

방한 관광객의 온라인 리뷰에 대한 빅데이터 분석 기반의 감성분석 및 평점 예측모형¹⁾

Sentiment Analysis and Star Rating Prediction Based on Big Data Analysis
of Online Reviews of Foreign Tourists Visiting Korea

홍태호 (Taeho Hong)

부산대학교 경영대학²⁾

〈 국문초록 〉

관광객이 작성한 온라인 리뷰는 관광산업의 관리 및 운영에 중요한 정보를 제공한다. 평점은 제품이나 서비스에 대한 정량적인 평가로 간편하지만 관광객의 진실한 태도를 반영하기 어려우며 평점과 리뷰내용에 대한 불일치 문제도 발생하고 있다. 불일치 문제는 잠재고객에게 혼동을 줄 수 있으며 구매의사결정에도 영향을 미칠 수 있다. 본 연구에서는 온라인 리뷰기반의 평점 예측모형을 통해 평점과 리뷰내용의 불일치 문제를 해결하고자 한다. 한국을 방문한 외국인 관광객이 작성한 관광지과 호텔에 대한 리뷰의 감성분석을 통해 평점과 감성의 차이를 비교하고 TF-IDF vectorization과 감성분석 결과로 변수를 선정하였다. 로짓, 인공신경망, SVM(Support Vector Machine)을 적용하여 평점을 분류하고, 인공신경망, SVR(Support Vector Regression)을 통해 평점을 예측하였다. 평점 분류모형과 예측모형 모두 불일치한 리뷰를 제거하고 감성분석을 반영한 모형에서 우수한 성과를 보여주었다. 본 연구에서 제안한 온라인 리뷰 기반의 평점 예측모형은 평점과 리뷰내용에 대한 불일치 문제를 해결하여 신뢰할 수 있는 정보를 제공하였으며 평점이 없는 온라인 리뷰에도 활용할 수 있을 것이다.

주제어: 온라인 리뷰, 텍스트마이닝, 감성분석, 빅데이터, 예측모형

1) 이 논문은 2010년도 부산대학교 인문사회연구기금의 지원을 받아 연구되었음

2) hongth@pusan.ac.kr

1. 서론

인터넷의 발전과 더불어 소셜 미디어와 전자상거래의 비약적인 발전은 온라인 플랫폼에서 생성되는 데이터의 폭발적인 증가를 초래하였다. 이에 따라 방대한 데이터에서 필요한 정보만을 선별하기 위해 빅데이터 분석의 필요성이 대두되고 있으며 최근 관광산업에서도 온라인 리뷰를 활용한 빅데이터 분석이 각광받고 있다. 관광객들은 여행을 계획할 때 다른 관광객들이 온라인 플랫폼상에 작성한 온라인 리뷰를 통해 여행정보를 검색하며 미지의 관광지에 대한 불확실성을 줄이고자 한다. 관광객의 지각된 위험과 지각된 불확실성은 관광 목적지에 대한 태도와 방문의도에 영향을 미친다(Quintal et al., 2010). 특히, 해외여행의 경우 온라인 리뷰를 통한 정보검색과 이를 참고하고자 하는 경향이 더욱 높게 나타난다. 하지만 다양한 온라인 댓글과 수많은 양의 온라인 리뷰는 정보 과부하의 문제를 발생시키며 관광객들은 필요한 정보를 검색하는 것에 어려움을 겪는다. 정보 과부하는 정보를 추구하는 사람이 정보검색을 중단하려는 행동의도에 영향을 미칠 수 있다(Swar et al., 2017). 즉, 온라인 리뷰가 많아질수록 정보량이 늘어나며 소비자의 신뢰와 구매의도를 높일 수는 있지만 너무 많은 정보는 소비자의 신뢰와 구매의도를 오히려 감소시킬 수도 있다(Furner & Zinko, 2017). 따라서 정보 과부하의 문제를 해결하기 위해 빅데이터 분석을 적용하여 유용한 정보를 선별하여 제공할 필요가 있다.

관광객들은 자신의 여행경험을 온라인 플랫폼에 작성하여 많은 사람들과 공유하고 있으며, 이러한 여행경험은 잠재관광객들에게 새로운 관광 정보로 인식되고 있다. 관광객이 작성한 온라인 리뷰에 대한 빅데이터 분석은 소비자의 실질적인 관광니즈, 만족도 등에 대한 정보를 파악할 수 있게 한다. 예를 들어, 온라인

리뷰에 대한 감성분석을 통해 호텔 서비스에 대한 관광객의 태도를 파악할 수 있으며(Hu et al., 2017), 온라인 리뷰에 대한 토픽 모델링은 호텔 고객 만족도에 영향을 미치는 요인을 찾을 수 있다(Guo et al., 2017). 온라인 리뷰 플랫폼에 게시된 관광지에 대한 정보는 관광지 선택이나 여행 계획에 중요한 역할을 하며, 아울러 관광지의 이미지에도 큰 영향을 미칠 수 있어 온라인 리뷰 분석에 대한 필요성이 대두되고 있다. 온라인 리뷰를 활용한 빅데이터 분석은 관광산업에 유용한 정보를 제공하며 관광산업의 활성화 전략을 위한 의사결정에도 도움을 준다. 관광객의 오피니언 분석을 위해서는 관광지, 호텔 및 레스토랑 등의 온라인 리뷰에 대한 감성분석이 많이 활용되고 있다. Prameswari et al. (2017)은 호텔에 대한 온라인 리뷰를 기반으로 텍스트 마이닝과 감성분석을 결합한 방법을 사용하였으며 소비자의 의견을 분류하여 호텔 산업의 품질을 개선하고자 하였다. Kim et al.(2017)은 스마트 관광의 한 측면으로 온라인 리뷰를 분석하기 위해 감성분석을 적용하였으며 서비스를 다양하게 분류하여 관광객의 지각된 품질을 측정하였다. 그러나 이들은 온라인 리뷰에 대한 문서수준의 극성분류 혹은 주관성 분류 등과 같은 감성분류 모형을 개발하는데 그치고 있어 한국을 방문한 외국인 관광객에 대한 감성분석과 예측모형 개발에 관한 연구는 미비하다.

