KNN 기반의 가짜 리뷰 계정 분류 모델 및 서비스 개발

한철현 ○, 권세빈, 박상근
경희대학교 소프트웨어융합학과
hch2454@khu.ac.kr, sebin4548@khu.ac.kr, sk.park@khu.ac.kr

Developing an Account-Level Fake Reviewer Detection Model and Service using KNN

Cheolhyeon Han○, Sebin Kwon, Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

온라인 쇼핑에서 리뷰는 소비자가 수많은 제품을 선택하는 데 중요한 기준이다. 많은 업체들이 리뷰를 관리하기 시작했으며, 리뷰를 전문적으로 조작하는 업체도 등장했다. 가짜 리뷰를 구분하기 위한 여러 연구가 진행되었지만, 일반 리뷰의 특성을 파악하고 이를 따라한 가짜 리뷰 분류의 어려움 등의 한계점이 존재한다. 본 연구에서는 이런 한계점을 극복하기 위해 가짜 리뷰를 작성하는 계정을 식별하는 알고리즘을 개발한다. 이를 위해, 쿠팡에서 가짜 리뷰 계정과 일반 리뷰 계정을 수집하고 두 계정 간 유의미한 차이가 있는 다섯 가지 속성을 찾아냈다. 이 다섯 가지 속성을 기반으로 가짜 리뷰 계정과 일반 리뷰 계정을 분류할 수 있는 KNN 분류 모델을 개발했다. 그리고 사용자가 실제 쇼핑몰 웹사이트에 접속했을 때 이 분류 모델을 사용할 수 있도록 크롬 확장 플러그인을 제작하여 가짜 리뷰 계정 분류 모델의 활용 가능성을 확인했다.

1. 서 론

온라인 쇼핑몰에서 리뷰는 소비자가 상품을 선택하는 중요한 판단 기준이다[1]. 이는 사회적 디폴트 효과로 인해 특정 선호를 정해두지 않은 상황에서 다른 사람의 선택을 따라가는 경향이 높기 때문이라고 볼 수 있다[2]. 따라서 리뷰의 개수와 담긴 정보에 따라 상품의 판매량이 영향을 받을 수 있으므로 많은 판매자가 리뷰 관리에 각별한 신경을 쓰고 있다. 이렇게 리뷰의 중요성이 커지자, 돈을 받고 리뷰를 전문적으로 관리하는 업체가 등장하였고, 이 업체들로 인한 가짜 리뷰가 사회적 문제로 떠오르고 있다.

이런 조작된 가짜 리뷰는 소비자의 잘못된 판단을 유도하기 때문에 경쟁 질서를 해치는 악의적인 행위이다. He et al.[3]은 가짜 리뷰가 제품의 평점 및 순위에 영향을 줄 수 있음을 밝혔다. 가짜 리뷰가 양산되면 크게 두 가지 문제가 발생한다. 첫 번째로, 소비자들은 상품에 등록된 리뷰가 정직하게 작성된 리뷰인지, 돈을 받고 가짜로 작성된 리뷰인지 알 수 없어서 리뷰의 신뢰도가 낮아진다. 두 번째로, 돈을 받고 가짜 리뷰를 작성해 주는 업체를 이용하는 판매자가 많아질수록, 리뷰를 조작하지 않는 상품 판매자가 검색 순위에서 밀리는 등의 피해를 보게 된다.

가짜 리뷰를 식별하기 위한 다양한 연구가 진행되었다. 각각의 리뷰 텍스트를 분석하여 가짜 리뷰인지 식별하는 방법[4, 6]뿐 아니라 리뷰의 길이, 리뷰 작성 날짜 등 다양한 데이터를 활용한 가짜 리뷰 식별 연구[5, 8] 등이 있다. 이런 방법을 통해 높은 정확도로 가짜 리뷰를 식별하는 것은 가능하지만, 기존 연구에서 제시한 일반 리뷰와 가짜 리뷰를 구분하는 특징을 찾아낸 다음에 일반 리뷰의 특징을 흉내 낸 가짜 리뷰를 작성하면 구분하기가 어려워진다는 한계가 존재한다.

