빅데이터 프로젝트

광고성 리뷰 파악 데이터 수집

Prepared by: Chanyoung Park

Date of Report Creation: May 13, 2025

연구 목표

많은 사용자들이 커머스 서비스를 이용할 때 구입할 물품에 대한 리뷰를 보고 실제 구매로 이어집니다.
그만큼 상품에 대한 '리뷰'는 사용자들이 물품을 구매하게 만드는 중요한 척도로 작용합니다.

하지만 이러한 특성을 이용하여 유명인이나 업체에 **실제 사용자인 것 처럼 조작 된 리뷰 작성을 위탁하고 금전적 이득을 취하는 케이스**도 다수 발생하고 있어 사용자가 리뷰를 온전히 신뢰하기 힘든 상황도 자주 발생합니다.

문제점	발생 비율
영국 전자상거래 플랫폼의 허위리뷰 비율 (가전, 주방용품, 스포츠용품 카테고리)	11 ~ 15%
소비자 결정 영향력	3.1%
가짜 리뷰로 인한 소비자 피해 추정액	50M EUR ~ 300M EUR

- 소비자의 ***82.4%***가 리뷰 누적 수에 따라 구매를 결정합니다.
- 리뷰가 없거나 부정적일 경우 **구매를 포기하는 비율**이 높아 리뷰의 중요도도 매우 높습니다.
- 긍정적인 리뷰가 많은 상품의 경우 **플랫폼 알고리즘에서 상위에 배치될 수도 있기에 다른 선량한 자영업자의 금전적인 피해로 발전할 위험**이 존재합니다.

이번 연구의 목표는 다음과 같습니다.

- 상품의 *URL*을 기입하면 해당 상품의 리뷰를 수집할 수 있는 크롤링 프로그램
- 수집 된 리뷰를 *ML* 학습용 데이터로 가공하는 프로그램

[레퍼런스]

- *https://www.gov.uk/government/publications/investigating-the-prevalence-and-impact-of-fake-reviews*
- *https://www.dailycnc.com/news/articleView.html?idxno=209683*

데이터 수집 방안

국내에서 가장 많이 사용 되는 온라인 커머스 서비스는 1위가 쿠팡, 2위가 네이버입니다.
이 두개의 서비스에 업로드 된 상품의 리뷰를 수집하여 아래와 같은 데이터 구조로 정리합니다.

Key	Description
상품명	상품의 이름
작성자	유저 이름 (마스킹 되어 있을 가능성 존재)
작성일	리뷰 작성 시점
별점	1 ~ 5점
리뷰 내용	실제 리뷰 텍스트
상품 ID	상품 URL에서 추출 가능한 고유 ID

크롤링 방지 모듈이 적용된 사이트의 경우

- 크롤러의 User-Agent 값을 조정하여 구글 크롤러 봇이나 일반 사용자로 위장합니다.
- 요청간 고의적인 지연을 랜덤으로 설정하여 방지 모듈이 봇으로 판단하지 못하게 합니다.

조작 된 리뷰를 판단하는 척도

조작 된 리뷰는 다음과 같은 특성을 가질 수 있다고 가정합니다.

특성	설명
과도하게 긍정적	“완전 강추!! 인생템 ㅠㅠㅠ”과 같은 불균형 표현이 많은지 판단합니다.
반복 문구	“배송 빠르고 좋아요”와 같은 형식적인 문구가 다수의 리뷰에서 발견이 되는지 판단합니다.
짧은 리뷰	리뷰가 해당 포털 사이트의 최소 글자수 기준에만 부합하는지 판단합니다.
리뷰 폭주 현상	특정 시간대에 밀집하여 작성 된 리뷰인지 판단합니다.
사용자 패턴 이상	같은 사용자가 모든 제품에 항상 5점을 부여하는지 확인합니다.

ML모델 학습용 데이터 가공 처리

위에서 수집 된 상품별 데이터를 기반으로 ML 모델 학습용 데이터시트를 제작합니다.

Key	Description
review_text	리뷰 본문
rating	별점
length	리뷰 길이
emotion_score	감정 점수 (자연어 감성 분석기 사용)
duplicate_flag	반복 문구 여부
user_review_count	해당 사용자의 총 리뷰 수
user_avg_rating	사용자의 평균 평점
product_review_count	해당 상품의 총 리뷰 수
review_time	작성 시점
dense_time_flag	특정 시간대에 몰려 작성 된 리뷰가 맞는지

- 리뷰 본문은 전처리 후 KoBERT 기반의 감성 분석기로 긍정 점수를 추출하여 `emotion_score`에 저장합니다.
- 반복 문구 여부는 TF-IDF 기반 유사도 분석으로 판단하며, 일정 유사도 이상일 경우 `duplicate_flag = 1` 처리합니다.
- `dense_time_flag`는 리뷰 작성 시간의 시간대 밀집도를 판단하여, 30분 이내 10건 이상 집중된 경우를 기준으로 1로 처리합니다.

데이터 사용 예상 시나리오

본 데이터는 리뷰의 조작 여부가 확실하지 않아 **해당 내용을 별도로 기재하지 않습니다.**
그렇기에 실제 데이터로 활용하기 위해서는 비지도학습을 사용해야 합니다.

✅ 활용 가능 모델

- KMeans
- DBSCAN
- Hierarchical Clustering

💡 활용 피처

- length, emotion_score, rating, user_avg_rating, duplicate_flag
- review_text는 TF-IDF 기반 벡터로 변환하여 결합

🎯 목적

- 특정 클러스터에 **조작 리뷰 의심 특징이 다수 포함되면**, 해당 클러스터 전체를 의심 대상으로 분류
- ML 모델 없이 **탐색적 분석(EDA) 및 Rule 정제**에도 유용