또한 외국인 관광객은 언어와 문화적 차이로 인해 만족도 조사에 대한 정확한 분석이 어렵다는 문제가 존재한다. 이를 해결하기 위해 관광 리뷰 플랫폼에 공유되어 있는 다양한 국적의 관광객에 의해 작성된 리뷰 데이터를 활용한다면 관광객의 의견 및 만족도에 대한 파악이 가능하다. 특히 개별 관광객은 단체 관광객보다 온라인 플랫폼을 통해 관광 후기를 활발하게 공유하고 있으며, 이는 다른 관광객의 의사결정에 영향을 미칠 수 있어 높은 관심을 받고 있다. 4차 산업혁

명과 함께 온라인 리뷰 플랫폼 시장도 급격하게 성장하고 있으며, TripAdvisor, Yelp, Airbnb 등과 같은 세계적인 관광 리뷰 플랫폼에서는 여러 국가의 리뷰가 동시에 게시되고 있다. 리뷰 플랫폼에서 제공하는 정략적인 평점 정보는 간편하게 사용하는 대신 관광지에 대한 구체적인 정보를 전달하기 어렵다. 또한 리뷰 플랫폼은 다양하고 풍부한 정보의 제공도 중요하지만 정확한 정보를 제공할 필요가 있으나 평점과 리뷰내용에 대한 불일치 문제도 발생하고 있다. 리뷰의 텍스트에 표현된 의견과 평점 간의 불일치 문제는 리뷰를 읽는 잠재고객에게 혼동을 줄 수 있으며 리뷰 정보에 대한 신뢰도에도 영향을 미칠 수 있다. 따라서 평점과 리뷰내용이 일치하는 평점 정보를 제공할 수 있도록 온라인 리뷰를 기반으로 한 평점 예측모형을 제안하고자 한다.

본 연구는 TripAdvisor에서 외국인 관광객이 작성한 서울의 관광지과 호텔에 대한 리뷰를 활용한다. 온라인 리뷰에 대해 감성분석을 수행하여 한국 관광에 대한 외국인 관광객의 태도를 파악하고, 감성분석을 통해 리뷰의 텍스트에 표현된 의견과 평점이 불일치한 리뷰를 선별한다. 불일치한 데이터를 제외하여 데이터의 신뢰성을 확보한 다음 로짓, 인공신경망, SVM(Support Vector Machine)을 적용하여 평점 분류모형을 개발하며 인공신경망과 SVR(Support Vector Regression)을 적용하여 평점 예측모형을 구축한다. 이는 정보 과부화의 문제를 해결하며 관광객들에게 신뢰할 수 있는 정보를 제공할 수 있다.

2. 이론적 배경

2.1. 온라인 리뷰와 감성분석

온라인 리뷰는 제품이나 서비스에 대한 사용경험을

텍스트나 이미지 등으로 자유롭게 표현하여 평점으로 파악하기 어려운 유용한 정보를 제공한다. 소비자들은 제품이나 서비스에 대한 불확실성을 해소하기 위해 온라인 리뷰를 활용하고 있으며 관광산업에서도 온라인 리뷰에 대한 영향력이 점차 커지고 있다. 관광상품은 경험재(experienced good)이기 때문에 실제 경험한 사용자들이 제공하는 정보가 소비자들에게 중요한 영향을 미치고 있다(Sotiridais et al., 2013). 관광지, 호텔 및 레스토랑 등에 대한 온라인 리뷰를 활용한 연구가 활발하게 이루어지고 있으며, Xiang et al.(2017)은 TripAdvisor, Expedia 및 Yelp 등에서 호텔에 대한 정보 품질을 비교하기 위해 텍스트 분석을 적용하였다. 야오즈옌 등(2021) 온라인 리뷰를 통해 감성과 감정이 리뷰 유용성에 영향을 주는 점을 레스토랑의 데이터를 이용하여 주장하였다. 온라인 리뷰에 대한 언어적 특성, 의미적 특성, 감성, 평점, 유용성 및 특성들 간의 차이를 비교하여 호텔 산업에 대한 온라인 플랫폼에서의 불일치한 표현을 보여주었으며, TripAdvisor의 정보 품질이 전체적으로 우수하며 유용한 정보를 제공한다고 인식되고 있다고 하였다.

감성분석(Sentiment analysis)은 텍스트 마이닝 기법 중 하나로 텍스트 문서에 포함되는 다양한 극성과 감성을 추출하는 방식이다. 일반적으로 텍스트에 대한 긍정 혹은 부정적 언어를 식별하고 분류하는 텍스트 분류 문제에 많이 활용되고 있다(Hu et al., 2012). 감성분석 방법은 어휘기반, 기계 학습 및 하이브리드 방법의 세 가지 유형으로 분류 할 수 있다(Ravi & Ravi, 2015). 어휘기반 감성분석은 감성사전에 의해 이루어지며 각 문서의 단어를 감성사전의 어휘와 매칭한다. 기계 학습 방법은 텍스트에 수동으로 레이블을 지정하는 과정이 필요하며 모델과 데이터에 크게 의존하기 때문에 어휘기반 감성분석이 보다 효율적인 방법이라 할 수 있다(Zhu et al., 2020). 어휘기반의 감성분석을

위해 HowNet, SentiWordNet, LIWC(Linguistic Inquiry and Word Count), AFINN, GI(General Inquirer) 등의 감성사전이 널리 사용된다. 감성분석은 텍스트로 표현된 의견과 감성을 식별할 수 있고 온라인 리뷰를 통한 소비자의 긍정이나 부정적인 제품 및 서비스에 대한 평가를 구분할 수 있다. Sharma et al.(2020)은 감성분석을 통해 리뷰 평점과 리뷰 감성 간의 관계를 분석하였으며 소비자가 제공받은 호텔의 서비스와 기대한 서비스의 차이가 감성에 영향을 미친다는 것을 밝혔다. Tsai et al.(2020)은 호텔 리뷰 중 유용한 정보를 정확하게 추출하기 위하여 감성분석을 기반으로 자동 리뷰 요약 시스템을 구축하였다. 이처럼 감성분석은 온라인 리뷰의 평가뿐만 아니라 암호화폐 수익률 변동의 예측과 화장품 회사의 브랜드컨셉 개발과 같이 폭 넓은 분야에서 적용되고 있다. (김은미, 2021; 이주민, 방정혜, 2020)

본 연구는 TripAdvisor의 관광 리뷰를 활용하여 어휘기반 감성분석을 적용하여 한국을 방문한 외국인 관광객의 긍정 또는 부정적인 의견을 파악하고자 한다. 온라인 리뷰에 대한 감성분석은 리뷰의 텍스트 콘텐츠에 포함된 긍정적, 부정적 또는 중립적인 태도나 오피니언을 분류하는 과정으로 SentiWordNet 감성사전을 사용하여 관광지과 호텔에 대한 온라인 리뷰의 감성을 탐색한다. SentiWordNet 감성사전은 감성분류 및 오피니언 마이닝을 위해 개발된 사전으로, 최근에는 제품, 영화, 스포츠, 정치 등 다양한 영역에서의 텍스트 분석에 적용되고 있다(Mahadzir et al., 2021).