이 한계를 극복하기 위해, 본 논문에서는 가짜 리뷰를 식별하는 대신 가짜 리뷰어를 식별해 내는 새로운 접근 방식을 제시한다. 이를 위해 가짜 리뷰어를 식별하기 위한 새로운 특징들을 찾아내고, 이를 기반으로 머신러닝 기법을 활용해 가짜 리뷰어와 일반 리뷰어를 구분한다. 나아가 실제 온라인 쇼핑몰에 접속했을 때, 사용자가 보는 제품의 리뷰 중에서 특정 리뷰는 가짜 리뷰를 주로 작성하는 가짜 리뷰어에 의해 작성된 리뷰임을 알려줄 수 있는 기능을 구현하여 소비자의 올바른 상품 판단을 돕고 리뷰 조작의 영향력을 감소시키는 서비스를 구현해 그 활용성을 확인한다.

2. 관련 연구

리뷰 텍스트 분석을 활용해서 가짜 리뷰를 찾아내기 위한 다양한 연구가 수행되었다. Alsubari et al. [4] 은 TF-IDF 와 N-gram 기법을 활용해서 호텔 리뷰 텍스트에서 특징을 추출하고, 여러 가지 머신러닝을 활용해 높은 정확도로 가짜 리뷰를 식별할 수 있음을 보였다. 강지우 등[6]은 한국어 형태소 분석 라이브러리를 활용해 음식점 리뷰 텍스트를 어간 단위로 재구성하고, 이를 기반으로 머신러닝을 활용해 가짜 리뷰를 판별했다.

가짜 리뷰를 판별하기 위해, 리뷰 텍스트 외에 다양한 특징을 활용하는 연구도 있다. 이민철 & 윤형식[8]은 블로그에서 가짜 리뷰(광고 포스팅)를 식별하기 위해 텍스트뿐 아니라 해당 포스트의 길이, 작성 날짜, 사용된 이미지 수 등의 다양한 데이터를 활용했다. 이를 통해, 텍스트 외의 다양한 데이터도 가짜 리뷰를 식별하는 것에 중요한 요인이 될 수 있음을 확인하였다. Mohawesh et al. [5] 또한 리뷰 텍스트의 특징뿐 아니라 리뷰 작성자의 행동 패턴(예: 긍정 리뷰 비율, 평균 리뷰 길이 등)을 활용해 딥러닝으로 가짜 리뷰를 식별할 수 있음을 보였다.

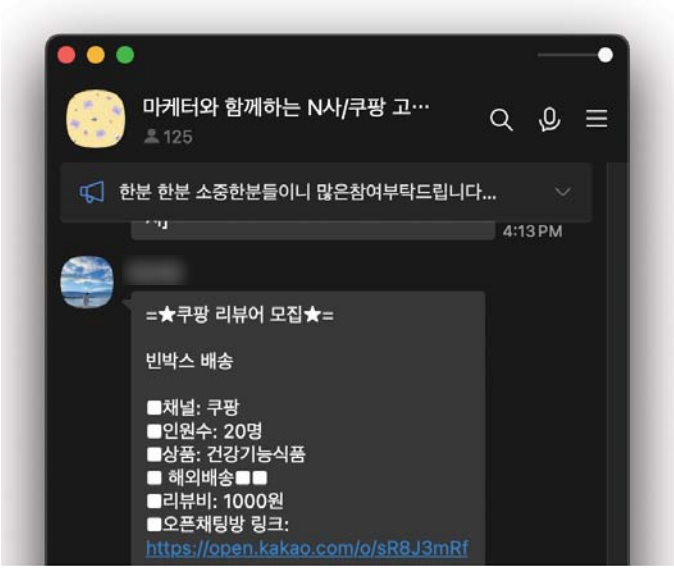
Patel & Patel [7]은 다양한 알고리즘을 비교 분석하면서 기존의 가짜 리뷰 식별 연구의 한계점을 지적했다. 가짜 리뷰와 일반 리뷰를 구분하고, 각 리뷰의 특징을 추출하여 학습시키는 방법으로는 일반 리뷰를 흉내 낸 가짜 리뷰를 식별하기가 매우 어렵다는 점이다. 예를 들어, 특정 가짜 리뷰가 일반 리뷰와 비슷한 길이로 작성되고, 비슷한 이미지 개수를 포함하며, 비슷한 단어를 사용하면 일반 리뷰인지 가짜 리뷰인지 식별이 어렵게 된다.

