한국빅데이터학회지
제10권 제1호, 2025, pp. 47-62 https://doi.org/10.36498/kbigdt.2025.10.1.47

유튜브 기반 사용자 콘텐츠에서의 리뷰 이상 탐지: 사용자 생성 콘텐츠에서의 극단성과 허위성 분류

Decoding Review Anomalies: Classifying Extremity and Falsity in User-Generated Content

가중정[1] · 김엘레나[2] · 최재원[3†]

순천향대학교 경영학과[1], 순천향대학교 경영학과[2], 순천향대학교 경영학과[3]

요 약

본 연구는 YouTube 플랫폼의 호텔 리뷰를 대상으로 머신러닝과 자연어처리(NLP) 기법을 활용해 극단적 및 조작된 리뷰를 식별·필터링하고자 한다. 소셜 미디어는 소비자 구매 결정에 중요한 영향을 미치며, 사용자 생성 리뷰는 마켓플레이스 신뢰도의 핵심 요소로 작용한다. 그러나 일부 판매자들의 평점 조작과 조작된 리뷰 확산은 플랫폼의 신뢰성을 저해하고 있다. 본 연구는 리뷰의 진정성과 신뢰성을 제고함으로써 소비자의 합리적 의사결정을 지원하는 것을 목표로 한다.

■ 중심어 : 극단적 리뷰, 조작된 리뷰, 자연어 처리, 사용자 생성 콘텐츠, 리뷰 신뢰성

Abstract

This study aims to identify and filter extreme and fake hotel reviews on the YouTube platform using machine learning and natural language processing (NLP) techniques. As social media increasingly influences consumer decisions, user-generated reviews have become a key factor in online marketplace credibility. However, review manipulation and fake reviews threaten platform trust. This research contributes to enhancing review authenticity and supporting rational consumer decisions.

■ Keyword : Extreme Reviews, Fake reviews, Natural Language Processing, User-generated Content, Review Credibility

2025년 05월 12일 접수; 2025년 05월 24일 수정본 접수; 2025년 06월 01일 게재 확정.
* 본 연구는 순천향대학교 학술연구비 지원으로 수행하였음. 이 논문 또는 저서는 2024년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2024S1A5A2A01020049).
†교신저자 (jaewonchoi@sch.ac.kr)

Ⅰ. 서론

인터넷과 소셜 미디어의 광범위한 사용으로 인해 이러한 플랫폼은 여론 형성에 중요한 도구가 되었으며, 사회와 경제를 포함한 여러 분야에서 깊은 영향을 미치고 있다[1]. 현대 소비자들은 구매 결정에 대한 정보를 얻기 위해 소셜 미디어에 점점 더 많이 의존하고 있으며, 이러한 플랫폼은 사용자가 개인적인 경험, 리뷰, 추천 및 공유할 수 있는 공간을 제공한다. 실제 사용자들의 평점과 리뷰는 온라인 마켓플레이스에서 중요한 부분이 되었으며 소비자의 구매 선택에 큰 영향을 미친다. 한 연구에 따르면 대다수의 소비자는 상품 구매 전에 자신의 결정이 정당한지 확인하기 위해 리뷰를 참조하는 것으로 나타났다[2,3]. 또한 평판 시스템은 시장 경쟁에서 중요한 역할을 하며, 그 영향은 전자상거래에만 국한되지 않고 케이터링(음식 공급), 호텔경영, 헬스케어 등 여러 산업 분야에 걸쳐 널리 퍼져 있다[4]. 온라인 소비자 리뷰(OCR)는 제품과 서비스의 품질을 평가하는 데 중요한 정보원이 되었다. 기존 연구에 따르면 25~34세 미국 소비자 인구의 약 72%가 상품과 서비스를 구매하기 전에 소셜 미디어에서 연락처로부터 추천과 의견을 구하는 것으로 나타났다[5]. 그러나 리뷰의 광범위한 영향으로 인해 판매자들은 자신에게 유리한 시장 이미지를 묘사하기 위해 평점을 조작하려고 시도하기도 했다. 최근 몇 년 동안 조작된 리뷰의 확산은 점점 더 큰 문제가 되고 있으며, 일부 판매자들은 조작된 리뷰를 구매하는 등 소비자의 구매 결정에 영향을 미치고 있다[6].

조작된 리뷰는 소비자를 기만할 뿐만 아니라 시장의 공정한 경쟁을 저해함으로써 온라인 마켓플레이스에서 주요 문제가 되고 있다. 판매자들은 종종 이러한 사기 행위로 자신의 평판을 높이거나 경쟁자를 억제하기 위해 조작된 리뷰를 개시하여 소비자 인식을 조작하려고 시도한다[7]. 전자상거래 플랫폼의 관점에서 보면, 조작된 리뷰의 존재는 정보의 질을 저하시키고 소비자가 정보에 입각한 결정을 내리기 어렵게 만든다. 판매자의 관점에서 보면, 조작된 리뷰의 생성과 대응 모두 운영 비용을 증가시킬 것이다. 소비자의 관점에서 보면, 조작된 리뷰는 온라인 평가의 신뢰성과 유용성을 저해하여 궁극적으로 쇼핑 경험을 해치는 요소로 인식된다[8].

이 문제는 한 시장에만 국한된 것이 아니라 전 세계적인 도전 과제다. [9]는 미국의 대표적인 레스토랑 리뷰 사이트인 Yelp에서 리뷰의 약 16%가 가짜인 반면, 아마존 플랫폼에서는 그 비율이 42%로 훨씬 더 높다는 사실을 발견했다. 조작된 리뷰의 광범위한 영향에 대응하기 위해 연구자들은 다양한 탐지 방법을 제안해 왔다. 특히 많은 조작된 리뷰에서 극단적인 리뷰가 두드러지는 것으로 관찰되었다. 이러한 리뷰는 할당되지 않은 긍정적 또는 부정적 감정의 표현으로 대표된다. 이러한 리뷰는 소비자의 진정한 의견에 뿌리를 두고 있거나 경쟁업체가 불공정 경쟁의 수단으로 사용할 수 있다. 예를 들어, 일부 판매자는 경쟁업체의 시장 성과를 손상시켜 결과적으로 매출에 영향을 미칠 수 있는 극단적인 부정적 리뷰를 의도적으로 게시할 수 있다. 연구에 따르면 극단적인 리뷰는 소비자가 구매 전 결정을 내릴 때 극단적인 의견에 영향을 받는 경향이 있기 때문에 제품 판매에 특히 큰 영향을 미치는 것으로 나타났다[1]. [1]은 극단적인 의견을 찾기 위해 가장 부정적이고 긍정적인 단어가 포함된 새로운 “Thesaurus”를 자동으로 구성하여 리뷰의 극단적인 감정 표현을 탐지하는 비지도 학습 기반 접근 방식을 제안했다. 그 결과는 조작된 리뷰를 탐지하는 기술이 진화하고 있음을 보여주지만, 조작된 리뷰가 시장에 미치는 부정적인 영향을 완전히 완화하기 위해서는 추가 연구가 필요하다. 이러한 근거로 연