2.2. 온라인 리뷰의 예측모형

온라인 리뷰는 제품 및 서비스에 대한 소비자의 다양한 의견을 포함하고 있어 소비자의 행위에 영향을 미치는 요인에 대한 탐색이 가능하다. 판매자도 온라

인 플랫폼을 통한 판매 및 마케팅 전략을 위해 활용할 수 있기에 온라인 리뷰를 기반으로 다양한 목적의 예측모형이 개발되고 있다. Al Ajrawi et al.(2021)은 온라인 리뷰에 대한 감성분석을 수행하고 Linear Regression, SVM, Random Forest 등의 지도학습 알고리즘을 적용하여 평점 예측모형을 제안하였다. Krishnamoorthy (2015)는 유용성에 영향을 미치는 요인을 분석하여 유용성 예측모형을 개발하였다. 온라인 플랫폼에서 제공하는 리뷰의 유용성은 투표를 통해 랭킹 순서대로 제시하여 소비자의 정보 과부하 문제를 최소화하고자 한다. Lee et al.(2018)은 리뷰 품질, 감성, 리뷰어 특징 등을 활용하여 리뷰 유용성 예측모형을 구축하였으며 Ahani et al.(2019)은 온라인 리뷰 및 평점을 통해 마케팅 전략 수립에 유용한 정보를 제공하고 마케팅 비용을 절감할 수 있도록 머신러닝 기법을 통한 고객 세분화를 제안하고 스파 호텔 세분화 및 여행 선택 예측을 실현할 수 있는 하이브리드 모형을 개발하였다.

최근 관광분야에서는 온라인 리뷰의 품질을 반영할 수 있는 콘텐츠의 감성과 리뷰 평점 간의 일치성 정도에 대한 예측연구를 많이 진행하였다. Luo & Xu(2021)은 코로나 시대에 소비자의 안전한 레스토랑 선정을 위해 레스토랑의 리뷰를 대상으로 딥러닝 알고리즘을 사용하였으며 리뷰의 감성 분류와 평점예측을 수행하여 감성과 평점이 불일치하는 리뷰를 탐지하였다. 또한, Zheng et al.(2021)은 온라인 리뷰의 평점을 기반으로 딥러닝 기법을 사용하여 평점 예측모형을 개발하였으며, 100만 개 리뷰에 대한 예측 분석을 통해 높은 예측성능을 확보하였다. 리뷰와 평점에 대한 불일치는 모호성 문제를 유발할 수 있으며 편향된 평점과 신뢰성 낮은 리뷰를 분류하면 리뷰 품질에 대한 이해에도 도움이 될 수 있다.

온라인 소비자는 자신의 정보 처리능력과 검색비용으로 인해 온라인 플랫폼에 제시되어있는 수백 개의

리뷰를 모두 읽지 않기 때문에 리뷰의 평점은 소비자의 구매의사결정에 중요한 역할을 한다. 또한 온라인 리뷰의 텍스트 콘텐츠를 기반으로 한 평점 예측은 관광객이 부여한 평점의 편차를 식별하여 리뷰의 신뢰도 문제를 해결하는 데 도움이 될 수 있다. 따라서 본 연구는 기존연구에서 주장한 리뷰 내용과 평점 간의 불일치 문제를 해결하고자 한국을 방문한 외국인 관광객의 한국 관광 산업에 대한 진실한 태도를 파악하기 위해 관광지과 호텔 리뷰에 대한 감성분석 결과를 기반으로 머신러닝을 활용한 평점 분류 및 예측모형을 제안한다.

2.3. 온라인 리뷰의 예측기법

예측모형은 데이터의 종류에 따라 다양한 데이터마이닝 기법들이 적용되고 있으며 분류를 위해 로짓, 인공신경망, SVM 등의 머신러닝 기법이 많이 사용되고 있다. 로짓분석은 종속변수가 이분형 데이터일 때 주로 사용되며 기업부도 예측, 주가지수 예측 등 다양한 경영학 분야에 적용되어왔다. 온라인 리뷰에서도 평점의 좋고 나쁨, 유용한 리뷰인지 아닌지에 대한 이진 분류를 위한 문제에 적용되고 있다(Lee et al., 2018; Wu et al., 2020; Al Ajrawi et al., 2021).

인공신경망은 뇌에서 뉴런들이 상호작용하고 학습하는 것을 모형화한 방법으로(Zang et al., 1999), 분류 및 예측을 위해 많이 적용되어왔다. 인공신경망은 입력층, 은닉층, 출력층의 3개의 층으로 구성되어 있으며 모두 가중치로 연결되어 있어 모형이 복잡하고 계산이 복잡하다. 그러나 기존의 분류 및 예측모형보다 우수한 예측성능을 보여주며 다양한 분야에서 활용되고 있다. Bilal et al.(2021)은 온라인 리뷰의 유용성을 예측하기 위해 인공신경망을 적용하였으며 소셜네트워크의 특징을 적용하여 친구 및 팔로워가 리뷰유용

