상헌・정철, 2017), 최근에는 키워드 검색을 통하여 수집된 리뷰를 토대로 연관규칙분석을 수행하여 관광객의 변화를 유형화하기 위한 접근법 등이 제안되고 있다(전효재, 2018).

이와 같이 현재까지의 연구들은 관광분야에서 온라인 리뷰 빅데이터를 실증적으로 활용할 수 있는 접근법과 방법론적 대안을 제시하였지만 한계점이 존재한다. 우선 온라인 리뷰는 소비자의 질적·양적 정보가 동시에 존재함에도 불구하고 텍스트 또는 평점의 단일 측면만 고려한 연구들이 이루어지고 있어 현상에 대한 소비자의 패턴을 파악한 이후 여행객의 선호행동, 또는 관광지 특성 등의 영향관계까지 파악하기에는 제한적일 수밖에 없다. 따라서 온라인 리뷰의 질적, 양적 정보를 다각적으로 활용한다면 현상에 대한 탐색적인 조명과 인과적 영향관계까지 규명할 수 있을 것이다.

또한 온라인 리뷰의 텍스트 분석 과정에서 데이터가 가진 문맥의 의미를 사용하는 대신, 특정 단어가 얼마나 사용되는지 또는 다른 데이터와 얼마나 자주 연관성이 있는지 등의 빈도수 기반

의 텍스트 마이닝 기법이 주로 활용되었다. 이러한 접근법은 직관적으로 주제 및 관심분야를 탐색적으로 밝힐 수는 있지만, 데이터에 내재되어 있는 소비자들의 심리를 파악하기에는 한계점이 존재할 수 있다. 따라서 여타 산업보다 온라인구전의 의존성과 파급효과가 큰 관광의 경우, 의미론적 맥락에서 분석이 이루어진다면 여행객들의 평가를 보다 심층적으로 조명할 수 있을 것이다. 최근 자연어 처리 분야에서는 텍스트에 대하여 형태소 분석에서 나아가 기계학습, 딥 러닝 기반의 신경망 모델링 기법을 활용하여 단어의 앞, 뒤 문맥에서 텍스트를 분석하기 위한 의미론적 접근법인 신경망 언어 모델(NNLM)을 통한 분석의 시도가 이루어지고 있다(Chakraborty, Bhattacharyya, Bag & Hassanien, 2018; Shao, Chen & Chen, 2018). 따라서 이러한 모델링 기법이 관광분야에서 적절히 활용된다면 여행객들의 의견에 내재되어 있는 심층적이고 구체적인 정보를 보다 효과적으로 파악하는 적절한 대안이 될 수 있을 것이다.

이와 같이 빅데이터를 분석하는 방법론적 접근법 외 이론적인 관점에서 온라인 리뷰의 궁극적인 활용은 여행객들의 경험에 내재되어 있는 표현들을 분석하고 해석하는 과정을 거쳐 관광지가 차별화되기 위한 고유한 정체성을 확립함으로써 잠재 관광시장의 자극을 통한 관광 수요, 여행객 행태 등을 설명할 수 있는 근거가 될 수 있다. 특히 관광지가 경쟁력을 가지고 하나의 장소로서 차별화되기 위해서는 관광지 평가에서 가장 선행되는 요인인 개성에 대한 이해가 필수적이다(Tasci, Gartner & Cavusgil, 2007).

따라서 본 연구의 목적은 온라인 리뷰 빅데이터에 기반하여 자연어 처리 분야에서 최근 화두가 되고 있는 신경망 언어 딥 러닝 모델인 Word2Vec을 통하여 관광지 평가에 중요한 개념인 관광지 개성의 구성개념을 의미론적 맥락에서 확장하고, 여행객 평점과의 연계를 통하여 텍스트 데이터와 정량 데이터 간 영향관계를 실증적으로 규명하는 것이다. 이러한 본 연구의 시도는 관광분야에서 신경망 언어 모델의 실질적 활용과 방법론적 체계를 마련하고, 아울러 평점 데이터와 연계하여 모형을 측정하는 선험적 연구로서 그 의의가 클 것으로 사료된다.

Ⅱ. 이론적 배경

1. 관광지 개성의 측정

관광지 개성은 마케팅 분야의 브랜드 개성 이론이 근간이 된 것으로, 브랜드 개성은 소비자가 상품에 대하여 인간의 특성(human trait)을 토대로 상징적, 자아 표현적으로 평가한다는 개념이다(Kumar & Nayak, 2018). 브랜드 개성 이론을 정립한 Aaker(1997)는 브랜드 개성 측정을 위해 최초 114개의 측정문항을 구성하였으며, 실증분석을 통하여 진실성(sincerity), 흥미로움(excitement), 유능함(competence), 세련됨(sophistication), 강인함(ruggedness)의 5가지의 차원과 15개의 속성, 42개의 측정항목의 구성개념을 바탕으로 상품에 대하여 소비자가 인식하는 개성의 개념을 구체화하였다.

관광분야에서는 Ekinci and Hosany(2006)가 Aaker(1997)의 브랜드 개성의 개념적 틀을 적용하여 여행객이 관광지에서의 경험을 통해 인식하는 개성을 측정함으로써 브랜드 개성의 개념을 관광분야까지 확장하였다. 특히 이 연구에서는 관광지 개성에 대하여 기존의 브랜드 개성 차원인 진실성(sincerity), 흥미로움(excitement)외 유쾌함(conviviality)이라는 차원을 도출하였을 뿐만 아니라 나아가, 관광지가 지니는 개성이 여행객의 관점에서 지각되는 이미지의 선행요인으로 긍정적 이미지를 강화시키고 관광객들의 방문의도에도 영향을 준다는 것을 실증적으로 규

명하였으며, 여행객들은 관광지에 따라 상이한 개성의 속성(personality traits)을 부여한다는 것을 입증하였다.

따라서 개념적 관점에서 브랜드 개성과 관광지 개성 간에는 소비자의 경험재에 따라 상이한 개성의 속성을 부여한다는 점에서는 맥을 같이 하지만 유·무형의 상품이 결합된 관광상품의 특성상, 각 관광지에 적합한 개성을 강조함으로써 긍정적 이미지 창출을 극대화할 수 있으며, 이는 곧 관광객들의 방문, 충성도, 재방문, 추천의도 등 행동의도에 직접적인 영향을 미친다는 것을 알 수 있다. 그러므로 관광지 개성은 여행객이 관광지를 평가하기 위한 구조적 관계에서 중요한 선행요인으로서 그 기능을 담당한다는 점이 브랜드 개성과 가장 큰 차이점이라 할 수 있다.

이처럼 관광지 개성이 소비자의 인식에서 비롯되는 관광지 평가의 시작점이 되는 변수로 그 중요성이 강조되면서, 관광지 개성 관련 후속 연구에서는 Aaker(1997)가 제안한 5가지 브랜드 개성의 차원 및 구성개념을 활용하여 관광분야에서 활용 가능한 개성의 요인을 도출하기 위해 연구들이 이루어졌다. 이러한 연구들은 관광지에 대한 여행객의 평가와 통찰을 정서적인 차원에서 규명하기 위한 단초를 마련하였을 뿐만 아니라 브랜드 개성의 개념을 관광지 개성으로 확장하는 이론적 성과에 기여하였다(Murphy, Benckendorff & Moscardo, 2007; Sahin & Baloglu, 2011).

더욱이 최근에는 관광지 개성에 대하여 관광분야에 적합한 척도개발을 통하여 여행객의 행동의도 등과의 세부적인 영향관계까지 밝히는 이론적, 학술적 성과를 제고하였지만(Chen & Phou, 2013; Hultman, Skarmeas, Oghazi & Beheshti, 2015; Papadimitriou, Apostolopoulou & Kaplanidou, 2015), 〈표 1〉에서 정리된 바와 같이 Aaker(1997)가 제안한 브랜드 개성의 구성개념을 선택적으로 활용하였다는 점에서 본원적으로 제품주의 개념적 틀을 크게

〈표 1〉 관광지 개성의 측정

연구자	구성개념의 활용	측정 및 연구내용
Murphy *et al.* (2007)	Aaker(1997)의 BPS 27개 항목 활용	Aaker(1997)의 BPS 27개 항목 활용하여 진실성, 흥미로움, 유능함, 세련됨, 강인함의 5개 차원을 도출하였으며, 관광지 개성이 자아일치성과 관광지 이미지, 여행객 행동에 미치는 영향관계를 규명함
Sahin & Baloglu(2011)	Aaker(1997)의 BPS 20개 항목 활용	Aaker(1997)와 Hosany, Ekinci and Uysan(2006)이 제안한 28개의 BPS 항목을 활용하여 이스탄불의 개성과 이미지, 행동 의도간의 관계를 사회연결망분석, ANOVA 등을 통해 밝힘
Usakli & Baloglu(2011)	Aaker(1997)의 BPS 29개 항목 활용	Aaker(1997)의 BPS 29개 항목을 토대로 내용분석을 통해 활기참, 세련됨, 유능함, 현대적인(contemporary), 진실성의 5개 차원을 도출하였으며, 라스베가스에 대한 개성과 자아일치성, 행동의도 간의 관계를 규명함
Kim & Lehto(2013)	Aaker(1997)의 BPS 42개 항목 활용	Aaker(1997)의 BPS 42개 항목과 5가지 차원인 흥미로움, 진실성, 유능함, 세련됨, 강인함을 활용하여 내용분석을 통하여 아웃바운드 여행객이 인식하는 한국의 관광지 개성을 측정함
Papadimitriou *et al.*(2015)	Aaker(1997)의 BPS 42개 행목 활용	Aaker(1997)의 BPS 8개 항목을 활용하여 Excitement, Sincerity의 2개 차원을 도출하였으며, 구조방정식 모형을 통해 그리스 도시 관광지의 개성과 관광지 이미지, 추천의도 및 방문의도에 대한 영향관계를 밝힘

주 : 기존연구를 바탕으로 연구자 정리

벗어나지 못하여 유·무형자산이 결합된 관광지가 가지는 본질적 차이를 반영하지 못하였다는 한계점이 존재한다(Pan, Zhang, Gursoy & Lu, 2017; Usakli & Baloglu, 2011).

그럼에도 불구하고 선행연구에서는 관광지 개성의 개념이 자아일치성, 관광지 이미지, 행동의도 등과의 보다 세분화된 구조적 관계에서도 유의미한 영향관계를 가지는 주요한 변수임을 여러 방법론을 통해 실증적으로 밝혔으며, 관광지 개성은 특정 관광지가 경쟁적 우위를 가지기 위해 가장 선결되어야 하는 핵심변수로 그 기능을 담당하고 있다는 것을 증명하였다.

정리하면, 관광지 개성은 여행객이 특정 관광지에서의 경험과 인식에 기반하여 인간의 특성을 바탕으로 상징적, 자아 표현적인 긍정, 부정의 정서와 연관시킴으로써 관광지 평가의 다차원적 구조에서 가장 선행하는 개념으로 정의할 수 있다.

