NLP 기반 여행 리뷰 분류 및 추천 시스템 설계

홍영민, 박영덕1)
영남대학교 컴퓨터공학과
mjdal0523@gmail.com, ydpark@yu.ac.kr

NLP-based Travel Review Classification and Recommendation System Design

Hong Youngmin and Young Deok Park
Dept. of Computer Engineering, Yeungnam University

요 약

Covid19의 세계적 유행 이래로 긴 일정의 해외여행이 감소하고 국내 여행의 수요가 꾸준히 증가하는 추세이다. 현재 다수의 국내 여행 숙박 플랫폼은 가성비 측면으로 이용자가 숙박업소를 선택하고 소비자와 업체를 연결해주는 과정에서 수수료를 얻는 상업적 모델이다. 본 논문에서는 가격 경쟁 중심의 기성 시스템이 아닌, 여행자 개인의 가치를 맞춤화하고 공익의 목적으로 업체를 홍보하는 시스템을 제안한다. 이 시스템은 웹 기반의 시스템을 구현하여 여행자에게 개인 가치에 맞는 업소를 맞춤형으로 추천하고 해당 업소에 대한 평가 지표를 시각화하여 제공한다. 본 시스템은 맞춤형 업소 추천과 평가 지표 제공을 위해 소비자의 리뷰 데이터를 사용한다. 텍스트 데이터를 분석하고 해당 데이터를 다중 분류를 통해 업소에 대한 평가 지표별 점수를 산정한다. 본 시스템은 여행자에게 다양한 관광지와 관광 업소를 추천함으로써 지역 관광을 유도하고 해당 여행지 업소와 지역 경제에 도움을 줄 것이라고 기대된다. 본 논문에서 제안된 기법은 오픈소스로 공개되었다[1].

1. 서론

사회적 거리두기로 인해 여행 수요가 많이 억압받다가 점차 국내 여행의 수요가 꾸준히 증가하는 추세이다[2]. 국내 OTA 서비스에서는 대부분 숙박업소에 대한 소비자 리뷰 시스템을 운영하고 있으며 인기 있는 업소의 경우 수천, 수만 개의 리뷰가 데이터로 쌓인다. 이 리뷰 데이터는 자연어로 구성되어 있으며 데이터의 활용 방안에 따라 서비스 활용 가치는 무궁무진하다[3].

본 논문에서는 텍스트 다중 분류를 사용해 숙박업소에 대한 소비자 리뷰를 6가지 범주[안전, 가격만족도, 재방문/지인추천, 서비스, 시설/편의, 위생/청결]로 분류하는 기법을 제안한다[4]. 해당 기법을 이용하여 소비자는 관광 업소별 평가 지표를 점수로 확인할 수 있으며 본인의 여행 가치에 맞게 추천받을 수도 있다. 제안하는 기법은 다음 다섯 가지의 과정으로 구성된다. 첫 번째, 소비자 리뷰 데이터를 형태소 분석하고 단어 간 유사도를 구한다. 두 번째, 단어들의 6가지 범주에 따른 감성 사전을 제작한다. 세 번째, 감성 사전을 바탕으로 리뷰에 라벨링을 하고 LSTM(Long Short Term Memory) 기반 딥러닝 모델에 학습시킨다. 네 번째, 학습된 모델은 새로운 소비자 텍스트 리뷰의 입력에 대해 6가지 범주로 예측하고 모델을 평가한다. 마지막으로, 해당 범주로 예측된 소비자 리뷰를 통해 각 범주별 점수를 도출해 소비자에게 숙박업소에 대한 시각적 자료를 제공한다.

2. 시스템 설계 및 구현

2.1. 리뷰 데이터 수집 및 전처리

학습 데이터로 활용하기 위해 국내 숙박 업소에 대한 소비자 리뷰 정보를 약 94,000건을 크롤링하여 수집하였다. 해당 데이터는 숙박업소, 소비자 리뷰, 별점 Attribute를 포함한다. 표 1은 수집된 소비자 리뷰 데이터의 예이다. 소비자 리뷰 Column의 텍스트 데이터를 살펴보면 “비상시”, “대피” 등의 안전에 대한 정보, “가격대비”, “가성비” 등의 가격만족도에 대한 정보, “재방문”, “강추” 등의 재방문/지인추천에 대한 정보, “직원”, “친절” 등의 서비스에 대한 정보, “수영장”, “주차장” 등의 시설/편의에 대한 정보, “깨끗”, “더럽” 등의 위생/청결에 대한 정보 등을 살펴볼 수

1) Corresponding author: 박영덕
이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1G1A1095238). 이 연구는 2023년도 영남대학교 학술연구조성비에 의한 것임.

있다.