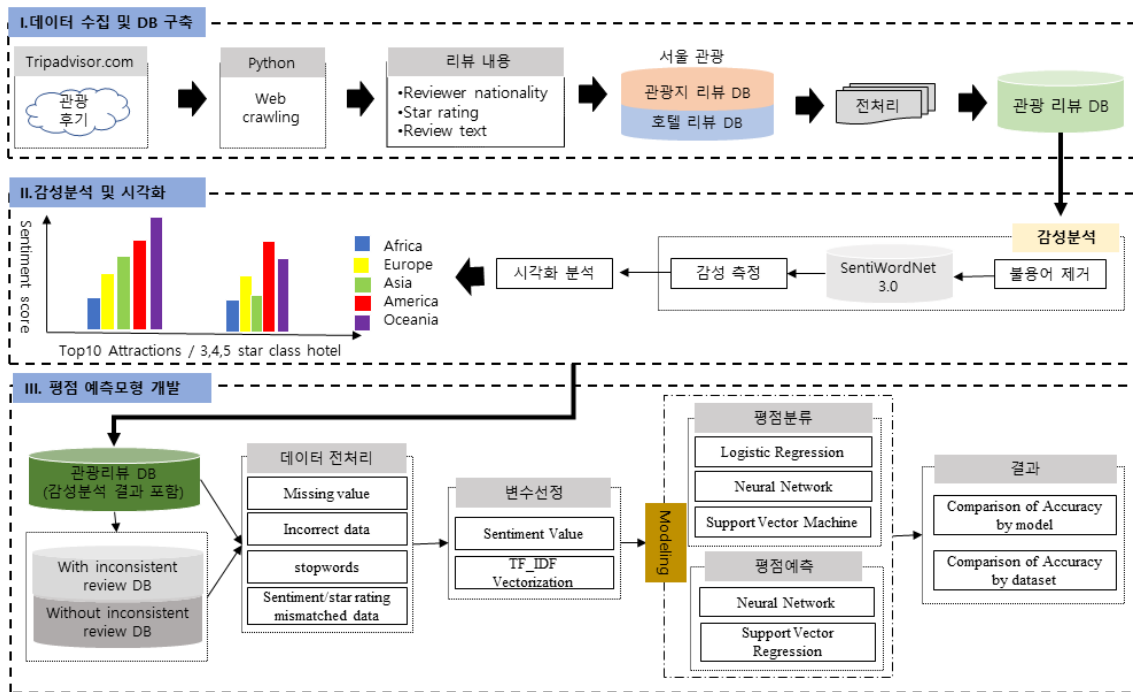
성에 미치는 영향을 조사하였다.

SVM(Support Vector Machine)은 Vapnik(1995)에 의해 제안된 통계기반의 학습기법이다. 분류문제에 사용되는 대부분의 방법들이 경험적 위험 최소화에 기반한 것이라면 SVM은 구조적 위험 최소화에 중점을 두어 일반화가 용이하며 우수한 성능을 보인다(Tay and Cao, 2001). SVR(Support Vector Regression)은 SVM의 회귀모형에 ϵ -무감도 손실함수(ϵ -insensitive loss function)를 도입하여 회귀문제 영역까지 확장한 모형이다(Vapnik et al., 1996). 온라인 리뷰의 유용성과 평점을 예측하고 분류하기 위해서도 많이 활용되고 있다. Zhang & Lin(2018)은 온라인 리뷰에 대한 유용성을 예측하기 위해 SVM을 사용하였으며 우수한 성과를 확인하였다. Al Ajrawi et al.(2021)은 온라인 리뷰에서 평점을 예측하기 위해 로지스틱 회귀분석, SVM 등을 적용하였으며 우수한 예측성능을 보여주었다.

3. 연구 프레임워크

본 연구에서는 평점과 리뷰내용의 불일치성 문제를 해결하고자 온라인 리뷰를 기반으로 한 평점 예측모형을 제안한다. 온라인 리뷰는 평점으로는 파악할 수 없는 구체적인 정보를 포함하고 있어 관광객들의 진실한 의견을 반영할 수 있다. 외국인 관광객이 작성한 온라인 리뷰에 대한 감성분석을 통해 한국관광산업에 대한 관광객의 태도를 파악하며 리뷰의 감성값과 평점을 비교하여 불일치한 리뷰를 제외함으로써 데이터에 대한 신뢰성을 확보한 다음 평점 예측모형을 구축하여 예측성능을 향상시키고자 한다.

연구 프레임워크는 <그림 1>과 같이 데이터 수집과 전처리, 분야별 및 국가별 감성분석, 감성분석 기반의 평점 예측모형 개발의 3단계로 구성된다. 먼저 데이터



〈그림 1〉 연구 프레임워크

수집을 위해 TripAdvisor에서 서울의 관광지과 호텔에 대한 리뷰를 수집한다. TripAdvisor에서 영어로 작성된 리뷰의 내용과 리뷰어의 정보를 수집하고 대륙별로 구분하여 관광 리뷰 데이터베이스를 구축한다. 다음은 불용어 제거와 같은 데이터 전처리 후 SentiWordNet 감성사전을 사용하여 어휘기반의 감성분석을 수행하고, 감성분석 결과값을 통해 평점과 불일치한 리뷰를 선별한다. 만약 긍정적인 리뷰로 분류되었으나 평점이 1 또는 2로 낮게 평가되었다면 리뷰내용과 평점이 불일치한 리뷰로 분류한다. 리뷰어의 국적을 기준으로 대륙별로 감성값과 평점을 시각화하여 비교함으로써 대륙별로 관광지 및 호텔에 대한 태도를 파악한다. 마지막으로 감성분석 결과와 TF-IDF 벡터화의 결과를 통해 입력변수를 선정하고 4가지 모형을 구축한다. 모형1은 불일치한 리뷰를 제외하지 않고 감성분석을 반영한 모형이며, 모형2는 불일치한 리뷰를 제외하지 않고 감성분석을 반영하지 않은 모형, 모형3은 불일치한

리뷰를 제외하고 감성분석을 반영한 모형, 모형4는 불일치한 리뷰를 제외하고 감성분석을 반영하지 않은 모형으로 구축하였다. 평점 분류 모형을 위해서는 로짓, 인공신경망, SVM(Support Vector Machines)을 사용하며 평점 예측을 위해 인공신경망, SVR(Support Vector Regression)으로 모형을 구축한다.