이러한 관광지 개성을 통해 여행객의 경험을 바탕으로 관광지를 하나의 브랜드로 인식할 수 있도록 형상화됨으로써 특정 관광지가 차별화되는 정체성을 확립할 수 있으며, 이는 관광수요를 유발하고 나아가 여행객과 관광지 간 지속적이고 장기적인 관계 구축의 단초 역할을 한다. 그러므로 관광지가 경쟁력을 가지고 하나의 장소로 여행객들의 인식에서 경쟁우위를 갖추기 위해서는 관광지 개성에 대한 심층적인 이해가 필요하며, 이러한 이해는 유·무형의 상품과 서비스를 제공하는 관광지의 내재적 본질을 반영한 구성개념의 탐색적인 고찰 및 측정이 수반되어야 가능해질 수 있다.

따라서 본 연구에서는 Aaker(1997)가 제안한 5가지 개성 차원을 바탕으로 온라인 여행 커뮤니티의 리뷰 데이터를 통하여 여행객의 관점에서 표현되고 있는 개성의 구성개념을 의미론적 차원에서 확장함으로써 관광의 맥락에서 개성의 구성개념을 이해하고자 하며, 관광지에 대하여 소비자의 내재적 의미가 반영된 개성의 차원과 소비자의 행동의도 간 구조적 관계를 규명하고자 한다.

2. 관광분야 온라인 리뷰의 양적·질적 데이터의 활용

온라인 리뷰는 소비재 시장에서 제품, 서비스 등과 관련된 정보나 이용후기 및 평점 등을 토대로 소비자의 만족에 대한 경험을 온라인 플랫폼이라는 매개체를 통하여 소비자들과 손쉽게 커뮤니케이션을 할 수 있도록 하는 정보의 교환 과정이다(박정환・이병철, 2014). 특히 관광분야에서는 관광정보 탐색, 여행 계획수립, 소비에 이르기까지의 일련의 관광행돌 의사결정단계에서 온라인 리뷰정보는 구전효과를 통해 수요창출의 원천으로 안착하였다(심영석・김홍범, 2016). 최근에는 웹상의 온라인 리뷰정보의 양적증가로 인하여 여행객들이 방문한 관광지에 대해 어떤 제약에도 얽매이지 않고 그들의 의견을 솔직담백하게 표현한 정보를 과학적 증거기반의 빅데이터 분석을 통해 산업의 전반적인 흐름을 파악하는 유용한 자료로서 활용되고 있다(Xiang, Gretzel & Fesenmaier, 2009).

이러한 현상은 소비보다 구매가 선결되는 관광분야의 특성상 온라인 리뷰를 받아들이는 수용자가 정보에 대한 신뢰성과 유용성을 인정하고 의사결정에 확신을 가짐으로써 소비로까지 이어진다(Schuckert *et al.*, 2015). Baek, Ahn and Choi(2012)는 잠재 고객들이 이미 소비재를 경험한 고객들이 게재한 온라인 리뷰를 바탕으로 구매에 대한 확신을 가지는 등 의사결정 과정에서 온라인 리뷰하게 된다는 연구결과 또한 이를 뒷받침한다. 뿐만 아니라 Gretzel, Yoo and Purifoy(2007)는 온라인 리뷰 데이터의 파급력과 역할에 대한 연구에서 여행계획 단계에서 잠재 여행객이 가장 선호하고 영향력이 큰 행

동은 타인이 게재한 여행 후기나 블로그를 읽는 것으로 나타났다.

관광분야에서 온라인 리뷰 데이터를 활용한 연구를 살펴보면, 해외의 경우 온라인 리뷰상의 정량적 평가인 여행객 평점, 국내의 경우는 정성적 평가에 해당하는 글 단위를 주로 분석단위로 활용하고 있다. O'Connor(2010)는 영국의 1,042개소의 호텔에서 100개 호텔을 표본으로 하여 여행객 평점, 리뷰, 호텔의 가격정보를 활용함으로써 호텔의 평판을 관리하는데 사용자 제작 콘텐츠(UGC, user-generated contents)가 유의미한 긍정적 역할을 한다는 것을 규명하였다. 또한, Lee, Law and Murphy(2011)는 글로벌 온라인 여행 플랫폼인 트립어드바이저의 리뷰 데이터를 활용하여 리뷰어의 행태를 파악하였으며, 잠재여행객에게 도움이 되는 리뷰어는 활발한 리뷰 게재와 더불어 관광지에 대하여 낮은 평점을 부여하는 특성이 있다는 것을 밝혔다. 또한, Liu, Schuckert and Law(2015)은 홍콩의 185개 호텔에 대한 여행객 평점을 수집하여 호텔단위의 평점과 개별 소비자가 남긴 평점 간의 차이를 분석하였으며, 낮은 등급의 호텔이 전체 평점과 개인 평점 간 유의한 차이가 있는 것으로 나타났다. 국내의 경우, 오익근 외(2015)는 포털 사이트에 게재된 정보 가운데 관광/여행관련 키워드가 포함된 웹 페이지의 리뷰만을 수집하여 사회연결망 분석을 통해 한국관광에 대한 인식을 실증 분석하였다. 박득희・김태구・이계희(2016)는 서울과 제주도의 공식 소셜미디어인 페이스북 페이지에서 리뷰를 수집하여 사회연결망 분석을 수행하였으며, 소셜미디에서의 정보흐름의 확산에 대하여 탐색적으로 조사하였다. 또한, 박수지 외(2017)는 네이버 연관키워드조회를 통해 관광관련 키워드를 선정하여 리뷰데이터를 수집한 후 텍스트 마이닝 분석을 통하여 키워드를 도출한 후, 관광정보지식시스템의 관광객 수 추이와 유의성 판단을 통한 수요예측을 조사하였다.

이처럼 관광분야에서 온라인 리뷰를 다각적으로 활용함으로써 방법론적 대안을 제시하였지만, 온라인 리뷰는 관광지에 대하여 글로 표현된 질적 데이터와 평점으로 표현된 양적 데이터가 동시에 존재하기 때문에 이를 연계한 방법론적 접근이 필요하다는 것을 알 수 있다. 더불어, 여행객들의 심리를 보다 심층적으로 이해하기 위해 텍스트 분석에 있어서 현재까지 이루어졌던 단어를 토대로 한 빈도 기반의 키워드 분석이 아닌 문맥을 고려하여 도출된 단어들을 활용할 필요성 또한 제기된다. 실제로 빈도기반의 분석기법들은 연구주제에 부합하는 키워드를 직관적으로 이해하는데 기여하였지만, 최근 딥 러닝 기법을 통한 신경망 언어 모델을 활용하여 의미론적 차원에서 텍스트를 분석하는 것이 가능해졌으므로 이에 대한 활용방안도 제고해 볼 필요성이 있다.

3. Word2Vec의 개념 및 특징

Word2Vec(word embedding to vector)은 구글(google)의 연구원인 Mikolov, Chen, Corrado and Dean(2013)이 제안한 딥 러닝(deep learning)방법을 이용한 단어 임베딩 학습 모델로써 NNLM(neural network language model)에 기반을 둔 방식을 통하여 단어들의 의미를 특정 차원의 벡터공간 모델에서 값으로 계산, 표현하는 학습 기법이다. 특히 벡터공간 모델(vector space model)은 의미상 유사한 단어들은 가까운 지점으로 매핑(mapping)되어지는 연속된 벡터 공간들에서 단어들로 표현되는 것이며, 이러한 단어를 벡터로 표현하는 과정을 단어 임베딩(word embedding)이라고 한다(Collobert *et al.*, 2011; Turian, Ratinov & Bengio, 2010). 이러한 Word2Vec은 기본적으로 비슷한 분포를 가진 단어들은 비슷한 의미를 가진다는 언어학의 '분산 가설(distributional hypothesis)'

을 바탕으로 하기 때문에(Harris, 1954) 단어 간의 관계 속에서 각 단어의 출현 빈도를 고려하여 단어마다 유의미한 벡터 값을 부여하게 된다. 다시 말해 Word2Vec은 입력된 학습문장 내에서 단어의 분포가 가까운 단어 간 산출되는 벡터 값의 결과는 유사해지며, 각 단어를 계량화된 숫자로 나타내는 개념이라 할 수 있다(Mikolove *et al.*, 2013).

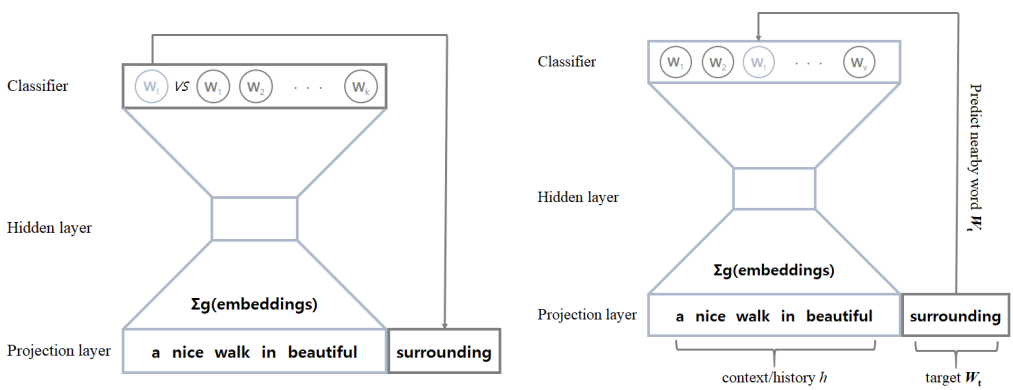
이러한 Word2Vec은 심층 신경망은 아니지만 텍스트를 처리하는 두 개의 층으로 구성된 신경망 모델로써 〈그림 1〉과 같이 CBOW(continuous bag of words)와 skip-gram이라는 두 가지 학습 모델이 존재한다. CBOW는 문서(document)를 자동으로 분류하기 위한 BOW(bag of words) 기법이 발전한 형태로, 문맥(context)에 포함된 단어들의 분포를 통하여 특정 단어를 예측하는 모델이며, skip-gram은 주어진 단어 하나를 바탕으로 주변에 등장하는 단어들을 예측하는 모델이다. 보다 직관적으로 CBOW 모델은 “a nice walk in beautiful”이라는 문맥이 주어졌을 때 “surrouding”을 예측하는 방식이며, skip-gram은 “surrounding”이라는 단어가 주어졌을 때 “a nice walk in beautiful”을 예측하는 방식이다.

Word2Vec 방법 외에도 SVM(support vector machine), 연관성 분석(association rules) 등과 같이 텍스트 데이터를 통해 연관 단어를 분류, 유추하는 데이터 마이닝 기법이 존재하지만, Word2Vec은 이러한 방식에서 좀 더 발전한 형태로 단어의 의미를 단어 그 자체가 아닌 의미를 벡터 공간에 표현함으로써 복잡한 개념뿐만 아니라 나아가 유사 또는 이질적인 단어까지 수치화된 벡터 값을 통하여 구현 가능하게 하였다(Rong, 2014).