본 논문에서는 기존 연구에서 제시한 다양한 특징 추출 방법을 참고하되, 가짜 리뷰를 리뷰 단위로 분석하는 것이 아니라 가짜 리뷰를 작성하는 리뷰어 계정을 찾아내는 새로운 접근 방법을 제시한다. 리뷰 하나하나의 특징에 크게 의존하지 않기 때문에, 특정 리뷰가 일반 리뷰처럼 보인다고 하더라도 가짜 리뷰어의 리뷰라는 것을 알아내는 것 자체만으로 해당 리뷰를 의심할 수 있는 합리적 근거를 제시할 수 있다. 그뿐만 아니라, 가짜 리뷰어 식별 알고리즘을 활용하여 실제 온라인 쇼핑몰에서 작동하는 서비스를 구축했고, 이를 통해 이 알고리즘의 활용 가능성을 확인했다.

3. 학습 데이터 수집

3-1 가짜 리뷰 계정 수집

가짜 리뷰 계정과 일반 리뷰 계정을 분류하기 위한 데이터를 확보하기 위해, 가짜 리뷰어를 모집하는 카카오톡 오픈채팅방에 직접 접속해서 어떤 과정으로 가짜 리뷰가 양산되고 있는지 관찰했다 [그림 1]. 쿠팡에서 '빈 박스 배송'을 활용해 가짜 리뷰를 작성하는 경우가 많음을 확인했다. 빈 박스 배송이란, 가짜 리뷰 작성을 위해 고용된 가짜 리뷰어가 가짜 리뷰 작성 대상 제품을 구매하면, 판매자는 가짜 리뷰어 구매자에게 실제 제품을 배송하지 않고 빈 박스만 배송하는 것을 말한다. 판매자는 실제로 제품을 보내지 않고 빈 박스만 배송했지만, 쿠팡에는 마치 실제로 구매가 이뤄진 것으로 기록이 남는다. 그러면 빈



[그림 1]가짜 리뷰 작성자 모집 모습

박스를 배송받은 가짜 리뷰어는 마치 실제 제품을 구매한 것처럼 쿠팡에 가짜 리뷰를 작성할 수 있다.

가짜 리뷰어 모집 오픈채팅방을 통해 가짜 리뷰 작성 대상 제품 리스트 52 개와 제품별 빈 박스 배송 일정을 확인했다. 제품별 리뷰 중 빈 박스 배송 시작 이후 3 일 이내에 리뷰를 작성한 계정의 ID 를 모두 수집했다. 이렇게 수집한 리뷰어 ID 중 15 개 이상의 가짜 리뷰 작성 대상 제품에 리뷰를 작성한 ID 를 가짜 리뷰어 ID 로 가정하였다.

3-2 가짜 리뷰 계정 분류를 위한 가설 설정

특정 계정이 가짜 리뷰를 작성하는 계정인지 구분하기 위한 분류 모델을 학습하기 위해서는 가짜 리뷰 계정과 일반 계정의 차이를 구분할 수 있는 특징을 찾아야 한다. 가짜 리뷰 계정과 일반 계정을 분류하는데 적절한 데이터를 찾기 위해 다음 다섯 가지 가설을 선정했다.

a. **리뷰수**: 가짜 리뷰 계정은 일반 고객보다 리뷰를 더 많이 작성한다.

b. **성실도**: 가짜 리뷰 계정에서 구입한 제품은 성실한 리뷰의 비중이 높다. (성실한 리뷰: 쿠팡에서 별점 리뷰(1~5 점) 외에 별도로 "예상보다 맛있어요", "괜찮아요", "예상보다 맛없어요" 등을 선택하는 리뷰)

c. **쿠팡비율**: 가짜 리뷰 계정의 구매 제품은 로켓배송 등 판매자가 쿠팡(주)인 제품인 비율이 낮다. (가짜 리뷰는 빈 박스 배송을 통해 이뤄지는데, 로켓 배송 제품은 빈 박스 배송이 불가능하기 때문)

d. **글자수**: 가짜 리뷰 계정은 일반 고객보다 리뷰를 더 길게 작성한다.

e. **작성일수**: 가짜 리뷰 계정은 리뷰를 작성한 날짜가 많다. (리뷰 아르바이트라는 특성상 가짜 리뷰 작성자는 일반 고객보다 더 많은 리뷰를 작성한다고 가정)

3-3 가설 검증

2023 년 10 월 05 일부터 10 월 22 일까지 총 18 일 동안 가짜 리뷰 작성 중개인에게 안내받은 가짜 리뷰어 모집 제품 52 개의 쿠팡 URL 을 확보했다. 구글에서 만든 Node.js 라이브러리인 Puppeteer[1]를 활용해 해당 제품들의 쿠팡 제품 페이지에 접속한 후, 리뷰를 작성한 계정의 ID 를 모두 수집했다. 이 중 빈 박스 배송 시작 이후 3 일 이내에 리뷰를 작성한 379 개 계정을 확보하고 이를 가짜 리뷰 계정 ID 로 가정했다.