3.1. 데이터 수집

TripAdvisor는 세계적인 여행 사이트로 관광지, 호텔, 레스토랑 등 여행과 관련된 다양한 정보를 제공하고 있다. 본 연구에서는 한국을 방문한 외국인 관광객이 작성한 리뷰 데이터를 수집하였다. 2019년 12월부터 2020년 1월까지 관광지와 호텔에 대해 영어로 작성된 리뷰의 리뷰어 정보, 평점, 리뷰 내용 등을 분석도구 R을 사용하여 크롤링하였다. 문화체육관광부에서 매년 발간하는 외래관광객조사(2020)에 따르면 한국을 방문한 관광객 중 서울 방문이 47.2%로 가장 높게 나타

<표 1> 대륙별로 분류된 한국 방문 외국인 관광객의 리뷰 데이터

대륙	관광지	호텔
아메리카	2,582 (29.92%)	5,344 (23.51%)
아시아	3,757 (43.54%)	11,968 (52.66%)
아프리카	79 (0.92%)	207 (0.91%)
오세아니아	794 (9.2%)	1,825 (8.03%)
유럽	1,417 (16.42%)	3,383 (14.89%)
합계	8,629	22,727

났으며 본 연구에서도 서울 관광지를 대상으로 하였다. TripAdvisor에서 선정한 서울의 상위 10곳에 대한 관광지 리뷰와 3, 4, 5성급의 호텔 리뷰를 대상으로 하였다. 관광지의 리뷰 수는 편차가 크기 때문에 선정한 10곳의 관광지에 대한 리뷰를 균형있게 반영하고자 관광지별로 약 1,000개씩 수집하여 총 11,010개의 리뷰 데이터를 수집하였다. 호텔은 많은 리뷰를 확보하기 위해 리뷰가 많은 호텔을 기준으로 하였으며 3성급 6,929개, 4성급 9,348개, 5성급 7,731개로 총 24,008개의 리뷰 데이터를 수집하였다.

서울은 세계적인 관광지로 다양한 국적의 관광객들이 방문하고 있으며 다양한 문화적 배경을 가진 관광객들은 온라인 리뷰에서도 유의미한 차이를 보인다(야오즈옌 외, 2020). 온라인 리뷰에 대한 관광객들의 차이를 비교하기 위해 리뷰어의 국적이 한국이거나 국적이 없는 리뷰를 제거하였다. 관광지에 대한 리뷰 2,381개와 호텔에 대한 리뷰 1,281개가 제거되어 최종적으로 관광지 8,629개, 호텔 22,727개의 데이터를 분석에 사용하였다. 대륙별로 작성된 리뷰는 <표 1>와 같으며 아시아 관광객이 높은 비중을 차지하고 있으며 아프리카 관광객이 가장 낮은 비중을 보여준다.

3.2. 온라인 리뷰 기반의 감성분석 및 예측모형 개발

온라인 리뷰에서 표현된 의견과 평점이 일치하지 않

는 경우를 종종 확인할 수 있으며 이러한 불일치 문제는 리뷰를 읽는 잠재고객에게 혼동을 줄 수 있고 구매 의사과정에 불확실성을 증가시키는 요인이 된다. 따라서 본 연구는 리뷰내용에 대한 감성분석을 통해 평점과 불일치한 리뷰를 선별하고 리뷰 기반의 평점 예측모형을 개발하고자 한다. 감성분석은 어휘기반의 감성분석을 적용하며 영어 감성사전으로 SentiWordNet 3.0을 적용하였다. 전처리를 통해 각 리뷰의 불용어를 제거하고 SentiWordNet 3.0의 감성 사전을 기반으로 각 리뷰의 감성 값을 계산하고 예측모형 구축에 적용하였다.

구체적인 예측모형 구축과정은 다음과 같다. 먼저 감성분석 결과 긍정적인 리뷰로 분류되었으나 평점이 1 혹은 2인 리뷰와 부정적인 리뷰로 분류되었으나 평점이 4 혹은 5로 나타난 리뷰는 리뷰 내용과 평점이 불일치한 데이터이다. 관광지 리뷰에서는 1,586개, 호텔 리뷰에서는 2,674개의 리뷰가 불일치한 것으로 나타났다. 불일치한 데이터를 제외한 모형에서는 관광지 7,043개, 호텔 20,053개의 데이터를 사용하여 모형을 구축하였다. 이진분류(binary classification)에 대한 모형 구축을 위해 평점이 1, 2인 부정적인 리뷰를 0으로 설정하고, 평점이 4, 5인 긍정적인 리뷰는 1로 설정하였으며 평점이 3인 경우는 제외하였다. 예측모형의 입력변수는 TF-IDF vectorization(단어의 최소빈도가 100 이상인 경우)과 감성분석 결과를 적용하였다. 평점 분류모형은 로짓, 인공신경망, SVM(Support Vector

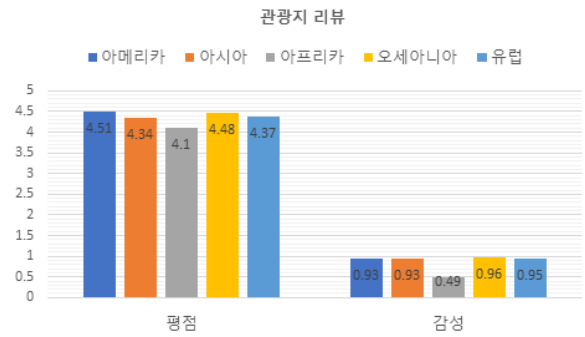
Machine)으로 구축하였으며 평점 예측모형은 인공신경망, SVR(Support Vector Regression)을 적용하였다.

로짓모형에서는 학습용 데이터와 평가용 데이터를 4:1로 분할하였으며 인공신경망은 학습용, 검증용, 평가용 데이터를 3:1:1로 하여 학습률과 모멘텀은 0.1로 설정하였다. 은닉노드의 수는 1에서 30까지 변화시켜가며 성과가 가장 우수한 모형을 선정하였다. SVM과 SVR 모형에서도 학습용, 검증용, 평가용 데이터를 3:1:1로 분할하였으며 RBF 커널함수를 사용하고 상한값을 나타내는 C와 파라미터 σ 와 ϵ 는 그리드 탐색 기법을 적용하여 성과가 가장 우수한 모형을 선정하였다. 예측모형은 python의 sklearn 라이브러리를 사용하였다.

