머신러닝 기반 외식 서비스(식당) 리뷰 감성 분석 및 만족도 지수 개발

AI융합학부 20243305 하정훈

1. 문제 정의

가. 최종 목표

본 프로젝트는 네이버 지도 등 지도 서비스에서 카페, 식당 등의 외식 장소를 선택할 때 사용자가 겪는 현실적인 불편함에서 시작한다.

1) 샘플링 편향

일본 고베 여행에서 방문한 별점 4.3의 야키니쿠 식당에서 불친절한 서비스로 불만족스러운 경험을 한 적이 있다. 리뷰를 조금 더 살펴보니 서비스에 불만을 토로하는 외국인의 후기가 심심찮게 있었다. 한편, 3,000여 개의 후기를 보유한 경기도 안산시의 한 정육식당은 몇몇 상위 노출된 긍정적인 후기를 제외하고는 부정적인 리뷰를 볼 수 있었다.

이처럼 사용자는 평균 별점과 상위에 노출된 몇 개의 리뷰만을 확인하게 되므로 예상과 다른 부정적인 경험을 하는 경우가 빈번하다. 또한, 긍정 후기 몇 개만을 보여주므로, 전체적인 부정 문맥을 반영하지 못하는 샘플링 시스템의 문제를 보여준다.

2) 정보 과부하

모든 리뷰 텍스트를 읽고 장소의 실제 만족도를 파악하는 것은 쉽지 않다. 인기 점포의 경우, 수백~수천 개의 리뷰가 존재한다. 외국의 특정 지점은 수만 개의 리뷰가 존재하기도 한다. 사용자는 모든 리뷰를 읽고 실제 만족도를 파악할 수 없다.

3) 객관성 확보의 어려움

별점 시스템은 별점의 평균만을 보여준다. 평균에는 리뷰의 개수, 협찬 댓글, 사용자의 주관 및 실수(반어법 등의 사용자 의도, 별점 클릭 실수 등의 노이즈 데이터)이 반영될 수 있다. 4~5점의 긍정적인 후기에도 부정적인 문맥은 숨어있을 수 있다. 긍정과 부정의 정도를 텍스트에서 확률 기반으로 파악하여 전반적인 후기의 긍정 정도를 파악할 필요가 있다.

본 프로젝트는 이러한 문제를 머신러닝으로 해결하는 데 최종 목표가 있다. Kaggle의 대규모 리뷰 데이터(zomato, yelp 등)를 활용하고 텍스트의 실제 문맥을 기반으로 한 객관적인 만족도 지수를 개발한다. 이를 평균 별점이 놓칠 수 있는 속뜻을 정량화하여 사용자의 의사결정을 돕는 신뢰할 수 있는 지표를 제시할 수 있다.

나. 연구 가설

1) 별점은 일부 노이즈를 포함하지만, 리뷰 텍스트 본문의 긍/부정 문맥과는 매우 강한 상관관계를 가질 것이다. 리뷰 텍스트 본문의 긍/부정 레이블 데이터로서 별점을 활용할 수 있다.
2) 고전적인 방법의 TF-IDF와 같은 방식보다 BERT와 같은 딥러닝 임베딩을 특징 추출기를 사용하고 머신러닝 분류기에 결합한 하이브리드 모델이 텍스트 감성 분류에서 더 높은 성능을 보일 것이다.

다. 연구 문제 설정

1) Kaggle의 대규모 다국어(영어/일본어) 텍스트를 머신러닝이 학습할 수 있도록 어떤 정제·전처리 전략을 세울 것인가?(인코딩 문제, 별점 형식 통일 문제, 데이터 병합 문제 등)
2) 별점 3점과 같은 모호한 데이터는 어떻게 처리할 것인가?
3) 텍스트와 별점이 일치하지 않는 노이즈 데이터는 어떻게 처리할 것인가?
4) 텍스트를 벡터로 변환하는 주요 방식(TF-IDF, 딥러닝 임베딩 등) 간 성능 차이는 어떠한가?
5) 다양한 머신러닝 분류기 중 어떤 모델이 텍스트 기반 감성 분류에 가장 적합한가?
6) 최종 모델을 활용하여 별점보다 객관적인 텍스트 기반 만족도 지수를 산출할 수 있는가?

라. 머신러닝 Task

1) 이진 분류

리뷰 텍스트 데이터를 기반으로 해당 리뷰의 별점을 가공하여 만든 긍정(1) 또는 부정(0) 레이블을 예측한다.

2. 데이터 정의

가. 필요 데이터

1) 모델 훈련용 데이터
- 리뷰 텍스트 본문(X, 피처) : 모델이 학습할 텍스트 데이터
- 별점(y, 타겟) : 긍/부정 레이블을 생성할 데이터

나. 데이터 확보 방안

1) 출처 : Kaggle 오픈 데이터셋

2) 세부 프로젝트
- 영어권 외식 서비스 리뷰 감성 분석

Zomato Bangalore Restaurants 데이터셋을 메인 분석 대상으로 선정하였다.

- 일어권 외식 서비스 리뷰 감성 분석

추가적인 프로젝트로 메인 프로젝트(영어권)에서 구축한 모델이 비영어권인 일본어 데이터에도 적용이 되는지 확장 실험을 진행할 예정이다. 데이터로는 Tokyo Restaurant Reviews on Tabelog를 활용할 예정이다.

다. 데이터 상세 명세

1) Zomato Bangalore Restaurants

컬럼명	데이터 타입	설명
url	object	zomato 사이트 내의 식당 url
address	obejct	식당 주소

name	obejct	식당 이름
online_order	obejct	온라인 주문 가능 여부
book_table	obejct	좌석 예약 가능 여부
rate	obejct	식당의 전체 평점(5점 기준)
votes	int64	해당 기간의 평점 개수
phone	obejct	식당 전화번호
location	obejct	식당이 위치한 동네 정보
rest_type	obejct	식당 종류
dish_liked	obejct	개별 고객 선호 메뉴
cuisines	obejct	식당 취급 요리 종류
approx_cost (for two people)	obejct	2인 기준 적정 비용
reviews_list	obejct	개별 후기의 평점과 텍스트 본문
menu_item	obejct	메뉴 항목
listed_in(type)	obejct	웹사이트 내 분류 : 식당 종류
listed_in(city)	obejct	웹사이트 내 분류 : 소재 도시

2) Tokyo Restaurant Revies on Tabelog

컬럼명	데이터 타입	설명
store_id	int64	URL에서 얻어지는 고유 식별 ID
name	object	식당 이름
nearest_station	object	인근 역
nearest_distance	float64	인근 역에서의 거리(m)
genre	object	식당의 종류
rating_val	float64	타베로그 내 전체 평점
review_cnt	float64	식당 후기 개수
save_cnt	float64	식당 '찜(저장)' 개수
budget_dinner	object	저녁 식사 평균 예산
budget_lunch	object	점심 식사 평균 예산
holiday	object	식당 휴무일 정보
address	object	식당 주소
prefecture	object	식당 지방 행정 구역(도쿄)
municipalities_1	object	지방 행정 구역(구)
municipalities_2	object	지방 행정 구역(시)
municipalities_3	object	지방 행정 구역(해당 데이터셋 내 없음)