가짜 리뷰 계정이 아닌 일반 계정 확보를 위해, 쿠팡에서 애플(Apple)의 맥북 제품 판매 페이지에 접속하여 리뷰를 작성한 계정의 ID 380 개를 수집했다. 애플 같은 유명 제조사 들은 리뷰 조작을 하지 않고도 높은 품질과 평판이 보장되었기에 가짜 리뷰를 사용하지 않기 때문이다. 유명 제조사의 제품 리뷰 페이지 속에서 평점별로 동일한 수의 ID 를 랜덤 추출하여 일반 계정 380 개를 선정했다.

Puppeteer 를 다시 한번 활용해서, 수집된 모든 가짜 리뷰 계정과 일반 계정의 리뷰 이력 페이지에서 계정별로 '리뷰수', '성실도', '쿠팡비율', '글자수', '작성일수' 데이터[2]를 수집했다.

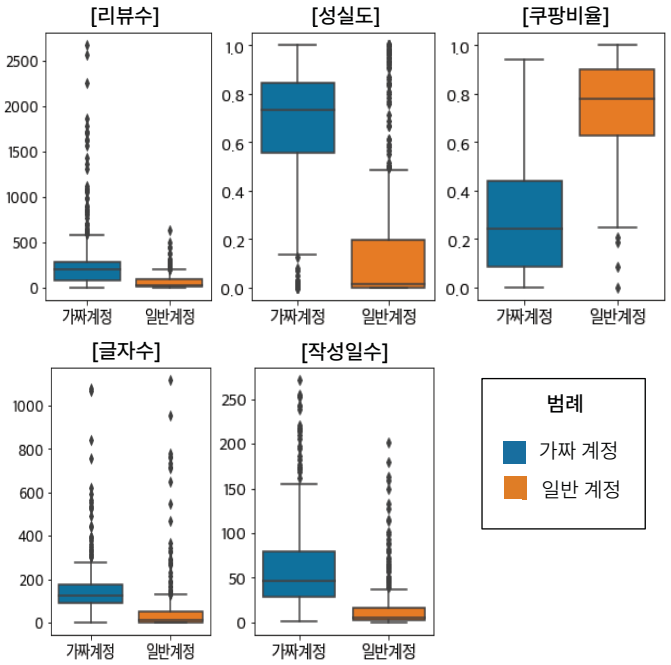
- 리뷰수 : 해당 계정이 작성한 리뷰 개수
- 성실도 : 전체 리뷰 중 '성실한 리뷰' 비율
- 쿠팡비율 :전체 리뷰 중 판매자가 쿠팡(주)로 표기된 제품에 작성한 리뷰의 비율
- 글자수 : 평균 리뷰 글자수
- 작성일수 : 리뷰를 작성된 날짜의 수

[표 1] 속성별 가짜/일반 계정의 평균 및 표준편차

	리뷰수	성실도	쿠팡비율	글자수	작성일수
가짜 계정	293 (SD=372)	0.66 (SD=0.25)	0.29 (SD=0.24)	153 (SD=127)	62 (SD=53)
일반 계정	69 (SD=94)	0.17 (SD=0.29)	0.75 (SD=0.20)	56 (SD=137)	16 (SD=28)

수집된 데이터를 기반으로, 가짜 리뷰 계정과 일반 계정의 차이를 확인하였다 [표 1]. 가짜 리뷰 계정과 일반 계정을 분류하는데 해당 데이터를 사용하는 것이 적절한지 확인하기 위해, 속성별로 t-test 를 수행했다. 모든 속성에서 p-value 가 유의수준(0.05) 보다 작음을 확인했으며, 이를 통해 각 속성의 분포는 가짜 리뷰 계정과 일반 계정 간 유의미한 차이가 있음을 알 수 있었다.

이렇게 통계적으로 검증된 일반 계정과 가짜 계정 간의 분포 차이를 시각화하면 가짜 리뷰 계정과 일반 계정 간에 차이가 있음을 한눈에 확인할 수 있다 [그림 2].