Ⅲ. 연구방법

1. 연구의 절차 및 체계

여행객들이 관광지를 평가하는 개성의 구성개념 확장 및 평점(rating score)과의 관계를 실증분석하기 위해 온라인 여행 플랫폼의 리뷰 데이터를 수집하여 자연어 처리 신경망 모델(NNLM)인 Word2Vec 모델을 바탕으로 관광지 개성의 차원을 바탕으로 유사단어 분류를 통한 구성개념을 확장한 후, 정량 데이터에 해당하

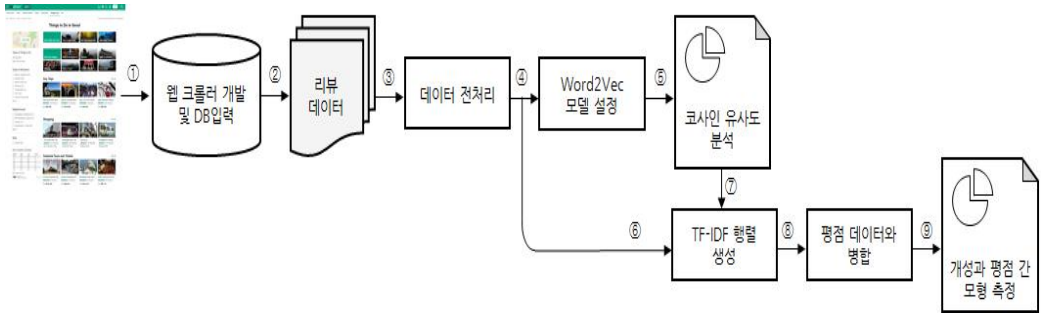


〈그림 1〉 Word2Vec의 CBOW모델과 Skip-gram모델의 구조

는 평점과의 인과모형을 추정하였다. 세부적으로 웹 크롤러를 개발하여 자료를 수집한 후 데이터베이스(DB)에 저장하였으며, DB에 입력된 리뷰 데이터를 추출하여 텍스트 전처리 및 형태소 분석을 수행하였다. 이후 텍스트 데이터의 가중치 산출법 중 하나인 TF-IDF를 적용하여 가중치를 산출하였으며, 비구조화된 텍스트 데이터를 분석이 가능하도록 구조화된 형태인 문서-단어 행렬(document-term matrix)로 변환하였다. 또한, 최초 형태소 분석을 통하여 전처리된 텍스트를 바탕으로 Word2Vec 모델을 설정하여 텍스트의 벡터화(vectorization)를 통한 단어 임베딩과정을 거침으로써 텍스트 각각의 벡터 값을 산출하였으며, 텍스트의 벡터 값을 토대로 관광지 개성의 5가지 차원에 대한 각 항목별 코사인 유사도 분석을 통해 관광지 개성차원별 10개의 확장된 구성개념을 도출하였다. 관광지 개성의 각 차원별로 확장된 구성개념을 최초 5가지 개성 차원으로 축소하였으며, 이러한 일련의 과정을 거쳐 재생성된 5가지 관광지 개성은 여행객이 관광지에 대하여 평가하는 내재적 의미를 포함하는 차원으로서 관광분야의 맥락에 부합하는 기능적 차원이라 할 수 있다. 최종적으로, 재생성된 5가지 개성 차원을 여행객 평점과 연계하여 회귀모형을 추정하여 관광지 개성과 여행객 평점 간 영향관계를 파악하였으며, 세부적인 연구 절차는 자료수집, 텍스트 마이닝을 통한 리뷰 데이터의 전처리, Word2Vec 및 모형 측정 등 총 세 단계의 체계로 〈그림 2〉와 같이 진행하였다.

첫 번째, 자료 수집 단계에서는 온라인 여행 플랫폼 중 '트립어드바이저(tripadvisor.com)'에서 2004년 7월10일부터 2017년 2월28일까지 누적된 여행객 리뷰를 웹 크롤러를 통해 수집하여 데이터베이스에 저장하였다. 이 과정에서는 리뷰 데이터의 분석 이후 평점 데이터와의 데이터 연계를 위해 필요한 리뷰별 식별번호를 부여하였다[①,②]

두 번째, 텍스트 마이닝을 통한 리뷰 데이터의 전처리 단계에서는 데이터베이스에 저장된 리뷰 데이터에 대하여 기능어에 해당하는 관사, 전치사, 조사, 접속사 등과 불용어, 구두점, 숫자 등을 우선적으로 제거하였으며 텍스트를 형태소 분석이 가능한 단위로 분리하기 위한 방법인 토큰화(tokenization)과정을 거쳤다. 특히, 이 단계에서는 토큰화된 텍스트를 토대로 어간추출(stemming words)을 수행함으로써 텍스트 처리의 효율성을 높이도록 하였다. 어간추출은 예컨대, exiting, excited와 같이 단어의 어근을 추출하여 접사를 제거하는 것으로 동일한 의미의



〈그림 2〉 연구의 세부 절차

다른 형태의 단어를 하나의 단어로 처리하는 방법이다(한지연・김홍범, 2017)〔③〕. 이와 같이 자연어 처리된 텍스트 데이터에 대하여 텍스트 가중치 산출법인 TF-IDF 기법을 적용하여 해당 텍스트의 가중치를 산출하였다. TF-IDF는 하나의 문서에서 특정 단어의 중요도는 그 문서 내의 출현 빈도에 비례하고, 전체 문서에서 나타난 단어의 출현 빈도와는 반비례하도록 가중치를 산출하는 방법으로 개별 문서에서 단어의 중요도를 표현할 수 있는 기법이다(심영석・김홍범, 2016; 이태원・홍태호, 2015)〔⑥〕. 최종적으로, 가중치가 부여된 텍스트 데이터는 비구조화된 형태이기 때문에 이후 분석에 적합하도록 구조화된 자료 형태인 문서-단어 행렬로 변환하였으며, 각 리뷰별로 할당한 식별번호를 토대로 리뷰데이터와 연계작업을 수행하였다〔⑧〕.

세 번째, 전처리 과정〔③〕을 거친 데이터에 대하여 관광지 개성에 해당하는 5가지 차원에 대하여 여행객의 관점에서 표현되는 개성의 차원으로 구성개념을 확장하기 위해 자연어를 처리하는 신경망 언어 모델(NNLM)인 Word2Vec 모델 중 Skip-gram 모델을 오픈소스 딥러닝 프레임워크인 텐서플로우 백엔드(TensorFlow-backend) 기반의 케라스(Keras) 라이브러리를 통해 구현한 후, 코사인 유사도 분석을 수행하였다. Word2Vec의 skip-gram 모델은 학습된 문장(training set)에서 특정 단어를 추출하여 주변 단어들과 중심단어 간의 거리를 계산하여 학습(learning)을 진행하고, 단어 간의 의미를 파악하여 다차원의 벡터 공간 내에서 단어들을 효율적으로 표현하는 자연어처리 머신 러닝 기법이다(전근식・공성언・최용석, 2017)〔⑤〕. 따라서 Skip-gram 모델에서 학습된 단어들은 벡터 공간 내에서 벡터 값을 기반으로 문장의 앞, 뒤를 고려하여 문맥상 유사한 단어끼리 위치하게 되고, 이러한 벡터 값을 토대로 코사인 유사도 분석을 수행함으로써 관광지 개성의 5가지 차원과 의미적으로 유사한 단어(구성개념) 추출이 가능하다. 구체적으로, 코사인 유사도 분석 단계에서는 관광지 개성에 해당하는 5가지 차원(sincerity, excitement, competence, sophistication, ruggedness)에 대해 각 차원별로 10개의 유사 단어를 의미론적인 접근법에 따라 분류하여 관광지 개성의 세부 구성개념(components)을 도출 및 확장하였으며, 기존의 브랜드 개성 측정항목과 비교하였다〔⑦〕. 이를 토대로 〔⑧〕과정을 통하여 생성된 TF-IDF 문서-단어 행렬에서 최초 관광지 개성에 해당하는 5가지 차원을 재생성한 후 여행객 평점과 영향관계를 파악하였으며〔⑨〕, 연구모형은 식(1)과 같다. 연구모형의 관광지 개성 차원은 유사도 분석에서 각 차원별로 도출된 10개의 구성개념의 평균값을 이용하여 하나의 차원으로 축소한 것으로 해당 변수는 의미론적 관점에서 관광의 맥락에 부합하도록 내재적 의미를 반영한 개성의 차원이라 할 수 있다. 해당 과정을 거쳐 재생성된 관광지 개성의 차원은 온라인 리뷰의 시계열 자료에 해당하므로 자기상관성 검증을 위한 Durbin-Watson test와 다중공선성 진단을 실시하였다. 또한, 해당 설명변수는 질적 데이터인 리뷰 데이터를 수치화하였기 때문에 설명변수와 종속변수 간 분산의 분포를 파악하기 위한 등분산성(homoscedasticity) 검정인 Breusch-Pagan Test를 수행하였으며, 최종적으로 최소제곱법(OLS)과 가중최소제곱법(WLS)을 통하여 추정된 모형을 비교하였다.

$$\begin{aligned} RATING_i = {} & \beta_0+\beta_1 SINCERITY_i+ \\ & \beta_2 EXCITEMENT_i+ \\ & \beta_3 COMPETENCE_i + \\ & \beta_4 SOPHISTICATION_i+ \\ & \beta_5 RUGGEDNESS_i+\varepsilon_i \end{aligned} \quad (1)$$

2. 조사 설계

1) 변수 측정항목의 구성

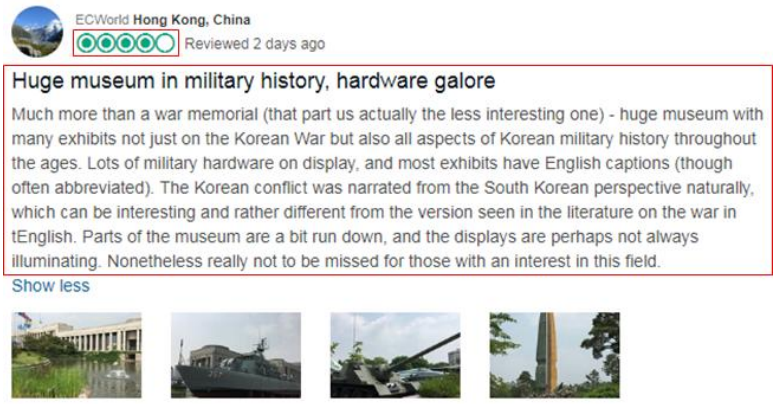
온라인 리뷰 데이터를 통한 키워드 추출 과정에서는 특정 단어를 문맥에 대치시키기 위해 단어가 가지는 개념의 범주를 선정하는 것이 필수적이다(Dickinger & Koltringer, 2012). 그동안 관광지 개성과 관련한 연구 중 개성의 차원을 통해 여행객의 행동의도, 만족도 등과의 영향관계를 규명하거나 구조적 관계를 검증한 연구는 Hosnay, Ekinci and Uysal(2006), Hultman *et al.*(2015), Papadimitriou *et al.*(2015) 등이 있다. 최근에는 제품범주의 브랜드 개성의 차원을 관광지를 평가하는데 더욱 적합하도록 척도개발 등의 노력 또한 이루어지고 있다(Kumar & Nayak, 2018; Pan *et al.*, 2017). 이러한 학문적 성과를 통해 관광산업에 적합한 '공손함(courteousness)', '활기참(vibrancy)', '여성스러움(femininity)' 등의 개성 차원들이 도출되고 있지만, 앞선 문헌연구에서도 살펴본 바와 같이 그 근간에는 Aaker(1997)가 제시한 5가지 차원을 단일항목으로 측정하거나 이에 상응하는 15개의 속성, 42개의 측정항목이 선택적으로 활용되고 있다는 것을 알 수 있었다.