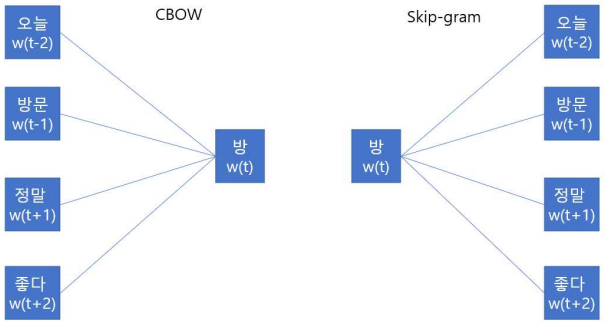
<표 1> 소비자 리뷰 크롤링 데이터

name	consumer reviews	rating
A 업소	아주 늦은시간 도착했는데 직원분 친절하시고, 객실정비도 잘...	4.50
A 업소	호텔도 깨끗하고 비상시 대피 안내문도 잘 붙여져 있네요...	4.70
B 업소	가격대비 만족합니다. 주차장과 수영장, 헬스장 시설이...	4.80
B 업소	유리창이 너무 더러워서 오션뷰는 포기해야했습니다. 재방문...	3.00
C 업소	호텔도 깨끗하고 비상시 대피 안내문도 잘 붙여져 있네요. 강추...	5.00
C 업소	기대없이 갔는데 가성비가 좋았어요..	4.10

본 연구에서는 수집된 리뷰를 정보 추출에 용이한 형태로 바꾸기 위해 맞춤법 수정, 형태소 분석, 불용어 제거 과정을 진행하였다.

2.2. 단어 유사도 측정

본 연구에서는 Word Embedding(워드 임베딩) 기법을 이용해 단어 내 유사도를 도출하였다. 워드 임베딩은 단어를 벡터로 바꾸는 기법 중 하나로 특정 단어를 주변 단어를 이용하여 유사도로 대체한다. 워드 임베딩은 대량의 문서를 학습하여 모든 단어 간 유사도를 복수의 차원으로 기록하고 컴퓨터가 특정 단어를 숫자의 나열 즉, 벡터로 인식할 수 있게 한다. Word2Vec은 워드 임베딩의 대표적인 기법으로 두 가지 하위 모델(CBOW, Skip-gram)로 구분된다.



(그림 1) Word2Vec 기법

그림 1은 "오늘 방문한 방 정말 좋다" 데이터에 대해 CBOW와 Skip-gram의 동작상 차이를 보여준다. CBOW는 주변에 존재하는 문맥 단어(Context Word)을 이용해 타겟 단어(Target Word)를 예측하는 반면, Skip-gram은 타겟 단어를 이용해 주변 문맥 단어를 예측한다. Word2Vec은 문장 맥락에 따라 가깝게 등장하는 단어들을 벡터 공간상에서 가깝게 위치시킨다. 본 연구에서는 Skip-gram 기법을 이용하여 리뷰에 나타난 단어들을 다차원 공간에 벡터화한다. 표 2는 타겟 단어에 대해 문맥 단어와의 유사도를 측정한 예를 보여준다. 도출된 유사도를 이용하여 유사도가 높은 단어끼리 묶어 범주화하고 감성 사전을 제작한다.

<표 2> 단어 간 유사도 측정

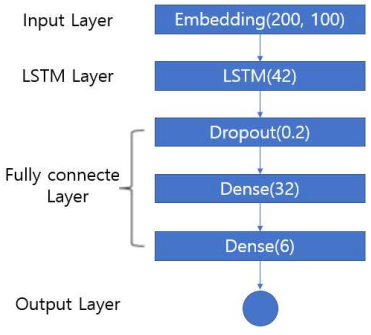
target word	context word	similarity
위생	청결	0.63
	상태	0.58
서비스	편의	0.54
	친절	0.63

2.3. 감성 사전 제작 및 감성 분류

감성 분류는 텍스트마이닝의 한 분야로, 문서에서 나타나는 텍스트를 통해 유용한 정보를 얻기 위해 문서상의 생각과 의견을 분류하기 위한 기법이다[5]. 본 논문에서는 유사도가 높은 단어별로 묶어 국내 숙박 업소에 대한 6가지 범주[안전, 가격만족도, 재방문/지인추천, 서비스, 시설/편의, 위생/청결]로 감성 사전을 구축하였다.

2.4. LSTM 모델 학습 및 평가

Long Short Term Memory(LSTM)은 RNN 모델에 Long-Term Memory를 관리하는 게이트를 추가하여 장기의존성 문제를 보완한 모델이다. LSTM 모델은 시계열 데이터를 학습하고 예측할 때 주로 쓰이며 문장 속 과거 단어는 미래 단어를 예측하는 데 큰 영향을 준다. 본 연구에서는 리뷰 데이터를 워드 임베딩을 통해 벡터화시켜 입력 데이터로 사용하고 6개의 범주에 대해 출력하는 LSTM 모델을 사용한다.



(그림 2) LSTM 모델 정의

그림 2는 본 논문에서 사용된 LSTM 모델의 구성을 보여준다. 임베딩층, 순환층, 은닉층, 출력층으로 구성되며 출력층에서의 활성화 함수는 Softmax를 이용하여 정규화하여 결과값을 출력하였다. 학습된 해당 모델을 통해 새로운 리뷰를 입력하면 6가지 범주에 대한 예측을 수행한다.