[그림 2] 5 가지 속성별 전체 데이터 Boxplot

4. 가짜 리뷰 계정 & 일반 계정 분류 모델

4-1 학습 데이터 전처리

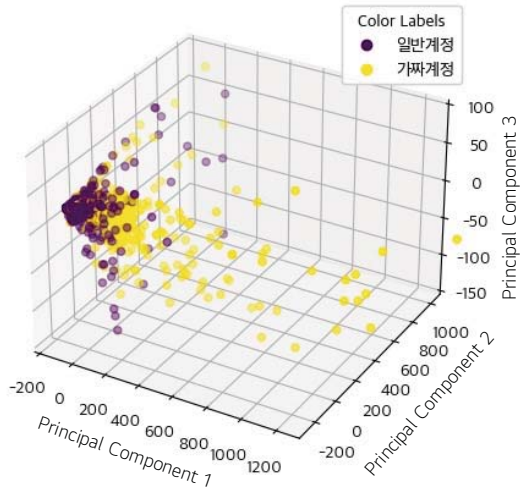
[표 2] 5 가지 속성 간 상관관계

	리뷰수	성실도	쿠팡비율	글자수	작성일수
리뷰수	1.00	0.38	-0.33	0.15	0.80
성실도	0.38	1.00	-0.57	0.56	0.53
쿠팡비율	-0.33	-0.57	1.00	-0.12	-0.36
글자수	0.15	0.56	-0.12	1.00	0.34
작성일수	0.80	0.53	-0.36	0.34	1.00

모델을 학습하기 전 다중공선성 문제를 확인하기 위해, 속성 간 상관관계를 계산했다 [표 2]. 리뷰수와 작성일수 사이에는 0.8 의 강한 상관관계가 있음을 확인했다. 강한 상관관계를 가진 속성 들은 다중공선성의 우려가 있기에, PCA 를 이용하여 차원을 축소했다. PCA 결과로 3 개의 주성분이 누적 분산 99.99% 로 설명할 수 있다는 결과를 확인할 수 있었다. [그림 3]은 PCA 를 수행한 후 일반 계정과 가짜 리뷰 계정 집단 간 주성분 분포를 확인한 결과이다. 두 집단이 비교적 명확하게 분리된 것을 확인할 수 있다.

[1] https://pptr.dev/

[2] https://github.com/festring/coupang_review_dataset



[그림 3] PCA 시각화

4-2 모델 구축 및 평가

앞에서 차원 축소를 통해 얻은 주성분을 기반으로, 가짜 리뷰 계정과 일반 계정을 분류하기 위해 노이즈에 강한 K-Nearest Neighbors (KNN) 알고리즘을 사용했다. K 값을 적절히 선택하여 노이즈에도 일관된 결괏값을 얻기 위해서이다. 데이터의 총개수가 적은 부분을 보완하기 위해 K-Fold 교차 검증을 사용했고, 데이터의 개수를 고려하여 Fold 값은 3 으로 설정했다. scikit-learn 1.3.2 [3] 의 KNeighborsClassifier 클래스와 cross_val_score 클래스를 활용했고, 파라미터는 모두 기본값으로 설정했다.

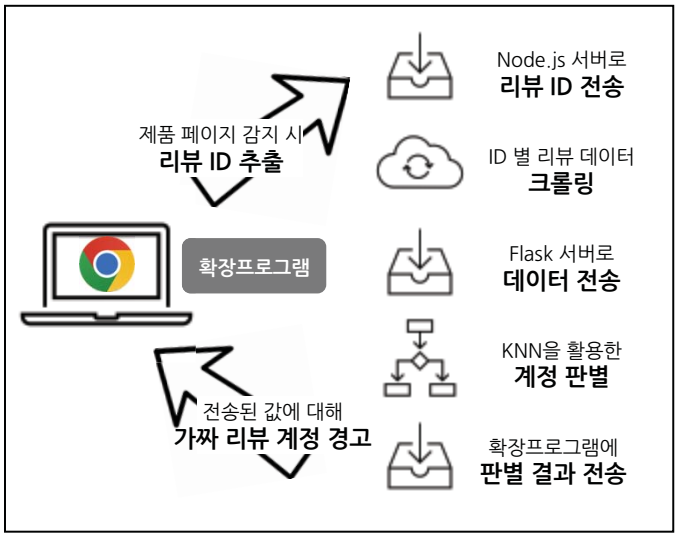
[표 3] K 값에 따른 모델의 3-fold 교차검증 결과

	k = 1	k = 3	k = 5	k = 7	k = 9
F1 Score	0.83 (SD=0.05)	0.86 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)
Recall	0.83 (SD=0.05)	0.86 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)
Precision	0.84 (SD=0.05)	0.86 (SD=0.02)	0.86 (SD=0.02)	0.85 (SD=0.02)	0.86 (SD=0.01)
Accuracy	0.83 (SD=0.05)	0.86 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)	0.85 (SD=0.02)