구자들은 가짜 정보가 소비자 의사 결정에 미치는 영향을 줄이기 위해 조작된 리뷰 탐지를 개선하기 위한 머신러닝 기술을 통합하는 방법을 탐구한다. 여러 조작된 리뷰 인식 방법이 제안되었지만, 현재 연구는 여전히 여러 가지 도전 과제에 직면해 있다. 예를 들어, 다중 플랫폼 및 다중 모달 콘텐츠 인식의 제한된 효과와 극단적인 감정 인식의 정확성 부족 등이 있다. 따라서 본 논문에서는 유튜브 플랫폼에서 호텔 리뷰를 연구하여 가짜 및 극단적인 리뷰를 인식하는 문제에 초점을 맞추기로 한다. 전통적인 텍스트 플랫폼과 비교할 때, 소셜 비디오 플랫폼에서의 리뷰는 더 다양하고 더 많은 노이즈를 포함하고 있어 인식 난이도가 높아지며, 이 문제를 해결하기 위해서는 더 정교한 모델링 방법이 절실히 필요하다. 본 연구는 머신러닝과 자연어 처리 기술을 결합하여 리뷰의 텍스트 특징과 감정 경향을 분석하여 잠재적으로 허위 또는 조작 가능한 리뷰를 식별하는 효율적이고 해석 가능한 탐지 프레임워크를 구축하는 것을 목표로 한다. 그에 따라 본 연구는 YouTube라는 특정 영상 기반 플랫폼에서 수집된 사용자 댓글 데이터를 중심으로 감정 이상성과 허위성 패턴을 분석하였으며, 이는 비디오 소셜 플랫폼에서 리뷰 인식 분야의 현재 격차를 메울 뿐만 아니라 플랫폼 콘텐츠의 신뢰성을 높이고 사용자 결정의 품질을 향상시키는 데 도움이 될 수 있다.

조작된 리뷰와 극단적인 리뷰가 점점 더 흔해지고 있으며, 본 연구는 이를 발견할 수 있는 방법을 제안한다. 본 연구는 감성 분석, 기계 학습, 주제 모델링이라는 세 가지 방법을 결합하여 사용자가 생성한 리뷰에서 기만적이거나 조작적인 내용을 식별한다.

본 연구는 다음과 같이 구성되어 있다: 2장에서는 조작된 리뷰 탐지 및 극단적인 리뷰 탐지에 대한 기존 연구에 대해 설명한다. 3장에서는 연구방법을 설명하며, 4장에서는 연구결과를 설명한다. 5장에서는 논의와 결론 및 향후 연구의 방향에 대한 제언을 다룬다.

Ⅱ. 이론적 배경

2.1 조작된 리뷰 탐지

2.1.1 조작된 리뷰 탐지를 위한 학습 기반 접근법

조작된 리뷰의 확산은 온라인 플랫폼에서 큰 문제가 되었고, 그 결과 조작된 리뷰를 식별하는 여러 가지 방법이 등장했다. 연구자들은 조작된 리뷰를 감지하기 위해 다양한 방법을 사용해 왔다. 비지도 학습 기법은 라벨링 된 데이터가 부족하거나 사용할 수 없는 경우 조작된 리뷰 탐지에 특히 유용할 수 있다. [13]은 빈번한 아이템 마이닝 기법을 사용하여 의심스러운 후보 리뷰 클러스터를 식별하기 위해 비지도 학습 알고리즘을 사용하는 선례를 제시했다. 그런 다음 이러한 클러스터를 조작된 리뷰 클러스터를 탐지하기 위한 모델 구축을 통해 추가 분석을 수행했다. 반면, 지도 학습은 라벨이 붙은 데이터를 사용하여 진짜 리뷰와 조작된 리뷰를 구분하는 분류 모델을 학습시켜야 한다. [10]은 특정 단어나 구문의 빈도와 같은 특정 언어적 특징을 기반으로 조작된 리뷰를 식별하기 위해 서포트 벡터 머신(Support Vector Machine, SVM)과 나이브 베이즈(Naive Bayes) 분류기를 구축했다. 준지도 학습은 비지도 학습과 지도 학습의 장점을 결합한 기법이다. [15]와 [14]는 협력 학습이 리뷰 텍스트 특징과 리뷰 작성자 행동 특징을 결합하여 분류사를 더 효율적으로 학습시키는 방법을 보여준다.

2.1.2. 복합 속성 기반 리뷰 탐지 방법론

조작된 리뷰 탐지에 대한 기존 연구는 주로 텍스트, 평점, 시간, 네트워크 구조 등 다양한 관점에서 접근되어 왔다. 텍스트 기반 분석에서는

[16]이 대표적인 초기 연구를 수행하였다. 이들은 아마존(Amazon) 리뷰 데이터를 활용하여, 조작된 리뷰에는 중복되는 특징이 존재하며, 일반 리뷰에 비해 텍스트 유사도가 높다는 점에 주목하였다. 특히, 이러한 유사성은 Jaccard 유사도와 같은 정량적 지표를 통해 효과적으로 측정될 수 있음을 제안하였다. 이 연구는 이후 텍스트 유사성 기반의 조작 리뷰 탐지 연구에 중요한 토대를 제공하였다.

평점 기반 분석에서는 [17]이 Flipkart 플랫폼에서 사용자 평가 데이터를 분석하여 조작 리뷰 탐지 가능성을 제시하였다. 이들은 긍정적 리뷰어와 부정적 리뷰어의 평가 분포를 통계적으로 분석하고, 이를 일반 리뷰어의 패턴과 비교하였다. 분석 결과, 특정 집단에서 평가 분포의 불균형이나 극단성이 뚜렷하게 나타났으며, 이는 조작 리뷰의 가능성을 시사하는 중요한 정량적 지표로 활용될 수 있음을 입증하였다.

시간 기반 분석의 대표적인 연구로는 [18]의 연구가 있다. 이들은 동일 사용자가 특정 앱에 여러 차례 리뷰를 작성하는 행위에 주목하였다. 동일한 사용자가 시간 간격을 두고 반복적으로 리뷰를 남기는 경우, 리뷰의 내용과 게시 시점 간의 불일치가 조작 가능성을 높이는 요인임을 제안하였다. 이와 같은 반복적이고 비정상적인 시간 패턴은 조작된 리뷰 탐지의 보완적 기준으로 기능할 수 있다.

마지막으로, 그래프 구조 기반 분석에서는 [19]이 제안한 소셜 네트워크 접근이 주목할 만하다. 이들은 리뷰 작성 행위를 기반으로 사용자 간의 관계를 네트워크 형태로 모델링하였으며, 조작된 리뷰 작성자들은 종종 일관되고 협력적인 행동을 통해 밀집된 하위 그래프(community)를 형성한다는 사실을 발견하였다. 이와 같은 관계 기반 구조 분석은 사용자 간 상호작용을 정량화하여 허위 리뷰 작성 패턴을 탐지하는 데 효과적으로 활용될 수 있다. 이처럼 다양한 분석 기법들은 조작된 리뷰의 탐지와 식별을 위한 기반이 되어 왔으며, 본 연구 역시 이러한 접근들을 통합적으로 검토하고 YouTube 기반 리뷰 데이터에 적합한 탐지 모델을 제안하고자 한다.

2.1.3. 준지도 학습과 행동 밀도 분석을 통한 리뷰 탐지

본 연구는 조작된 리뷰 탐지를 위해 Positive-Unlabeled(PU) 학습 알고리즘을 적용하였다. PU 학습은 이진 분류를 위한 머신러닝 기법으로, 긍정적인 레이블이 부여된 샘플(P)과 라벨이 없는 샘플(U)만으로 학습이 가능하다는 특징을 지닌다[20]. 기존의 지도 학습(supervised learning) 방법과는 달리, PU 학습은 부정 샘플(Negative)에 대한 사전 라벨링이 필요 없기 때문에, 실제 조작 여부를 명확히 판별하기 어려운 온라인 리뷰 환경에서 특히 유용하다. 대부분의 리뷰는 조작 여부에 대한 명확한 정답이 없고, 조작으로 확인된 소수의 리뷰만이 레이블링 가능한 현실에서 PU 학습은 잠재적으로 조작된 리뷰를 효과적으로 식별할 수 있는 대안으로 주목받고 있다. 최근 연구에 따르면 PU 학습 알고리즘은 온라인 마켓플레이스 환경에서 조작된 리뷰 탐지의 정확도를 유의미하게 향상시킨 것으로 나타났다[21].