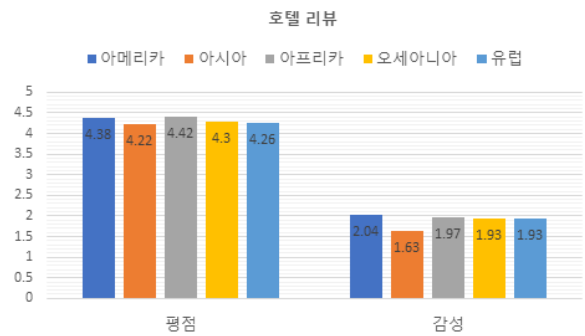
4. 분석결과

4.1. 감성분석

온라인 리뷰에 포함된 감성을 추출하고 평점과 불일치한 리뷰를 선별하기 위해 감성분석을 적용하였다. <표 2>는 한국을 방문한 외국인 관광객을 대륙별로 구분하여 평점 평균값(표준편차)과 감성 평균값(표준편차)의 분석결과를 제시하였으며 <그림 2>와 <그림 3>은 <표 2>에 대한 결과를 시각화한 결과이다. 관광지에 대한 리뷰를 통해 아프리카 관광객이 가장 낮



<그림 2> 대륙별 관광지 리뷰의 평점과 감성



<그림 3> 대륙별 호텔 리뷰의 평점과 감성

은 평점을 주며 아메리카 관광객이 가장 높은 평점을 주는 것을 확인할 수 있다. 또한 텍스트에서도 평점과 동일하게 아프리카 관광객의 감성값이 가장 낮게 나타났다. 다른 대륙과의 차이도 크게 나타났다. 호텔에 대한 리뷰에서는 아시아 관광객이 가장 낮은 평점을 주며 아프리카 관광객이 가장 높은 평점을 주는 것으로 나타났다. 텍스트에서도 아시아 관광객의 감성이 가장 낮았으며 아메리카 관광객의 감성이 가장 높

<표 2> 한국 방문한 외국인 관광객의 리뷰 평점과 감성에 대한 기술통계량

대륙	관광지 (N=8,629)		호텔 (N=22,727)	
	평점(SD)	감성(SD)	평점(SD)	감성(SD)
아메리카	4.51(0.70)	0.93(1.16)	4.38(0.89)	2.04(2.19)
아시아	4.34(0.76)	0.93(1.16)	4.22(0.85)	1.63(1.90)
아프리카	4.10(0.50)	0.49(0.27)	4.42(0.76)	1.97(1.78)
오세아니아	4.48(0.69)	0.96(1.16)	4.30(0.84)	1.93(2.13)
유럽	4.37(0.80)	0.95(1.17)	4.26(0.88)	1.93(2.13)

은 것으로 나타났다. 아프리카 관광객은 관광지와 호텔에 대한 인식 차이가 분명함을 알 수 있다.

4.2. 예측분석결과

4.2.1. 분류모형

본 연구는 평점 분류를 위해 4가지 모형을 구축하였다. 모형1은 불성실한 리뷰를 포함하며 감성점수를 포함한 모형, 모형2는 불성실한 리뷰를 포함하며 감성점수를 포함하지 않는 모형, 모형3은 불성실한 리뷰를 제외하고 감성점수를 포함한 모형, 모형4는 불성실한 리뷰를 제외하고 감성점수를 포함하지 않는 모형으로 로짓, 인공신경망, SVM의 분류기법을 적용하였다. <표 3>은 관광지 리뷰를 이용한 분류 결과이며 <표 4>은 호텔 리뷰를 이용한 분류 결과이다. 관광지 리뷰의 분류 결과에서는 리뷰 내용과 평점이 일치하지 않는 불성실한 데이터 셋을 제외하고 감성점수를 포함한 모형3의 SVM모형에서 예측성도가 99.6%로 가장 높은 것으로 나타났다. 호텔 리뷰의 분류에서도 불성실한 데이터 셋을 제외하고 감성점수를 포함한 모형3의 SVM모형에서 98.0%의 예측성도를 보이며 가장 우수한 것으로 나타났다. 관광지와 호텔 모두 리뷰의 감

성점수와 평점이 일치하지 않은 불성실한 데이터 셋을 제외한 모형3과 모형4의 예측성도가 불성실한 데이터를 포함한 모형1과 모형2보다 예측성도가 우수함을 확인할 수 있었다. 또한 감성점수를 모형에 반영한 모형1과 모형3의 예측성도가 감성점수를 사용하지 않은 모형2와 모형4의 예측성도보다 모두 높은 것을 확인할 수 있다.

4.2.2. 회귀모형

평점을 예측하기 위해서도 분류모형과 동일하게 4가지의 모형을 적용하였으며 인공신경망과 SVR(Support Vector Regression)을 적용하고 RMSE, MAE, MAPE로 모형의 예측성도를 평가하였다. <표 5>는 관광지 리뷰를 사용한 분석결과이며, 불성실한 리뷰를 제외하고 감성점수를 반영한 모형3의 인공신경망과 SVR의 예측성도가 우수하였으며 감성점수를 반영한 모형1과 모형3이 감성점수를 반영하지 않은 모형2와 모형4보다 우수한 것으로 나타났다. <표 6>는 호텔의 리뷰를 사용한 분석결과이며, 불성실한 리뷰를 제외하고 감성점수를 포함한 모형3의 성과가 가장 우수한 것으로 나타났다. 또한 관광지 리뷰와 동일하게 감성점수가 반영된 모형1과 모형3이 감성점수를 반영하지 않은 모형2와

<표 3> 관광지 리뷰를 이용한 분류 결과

	모형1			모형2			모형3			모형4		
	Logit	NN	SVM	Logit	NN	SVM	Logit	NN	SVM	Logit	NN	SVM
학습용	83.0	100.0	76.9	79.9	100.0	74	97.4	100.0	100.0	86.3	100.0	85.9
검증용		80.7	90.7		80.7	89.7		96.1	98.6		92.6	95.4
평가용	97.5	83.3	98.2	96.7	81.2	97.5	98.9	97.5	99.6	97.5	92.4	97.5

<표 4> 호텔 리뷰를 이용한 분류 결과

	모형1			모형2			모형3			모형4		
	Logit	NN	SVM	Logit	NN	SVM	Logit	NN	SVM	Logit	NN	SVM
학습용	94.8	99.8	98.2	94.9	99.8	99.2	99.6	99.8	99.8	98.1	99.8	93.3
검증용		83.1	85.3		83.1	86.2		96.8	98.5		94.1	93.8
평가용	77.2	78.0	83.0	70.8	77.8	81.2	97.4	95.1	98.0	91.3	92.5	91.6

〈표 5〉 관광지 리뷰를 이용한 예측 결과

	모형1						모형2					
	NN			SVM			NN			SVM		
	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse
학습용	0.42	10.95	0.54	0.46	13.55	0.65	0.45	12.30	0.58	0.24	7.84	0.42
검증용	0.61	16.36	0.82	0.53	15.10	0.72	0.58	15.89	0.76	0.53	14.98	0.70
평가용	0.61	15.13	0.77	0.53	13.90	0.68	0.59	14.88	0.74	0.54	14.03	0.68
	모형3						모형4					
	NN			SVM			NN			SVM		
	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse
학습용	0.30	9.38	0.47	0.19	5.29	0.30	0.43	10.76	0.54	0.24	7.11	0.39
검증용	0.57	14.62	0.73	0.51	13.24	0.63	0.62	15.65	0.78	0.53	14.20	0.68
평가용	0.58	14.64	0.74	0.52	13.13	0.64	0.63	15.41	0.77	0.54	14.03	0.68