[표 3]은 홀수인 K 값에 따른 KNN 모델 평가 지표들이다. K 값이 1 일 때 값이 가장 낮고, 3 이상에서는 큰 변화 없이 일정한 값들이 나타난다. 이를 통해 현재 데이터셋에서 모델은 K 값이 3 이상 일 경우 안정적으로 작동함을 알 수 있다. 본 연구에서는 K 값을 7 로 설정하고 온라인 쇼핑몰 가짜 리뷰 계정 판별 서비스 구현에 활용하기로 결정했다.

5. 온라인 쇼핑몰 가짜 리뷰 계정 판별 서비스 개발

5-1 작동 구조



[그림 4] 가짜 리뷰 계정 판별 서비스 작동 구조

본 논문에서 구현한 가짜 리뷰 계정 판별 알고리즘을 사용자가 직접 사용해 볼 수 있는 서비스를 개발했다. 서비스의 작동 구조는 [그림 4]와 같다. JavaScript 를 기반으로 한 크롬 플러그인 형태로 개발하여 사용자가 간편하게 설치하고 사용할 수 있다. 사용자가 본 크롬 플러그인의 모드를 ON 으로 해놓고 쿠팡의 특정 제품 페이지에 접속하면, 해당 리뷰 페이지에 리뷰를 작성한 계정의 ID 가 모두 Node.js 로 구현한 서버로 전송된다. 전송된 ID 를 서버에서 각 리뷰 계정 ID 가 작성한 모든 리뷰를 수집한 다음 [표 1]에서 언급한 다섯 가지 속성을 Flask [4] 로 구현한 서버로 전송한다. 이 과정이 별도의 서버에서 이뤄지기 때문에, 사용자는 다른 방해를 받지 않고 계속 쇼핑을 진행할 수 있다.