탐지의 정밀도와 신뢰성을 더욱 높이기 위해, 본 연구는 행동 밀도 분석(Behavioral Density Analysis)을 2차 필터링 메커니즘으로 도입하였다. 행동 밀도란 특정 범위 내에서 비정상적인 사용자 행위가 집중되는 현상을 의미하며, 이는 조작된 리뷰 패턴을 식별하는 데 효과적인 지표로 활용될 수 있다. 본 연구에서는 사용자 행동 밀도와 애플리케이션 행동 밀도를 각각 계산하여 분석하였다. 특정 리뷰의 행동 밀도가 사전에 설정된 임계값을 초과할 경우, 해당 리뷰는 조작 가능성이 높은 것으로 간주된다[21]. 이와 같은 다단계 탐지 접근은 정상 리뷰 사이에 은

〈표 1〉 조작 리뷰 분류기법 사례

분류	분류기법	주요 발견	저자
1	비지도 학습	F-통계를 통해 향상된 K-평균을 사용하여 리뷰를 적응적으로 클러스터링하고 비정상 그룹을 감지하여 가짜 리뷰 식별 기능을 향상	[39]
2	지도 학습	리뷰 텍스트, 사용자 행동, 판매자 데이터의 기능을 사용하여 가짜 리뷰 탐지의 정확성과 효율성을 균형 있게 유지하는 데 강력한 성능을 보인 DDAG-SVM 모델 적용	[40]
3	반지도 학습	먼저 일반 베이지안 모델과 수동 데이터 어노테이션 결과를 기반으로 여러 특징 조합의 효과를 평가하고, 이를 바탕으로 다수의 어노테이션이 없는 텍스트를 최대한 활용하기 위해 Co-training과 Tri-training이라는 두 가지 반지도 학습 전략을 도입하여 탐지 성능을 효과적으로 향상	[20]
4	텍스트 기반 분석	댓글 텍스트의 언어적 내용과 의미적 특징이 댓글의 진위를 반영할 수 있으며, 따라서 텍스트 분석 기법은 허위 댓글을 구별하는 가장 중요한 수단이라고 강조	[16]
5	비판 기반 분석	가짜 리뷰어들이 종종 비정상적인 평가 패턴을 보이며, 실제 사용자와 크게 다른 분포를 보인다는 사실을 발견했습니다. 이는 탐지의 핵심 단서로 작용합니다.	[41]
6	시간 기반 분석	가짜 리뷰와 실제 사용자 리뷰의 차이를 발견. 가짜 리뷰는 짧은 시간 내 집중적으로 게시. 시간적 패턴이 뚜렷한 것이 특징. 실제 사용자 리뷰가 구매 후 무작위로 작성	[16]
7	차트 구조 기반 분석	시간적 밀집성과 그래프 구조를 결합해 리뷰 그룹 정의 확대. 사기 사용자들은 시간 일관성과 높은 연결성을 보임.	[19]
8	PU 학습 알고리즘 및 행동 밀도	PU 학습과 행동 밀도를 결합하여 최소한의 라벨링된 데이터를 사용으로 가짜 댓글을 감지함으로써 실제로 높은 효과를 보여줌.	[20]

폐된 조작된 리뷰 클러스터를 효과적으로 분리할 수 있으며, 탐지 시스템의 신뢰성과 견고성을 크게 향상시키는 데 기여한다. 따라서, 이전 연구에서 사용한 리뷰 분류기법의 주요 유형들을 고찰하였으며, 서술된 모든 방법은 <표 1>에 제시했다.

2.2 극단적 및 조작된 리뷰의 식별

온라인 사용자 리뷰는 소비자 의사결정에 강력한 영향을 미치는 요소로 작용하며, 그 신뢰성과 진정성 확보를 위한 탐지 기술의 중요성이 점차 강조되고 있다. [22]는 형용사의 극성을 판별하기 위해 코퍼스 기반의 비지도 학습 접근 방식을 제안하였다.

이들의 방법은 텍스트 내에서 접속사(예: “and”, “but”)로 연결된 형용사들이 일반적으로 유사한 감성 극성을 지닌다는 언어학적 가정을 기반으로 한다. 이러한 원칙은 이후 감성어 사전 구축 및 감성 분석 자동화의 핵심 기초가 되었으며[23] 규칙 기반 감성 분류에서 텍스트의 감정 성향을 결정하는 데 널리 활용되었다. 다만, 해당 방식은 비속어나 신조어, 반어법 등 문맥에 따라 감정이 변하는 표현을 정교하게 처리하는 데는 한계가 존재한다.

문서 단위에서의 감성 분류는 특히 극단적인 리뷰 탐지에 효과적인 접근법으로 여겨진다.

[1]은 극단적 견해와 비극단적 견해를 이진 분류하는 연구를 통해, 감정적으로 과장된 콘텐츠가 온라인 담론의 진정성을 왜곡할 수 있다는 점을 지적하였다. 이는 조작적이거나 편향된 정보가 소비자 판단에 미치는 영향을 최소화하기 위한 기

술적 기반으로서 중요한 시사점을 제공한다.

조작된 리뷰 식별을 위한 기준으로는 콘텐츠 유사성, 반복 행동, 극단적 평점 부여, 그리고 과도한 긍정적 감성의 사용 등이 주요 지표로 제시된다. [25]는 조작된 리뷰들이 서로의 리뷰를 복사하거나 유사한 문장을 반복하여 작성하는 경향이 있으며, 실제로 조작된 리뷰의 70% 이상이 텍스트 유사도 0.3 이상을 기록하였다고 보고하였다. 이는 조작된 콘텐츠가 동일한 서술 패턴을 보이는 경우가 많음을 시사한다.

[24]은 동일 리뷰어가 동일 제품에 대해 여러 개의 유사한 리뷰를 남기는 반복 행동 패턴이 조작 가능성을 나타내는 주요 특징 중 하나라고 지적하였다. 이러한 반복은 비정상적인 활동의 명백한 지표로 간주된다. 또한 [24] 등의 연구에 따르면, 별점 1점 혹은 5점 등 극단적 평점을 반복적으로 부여하는 리뷰는 감정적 과장 또는 의도적 왜곡을 통해 소비자 판단에 영향을 미치려는 의도가 있을 수 있다고 분석하였다.

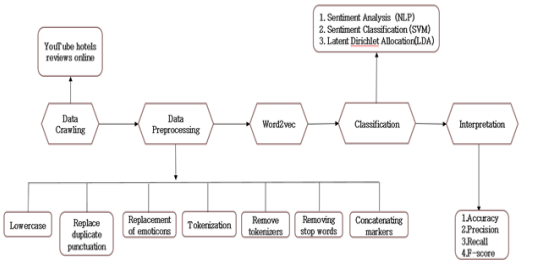
이와 함께, [26]은 조작된 리뷰들이 주로 긍정적 감정을 표현하는 리뷰를 반복적으로 게시하고, 제품에 대한 인위적 입소문 효과를 유도한다고 주장하였다. 이는 소비자에게 제품이 널리 찬사를 받고 있다는 인식을 심어주기 위한 전략으로 활용되며, 실제 제품 품질과 무관한 과도한 신뢰를 형성할 수 있다.

이와 같은 극단성, 유사성, 반복성, 과잉 긍정성은 모두 조작된 리뷰 탐지 알고리즘에서 핵심적인 탐지 기준으로 활용될 수 있으며, 본 연구에서는 이러한 요소들을 기반으로 탐지 모델을 설계하고 실증 분석을 수행하고자 한다.