정리하면, 관광지 개성 측정을 위해서도 Aaker(1997)가 제시한 개념적 틀은 충분히 활용이 되어야 하지만 제품범주의 브랜드 개성을 평가하는 세부속성 및 측정항목을 활용하는 것은 관광지의 특성을 반영한 개성측정에는 이론적, 학술적 격차가 존재한다.

따라서 본 연구에서는 온라인 여행 커뮤니티에서 소비자들이 여행 후 관광지에 대한 평가와 경험을 공유한 리뷰 데이터를 활용하여 Aaker(1997)가 제시한 5가지 차원을 토대로 유사도 분석을 수행함으로써 의미론적 관점에서 관광분야의 맥락에 부합하는 세부 구성개념을 도출하였다. 더불어 도출된 구성개념을 기존 Aaker(1997)가 제시한 5가지 차원으로 축소하였으며, 이러한 일련의 과정을 거쳐 재생성된 5가지 차원은 관광지에 대하여 여행객의 경험과 평가가 내재된 변수이므로 현재까지 관광분야에서 활용되었던 개성의 차원보다 더욱 심층적이고 구체적인 의미를 포함하는 변수라 할 수 있다.

2) 표본 및 자료수집 방법

본 연구의 자료수집대상은 '트립어드바이저(tripadvisor.com)'로 설정하였으며, 서울의 관광지 및 관광 상품에 대한 온라인 리뷰를 조사·분석에 활용하였다. 최근 웹상의 블로그, SNS, 온라인여행 커뮤니티 등의 플랫폼이 확대됨에 따라 여행객들이 관광지와 관광지내의 제반활동 등의 경험이나 기대, 만족/불만족 등의 정보를 실시간으로 접근할 수 있게 되었으며(Llodrà-Riera, Martínez-Ruiz, Jiménez-Zarco & Izquierdo-Yusta, 2015), 이러한 온라인구전(eWOM) 정보가 소비자 행동에 영향을 미친다는 연구결과가 계속적으로 보고되고 있다(Herrero, San Martín & Hernández, 2015; Eriksson & Fagerstrøm, 2017). 이에, 학계에서는 온라인 관광정보에 내재된 유의미한 정보를 추출하고 탐색, 분석할 수 있는 다양한 방법들이 논의되고 있으며(Lee *et al.*, 2011; O'conner, 2010; Schuckert *et al.*, 2015), 특히 관광정보를 공유하는 웹 플랫폼 중 글로벌 여행 커뮤니티인 '트립어드바이저'가 주는 정보 그 자체의 유용성 및 신뢰성에 대한 연구(Ayeh, Au & Law, 2013)와 실제 온라인 리뷰 데이터를 활용한 연구들이 다각적으로 이루어지고 있다(Chang, Ku & Chen, 2017; Schuckert, Liu & Law, 2016). 이와 같이 온라인 리뷰가 활용가능성이 높은 이유는 〈그림 3〉과 같이 여행객이 특정 관광지를 경험한 후의 정량적인 평가와 정성적인 평가가 동시에 공유되므로 여행객들이 경험한 관광지에 대한 직접적인 인식과 감



〈그림 3〉 트립어드바이저의 온라인 리뷰 예

정 등의 정보를 획득할 수 있기 때문이다.

따라서 자료 수집을 위해 웹 크롤러를 개발하여 서울의 관광지 및 관광상품에 대한 여행객 리뷰를 수집하였다. 최종적으로 수집된 양은 데이터는 최초 리뷰작성일자인 2004년 7월10일부터 2017년 2월28일까지 누적된 36,969개의 리뷰이며, 해당 리뷰 전체를 분석단위로 활용하였다. 웹 크롤러(web crawler)는 빅데이터 분석의 자료수집 단계에서 방대한 양의 웹 문서를 자동으로 수집가능하게 하는 기법이다(김광영 외, 2011). 웹 크롤러는 수집 대상이 되는 웹 페이지의 네트워크 전송방법, url의 구조 등을 사전에 분석하여 해당 프레임에 맞도록 설계함으로써 해당 웹 페이지에 웹 크롤러가 자동적으로 접근하여 사용자가 원하는 데이터를 정확하게 수집할 수 있다(Kang, Yoo & Han, 2009). 본 연구에서 활용된 웹 크롤러는 통합 개발 환경(IDE) 플랫폼인 이클립스(eclipse)에서 자바(JAVA)언어를 기반으로 트립어드바이저(tripadvisor.com) 웹 페이지의 HTML Source를 다운로드한 후 태그(tag)분석을 통하여 트립어드바이저상에서 분류하고 있는 관광지 및 관광상품에 대한 12개 카테고리(명소/랜드마크, 박물관, 자연/공원, 쇼핑, 나이트라이프, 야외활동, 투어, 콘서트/쇼, 음식/음료, 위락시설, 스파/웰니스, 여행자리소스)의 117개소에 작성된 리뷰, 리뷰작성일자, 평점 등 특정 데이터만 추출하는 파싱(parsing)과정을 거치는 구조로 설계 및 구현하였으며, 세부적인 목록은 〈표 2〉와 같이 정리된다.

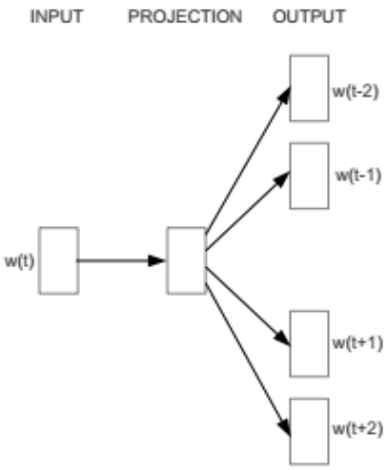
신경망 모델링 과정에서는 데이터의 학습량이 정확한 결과물(output) 출력에 중요한 영향을 미치므로 가능한 다량의 데이터를 확보할 수 있도록 설계할 필요가 있다(이슬기・정성관・이우성・박경훈, 2011). 일반적으로 자료의 시간적 범위가 넓을 경우 가능한 외생변수의 영향을 고려해야 하지만, 본 연구에서 활용된 Word2Vec의 skip-gram 모델 또한 신경망 언어 모델로서, 학습문장 내 단어의 분포가 가까운 단어 간의 벡터 값을 산출하는 '분산 가설'이 바탕이 되므로 시기적으로 나타날 수 있는 특정 표현들 간 빈출 현상을 다량의 데이터 학습을 통하여 통제할 수 있다는 장점이 있다.

따라서 본 연구에서는 이와 같은 점을 고려하여 해당기간의 데이터를 모두 분석단위로 활용하였으며, 특정기간 내 사회적, 기후적, 정치적 변화 등 외생적인 요인에 의하여 탄력적인 관광수요에서 나타날 수 있는 여행객의 표현을 다량의

〈표 2〉 온라인 자료 수집 대상지(서울의 관광지 및 관광상품) 목록

구분	관광지 및 관광상품		
명소/랜드마크	경복궁	창덕궁	덕수궁
	남산N서울타워	동대문디자인플라자	홍익앞거리
	명동	인사동	북촌
	봉은사	조계사	-
박물관/전시관	한국전쟁기념관	국립중앙박물관	국립민속박물관
	서울역사박물관	서대문형무소역사관	-
	삼성미술관 리움	국립현대미술관서울관	-
	트릭아이 박물관	한가람박물관	-
자연/공원	북한산국립공원	동대문역사문화공원	남산공원
	한강공원	서울숲	청계천
	여의도한강공원	여의도공원	-
	낙산공원	서울어린이대공원	-
쇼핑/마켓	남대문시장	타임스퀘어몰	강남지하상가
	스타필드코엑스몰	홍대프리마켓	이태원
	노량진시장	동대문쇼핑몰	-
	신세계백화점본점/면세점	롯데백화점본점/면세점	-
나이트라이프	올댓재즈	바우드스탁	서울곤와일드
	르챔버	오케이퐁퐁	믹스&몰트
	클럽옥타곤	리솔베르시가샵&시가라운지	-
	마익스캐빈	골목바이닐&펍	-
야외활동	올림픽공원	이랜드크루즈	메가박스 코엑스몰
	노이스케이프	잠실야구장	명동사격장
	서울에스케이프룸홍대점	원데이코리아	-
	아띠인력거	스윗트래블코리아	-
투어	서울시티투어	투어스바이아론	DMZ스파이투어
	얼티미트코리아투어	코스모진투어	탑코리아투어
	컬러오브코리아	노바랜드투어스코리아	-
	익스클루시브투어코리아	코리도어	-
콘서트/쇼	명동난타극장	세종문화회관	점프!코믹 무술 퍼포먼스
	홍대난타극장	오페라하우스자유소극장	클라이브
	충정난타극장	예술의전당	-
	정동극장	드럼캣	-
음식/음료	온고코리아컬리널리	디스이즈코리아	서울푸드투어
	오미요리연구소	온고푸드커뮤니케이션즈	히어코리아
	젠김치코리안푸드	딜렉터블트래블스	-
	쟈넷쿠킹스튜디오	HaB코리아	-
위락시설	롯데월드	디큐브시티	-
	테지움서울	코엑스아쿠아리움	-
	롯데월드스타에비뉴	롯데월드아쿠아리움	-
	63시월드	디보빌리지	-
스파/웰니스	스파렉스	해피데이스파	실로암스파
	드래곤힐	더스파그랜드하얏트서울	캣카페고양이놀이터
	드래곤힐스파&리조트	스파레이	-
	SK-Ⅱ부띠끄스파	레비쉬스파	-
여행자 리소스	한국관광공사관광안내소	서울글로벌문화정보센터	서울도서관
	명동관광정보센터	공항철도트래블센터	-
	인사동홍보관광안내소	K-Style Hub	-
	한국지역정보센터	코엑스센터	-

주: 관광지 및 관광상품에 대한 12개의 기준은 트립어드바이저(tripadvisor.com)상 분류되어 있는 체계임



〈그림 4〉 Word2Vec의 Skip-gram모델의 구조[1)]

데이터의 학습시킴으로써 보다 안정화된 결과물(output)을 산출하도록 설계하였다.