〈표 6〉 호텔 리뷰를 이용한 예측 결과

	모형1						모형2					
	NN			SVM			NN			SVM		
	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse
학습용	0.44	12.83	0.57	0.51	16.23	0.68	0.34	9.10	0.45	0.35	12.08	0.54
검증용	0.57	17.16	0.76	0.53	16.66	0.70	0.60	17.48	0.82	0.53	16.50	0.69
평가용	0.55	14.77	0.71	0.50	13.71	0.63	0.58	15.66	0.78	0.51	14.04	0.64
	모형3						모형4					
	NN			SVM			NN			SVM		
	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse	mae	mape	rmse
학습용	0.42	10.51	0.51	0.42	11.68	0.55	0.40	10.97	0.51	0.34	10.15	0.49
검증용	0.53	13.99	0.66	0.48	12.96	0.60	0.57	15.09	0.72	0.50	13.89	0.64
평가용	0.54	14.18	0.67	0.49	13.14	0.61	0.58	15.41	0.73	0.51	14.25	0.65

모형4보다 우수한 것으로 나타났다.

5. 결론

본 연구는 빅데이터 분석을 이용하여 한국을 방문한 외국인 관광객이 작성한 온라인 리뷰를 대상으로 감성 분석을 수행하였으며 감성 값을 기반으로 평점 예측모형을 구축하였다. 한국을 방문하는 관광객들의 서울 방문 비중이 가장 높게 나타나기 때문에 서울을 대상

으로 리뷰를 수집하였으며 특정 관광지가 아닌 서울 관광지에 대한 리뷰를 수집하기 위해 TripAdvisor에서 제공하는 서울의 상위 10개에 대한 관광지와 3, 4, 5성급의 호텔에 대해 외국인이 작성한 3만 개 이상의 온라인 리뷰를 수집하였다. 외국인 관광객이 한국 관광산업에 대한 태도를 조사하기 위해 리뷰어의 국적에 따라 대륙별로 감성분석 결과를 비교한 결과 아프리카의 관광객이 한국 관광지에 대해 가장 낮은 평점과 낮은 감성을 보였으며 호텔 리뷰에서는 아프리카 관광객의 평점과 감성값은 다른 대륙과 비슷하게 나타났으며 오

히려 아시아 관광객들이 낮게 평가한 것을 확인할 수 있었다. 관광지 리뷰에 대한 분석을 통해 한국문화와 아프리카 문화 간의 큰 차이가 존재한다는 것을 짐작할 수 있으며 향후 보다 만족할 수 있는 관광 서비스를 제공할 수 있도록 외국인 관광객에게 맞춤형 관광지에 대한 서비스를 제공할 필요가 있을 것이다. 그러나 전반적으로 한국의 관광지와 호텔에 대해 높은 평점을 주고 있으며 감성 값도 긍정적인 결과를 보여 외국인 관광객이 한국관광 산업에 긍정적인 태도를 보이고 있다는 것을 확인할 수 있었다. 다음으로 온라인 리뷰의 내용과 평점의 불일치로 인한 문제를 해결하기 위해 감성값을 기반으로 평점은 높지만 부정적인 리뷰와 평점은 낮으나 긍정적인 리뷰를 선별하였다. 감성점수와 평점 간의 차이가 큰 경우 불성실한 리뷰(*inconsistent review*)로 분류하였으며, 불성실한 리뷰를 포함한 예측모형(모형1, 모형2)과 불성실한 리뷰를 제외한 예측모형(모형3, 모형4)의 성과를 비교한 결과 불성실한 리뷰를 제외한 모형의 예측성도가 높은 것을 확인할 수 있었다. 또한 감성점수를 평점 예측 변수로 활용할 경우 보다 향상된 성과를 확인할 수 있었다.

본 연구의 시사점은 다음과 같다. 첫째, 관광산업의 정보 과부화 문제를 해결하기 위해 빅데이터를 적용하여 유용한 정보를 선별하여 제공할 수 있게 하였다. 온라인 플랫폼에서 너무 많은 관광정보가 제공되고 있으며 너무 많은 온라인 리뷰는 소비자의 신뢰와 구매의도에 부정적인 영향을 미치기 때문에 빅데이터 분석을 적용하여 정보 과부화의 문제를 해결하였다. 더불어서 관광산업의 지식경영의 방향성은 스마트 관광에 초점을 두고, 빅데이터를 활용한 관광지식의 발굴을 위한 방안을 모색할 수 있었다. 둘째, 온라인 리뷰의 감성과 평점을 비교하기 위해 리뷰어의 국적을 활용하여 한국의 관광산업에 대한 외국인 관광객들의 태도를 대륙별로 비교하였다. 다양한 국가에서 방문

하는 세계적인 관광지로 거듭나기 위해 대륙별로 한국의 관광산업에 대한 인식을 파악하였으며 이는 맞춤형 관광상품 개발에 유용한 정보를 제공할 수 있을 것이다. 셋째, 온라인 리뷰와 평점의 불일치 문제를 해결하기 위해 불성실한 리뷰는 제외하고 예측모형을 구축하였다. 온라인 플랫폼에 제시되어 있는 많은 리뷰를 모두 확인하기 어렵기 때문에 평점은 소비자의 구매의사결정에 중요한 역할을 하고 있지만 리뷰와 평점의 불일치 문제가 존재한다. 본 연구에서는 불일치한 리뷰를 제외하여 리뷰의 신뢰도 문제를 해결하여 예측모형을 구축하였다. 또한 리뷰의 감성값을 예측모형에 반영하여 관광객의 의견을 반영한 평점 예측모형을 구축하여 예측성도를 향상시키고자 하였다. 하지만 본 연구에서 제안한 평점을 예측하는 모형은 평점과 리뷰의 내용이 불일치한 경우에만 유용한 것으로 향후 연구에서는 관광지와 호텔의 세부적인 측면에 대한 평가를 예측할 수 있는 모형을 개발할 필요가 있다. 또한 본 연구에서는 3만 개 이상의 온라인 리뷰를 사용했지만 대륙별로 큰 편차를 보이며 아프리카 관광객의 리뷰가 너무 적기 때문에 평점과 감성 값에 영향을 미칠 수 있을 것이다. 따라서 향후 연구에서는 연구 결과에 대한 검증과 일반화를 위하여 보다 많은 온라인 리뷰를 수집하여 분석할 필요가 있으며 대륙별로 온라인 리뷰의 수를 균형있게 수집할 필요가 있을 것이다.