Ⅲ. 연구방법

3.1 연구 과정

전 세계적으로 20억 명 이상의 사람들이 소셜 미디어를 사용하며, YouTube는 페이스북 다음으로 큰 소셜 미디어 플랫폼이고 사용자들은 YouTube에서 매우 활발하게 활동하며, 하루 평균 40분의 동영상을 시청하고 자주 리뷰를 달기도 한다[27]. 이러한 사용자 행동은 누적 약 10엑사바이트의 데이터를 생성하는 YouTube가 연구에 유용한 플랫폼이 된다[28]. 텍스트 분석은 다양한 방법, 머신러닝 기법 및 기술을 활용하여 자연어 텍스트에서 비정형 데이터를 심층적으로 파싱(구문분석) 하는 것이다. 이는 비용을 줄이면서 효율적인 의견 측정 방법(Way of Opinion Measurement)을 제공한다. 본 연구는 YouTube의 비정형 데이터를 사용한다. 불용어와 기호를 제거한 후, word2vec로 리뷰 유사도를 계산하기 위해 단어를 벡터로 변환한다. 이러한 벡터 값은 벡터 값을 처리하는 제안된 모델의 입력 값이 된다. 마지막으로. 이후 콘텐츠의 분류를 위하여, 감정분석, SVM, 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기법을 활용하여 결과를 해석하였다. 이후 최종 모형의 성과에 대해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 및 F1-score을 사용하여 분류 결과를 검증하였다. <그림 1>은 전체 연구 과정을 보여준다.



〈그림 1〉 연구 방법

3.2 데이터 수집 및 전처리 과정

본 연구에서는 YouTube 플랫폼에서 제공되는

리뷰 데이터를 기반으로 조작된 리뷰 탐지 모델을 구축하였다. 데이터 수집을 위해 YouTube API를 활용하여 특정 동영상 URL을 기준으로 리뷰 데이터를 자동으로 추출하는 크롤러를 개발하였다. 해당 방식은 비정형 텍스트 수집에 효과적으로 작동하며, 총 37,040개의 리뷰 데이터를 확보하였다.

수집된 리뷰는 대부분 짧은 문장 구조를 가지며, 이모티콘, 철자 오류, 특수기호, 불용어와 같은 분석에 방해가 되는 요소들을 포함하고 있어, 분석의 정확도를 높이기 위해 체계적인 전처리 과정이 필수적이다[29]. 특히 소셜 미디어 리뷰는 문맥의 다양성과 비표준 언어 표현이 혼재되어 있기 때문에 전처리 단계에서의 정제 수준이 전체 모델 성능에 큰 영향을 미친다.

이에 따라 본 연구에서는 리뷰 텍스트의 분석 품질과 모델링 효과를 향상시키기 위해 원본 텍스트 데이터에 대해 일련의 전처리 작업을 수행했다. 먼저 모든 텍스트를 소문자로 일괄 전환하여 대소문자 차이로 인한 어휘 중복을 피하고, 이어서 텍스트의 중복 문장, 기호, 숫자, 공백 등 알파벳이 아닌 문자를 제거하여 노이즈 데이터를 세척하고, 이모티콘을 대체하여 대응하는 정서어로 변환하여 잠재적인 감정 정보를 유지한 후 텍스트를 분사 처리하여 문장을 독립어원으로 분할하고, 숫자나 특수문자가 포함된 어원을 추가로 제거하여 어휘를 순수하게 하고, 그 후 “the”, “is”, “in” 등 흔하지만 실질적인 의미 기여가 없는 중단어를 제거하며, 마지막으로 세척된 어원을 연속된 어원 분석 문자열을 제공한다.

Ⅳ. 연구결과

본 연구에서는 리뷰 텍스트에 내포된 감정을 효과적으로 탐지하기 위해 머신러닝 기반 감성 분석 기법을 적용하였다. 이를 위해 자연어 처리 기술과 결합하여 문장 또는 문서 내의 주관적인 정보, 예컨대 긍정 또는 부정 감정의 표현을 식별하고, 리뷰어의 감정적 성향을 분류하였다. 감성 분석(Sentiment Analysis, SA)은 자연어 처리(Natural Language Processing, NLP) 분야에서 핵심적인 역할을 수행하며, 텍스트로부터 감정과 의견을 추출하는 것을 주요 목표로 한다[29,30]. 이와 같은 접근은 의견 마이닝(Opinion Mining)으로도 불리며, 특정 대상이나 주제에 대해 사람들이 표현하는 정서적 반응과 주관적 견해를 체계적으로 식별하고 추출하는 데 중점을 둔다[31]. 감성 분석은 일반적으로 명시적인 감정 표현과 암시적으로 드러나는 정서적 경향성 또한 함께 분석의 대상으로 삼는다.

또한, 본 연구에서는 리뷰 텍스트의 감정 경향을 효과적으로 분류하기 위해 SVM 알고리즘을 채택하였다. SVM은 Vapnik(1999)이 개발한 지도 학습 알고리즘으로, 텍스트 분류, 스팸 필터링, 사기 탐지 등 다양한 자연어 처리(NLP) 작업에서 널리 활용되고 있다[34]. 본 연구에서는 해당 기법을 활용하여 리뷰 데이터에 내포된 감정적 극성(polarity)을 자동 분류하여 조작 가능성이 높은 텍스트의 탐지 및 정밀한 감성 분류를 확인하였다.

이와 더불어, 리뷰 텍스트의 주요 주제와 논조를 심층적으로 파악하기 위해 LDA 모델을 통한 토픽모델링을 활용하였다. 토픽 모델링(topic modeling)은 정렬되지 않은 비정형 문서 집합으로부터 주요 주제를 자동으로 발견하고, 텍스트의 숨겨진 의미 구조를 탐색하기 위해 활용되는 자연어 처리(Natural Language Processing, NLP)의 대표적인 기법이다[35]. 이 중 LDA는 통계적 토픽 모델링 기법의 대표적인 방법으로, 텍스트 마이닝, 정보 검색, 소셜 미디어 분석 등 다양한 분야에 성공적으로 적용되어 왔으며, 컴퓨터 과학 및 인공지능 기반 텍스트 분석에서 핵심적인 역할을 수행하고 있다. LDA 기반 토픽 모델링에서 토픽은 일반적으로 함께 등장하는 단어들

의 확률적 분포로 정의되며, 이는 텍스트의 핵심 의미와 주제를 설명하는 데 사용된다. 이러한 토픽은 이메일, 기사, 블로그, 리뷰 등 다양한 비정형 텍스트에 내재된 의미를 통계적으로 유의미한 단어군을 통해 표현하며, 문서 집합 내 주제 구조를 추론할 수 있게 해준다. 본 연구에서도 리뷰의 주제적 흐름과 감정 간의 관계를 보다 명확히 파악하고, 텍스트 내 이상 패턴을 탐지하기 위해 LDA 기반 주제 모델링을 적용하였다. 이를 통해 가짜 리뷰나 극단적 리뷰가 특정 주제 클러스터에 과도하게 집중되는 경향을 식별할 수 있으며, 이는 조작된 콘텐츠 탐지 및 감성 분석의 정밀도를 동시에 향상시키는 데 기여한다.

4.1 VADER와 SVM 기반 조작된 리뷰 탐지

본 연구에서는 리뷰 텍스트의 감정 극성을 효과적으로 분류하기 위해 VADER(Valence Aware Dictionary and Sentiment Reasoner)를 기반으로 한 감성 분석 기법을 활용하였다. VADER는 소셜 미디어 텍스트, 온라인 리뷰, 비격식 표현이 포함된 문서 분석에 특화된 사전 기반 감정 분석 도구로, 정교한 규칙 기반 접근을 결합하여 텍스트의 감정을 빠르고 효율적으로 분류할 수 있도록 설계되었다.