3) Word2Vec의 Skip-gram 모델

Word2Vec은 모델에 입력된 학습 문장(training set)에서 단어를 추출하여 맥락(context) 속에서 주변 단어들과 중심단어 간 벡터 공간의 거리 값을 학습함으로써 단어 간의 의미를 파악하여 다차원의 벡터 공간 내에 단어들을 효율적으로 표현하는 자연어처리(NLP, natural language processing) 머신 러닝 방식이다. 이러한 Word2Vec모델은 앞서 설명한 바와 같이 주변 단어들로부터 중심단어를 예측하는 CBOW와 주어진 중심단어를 통해 주위에 등장하는 유사 단어를 예측하는 skip-gram이라는 두 가지 방식의 학습 모델이 존재한다(Mikolove *et al.*, 2013).

본 연구가 관광지 개성에 대한 5가지 차원(중심단어)을 토대로 유사 단어를 확장하는 과정이 필요하므로 Word2Vec의 두 가지 학습 모델 중 skip-gram 모델을 활용하고자 한다.

Mikolove *et al.*(2013)이 제안한 skip-gram 모델은 〈그림 4〉와 같이 입력층(input layer), 투사층(projection layer), 출력층(ouput layer)로 구성된 신경망 아키텍처를 취하고 있다. 〈그림 4〉의 Skip-gram모델은 $W(t-2)$, $W(t-1)$, $W(t)$, $W(t+1)$, $W(t+2)$라는 단어집합을 학습대상(training set)으로 하여 $W(t)$라는 단어를 통해 $W(t-2)$, $W(t-1)$, $W(t+1)$, $W(t+2)$를 예측하는 모델을 학습하는 예이며, $W(t)$를 입력층(input)으로 입력층의 일정 범위 앞인 $W(t-2)$, $W(t-1)$와 입력층의 일정 범위 뒤인 $W(t+1)$, $W(t+2)$을 출력층(output)으로 학습한다.

skip-gram 모델을 수식으로 표현하면 식(2)와 같으며, 학습단어 t의 단어 w_t와 그 주변의 w_{t+j}번째 단어가 등장하는 조건부 확률을 최대화 하는 것을 목적으로 한다. 식(2)에서 T는 $w1, w2, w3, ..., w_T$로 표현되는 각 단어의 위치

1) 출처. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

이며, c는 현재 위치 단어인 w_t 주변에 등장할 것으로 예측하는 단어의 범위가 된다. c가 클수록 더 많은 단어가 현재 위치의 단어와 유사한 단어로 정의되며, $\frac{1}{T}$는 전체 단어에 대하여 위치의 수로 나누는 정규화 항이다. 따라서 θ는 식(2)에 의해서 최적화되는 모든 변수를 의미하게 되며, 확률 값을 최대화할 수 있도록 하는 각각의 단어 벡터 값들이 된다.

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c \le j \le c, j \neq 0} \log p(w_{t+j}|w_t) \quad (2)$$

일반적으로 Skip-gram은 현재 위치 t의 단어 w_t와 그 주변에 w_{t+j}가 나타나는 확률인 $p(w_{t+j}|w_t)$를 정의하기 위해 식(3)과 같이 소프트맥스 함수(softmax function)를 사용하여 0과 1사이의 확률 값으로 나타낸다. 소프트맥스는 입력 값을 0과1사이의 값으로 정규화하며, 출력 값들의 총합이 항상 1이 되는 특성을 가진 함수이다.

$$p(w_o|w_I) = \frac{\exp(v'_{w_o} \top v_{w_I})}{\sum_{w=1}^{W}\exp(v'_{w_w} \top v_{w_I})} \quad (3)$$

식(3)에서 W는 중복을 제거한 전체 단어를 저장한 단어장(vocabulary)내의 단어의 수에 해당한다. w_I는 현재 위치 t의 단어 w_t의 고유한 단어 인덱스에 해당하는 입력값(input)이며, 따라서 v_w는 단어 w의 입력 벡터 표현의 행렬이 된다. 또한, w_o는 주변 위치 단어인 w_{t+j}의 고유한 단어 인덱스에 해당하는 출력값(output)이며, 따라서 v'_w는 단어 w의 출력 벡터 표현의 행렬이 된다. Skip-gram 모델은 이러한 v와 v'라는 두 개의 단어 벡터 행렬 값을 조정하며, 이 두 행렬은 서로 다른 값을 가지고 한 개의 단어는 두 개의 벡터 표현을 가지게 된다. 이 때, 서로 다른 행렬의 두 단어 벡터인 v'_{w_o}와 v_{w_I} 거리가 확률을 결정하게 된다(Rong, 2014).

이에 따라 Word2Vec의 skip-gram 모델에서는 학습대상 내에서 주위 단어의 거리가 벡터 공간에서 멀어지거나 가까워짐에 따라 벡터 공간에서 0과 1사이의 값을 가지게 되며, 반복학습을 통하여 유사 단어의 군집화가 형성됨으로써 산출되는 단어의 값이 비슷한 단어 간에는 의미가 유사한 것으로 간주하게 된다(Mikolove *et al.*, 2013).

4) 코사인 유사도(cosine similarity)

두 벡터 간의 유사도를 측정하기 위한 대표적인 방법에는 유클리디안 거리, 확장 자카드 유사도, 코사인 유사도 등이 있다(권영빈・이승도・양현・주요한, 2011). 이 가운데 가장 빈번하게 사용되는 거리 측정 방법은 유클리디안 거리 측정 방법으로 두 점 사이의 거리를 측정하여 유사도를 파악하는데 활용된다. 확장 자카드 유사도 측정 방법은 두 표본 집합사이의 유사도를 비교하기 위해 사용되는 측정방법으로, 두 표본의 교집합의 수를 두 표본의 합집합의 수로 나눔으로써 유사도를 측정하는 기법이다. 코사인 유사도 측정 방법은 두 벡터 간의 코사인 각도(θ)를 측정함으로써 두 벡터가 얼마나 유사한지를 판단하는 측정법 중 하나이다. 이 측정법은 두 벡터가 같은 방향을 향하고 있는지를 측정함으로써 유사도의 유무를 판단하는 것으로, 텍스트 문서를 효율적으로 분류하기 위하여 주로 활용된다(Shum, Dehak, Dehak & Glass, 2010). 이와 같은 코사인 유사도 측정법은 식(4)와 같이 나타낼 수 있다.

$$Similarity(A,B) = \cos(\theta) = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{n=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

이를 단어 간 유사도의 범주에서 해석해 보면, 두 단어 벡터의 A와 B 간의 각도를 측정하여 코사인 유사도가 0이 되면 A와 B사이의 각도는 90도가 되면서 두 단어는 아무런 유사도가 없어지고, 두 벡터가 1에 가까워질수록 A와B는 일직선상에 위치하게 됨으로써 두 단어의 유사도는 높아진다고 할 수 있다. 이러한 유사도 측정법은 두 객체가 얼마나 유사한가에 대한 척도를 수치적으로 나타내는 것이며, 본 연구에서는 관광지 개성의 5가지 차원에 대하여 Word2Vec의 skip-gram 모델을 통해 생성된 벡터를 활용하여 코사인 유사도 분석을 수행함으로써 관광지 개성의 각 차원별 유사한 단어를 확장하고자 한다.

5) TF-IDF 가중치 산출

앞서 Word2Vec의 skip-gram 모델 기반의 코사인 유사도 분석을 통해 도출된 관광지 개성의 확장된 키워드를 토대로 여행객 평점 간 영향관계 규명을 위해 최초 텍스트 데이터의 가중치 산출 기법인 TF-IDF를 적용하였다. TF-IDF (term frequency-inverse document frequency)는 문서 내 단어의 중요도를 계량적으로 측정하기 위하여 활용되는 방법으로 행(document)과 열(term)의 수준에서 분석단위가 이루어지며, 산출하는 방법은 식(5)와 같이 정리된다(Armstrong, Freitag, Joachims & Mitchell, 1995; Salton & Buckley, 1988).

$$TFIDF(w,d) = TF(w,d) \times \log\left(\frac{N}{DF(w)}\right) \quad (5)$$

$TF(w,d)$: 문서 d에 단어 w가 나타나는 횟수
$DF(w)$: 단어 w가 들어가는 문서의 총 수
N : 전체 문서의 총 수

식(5)는 문서 d에서 단어 w에 대한 가중치 값을 산출하는 것으로, 여러 문서에서 특정 단어가 문서 내에서 얼마나 중요한지를 통계적인 기법으로 나타낼 수 있다(Manning, Raghavan & Schütze, 2008). TF(term frequency)는 단어의 출현 횟수인 빈도를 의미하며, IDF (inverse document frequency)는 단어가 출현한 문서 수의 역수를 의미한다. IDF는 식(5)에서 나타난 바와 같이, 전체 문서수를 해당 단어를 포함한 문서의 수를 나눈 후 로그를 취함으로써 계산된다. 로그를 취한 이유는 IDF의 값이 커지는 것을 보정하기 위한 방법으로(Han, Kamber & Pei, 2006), 산식을 통하여 계산된 TF-IDF 값은 단어가 특정 문서 내에서 빈도 수가 높고 전체 문서 중 해당 단어가 포함된 문서가 적을수록 높아지게 되며, 이를 통해 단어의 중요도를 계량적으로 표현할 수 있다(심영석・김홍범, 2016; 이태원・홍태호, 2015).

예컨대, skip-gram 모델을 통하여 생성된 단어에 대한 벡터 값을 토대로 관광지 개성의 차원 중 'sincerity'와 관련된 유사단어를 코사인 유사도 분석을 통하여 도출한다. 이후, 최초 텍스트 데이터에서 TF-IDF 값을 토대로 한 행렬을 생성하여 유사도 분석을 통하여 도출된 유사단어가 포함된 문서를 추출한다. 다음으로, 'sincerity'와 관련된 유사단어의 문서들의 평균값을 이용하여 하나의 차원으로 축소한 후, 이 값을 바탕으로 평점 간 관계분석에 활용한다.

이를 정리하면, 관광지 개성의 5가지 차원

(sincerity, excitement, competence, sophistication, ruggedness)에 대하여 코사인 유사도 분석을 통해 각 10개의 키워드를 도출한 후 문서-단어 행렬에서 해당 키워드가 포함된 문서를 추출하여 TF-IDF 값의 평균을 바탕으로 하나의 차원으로 축소한다. 최종적으로, 차원 축소를 통해 재생성된 5가지 관광지 개성과 여행객 평점 간 영향관계를 규명하였으며, 모형은 식(1)과 같다.