〈참고문헌〉

[국내 문헌]

1. 김은미 (2021). 감성분석을 이용한 뉴스정보와 딥러닝 기반의 암호화폐 수익률 변동 예측을 위한 통합모형. *지식경영연구*, 22(2), 19-32.
2. 문화체육관광부 (2021). 2020 외래관광객조사.
3. 야오즈옌, 김은미, 홍태호 (2020). 온라인 리뷰의 텍스트 마이닝에 기반한 외국인 관광객의 문화적 특성 연구. *정보시스템연구*, 29(4), 171-191.
4. 야오즈옌, 박지영, 홍태호 (2021). 레스토랑의 온라인 리뷰를 통해 감성과 감정이 리뷰 유용성에 미치는 영향에 관한 연구. *지식경영연구*, 22(1), 243-267.
5. 이주민, 방정혜 (2020). 화장품 회사의 브랜드컨셉 개발 사례분석. *지식경영연구*, 21(3), 215-228.

[국외 문헌]

6. Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. *International Journal of Hospitality Management*, 80, 52-77.
7. Al Ajrawi, S., Agrawal, A., Mangal, H., Putluri, K., Reid, B., Hanna, G., & Sarkar, M. (2021). Evaluating business Yelp's star ratings using sentiment analysis. *Materials Today: Proceedings*.
8. Bilal, M., Marjani, M., Hashem, I. A. T., Malik, N., Lali, M. I. U., & Gani, A. (2021). Profiling reviewers' social network strength and predicting the "Helpfulness" of online customer reviews. *Electronic Commerce Research and Applications*, 45, 101026.
9. Furner, C. P., & Zinko, R. A. (2017). The influence of information overload on the development of trust and purchase intention based on online product reviews in a mobile vs. web environment: An empirical investigation. *Electronic Markets*, 27(3), 211-224.
10. Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism*

Management, 59, 467-483.

11. Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3), 674-684.
12. Hu, Y. H., Chen, K., & Lee, P. J. (2017). The effect of user-controllable filters on the prediction of online hotel reviews. *Information & Management*, 54(6), 728-744.
13. Kim, K., Park, O. J., Yun, S., & Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change*, 123, 362-369.
14. Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7), 3751-3759.
15. Lee, P. J., Hu, Y. H., & Lu, K. T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*, 35(2), 436-445.
16. Luo, Y., & Xu, X. (2021). Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *International Journal of Hospitality Management*, 94, 102849.
17. Mahadzir, N. H., Omar, N. F., Nawim, N. M., Salameh, A. A., & Hussin, K. C. (2021). Sentiment analysis of code-mixed text: A review. *Turkish Journal of Computer and Mathematics Education*, 12(3), 2469-2478.
18. Prameswari, P., Surjandari, I., & Laoh, E. (2017, October). Opinion mining from online reviews in Bali tourist area. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, IEEE, 226-230.
19. Quintal, V. A., Lee, J. A., & Soutar, G. N. (2010). Risk, uncertainty and the theory of planned behavior: A tourism example. *Tourism Management*, 31(6), 797-805.
20. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-based Systems*, 89, 14-46.

21. Sharma, A., Park, S., & Nicolau, J. L. (2020). Testing loss aversion and diminishing sensitivity in review sentiment. *Tourism Management, 77*, 104020.
22. Sotiridais, M. D., & Van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: The use of twitter by tourists. *Electronic Commerce Research, 13*(1), 103-124.
23. Swar, B., Hameed, T., & Reychav, I. (2017). Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. *Computers in Human Behavior, 70*, 416-425.
24. Tay, F. E. H., & Cao, L. J. (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*(4), 309-317.
25. Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tourism Management, 80*, 104122.
26. Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.
27. Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems, 9*.
28. Wu, J. J., & Chang, S. T. (2020). Exploring customer sentiment regarding online retail services: A topic-based approach. *Journal of Retailing and Consumer Services, 55*, 102145.
29. Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51-65.
30. Zang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research, 116*, 16-32.
31. Zhang, Y., & Lin, Z. (2018). Predicting the helpfulness of online product reviews: A multilingual approach. *Electronic Commerce Research and Applications, 27*, 1-10.
32. Zheng, T., Wu, F., Law, R., Qiu, Q., & Wu, R. (2021). Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach. *International Journal of Hospitality Management, 92*, 102658.

● 저 자 소 개 ●



홍 태 호 (Taeho Hong)

현재 부산대학교 경영학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공하여 공학석사와 공학박사를 취득하였다. 주요 관심분야는 비즈니스 애널리틱스, 딥러닝, 오피니언 마이닝, CRM 등이다. 주요 논문을 Expert Systems, Expert Systems with Applications, Information Processing & Management, Asia Pacific Journal of Information Systems, 정보시스템연구 등에 게재하였다.

〈 Abstract 〉

Sentiment Analysis and Star Rating Prediction Based on Big Data Analysis of Online Reviews of Foreign Tourists Visiting Korea

Taeho Hong *

Online reviews written by tourists provide important information for the management and operation of the tourism industry. The star rating of online reviews is a simple quantitative evaluation of a product or service, but it is difficult to reflect the sincere attitude of tourists. There is also an issue; the star rating and review content are not matched. In this study, a star rating prediction model based on online review content was proposed to solve the discrepancy problem. We compared the differences in star ratings and sentiment by continent through sentiment analysis on tourist attractions and hotels written by foreign tourists who visited Korea. Variables were selected through TF-IDF vectorization and sentiment analysis results. Logit, artificial neural network, and SVM(Support Vector Machine) were used for the classification model, and artificial neural network and SVR(Support Vector regression) were applied for the rating prediction model. The online review rating prediction model proposed in this study could solve inconsistency problems and also could be applied even if when there is no star rating.

Key Words: Online review, Star rating, Sentiment analysis, Big data, Prediction model

* Pusan National University College of Business Administration