본 연구에서는 NLTK 기반의 VADER 도구를 사용하여 각 리뷰에 대한 감정 점수를 계산하였다. VADER는 감정 극성을 -1에서 +1까지의 연속적인 스코어로 제공하며, 양수 값은 긍정적 감정을, 음수 값은 부정적 감정을, 0에 가까운 값은 중립적 감정을 의미한다. 예를 들어, 감정 점수가 0.9263 또는 0.7691로 나타난 리뷰는 “긍정적”으로, 0에 근접한 점수(예: 0.0000)는 “중립적”으로, 그리고 -0.5273과 같은 음수 점수는 “부정적”으로 분류된다. <표 2>는 이러한 감정 점수에 따른 분류 기준 및 예시를 요약한 것이다.

〈표 2〉 VADER 감성 분석 결과

No.	Text	Sentiment Label
1	’impressed’, ‘Hilton’, ‘Tokyo’, ‘husband’, ‘worked’, ‘IBM’…	positive
2	’Hilton’, ‘Tokyo’, ‘Osaka’, ‘prefer’, ‘Conrad’	neutral
3	’absolutely’, ‘love’, ‘Conrad’, ‘Hotels’…	positive
4	’thanks’, ‘great’, ‘video	positive
5	’laundry’, ‘area’	neutral
…	…	…
37040	’comment’	positive

추가적으로, 감정 분석 결과를 기반으로 TF-IDF 벡터화 및 SVM 분류기를 결합하여 조작된 리뷰 탐지 모델을 구축하였다. 해당 모델은 감정적 경향성을 정량화하고, 이를 기반으로 조작 가능성이 높은 리뷰를 효과적으로 분류하는 데 활용되었다.

〈표 3〉 SVM을 이용한 감정 분류 분석 결과

	precision	recall	f1-score	support
Negative	0.858	0.883	0.870	4555
Positive	0.917	0.898	0.908	6557
accuracy	0.892	0.892	0.892	0.892
macro avg	0.888	0.890	0.889	11112
weighted avg	0.893	0.892	0.892	11112

<표 3>에 제시된 바와 같이, 해당 모델은 긍정 및 부정 감정 분류에서 모두 우수한 성능을 보였다. 부정 리뷰 분류의 경우 정밀도(Precision) 0.86, 재현율(Recall) 0.88, F1-score 0.87을 기록하였으며, 긍정 리뷰 분류는 정밀도 0.92, 재현율 0.90, F1-score 0.91로 보다 높은 정확도를 나타냈다. 각 감정 범주의 리뷰 수는 부

정 리뷰 4,555건, 긍정 리뷰 6,557건으로 확인되었다.

전체 모델의 분류 정확도는 0.89로, 전체 리뷰의 89%가 올바르게 분류되었음을 의미한다. 정밀도, 재현율, F1-score의 macro-average는 모두 0.89로 감정 범주 간의 균형 잡힌 성능을 나타냈으며, weighted average 역시 0.89로 표본 수의 비율을 반영한 평균에서도 일관된 성능을 보였다.

요약하자면, 본 모델은 긍정 리뷰 분류에서 특히 높은 정확도와 신뢰성을 보였으며, 특정 감정 범주로의 편향 없이 조작 가능성이 있는 리뷰를 안정적으로 탐지할 수 있는 성능을 입증하였다.

4.2 감성분석 및 SVM기반 극단적 리뷰 탐지

극단적인 감성 경향을 지닌 리뷰를 식별하기 위해, 본 연구는 감정 분석 도구인 VADER에서 도출된 감정 점수를 기반으로 각 리뷰의 감정 수준을 정량적으로 분류하였다. 이 점수는 [-1, 1] 범위의 연속적 값으로 제공되며, 감정의 극성을 정밀하게 반영한다.

예를 들어, 감정 점수가 0.9263에 가까운 리뷰는 '극단적으로 긍정적'으로 분류되며, 이는 강한 호의적 정서를 반영한다. 반면, 0.0000에 가까운 점수는 중립적인 감성을 나타내며, 이러한 리뷰는 '중립' 범주로 분류된다. 감정 점수가 0.7691처럼 양수이지만 극단적인 임계치에 도달하지 못한 리뷰는 일반적인 '긍정'으로 간주된다.

이와 동일한 원칙이 부정 감성에도 적용되며, 매우 낮은 음수 점수를 가진 리뷰는 '극단적 부정'으로 분류된다. 이러한 기준은 <표 4>에 정리되어 있다.

〈표 4〉 감정 점수 기반 극단적 리뷰 분류

No.	Text	Sentiment Label
1	'impressed', 'Hilton', 'Tokyo', 'husband', 'worked', 'IBM'…	extreme-positive
2	'Hilton', 'Tokyo', 'Osaka', 'prefer', 'Conrad'	neutral
3	'absolutely', 'love', 'Conrad', 'Hotels'…	neutral
4	'thanks', 'great', 'video	neutral
5	'laundry', 'area'	neutral
…	…	…
37040	'comment'	neutral

감정 점수에 기반한 극단 리뷰 탐지를 위해 SVM 분류기를 적용하였다. 해당 모델은 각 리뷰가 '극단 긍정', '극단 부정', '긍정', '부정', '중립' 중 어떤 범주에 속하는지를 예측한다. 모델의 성능 결과는 <표 5>에 제시되어 있다.

〈표 5〉 SVM 기반 극단적 감정 분류 분석

	precision	recall	f1-score	support
Neutral	0.950	0.985	0.967	9969
Extremely negative	0.000	0.000	0.000	43
Extremely positive	0.810	0.569	0.668	1100
accuracy	0.940	0.940	0.940	0.940
macro avg	0.587	0.518	0.545	11112
weighted avg	0.933	0.940	0.934	11112

예측 결과에 따르면, 중립 리뷰의 분류에서 매우 우수한 성능이 확인되었으며, 정밀도 0.95, 재현율 0.99, F1-score 0.97을 기록하였다. 이는 해당 모델이 감정적 편향이 없는 리뷰를 효과적으로 식별할 수 있음을 보여준다. 한편, SVM 기반 감정 분류 결과에서 극단적 부정 범주의 F1-score, 정밀도, 재현율이 모두 0으로 나타났

다. 이는 해당 범주에 대한 탐지 성능이 전혀 확보되지 않았음을 의미한다. 이러한 결과는 다음의 요인에 기인한 것으로 해석된다. 첫째, 극단적 부정 리뷰는 전체 데이터셋 내 비중이 극히 낮아, SVM과 같은 전통적인 분류기가 해당 범주의 패턴을 효과적으로 학습하지 못하였을 가능성이 크다. 둘째, 극단적 부정 표현은 풍자(satire), 과장(exaggeration), 반어(irony) 등 고차원적인 언어적 특성을 포함하는 경우가 많아, TF-IDF 기반 벡터화 방식으로는 이러한 뉘앙스를 충분히 포착하기 어려웠을 것으로 보인다. 이러한 언어적 특성은 종종 긍정 또는 중립 범주로 오분류되는 원인이 된다.

한편, 극단적 긍정 리뷰의 경우 정밀도는 0.81, 재현율은 0.57, F1-score는 0.67로, 중간 수준의 분류 성능을 보였다. 이는 모델이 강한 긍정 감정을 어느 정도 식별할 수 있으나, 여전히 탐지 정확도 개선이 필요함을 의미한다.