Ⅳ. 분석결과

1. 단어 간 의미적 연관성을 고려한 코사인 유사도 분석

전용 웹 크롤러를 통해 수집된 서울의 관광지 및 관광상품의 2004년 7월10일부터 2017년 2월 28일까지 누적된 36,969개의 여행객 리뷰에 대하여 자연어처리(기능어, 구두점 및 공백, 숫자 제거)를 수행하였으며 형태소 분석 결과 40,251개의 단어가 자연어처리 되었다. 그리고 어간추출(stemming words)을 수행하여 동일한 의미지만 형태가 다른 단어를 하나의 단어로 처리하였으며 그 결과 33,976개의 단어로 축소되었다. 이후, 토큰화 과정을 거쳐 동사, 명사, 형용사를 추출하여 생성할 임베딩 벡터에 존재하게 될 전체 단어 수를 설정하였다. 그리고 토큰화 과정에서 공통 단어를 제외한 단어들에 대하여 단어장을 생성하였다.

이후, Word2Vec의 skip-gram 모델을 생성한 후 학습을 통하여 텍스트 임베딩 과정을 수행하였다. Word2Vec 학습 시, 인접 키워드의 범위(window)와 벡터공간의 차원에 대한 정의는 필수적이다(Zhang, Xu, Su & Xu, 2015). 여기서 window는 문맥에서 단어를 스캐닝하는 범위에 해당하며, 일반적으로 5~10의 값을 사용하며, 차원의 크기는 학습 데이터의 규모를 고려하여 50이상의 값을 사용한다(장환석・장은영・정광용, 2017).

본 논문에서는 〈표 3〉과 같이 워드에 임베딩할 벡터공간은 128, window size는 5로 설정하였으며, 최소 10회 이상 빈출된 단어만을 학습에 활용하였다. 또한, 학습의 반복 횟수에 해당하는 epoch을 5로 설정하였으며, 반복학습 횟수(epoch) 1에 대하여 데이터를 분할(split)하는 수치에 해당하는 step per epoch을 100으로 설정함으로써 1회 학습 시 보다 세부적인 단위에서 임베딩이 이루어질 수 있도록 하였다.

학습 후 생성된 단어 벡터를 활용하여 관광지 개성의 5가지 차원인 진실성(sincerity), 흥미로움(excitement), 유능함(competence), 세련됨(sophistication), 강인함(ruggedness)에 대한 각각의 코사인 유사도 분석을 수행하여 각 차원별 상위 10개에 해당하는 유사단어를 도출하였다. 〈표 4〉에서 나타난 바와 같이 도출된 유사단어와 기존 브랜드 개성의 구성개념을 비교해 볼 시, 관광지를 평가할 때의 개성관련 유사단어 중 강인함(ruggedness)을 평가하기 위한 항목인 강한(tough)을 제외한 모든 단어가 상이하다

〈표 3〉 Skip-gram의 단어 임베딩(word embedding) 정보

단어 임베딩 파라미터					단어 임베딩 정보	
embedding vector	window size	min count	step per epoch	epoch	token	vocabulary
128	5	10	100	5	5,294	3,694

〈표 4〉 코사인 유사도 분석을 통해 도출된 관광지 개성과 브랜드 개성 간 구성개념의 비교

구분	진실성 (sincerity)		흥미로움 (excitement)		유능함 (competence)		세련됨 (sophistication)		강인함 (ruggedness)	
Word2Vec Skip-gram 모델 기반 관광지 개성 구성개념[a]	참된 genuin	0.5794	재미 fun	0.7496	훌륭한 respectabl	0.4911	찬사 compliment	0.5160	강한 tough	0.6426
	친절한 kind	0.5156	즐거움 entertain	0.6325	유쾌한 jolli	0.4970	대화의 conversatio	0.4844	험준함 steepnes	0.5864
	파노라믹 panoramic	0.4971	혼잡함 crowd	0.5910	희생 sacrif	0.4530	차별화된 differenti	0.4740	도전적인 challeng	0.5453
	냉소적인 cynic	0.4738	활기 넘침 buzz	0.5606	완고한 diehard	0.4476	배치 가능한 deploy	0.4535	위험한 danger	0.5077
	참을성 patienc	0.4726	굉장한 awesom	0.5521	내구성 있는 durabl	0.4450	놀라운 remarkabl	0.4475	상향의 upward	0.4867
	열광적인 enthusiasti	0.4685	건전한 wholesom	0.5243	이타적인 selfless	0.4219	강화 reinforc	0.4453	혼미한 disori	0.4757
	빛나는 shini	0.4550	대단한 super	0.5034	눈부심 resplend	0.4137	정교한 exquisit	0.4365	고단함 exhaust	0.4733
	안전한 safer	0.4452	기대할만한 expectabl	0.4985	카리스마 charisma	0.4096	온전한 intact	0.4155	겁을 먹은 intimidat	0.4590
	행복한 happi	0.4336	열광적인 crazi	0.4954	대단한 extraordi	0.4058	솜씨 좋은 nifti	0.4154	격렬한 strenuous	0.4079
	열정 passion	0.4143	활기참 blast	0.4946	모험심 강한 adventur	0.4014	혁신적인 innovati	0.4093	미결정의 undecida	0.4067
브랜드 개성 구성개념[b]	정직한 honest		과감한 daring		성공적인 successful		매력적인 charming		거친 tough	
	다정한 cheerful		멋진 cool		믿음직한 reliable		여성스러운 feminine		남성적인 masculine	
	건전한 wholesome		상상력의 imaginative		지적인 intelligent		우아한 upperclass		외향적인 outdoorsy	
	친근한 friendly		독특한 unique		안전한 secure		화려한 glamorous		튼튼한 rugged	
	진실한 sincere		유행선도적인 trendy		전문적인 technical		부드러운 smooth		서구적인 western	
	본래의 original		젊은 young		선두의 leader		좋게 보이는 goodlooking			
	사실의 real		현대적인 contemporary		자신감 있는 confident					
	현실적인 down-to-earth		최신의 up-to-date		열심히 하는 hardworking					
	소박한 small-town		생기 있는 spirited							
	감상적인 sentimental									

주: a. 어간추출(stemming words)을 통하여 동일한 의미의 다른 형태의 단어를 하나의 단어로 처리함으로써 그 결과 값을 반환함.
b. Aaker(1997)가 제안한 브랜드 개성 차원별 구성개념임.

는 것을 알 수 있다. 특히, 기존 개성의 항목 중 진실성(sincerity)의 측정항목에 해당했던 건전한(wholesome)은 관광지 개성을 평가할 때 흥미로움(excitement)의 유사단어로 나타났다. 도출된 유사단어를 살펴보면, 상품 또는 재화를 평가하는데 초점을 맞춘 브랜드 개성의 기존 차원과 달리 관광지 및 관광상품을 경험한 후의 정서적인 감정을 전달하는데 초점을 맞춘 내재적 표현인 개관적인(panoramic), 빛나는(shiny), 안전한(safer), 즐거움(entertainment), 혼잡

함(crowded), 활기 넘침(buzzy), 모험심 강한(adventurous), 차별화된(differentiated), 기대할만한(expectable), 고단함(exhausted) 등이 유사단어로 나타났음을 알 수 있다.

2. 관광지 개성과 여행객 평점 간 영향관계 분석결과

코사인 유사도 분석을 통해 도출된 관광지 개

〈표 5〉 TF-IDF 기반의 차원축소를 통해 재생성된 관광지 개성의 기술 통계량

변수	평균	표준편차	최대값	최소값	VIF
여행객 평점 (traveler rating)	4.3925	0.8032	5	1	-
진실성 (sincerity)	0.3955	0.3357	3.6094	0	1.026
흥미로움 (excitement)	0.0290	0.1528	3.1448	0	1.083
유능함 (competence)	0.0252	0.1399	1.7056	0	1.019
세련됨 (sophistication)	0.0794	0.3009	6.9617	0	1.012
강인함 (ruggedness)	0.1049	0.2579	3.6046	0	1.034

N: 10,952

〈표 6〉 관광지 개성과 여행객 평점 간 영향관계 추정결과

변수	종속변수: 여행객 평점			
	OLS (모형 1)		WLS (모형 2)	
	β (S.E)	t-statistic (p-value)	β (S.E)	t-statistic (p-value)
진실성 (sincerity)	0.2327*** (0.0300)	7.760 (0.000)	0.2243*** (0.0057)	39.397 (0.000)
흥미로움 (excitement)	0.0533* (0.0237)	2.250 (0.024)	0.0816*** (0.0163)	5.017 (0.000)
유능함 (competence)	0.2570*** (0.0505)	5.092 (0.000)	0.2932*** (0.0228)	12.873 (0.000)
세련됨 (sophistication)	0.2342*** (0.0549)	4.264 (0.000)	0.3510*** (0.0355)	9.878 (0.000)
강인함 (ruggedness)	-0.0491 (0.0258)	-1.903 (0.057)	-0.0782*** (0.0126)	-6.224 (0.000)
Residual Standard Error	0.7994		0.8113	
R^2	0.0101		0.1464	
Adjusted R^2	0.0097		0.1460	

***p<0.001; **p<0.01; *p<0.05; df:10,946

성의 확장된 항목에 대하여 가중치가 부여된 TF-IDF행렬에서 해당 키워드가 포함된 문서를 추출하여 평균값을 이용하여 하나의 차원으로 축소하였다. 세부적으로, 전체 36,969개의 리뷰에서 앞서 도출한 관광지 개성의 구성개념을 추출한 후, 전체 50개 단어의 행에서 모두 0값을 가지는 26,017개의 행은 제거하였고, 최종적으로 10,952개의 리뷰를 분석단위로 활용하였으며 해당 기술 통계량은 〈표 5〉와 같이 정리된다.

서울의 관광지 및 관광상품에 대한 5점 척도의 여행객 평점평균은 약 4.4점정도로 높은 수준인 것으로 나타났다. 또한, 리뷰 내 단어의 중요도를 표현하는 TF-IDF를 활용하여 산출된 관광지 개성별 평균값은 진실성, 강인함, 세련됨, 흥미로움, 유능함의 순으로 큰 것으로 나타났다. 이러한 결과는 여행객들이 관광지 개성을 표현할 경우, 진실성(sincerity)과 더불어 이와 유사도가 높은 구성개념 표현이 상대적으로 많다는 것을 의미한다.

다음으로, 관광지 개성과 여행객 평점 간의 영향관계를 실증분석하기 위해 연구모형에 해당하는 식(1)을 최소제곱법과 가중최소제곱법을 통하여 계수를 추정하였으며, 결과는 〈표 6〉과 같다.