<표 3> 새로운 입력 리뷰 데이터에 대한 예측

new input review	prediction
“휴게 시설이 편리하고 좋습니다”	4(시설/편의)
“객실 청소가 잘 되어 있습니다.”	5(위생/청결)
“가격이 너무 비싸요”	1(가격만족도)

표 3은 새로운 입력 리뷰를 입력했을 때, 도출되는 예측값의 예를 보여준다. 새로운 리뷰가 입력되었을 때, 형태소 분석을 통해 자동으로 [“휴게”, “시설”, “너무”, “편리”, “좋다”]와 같이 토큰화되고 이에 대해 6가지 범주 예측을 진행하는 파이프라인을 통해 숙박업소에 대한 리뷰를 분류한다. 이후 분류된 데이터를 이용해 소비자가 원하는 여행 가치에 맞는 숙박업소를 추천하고 해당 숙박업소에 대해 범주별 리뷰 점수를 도출해 소비자에게 시각적인 점수로 제공한다.

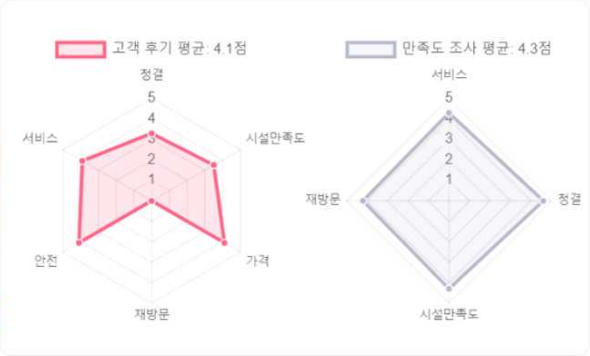
3. 시스템 구현 결과 및 활용

국내 여행 숙박업소에 대해 소비자들이 부여한 별점은 해당 업소의 세부 항목에 대한 평가가 아닌, 해당 업소에 대한 전체 평점을 의미한다. 따라서, 이러한 전체 평점만을 제공하는 경우 해당 숙박업소의 항목별 평가를 한눈에 파악하는 것이 어렵다. 이러한 한계를 해결하기 위해 이전 장에서 진행한 6가지 범주에 대한 리뷰 다중 분류를 이용하여 숙박업소에 대한 세부 평가 평점을 제공한다.

<표 4> 범주별 세부 평점

name	total	safety	revisit	price	service	facility	clean
A	4.5	4.5	3.7	4.2	4.6	3.9	4.1
B	4.8	4.8	4.5	4.9	4.7	4.4	4.7
C	3.5	3.7	3.9	2.3	3.7	3.5	3.2
D	4.0	4.1	4.1	4.2	4.1	3.8	3.9
E	3.7	4.1	3.7	3.3	3.5	4.0	3.9

표 4는 업소에 대한 전체 평점과 6가지 범주별 세부 평점을 나타내고 있다. 이를 기반으로 세부 평점 지표를 그림 3과 같이 레이더 차트로 시각화하여 시스템상에서 소비자에게 제공하여 소비자가 숙박업소에 대해 세부적으로 판단하고 본인의 여행 가치에 맞게 선택할 수 있게 한다.



(그림 3) 업소에 대한 범주별 점수 레이더 차트

4. 결론

본 논문에서는 AI 기반 사용자 맞춤형 여행 비서 시스템을 설계 및 구현하였다. 설계한 시스템은 텍스트 다중 분류를 통해 소비자 숙박업소에 대한 평가 지표별 점수를 산정하고 6가지 범주에 대해 세부 지표를 레이더 차트로서 시각화하여 소비자에게 제공한다. 또한, 소비자 개인 가치에 맞는 숙박업소를 추천하여 가격 중심의 선택지를 탈피할 수 있게 한다. 향후 연구로는 숙박업소 주변 관광지와 맛집에 대한 분석을 진행하고 숙박업소와 함께 추천할 수 있는 시스템을 설계할 예정이다. 또한, 실시간으로 입력되는 리뷰 데이터를 다시 학습데이터로 사용할 수 있게 파이프라인을 설계하는 연구를 수행할 예정이다.

참고문헌

[1] github.com/youngmin0523/KTO_Pjt_using_NLP, GitHub Open Source, 2023

[2] WISEAPP, 코로나 3년차, 국내 여행 시장 돌아보기: 야놀자, 여기어때 중심으로, 2022년 6월 29일

[3] Chevalier, J. A. and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," Journal of Marketing Research, Vol.43, NO.3, pp.345-354, 2006

[4] 전병길 and 허서영, "공유숙박 품질 측정 척도의 개발", 관광학연구, vol.46, no.4, pp.11-29, 2022

[5] Park, Hyun-jung, and Kyung-shik Shin. “Aspect-Based Sentiment Analysis Using BERT: Developing Aspect Category Sentiment Classification Models.” Journal of Intelligence and Information Systems, vol. 26, no. 4, 한국지능정보시스템학회, pp. 1 - 25, Dec. 2020