모델의 전체 정확도는 0.94로, 전체 리뷰 샘플의 94%를 올바르게 분류하였다. 하지만 macro-average 기준 정밀도(0.59), 재현율(0.52), F1-score(0.55)는 비교적 낮은 수치를 기록하였으며, 이는 극단적 부정 리뷰에 대한 탐지 성능이 전체 평균을 하락시킨 결과이다. 반면, weighted average 기준 정밀도(0.93), 재현율(0.94), F1-score(0.93)는 높게 나타나, 모델이 중립 리뷰와 긍정 리뷰처럼 표본 수가 많은 카테고리에서 안정적인 성능을 보이고 있음을 나타낸다.

이러한 결과는 본 모델이 극단 감정 탐지, 특히 중립 리뷰 분류에 매우 효과적이며, 데이터의 클래스 불균형 문제를 일정 부분 극복할 수 있는 가능성을 보여준다. 그러나 극단적 부정 감성 리뷰에 대한 탐지 성능을 개선하기 위해 추가적인 모델 구조 조정이 필요함을 시사한다.

4.3 토픽 모델링

본 연구는 감성 분석 결과를 기반으로, 극단적 또는 조작된 리뷰에 내재된 주제 흐름과 의미 패턴을 심층적으로 분석하기 위해 LDA 알고리즘을 부가적으로 적용하였다. LDA는 문서 집합으로부터 잠재적 주제를 자동으로 도출하는 확률 기반의 주제 모델링 기법으로, 복잡한 텍스트 구조를 최적화하고 의미의 혼란을 줄이는 데 효과적인 방법으로 널리 사용되고 있다[35]. 이 접근은 문서-단어 행렬을 기반으로 각 문서 내 단어의 출현 빈도를 분석하며, 결과적으로 각 문서에 내재된 주제 분포를 추정한다. 텍스트 데이터에 가장 자주 등장한 단어로 Word Cloud을 통해서 확인하였다. 도출된 Word Cloud는 그림과 같다.

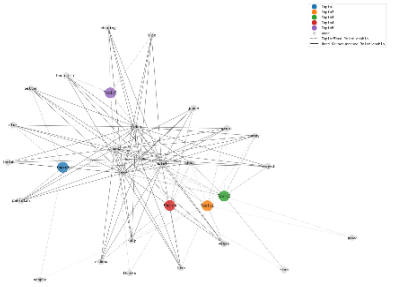
토픽 수는 사전 테스트와 비교 분석을 통해 적정성을 판단하였으며, 본 연구에서는 총 5개의 주제(topic)를 추출하였다. 이들은 리뷰 본문에서 명사를 중심으로 추출된 핵심 주제어를 기반으로 구성되었으며, <표 6>에 각 토픽에 대한 대표 키워드들이 제시되어 있다[37].

〈표 6〉 LDA 기반 토픽 분류 및 대표 키워드

Representative Keywords	Co-occurrence Score	Topic Label
hotel, pakistan, india, better, star, country, like, indian, people, bro, room, pakistani, don, come	0.44	Reviews of Hotels in Different Countries
like, vegas, hotel, great, ruby, love, stay, new, good	0.51	Staying Experience
room, stayed, hotel, time, pool, vegas, like, years, night, good, check, great, really, people, got, resort	0.53	Hotel Facilities and Rooms
video, japan, love, list, travel, youtube, thanks, like, mickey, series, time, thing, watching, best, know, great	0.52	Recommended Hotels for Traveling
thank, video, nice, beautiful, sharing, thanks, tour, day, hotel, watching, time	0.47	Visual Content of the Hotel

추출된 토픽 주제의 해석력을 높이기 위해, 본 연구는 동시출현 점수(co-occurrence score)를 계산하여 각 토픽의 의미적 일관성을 평가하였다. 이 점수는 특정 주제 내에서 키워드들이 동일한 문서 내에 얼마나 자주 함께 등장하는지를 수치화한 것으로, 주제어 간의 상호 연관성을 측정하는 지표로 활용된다. 수치화한 것으로, 주제어 간의 상호 연관성을 측정하는 지표로 활용된다. 동시출현 점수가 높을수록 키워드들이 밀접하게 연결되어 있는 일관된 주제를 형성하며, 예를 들어 '호텔 시설', '청결', '침대', '조식' 등이 함께 자주 등장하는 경우, 해당 리뷰는 특정 호텔의 물리적 조건에 집중된 논의임을 시사한다. 반면, 동시출현 점수가 낮은 토픽은 단어 간 맥락적 연결성이 약하거나, 리뷰 작성자들이 다양한 맥락에서 단어를 더 자유롭게 사용하는 경향이 있음을 나타낸다. 이는 주제가 분산되어 있거나, 단어가 서로 다른 상황에서 임의적으로 사용되고 있다는 가능성을 내포하며, 종종 감정적으로 과장된 표현이나 조작 가능성이 있는 리뷰와 관련된 주제로 연결될 수 있다.

따라서 LDA 기반 토픽 모델링과 동시출현 점수 분석은 텍스트 데이터 내 의미 구조를 해석할 수 있는 유용한 수단일 뿐 아니라, 극단적 혹은 비정상적 리뷰 패턴을 조기에 탐지할 수 있는 진단 도구로 기능할 수 있다. 이는 향후 조작 리뷰 자동 탐지 시스템 개발에 실질적인 기초자료로 활용될 수 있다.



〈그림 2〉 동시출현 네트워크

동시출현 네트워크는 사용자 리뷰의 다차원적인 토픽 구조를 보여준다. 다양한 색상의 토픽 노드는 여행 및 숙박 경험, 지역 리뷰, 동영상 콘텐츠 피드백 등과 같은 다양한 의미적 시나리오를 다루며, 고빈도 키워드는 여러 토픽을 동시 발생 관계를 통해 서로 연결하여 리뷰 콘텐츠의 다양성과 집중도를 보여준다. <그림 2>는 시각화 된 동시출현 네트워크를 나타낸다.

Word Cloud 시각화는 사용 시나리오에 대한 사용자의 태도, 의도, 콘텐츠 품질에 대한 인식 등 주제에 대한 다양한 관점에 대한 초기 인사이트를 제공한다. YouTube 데이터 세트에서 37040개의 리뷰를 분석하여 가장 빈번하게 사용되는 단어를 파악했다. <그림 3>는 데이터 세트에서 빈도가 높은 단어의 워드 클라우드를 보여준다.

〈그림 3〉 Word Cloud 시각화

Ⅴ. 결과

5.1 토론 및 결과

본 연구는 YouTube 플랫폼의 호텔 리뷰 데이터를 활용하여 극단적 및 조작된 리뷰를 탐지하기 위한 다단계 분석을 수행하였다. 리뷰의 감성 경향을 정량적으로 평가하기 위해 VADER 기반 감성 분석을 적용하였고, 텍스트 기반 특징을 바탕으로 SVM 분류 모델을 구축하여 감정 범주(긍정, 부정, 중립, 극단적 긍정, 극단적 부

정)별 탐지 성능을 분석하였다. 그 결과, 모델은 중립적 리뷰 분류에서 정밀도 0.95, 재현율 0.99, F1-score 0.97을 기록하며 매우 우수한 성능을 보였으나, 극단적 부정 리뷰에 대한 탐지는 미흡하였다(F1-score = 0). 반면, 극단적 긍정 리뷰 분류 성능은 F1-score 0.67로 중간 수준이었다. 이는 분류 성능이 감정의 강도와 분포 불균형에 따라 영향을 받을 수 있음을 시사한다.