먼저 OLS를 통해 추정된 모형1을 살펴보면, '강인함'을 제외한 모든 관광지 개성의 차원이 여행객 평점에 유의한 긍정적 영향을 미치는 것으로 나타났으며 '유능함', '세련됨', '진실성', '흥미로움' 순으로 여행객 평점에 긍정적인 영향을 미친다는 것을 알 수 있다. 세부적으로, '유능함', '세련됨', '진실성'의 경우 각 차원별 유사단어가 1회 언급될 경우 여행객 평점이 약 0.2점대의 상승이 나타난다는 것을 알 수 있으며, '흥미로움'의 경우 유사단어가 1회 언급될 경우 약 0.05점의 여행객 평점이 상승하는 것으로 나타났다. 참고로, '유능함'과 '세련됨'의 경우 '진실성'에 비해서 절대적인 평균값은 작지만 척도의 분산(설명력)값은 상대적으로 높게 나타나 회귀분석 결과 종속변수인 소비자평점에 미치는 영향은 '진실성'과 비슷한 정도의 크기로 나타났음을 알 수 있다. 이러한 경향은 이후 가중최소제곱(WLS) 방법으로 분석하는 경우 더 크게 나타나 관광지 개성 요인 중 '유능함'과 '세련됨'이 소비자평점에 미치는 영향이 다른 개성 요인들에 비해 가장 큰 것으로 나타났다.

OLS모형의 적합도 검증결과, F값은 22.35($p<0.05$)로 나타났으며 Durbin-Watson test의 값은 1.6546으로 나타나 잔차가 독립적인 것으로 나타났다. 또한, 다중공선성 검정결과 모든 설명변수의 VIF가 1.1이내의 값을 가져 보다 보수적인 관점인 2.0을 기준에서 보아도 변수간 상관관계는 없는 것으로 나타났다. 그러나 R2값이 0.01수준으로 나타나 모형설명력이 아주 낮은 것으로 나타났다. 이에 따라 오차항의 등분산성(homoscedasticity) 검정을 위해 BP test를 수행하였으며, BP값은 13.27($p=.021$)로 유의하게 나타나 오차항이 이분산성(heteroskedasticity)을 가지는 것으로 나타났다. 이러한 이분산성 문제를 해결하기 위해서는 일반적으로 가중최소제곱(WLS)을 통하여 이분산성을 해결하여 유효한(efficient) 추정 값을 얻을 수 있다(White, 1980).

따라서 WLS를 통해 추정된 모형의 적합도 검증 결과를 살펴보면, F값은 375.4($p<0.05$)로 나타났으며 R2값은 0.146으로 나타나 OLS모형에 비해 적합도가 우수한 것으로 나타났다. WLS를 통해 추정된 계수를 살펴보면, 관광지 개성의 모든 차원이 여행객 평점에 유의한 영향을 미치는 것으로 나타났다. 세부적으로, '세련됨'의 유사단어가 1회 표현될 경우 소비자의 평점이 가장 높게 (0.351) 증가하는 것을 알 수 있으며, 다음으로 '유능함'의 유사단어가 1회 표현될 경우 소비자의 평점은 약 0.29점 증가하는 것으로 나타났다. 또한, '진실성'의 유사단어가 1

회 언급될 경우 소비자의 평점은 약 0.22점 증가하는 것으로 나타났으며, '흥미로움'의 유사단어가 1회 언급될 경우 소비자의 평점은 약 0.08점 증가하는 것으로 나타났다. 이에 반해, '강인함'의 유사단어가 1회 언급될 경우는 소비자의 평점이 0.08점 낮아지는 것으로 나타났으며, '강인함'의 유사단어 중 부정적인 표현인 위험한(danger), 혼미한(disoriented), 겁을 먹은(intimidated), 격렬한(strenuous), 미결정의(undecidable) 등이 기인하여 평점에 부정적 영향을 미친 것으로 판단된다.

Ⅴ. 결 론

그동안 국내 관광분야에서 온라인 리뷰 데이터를 활용한 연구들이 대부분 질적 데이터에 해당하는 리뷰기반의 키워드분석, 감성분석, 사회연결망분석 등을 통한 여행객의 인식이나 목적지별 특성에 초점을 맞추고 있다(김민식・한학진・박병화, 2018; 오익근 외, 2015; 한지연・김홍범, 2017). 그러나 여행객이 관광지에서의 경험을 전달하는 리뷰 데이터는 글로 표현된 질적 데이터와 더불어 해당 관광지에 대한 정량적 평가에 해당하는 평점도 함께 부여하기 때문에 리뷰 데이터만을 활용한 연구에서 여행객 또는 관광지의 특성을 파악하기에는 한계가 있다. 이를 극복하기 위해 본 연구는 온라인 리뷰 빅데이터에 기반하여 자연어 처리 분야 중 텍스트 분석 분야에서 최근 화두가 되고 있는 신경망 언어 딥러닝 모델인 Word2Vec 기법을 활용하여 관광지 개성과 유사단어를 도출하였으며, 정량적 데이터인 여행객 평점과의 연계를 통해 인과관계를 실증적으로 규명함으로써 관광분야에서 온라인 리뷰 빅데이터를 다각적으로 활용 및 분석할 수 있는 접근법을 제시하였다. 특히, 본 연구는 관광분야에서 신경망 언어모델인 Word2Vec의 실질적 활용 및 체계를 조명하고 비정형 텍스트 데이터를 수치화하여 분석하는 일련의 과정과 정형데이터인 평점데이터와 연계하여 인과모형을 측정하는 선험적 연구로서 그 의의가 있다. 또한, 그동안 마케팅 분야에서 제품에 대한 상징적 자아표현 기능을 측정하였던 브랜드 개성의 차원을 토대로 관광분야에서 소비자의 관점에서 표현되는 관광지 개성의 구성개념을 확장한 점에서 학술적 의의를 가진다.

먼저, Word2Vec의 skip-gram 모델을 통해 임베딩된 값을 바탕으로 수행된 코사인 유사도 분석을 통해 관광지 개성과 기존 브랜드 개성 구성개념 간의 큰 차이를 확인할 수 있었다. 5가지 개성의 차원별 유사단어의 결과를 통해 기존 브랜드 개성의 구성개념에서 공통적으로 나타나는 항목은 '강한(tough)', '건전한(wholesome)'인 것으로 나타났으며, 관광지 개성과 관련된 주요 구성개념은 '참된(genuine)', '파노라믹(panoramic)', '빛나는(shiny)', '안전한(safer)', '재미(fun)', '즐거움(entertainment)', '혼잡함(crowded)', '활기 넘침(buzzy)', '훌륭한(respectable)', '유쾌한(jolly)', '모험심 강한(adventurous)', '차별화된(differentiated)', '기대할만한(expectable)', '도전적인(challenging)', '고단함(exhausted)' 등인 것으로 나타났다. 이러한 결과는 관광지 개성을 측정하기 위해서는 기존 브랜드 개성의 구성개념과 같이 유형의 상품에 대한 상징적인 표현이 아닌, 여행객의 정서적인 감정에 초점을 둔 구성개념의 고찰 및 측정항목 개발이 필요하다는 것을 의미한다. 따라서 딥 러닝 기법을 통해 도출된 본 연구의 관광지 개성관련 연관단어(구성개념)는 향후 관광지 개성의 측정과 척도 개발 연구에서 관광분야의 실정에 적합한 측정도구로 유용하게 활용될 수 있을 것이며, 또한 여행객의 관점에서 관광지에 대한 정서적인 평가를 위한 지표로써도 활용될 수 있을 것이다.

다음으로, 관광지 개성과 여행객 평점 간의 모형추정 결과를 통해 텍스트 데이터를 수치화하면서 회귀모형 가정에서 위배될 수 있는 부분을 점검하고 해소할 수 있는 방안을 제시하였다. 세부적으로, 텍스트 데이터가 수치화되면 이분산성이 나타날 수 있음을 확인할 수 있었고, 등분산성 가정의 위배는 단어×문서 행렬 구조인 텍스트 데이터의 독특한 특성상 하나의 문서에서 수많은 단어들로 구성된 행(raw) 객체의 산출 값들이 불규칙적인 비율로 나타날 수 있기 때문이라 추정할 수 있다. 이러한 텍스트 데이터의 이분산성 해결을 위해 WLS모형을 통해 계수를 추정하였으며, OLS모형에 비해 WLS의 모형이 보다 우수한 것으로 나타났다. 따라서 향후 질적 데이터를 수치화하고 양적 데이터와 연계한 인과모형을 설정할 경우에는 텍스트 데이터의 특성을 고려한 엄격한 사전검정을 통하여 모형설정이 필요할 것이라 사료된다.

가중최소제곱법(WLS)을 통한 모형 추정 결과, 관광지 개성의 5가지 차원 모두가 소비자의 평점에 유의한 영향을 미치는 것으로 나타났으며, '진실성', '흥미로움', '유능함', '세련됨'은 여행객 평점에 유의한 양의 영향을 미치는 반면 '강인함'은 여행객 평점에 음의 영향을 주는 설명 요인으로 확인되었다. 세부적으로 여행객 평점상승에 영향력이 큰 요인은 '세련됨' 〉 '유능함' 〉 '진실성' 〉 '흥미로움' 순인 것으로 확인되었다. 이는 본 연구의 코사인 유사도 분석결과와 연계하여, 여행객의 관광경험에 부정적인 영향을 줄 수 있는 '강인함'과 관련된 키워드인 '강한(touch)', '험준함(steepness)', '도전적인(challenging)', '위험한(dangerous)', '혼미한(disorienting)' 등이 나타날 수 있는 부분을 지양하고 상대적으로 평점 상승효과가 높은 '세련됨', '유능함', '진실성'과 관련된 키워드인 '찬사(compliment)', '대화의(conversational)', '차별화된(differentiated)', '훌륭한(respectable)', '유쾌한(jolly)', '참된(genuine)', '친절한(kind)' 등을 부각시킨다면, 여행객의 인식에서 관광지 개성을 통해 긍정적 이미지를 유도하는데 기인할 수 있다는 것을 의미한다. 나아가 본 연구의 결과는 향후 장소 브랜딩의 맥락에서 상징적 자아표현인 관광지 개성을 활용하여 웹(web)상에서 표현되는 여행객 평점, 선호도 등과의 영향관계를 분석하는 후속 연구들의 기초가 될 것이라 판단된다.

본 연구는 현재까지 관광분야에서 활용되고 있는 온라인 소비자 리뷰 빅데이터의 텍스트 마이닝 분석체계에서 보다 확장된 기법인 신경망 언어 모델(NNLM)을 통한 단어 임베딩과 유사도 분석, 그리고 리뷰 데이터와 평점 데이터와의 연계분석 등 빅데이터에 담긴 질적, 양적 데이터 모두를 활용하여 유의미한 정보를 도출하고자 하였다. 무엇보다 관광분야에서 텍스트 데이터에 대한 신경망 모델을 통한 딥 러닝 기법의 적용방법과 정형·비정형 빅데이터를 연계하여 분석할 수 있는 체계를 제시함으로써 관광분야의 학문적, 방법론적 성과를 제고하였다고 사료된다.