이와 더불어, LDA 기반 토픽 모델링을 통해 리뷰의 주제적 구조를 추출하고, 동시출현 점수를 통해 토픽의 의미적 일관성을 분석하였다. LDA는 대규모 비정형 리뷰에서 숨겨진 토픽을 감지하여 의미 패턴을 식별하는 데 도움을 주고 가짜 또는 극단적인 리뷰 뒤에 숨겨진 주제를 찾아내는 데 유용한데다 가짜 리뷰와 관련된 비정상적인 토픽 패턴(예: 과도한 칭찬 또는 편향적 태도)을 드러내어 모델 정확도를 높임으로써 이상 탐지를 지원하였다. 총 5개의 주요 토픽이 도출되었으며, 각 토픽에 대해 감성 분석을 병행한 결과, ‘Staying Experience’, ‘Hotel Facilities’, ‘Travel Recommendations’와 같은 실질적 경험 중심 토픽에서 긍정 감정이 높은 비중을 차지했다. 반면 ‘Visual Content’와 ‘International Hotel Comparisons’ 주제는 중립 및 부정 감성이 혼재되어 있어 보다 다양한 감정적 반응이 나타나는 것으로 분석되었다.

이러한 결과는 극단적 감성 표현이 특정 토픽에 집중되기보다, 감정 강도보다는 주제와 맥락의 복합적 상호작용에 의해 영향을 받는다는 점을 보여준다. 감성 분석과 주제 분석을 결합한 본 접근법은 조작 가능성이 있는 리뷰의 구조적 특성과 감정적 특이성을 다각도로 조망할 수 있도록 하였다는 데에 의의가 있다.

추가적으로, 본 연구는 골드 스탠다드 기반의 정답 레이블이 부재한 상황에서 PU 학습 기법을 활용하였으며, positive 클래스에 해당하는 리뷰는 전문가 판단과 텍스트 특징에 기반한 정성적 기준에 따라 구성되었다. 그러나 이와 같은 방식은 조작 여부에 대한 객관적·정량적 확증이 어려운 한계를 지닌다. 이러한 라벨링 신뢰도의 문제는 탐지 결과 해석의 신뢰성에도 영향을 줄 수 있기에, 향후에는 다중 평가자 기반의 독립 라벨링 및 리뷰 합치도 분석, 텍스트 메타데이터 기반의 보조 지표 활용 등을 통해 신뢰성 향상을 도모할 필요가 있다.

본 연구에서는 감정 분석을 위해 VADER를 활용하여 리뷰의 극단성을 정량화하였으나, 해당 도구는 규칙 기반(rule-based) 방식의 한계로 인해 문맥적 의미나 발화의 의도성을 충분히 반영하지 못한다는 제약이 존재한다. 특히, 조작적이거나 극단적인 리뷰는 반복적 문장 구조, 비정상적 어투, 반어적 표현 등 복합적인 언어적 특징을 수반하는 경우가 많으며, 이러한 특성은 단순한 감정 극성 점수만으로는 정확히 분류되기 어렵다. 따라서 향후 연구에서는 문장의 구조적 특징을 반영할 수 있는 구문 분석(syntactic parsing) 기법의 도입과 함께, BERT, RoBERTa와 같은 사전학습 기반 딥러닝 모델을 활용하여 감정 극단성에 대한 문맥 기반 탐지가 가능하도록 분석 체계를 확장할 예정이다. 이러한 보완을 통해 조작 가능성이 내포된 극단 리뷰의 판별 정확도를 보다 향상시킬 수 있을 것으로 기대된다.

5.2 시사점 및 향후 연구 방향

본 연구의 이론적 시사점은 다음과 같다. 본 연구는 기존의 조작 리뷰 탐지 연구가 단순한 감성 이분법(긍정/부정)에 의존했던 한계를 극복하고, 감정의 강도(극단성)와 주제 구조(LDA 기반)를 함께 고려한 분석 프레임워크를 제시했다는 점에서 이론적 기여가 있다. 특히 VADER 점수를 기반으로 감정의 연속성과 극단성을 정량화하고, SVM을 통해 이를 분류한 점은 감성

분석의 정밀도를 높이는 실증적 방법론으로 활용 가능하다. 또한 토픽 모델링과 감정 분포를 결합한 분석 방식은 비정형 리뷰 데이터에서 감정-주제 연관 구조를 밝히는 데 있어 새로운 접근을 제시하였다.

본 연구의 실무적 시사점은 다음과 같다. 실무적으로는 본 연구의 결과가 호텔 예약 플랫폼, 리뷰 기반 커머스, 유튜브 여행 콘텐츠 분석 등 다양한 분야에서 조작 리뷰 및 극단적 콘텐츠를 자동 탐지하고 모니터링하는 시스템 개발에 직접 활용될 수 있다. 감정 점수 기반 필터링과 토픽 기반 클러스터링을 통해 플랫폼은 신뢰성을 확보하고 소비자에게 보다 정확한 정보를 제공할 수 있다. 특히 중립 또는 과도하게 긍정적인 리뷰만 집중 노출되는 문제를 방지하기 위한 콘텐츠 검열 및 추천 알고리즘 설계에도 본 연구는 기여할 수 있다.

향후 연구에서는 다음과 같은 보완이 가능하다. 첫째, 본 연구는 감성 분석 및 분류에 주로 SVM 기반의 모델을 사용하였으나, 딥러닝 기반 모델(BERT, RoBERTa 등)을 적용하면 극단적 리뷰 탐지 성능을 더욱 향상시킬 수 있을 것이다. 둘째, 본 연구는 YouTube 플랫폼의 사용자 생성 콘텐츠를 대상으로 극단성과 허위성을 분류하는 데 초점을 맞추었으며, 분석 결과 역시 해당 플랫폼의 구조적·문화적 특성을 반영하고 있다. 따라서 본 연구의 결과를 다른 플랫폼(Amazon, TripAdvisor 등)의 사용자 리뷰 환경에 직접적으로 일반화하는 데에는 한계가 존재한다.

향후 연구에서는 다양한 플랫폼에서 수집된 리뷰 데이터를 통합 분석하거나, 플랫폼별로 비교 실험을 수행함으로써, 모델의 범용성과 플랫폼 특이성을 보다 정교하게 검증할 필요가 있다. 셋째, 리뷰의 감정 외에도 리뷰어 프로필, 리뷰 빈도, 시간 정보 등 행동적 메타데이터를 결합한 다변량 분석이 향후 조작 탐지의 정밀도를 높이는 데 도움이 될 것이다.

넷째, 본 연구는 전통적인 머신러닝 기반 감정 분류 모델이 소수 감정 범주, 특히 극단적 부정 범주에 대해 탐지 성능이 현저히 저하되는 한계를 보였다. 이는 데이터셋의 클래스 불균형과 극단 감성 표현의 언어적 복합성에 기인한다. 따라서 향후 연구에서는 다음과 같은 보완이 필요하다. 우선, SMOTE 등 오버샘플링 기법이나 텍스트 생성 기반의 데이터 증강(data augmentation) 기법을 활용하여 극단적 부정 범주에 대한 학습 샘플을 보완하는 접근이 필요하다. 또한, BERT, RoBERTa와 같은 사전학습 기반 언어모델을 적용하여, 극단 표현에 내포된 문맥 정보를 정밀하게 반영할 수 있는 분류 성능 향상이 기대된다. 마지막으로, 향후 연구에서는 정확도뿐 아니라 macro/micro F1-score, confusion matrix 등의 다양한 성능 지표를 병행하여 감정 분류 모델의 성능을 정밀하게 평가할 필요가 있다.

참 고 문 헌

[1] S. Almatarneh, P. Gamallo, “A lexicon-based method to search for extreme opinions,” PloS one, vol. 13(5): e0197816. 2018.

[2] L. Einav, C. Farronato, J. Levin, “Peer-to-peer markets,” Annual Review of Economics, vol. 8(1): 615-635. 2016.

[3] S. Tadelis, “19 two-sided e-commerce marketplaces and the future of retailing”, Handbook on the Economics of Retailing and Distribution, 2016.

[4] A. Gandhi, B. Hollenbeck, Z. Li, “Misinformation and Mistrust: The Equilibrium Effects of Fake Reviews on Amazon. Com,” 2024.

[5] Mintel, “Social Networking - Ireland,” [Online]. Available: http://academic.mintel.com/display/739944/, 2015.

[6] T. Zhang, L. He, Z. Wang, “Risk factors for death of follicular thyroid carcinoma: a systematic review and meta-analysis,” Endocrine, vol. 82(3): 457-466. 2023.

[7] A. Flostrand, L. Pitt, J. Kietzmann, “Fake news and brand management: a Delphi study of impact, vulnerability and mitigation,” Journal of Product and Brand Management, vol. 29(2): 246-254. 2020.

[8] L. Wang, Z. Zhang, X. Cui, “Identifying and Filtering Fake Reviews: Current Situation and Prospect,” J. of University of Electronic Science and Technology of China (SOCIAL SCIENCES EDITION), vol. 24(1): 31-41, 64. 2022.

[9] M. Luca, G. Zervas, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” Management Science, vol. 62(12): 3412-3427. 2016.

[10] N. Jindal, B. Liu, “Analyzing and detecting review spam,” IEEE International Conference on Data Mining, vol. 547-552. 2007.

[11] M. Ott, Y. Choi, C. Cardie, “Finding deceptive opinion spam by any stretch of the imagination,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, vol. 309-319. 2011.

[12] E. P. Lim, V. A. Nguyen, N. Jindal, “Detecting product review spammers using rating behaviors,” Proceedings of the 19th ACM International Conference on Information and Knowledge Management, vol. 939-948. 2010.

[13] A. Mukherjee, B. Liu, N. Glance, “Spotting fake reviewer groups in consumer reviews,” The 21st International Conference on World Wide Web, vol. 191-200. 2012.

[14] W. Zhang, Q. Wang, B. Chaoqi, “Co-training-based identification of online fake reviews,” Systems Engineering Theory and Practice, vol. 40(10): 2669-2683. 2020.

[15] N. Kumar, D. Venugopal, L. Qiu, “Detecting anomalous online reviewers: An unsupervised approach using mixture models,” Journal of Management Information Systems, vol. 36(4): 1313-1346. 2019.

[16] N. Jindal, B. Liu, “Opinion Spam and Analysis,” Proc. ACM Intl. Conf. Web Search and Web Data Mining (WSDM), vol. 219-30. 2008.

[17] N. Shah, A. Beutel, B. Hooi, L. Akoglu, S. Günnemann, D. Makhija, M. Kumar, & C. Faloutsos, “Edgecentric: Anomaly Detection In Edge-Attributed Networks,” Proc. IEEE 16th Intl Conf. Data Mining Workshops (ICDMW), vol. 327-34. 2016.

[18] N. S. Chowdhary, A. A. Pandit, “Fake Review Detection using Classification,” International Journal of Computer Applications, vol. vol. 180, no. 50, pp. 16-21. 2018.

[19] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, “Copy-Catch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks,” Proc. ACM 22nd Intl. Conf. World Wide Web (WWW), vol. 119-30. 2013.

[20] Y. Ren, D. Ji, “Research on False Comment Recognition Based on PU Learning Algorithm,” Computer Research and Development, vol. vol. 52, no. 3, pp. 639-48. 2015.

[21] D. He, M. Pan, K. Hong, et al., “Fake review detection based on Pu learning and behavior density,” IEEE Network, vol. 34(4): 298-303. 2020.

[22] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, vol. 174-181. 1997.

[23] B. Liu, “Sentiment Analysis and Subjectivity,” Handbook of natural language processing, 2010.

[24] V. V. Venkataraman, T. S. Kraft, and N. J. Dominy, “Tree climbing and human evolution,” Proceedings of the national academy of sciences, vol. 110(4): 1237-1242. 2013.

[25] A. Mukherjee, V. Venkataraman, and B. Liu, “Fake review detection: Classification and analysis of real and pseudo reviews,” Technical Report, UIC-CS-03-2013, vol. 2013.

[26] B. Qin, and T. Liu, “A language-independent neural network for event detection,” Science China Information Sciences, vol. 61: 1-12. 2018.

[27] Our World in Data, “The rise of social media,” [Online]. Available: https://ourworldindata.org/rise-of-social-media, 2020.

[28] Quora, “What is the total size (storage capacity) of YouTube and at what rate is it increasing?,” [Online]. Available: https://www.quora.com/..., 2020.

[29] S. Elbagir, J. Yang, “Sentiment analysis of twitter data using machine learning techniques and scikit-learn,” Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, vol. 1-5. 2018.

[30] B. Liu, “Sentiment analysis,” 2019.

[31] I. Chaturvedi, E. Cambria, R. E. Welsch, et al., “Distinguishing between facts and opinions for sentiment analysis: Survey and challenges,” Information Fusion, vol. 44: 65-77. 2018.

[32] M. Soleymani, D. Garcia, B. Jou, et al., “A survey of multimodal sentiment analysis,” Image and Vision Computing, vol. 65: 3-14. 2017.

[33] B. Agarwal, N. Mittal, “Machine learning approach for sentiment analysis,” Prominent feature extraction for sentiment analysis, vol. 21-45. 2016.

[34] V. Vapnik, *The nature of statistical learning theory,* Springer, 1999.

[35] D. M. Blei, M. I. Jordan, “Modeling annotated data,” Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, vol. 2003.

[36] B. Liu, L. Zhang, “A survey of opinion mining and sentiment analysis,” Mining text data, vol. pp. 415-463. 2012.

[37] L. L. Benites-Lazaro, L. Giattia, A. Giarollab, “Topic modeling method for analyzing social actor discourses on climate change, energy and food security,” Energy Research and Social Science, vol. 45: 318-330. 2018.

[38] H. X. Song, X. Yan, Z. T. Yu, L. B. Shi, & X. Su, “Fake comment detection based on adaptive clustering,” Journal of Nanjing University (Natural Science Edition), 49(4), 433-439, 2013.

[39] Y. Chen, L. F. Tan, “Research on the governance strategy of false review information of online commodities,” Modern Intelligence, 35(2), 150-153, 2015.

[40] B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, C. Faloutsos, “Birdnest: Bayesian inference for ratings-fraud detection,” In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 495-503), 2016.

[41] S. Rayana, & L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata,” In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 985-994), 2015.

저 자 소 개

가 중 정 (Zhongting Jia)

・2023년 8월 : 순천향대학교 자유전공학과(문학사)
・2023년 8월~현재 : 순천향대학교 경영학과 석사과정
<관심분야> : 디지털 마케팅, 소비자행동

김 엘 레 나 (Elena Kim)

・2020년 8월 : 순천향대학교 미디어커뮤니케이션 학과(미디어학사)
・2024년 2월 : 순천향대학교 경영학과(경영학석사)
・2025년 2월~현재 : 순천향대학교 경영학과 박사과정
<관심분야> : 디지털 마케팅, 소비자행동

최 재 원(Jaewon Choi)

・2004년 2월 : 가톨릭대학교 경영학과 (경영학사)
・2006년 2월 : 가톨릭대학교 경영학과 (경영학석사)
・2010년 8월 : 가톨릭대학교 경영학과 (경영학박사)
・2014년 3월~현재 : 순천향대학교 경영학과 교수
<관심분야> : 빅데이터 분석, 인공신경망, 소셜네트워크, 지능형의사결정시스템, 추천시스템