그럼에도 불구하고 본 연구는 학술적 의의에 비해 실무적 기여가 상대적으로 부족하다는 연구의 한계점을 가지고 있다. 다시 말해, 본 연구가 관광분야에서 빅데이터를 통한 언어신경망 딥 러닝 분석과 질적, 양적 데이터의 연계분석을 수행한 선험적 연구인만큼, 본 연구의 결과만을 통하여 구체적으로 관광지 개성을 통한 여행객의 행동 의도에 대하여 답하기는 어렵다. 따라서 향후 확장된 관광지 개성의 차원이 평점분포에 따라 어떻게 분류(classification)되는지 또는 리뷰 데이터에 포함된 여행객의 개별정보인 성별, 연령대, 여행유형 등을 활용하여 평점분포에서 어떻게 분류되는지를 밝힌다면 보다 의미 있는 연구로 확장될 수 있을 것이다. 또한, 본 연구에서 키워드 간 유사도 분석의 값이 충분히 높게 나타나지 않았는데, 이는 다소 부족한 학습 데이터의 양이 기인한 것이라 판단된다. 따라서 향후 연구

에서는 보다 많은 데이터를 확보하여 충분한 학습을 통한 Word2Vec의 성능측정과 이를 바탕으로 한 유사도 분석이 수반되어야 할 것이다.

참고문헌

권영빈・이승도・양현・주요한(2012). 키워드를 기반으로 마이너와 코사인 유사도를 이용한 컴퓨터 네트워크 관련 컨퍼런스 분석.『한국IT 서비스학회지』, 11, 223-238.

김광영・이원구・윤화묵・신성호・이민호(2011). 웹 자원 아카이빙을 위한 웹 크롤러 연구 개발.『한국콘텐츠학회논문지』, 11(9), 9-16.

김민식・한학진・박병화(2018). 트립어드바이저를 이용한 국내 목적지별 특성분석.『관광레저연구』, 30(2), 5-19.

노미진・이경탁. (2012)소셜커머스 수용에 있어서 지각된 위험의 영향력. 『경영학연구』, 41(1), 57-87.

박득희・김태구・이계희(2016). 소셜 빅데이터를 활용한 관광정보 네트워크 분석.『관광연구저널』, 30(8), 195-208.

박수지・신진옥・송상헌・정철(2017). 텍스트 마이닝을 통한 관광지 수요예측: 온라인 검색 엔진을 중심으로.『관광학연구』, 41(1), 13-27. http://dx.doi.org/10.17086/JTS.2017.41.1.13.27

박정환・이병철. (2014). 온라인과 오프라인 리뷰 비교분석: 박람회 서비스 품질을 중심으로.『관광경영연구』, 59, 61-79.

심영석・김홍범(2016). 텍스트 마이닝을 이용한 관광지 이미지 구성요인 및 측정에 관한 연구.『관광학연구』, 40(7), 221-245. http://dx.doi.org/10.17086/JTS.2016.40.7.221.245

오익근・이태숙・전채남(2015). 빅데이터 분석을 통한 한국관광 인식에 관한 연구.『관광학연구』, 39(10), 107-126.

이슬기・정성관・이우성・박경훈(2011). 인공신경망을 이용한 도시기온 예측모형 구축.『국토계획』, 46(1), 129-142.

이태원・홍태호(2015). Support Vector Machine을 이용한 온라인 리뷰의 용어기반 감성분류모형.『Information Systems Review』, 17(1), 49-64.

장환석・장은영・정광용(2017년 12월). Word2Vec를 이용한 감성어 분석 방법.『2017 한국소프트웨어종합학술대회 학술발표논문집』(pp. 661-663), 한국정보과학회, 부산시.

전근식・공성언・최용석(2017년 12월). 영화 평점 예측을 위한 Word2Vec 기반 협업 필터링.『2017 한국소프트웨어종합학술대회 학술발표논문집』(pp. 844-846), 한국정보과학회, 부산시.

전효재(2018). 숙박관광유형별 비정형데이터의 연관 규칙 분석과 활용에 관한 연구: IRTS 2008 의 수요와 공급 관점.『관광학연구』, 42(5), 137-150. http://dx.doi.org/10.17086/JTS.2018.42.5.137.150

한지연・김홍범(2017). 빅데이터 기반의 사회연결망 분석을 이용한 관광지 이미지 인식에 관한 연구.『관광학연구』, 41(8), 91-119. http://dx.doi.org/10.17086/JTS.2017.41.8.91.119

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of marketing research, 34*(3), 347-356.

Armstrong, R., Freitag, D., Joachims, T., & Mitchell, T. (1995, March). Webwatcher: A learning apprentice for the world wide web. In *AAAI Spring symposium on Information gathering from Heterogeneous, distributed environments* (pp. 6-12).

Ayeh, J. K., Au, N., & Law, R. (2013). "Do we believe in TripAdvisor?" Examining credibility perceptions and online travelers' attitude toward using user-generated content. *Journal of Travel Research, 52*(4), 437-452.

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: Readers' objectives and review cues. *International*

Journal of Electronic Commerce, 17(2), 99-126.

Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. E. (2018, February). Comparative Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Approach. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 311-318). Springer, Cham.

Chang, Y. C., Ku, C. H., & Chen, C. H. (In press). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2017.11.001

Chen, C. F., & Phou, S. (2013). A closer look at destination: Image, personality, relationship and loyalty. *Tourism Management, 36*, 269-278.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(Aug), 2493-2537.

Ekinci, Y., & Hosany, S. (2006). Destination personality: An application of brand personality to tourism destinations. *Journal of Travel Research, 45*(2), 127-139.

Eriksson, N., & Fagerstrøm, A. (2017). The Relative Impact Of Wi-Fi Service On Young Consumers' Hotel Booking Online. *Journal of Hospitality & Tourism Research*, 1096348017696844.

Gretzel, U., Yoo, K., & Purifoy, M. (2007). *Online travel review report: Role & impact of online travel reviews*. Laboratory for Intelligent Systems in Tourism.

Han, J., Kamber, M., & Pei, J. (2006). *Mining frequent patterns, associations, and correlations. Data Mining: Concepts and Techniques* (2nd ed., pp. 227-283). San Francisco, USA: Morgan Kaufmann Publishers.

Herrero, Á., San Martín, H., & Hernández, J. M. (2015). How online search behavior is influenced by user-generated content on review websites and hotel interactive websites. *International Journal of Contemporary Hospitality Management, 27*(7), 1573-1597.

Hultman, M., Skarmeas, D., Oghazi, P., & Beheshti, H. M. (2015). Achieving tourist loyalty through destination personality, satisfaction, and identification. *Journal of Business Research, 68*(11), 2227-2231.

Kang, H., Yoo, S., & Han, D. (2009). Modeling web crawler wrappers to collect user reviews on shopping mall with various hierarchical tree structure. In *a proceeding of the WISM 2009 International Conference* (pp. 69-73), Web Information Systems and Mining.

Kang, M., & Schuett, M. A. (2013). Determinants of sharing travel experiences in social media. *Journal of Travel & Tourism Marketing, 30*(1-2), 93-107.

Kim, S., & Lehto, X. Y. (2013). Projected and perceived destination brand personalities: The case of South Korea. *Journal of Travel Research, 52*(1), 117-130.

Kumar, V., & Nayak, J. K. (2018). Destination personality: Scale development and validation. *Journal of Hospitality & Tourism Research, 42*(1), 3-25.

Lee, H. A., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing, 28*(7), 675-688.

Liu, X., Schuckert, M., & Law, R. (2015). Can

response management benefit hotels? Evidence from Hong Kong hotels. *Journal of Travel & Tourism Marketing, 32*(8), 1069-1080.

Llodrà-Riera, I., Martínez-Ruiz, M. P., Jiménez-Zarco, A. I., & Izquierdo-Yusta, A. (2015). A multidimensional analysis of the information sources construct and its relevance for destination image formation. *Tourism Management, 48*, 319-328.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing, 77*(1), 87-103.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting and the vector space model. *Introduction to Information Retrieval, 100*, 2-4.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Morosan, C., & DeFranco, A. (2016). Modeling guests' intentions to use mobile apps in hotels: The roles of personalization, privacy, and involvement. *International Journal of Contemporary Hospitality Management, 28*(9), 1968-1991.

Murphy, L., Benckendorff, P., & Moscardo, G. (2007). Destination brand personality: Visitor perceptions of a regional tourism destination. *Tourism Analysis, 12*(5-6), 419-432.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management, 19*(7), 754-772.

Pan, L., Zhang, M., Gursoy, D., & Lu, L. (2017). Development and validation of a destination personality scale for mainland Chinese travelers. *Tourism Management, 59*, 338-348.

Papadimitriou, D., Apostolopoulou, A., & Kaplanidou, K. (2015). Destination personality, affective image, and behavioral intentions in domestic urban tourism. *Journal of Travel Research, 54*(3), 302-315.

Pike, S., & Ryan, C. (2004). Destination positioning analysis through a comparison of cognitive, affective, and conative perceptions. *Journal of Travel Research, 42*(4), 333-342.

Rong, X. (2014). Word2vec parameter learning explained. *arXiv preprint arXiv*:1411.2738.

Sahin, S., & Baloglu, S. (2011). Brand personality and destination image of Istanbul. *Anatolia-An International Journal of Tourism and Hospitality Research, 22*(1), 69-88.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Schuckert, M., Liu, X., & Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management, 48*, 143-149.

_____(2016). Insights into suspicious online ratings: direct evidence from TripAdvisor. *Asia Pacific Journal of Tourism Research, 21*(3), 259-272.

Shao, T., Chen, H., & Chen, W. (2018, April). Query Auto-Completion Based on Word2vec Semantic Similarity. In Journal of Physics: Conference Series (Vol. 1004, No. 1, p. 012018). IOP Publishing.

Shum, S., Dehak, N., Dehak, R., & Glass, J. R. (2010). Unsupervised speaker adapta-

tion based on the cosine similarity for text-independent speaker verification. In *Odyssey* (p. 16).

Tasci, A. D., Gartner, W. C., & Cavusgil, S. T. (2007). Measurement of destination brand bias using a quasi-experimental design. *Tourism Management, 28*(6), 1529-1540.

Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394). Association for Computational Linguistics.

Usakli, A., & Baloglu, S. (2011). Brand personality of tourist destinations: An application of self-congruity theory. *Tourism Management, 32*(1), 114-127.

Wang, Y., & Fesenmaier, D. R. (2004). Towards understanding members' general participation in and active contribution to an online travel community. *Tourism Management, 25*(6), 709-722.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*(4) 817-838.

Xiang, Z., Gretzel, U., & Fesenmaier, D. R. (2009). Semantic representation of tourism on the Internet. *Journal of Travel Research, 47*(4), 440-453.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. *Expert Systems with Applications, 42*(4), 1857-1863.

Zhang, Z., Ye, Q., & Law, R. (2011). Determinants of hotel room price: An exploration of travelers' hierarchy of accommodation needs. *International Journal of Contemporary Hospitality Management, 23*(7), 972-981.

2018년 6월 29일 최초투고논문 접수
2018년 8월 16일 최종심사완료 및 게재확정 통보
2018년 8월 22일 최종논문 도착
3인 익명심사 畢