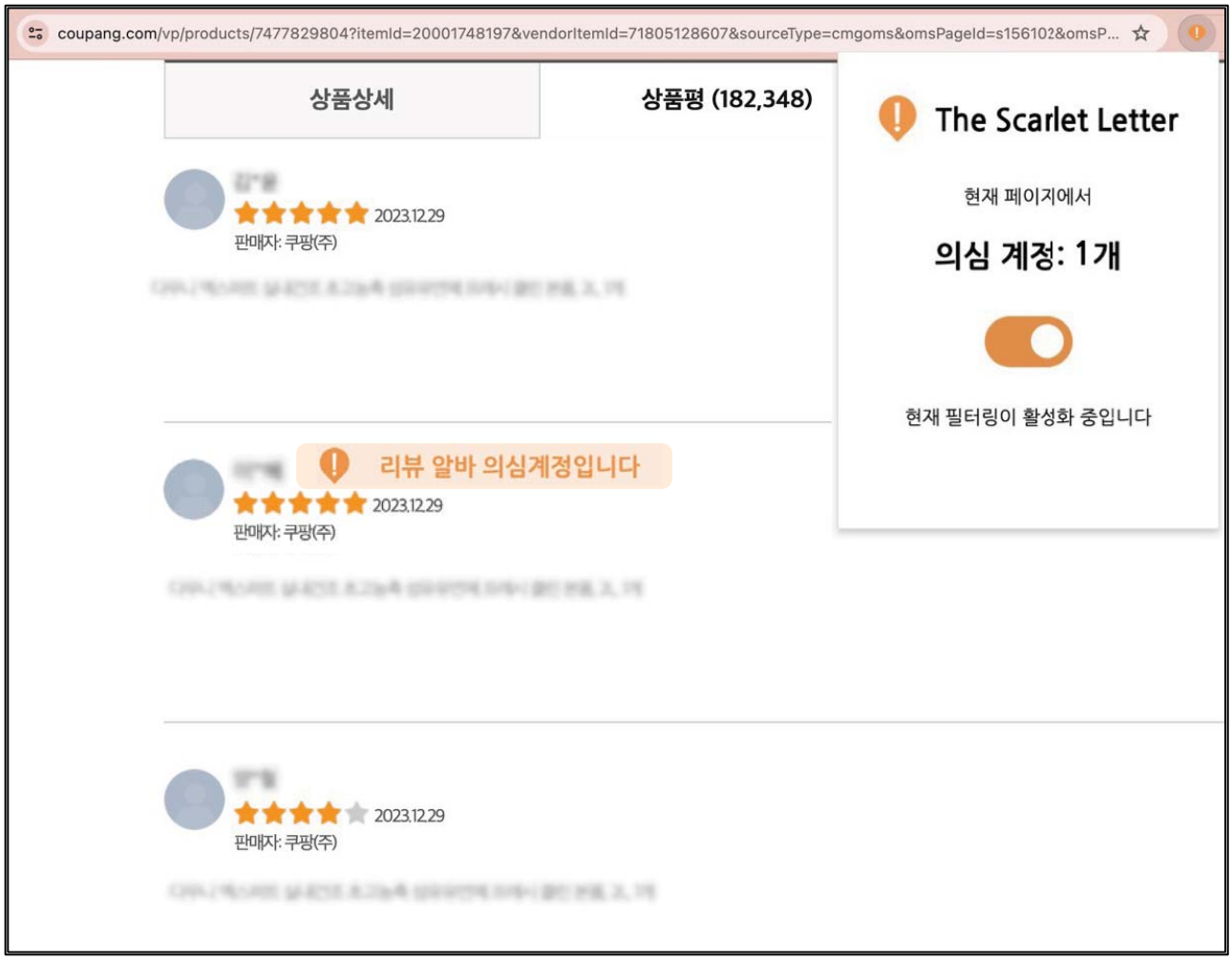
Flask 서버에서는 PCA 를 수행하여 속성을 3 개로 축소하고, 미리 학습된 KNN 모델에 입력하여 각 리뷰 계정 ID 가 가짜 리뷰 계정인지 일반 계정인지 분류한다. 분류된 결과는 다시 확장 플러그인에 전송되고, 확장 플러그인은 가짜 리뷰 계정으로 분류된 계정의 오른쪽에 경고 표시를 띄운다 [그림 5].

5-2 서비스 기능

최종 결과물의 핵심 기능으로 대시보드 기능과 의심 계정 표시 기능이 있다. 대시보드는 크롬 브라우저의 우측 상단에 존재하는 크롬 확장 플러그인 중 해당 플러그인 아이콘을 클릭하면 실행된다 [그림 5. 우측 상단]. 대시보드의 ON/OFF 기능으로 사용자가 원할 때만 가짜 리뷰 계정 경고 기능을 활성화/비활성화할 수 있다. 또한, 가짜 리뷰 계정으로 의심되는 계정이 몇 개인지 바로

[3] https://scikit-learn.org

[4] https://flask.palletsprojects.com/en/3.0.x/



[그림 5] 쿠팡 페이지에서 실제 작동 모습

표시함으로써, 현재 접속 중인 쿠팡 제품의 리뷰를 볼 때보다 주의를 기울일 수 있도록 구현하였다.

쿠팡 제품의 리뷰 페이지에서, 가짜 리뷰 계정으로 의심되는 계정에는 직접 HTML 요소를 삽입하여 “리뷰 알바 의심계정입니다”라는 문구를 표시한다 [그림 5. 가운데]. 이를 통해, 사용자는 가짜 리뷰 계정으로 추정되는 계정이 작성한 리뷰를 주의 깊게 살펴보면서 구매 여부를 결정하는 데 참고할 수 있다.

6. 결론 및 향후 연구

본 연구에서는 인터넷 쇼핑몰에 만연한 가짜 리뷰 문제를 해결하기 위해, 기존의 가짜 리뷰 판별이 아닌 가짜 리뷰 계정을 판별하는 머신러닝 모델을 개발하고, 이를 활용한 서비스를 직접 구현하여 그 활용 가능성을 확인했다. 본 연구의 아이디어를 확장하면, 가짜 리뷰 계정뿐 아니라 가짜 정보를 양산하는 계정을 찾아내는 등, 기존과 다른 방법으로 사용자의 정상적인 판단을 저해하는 거짓 정보를 차단하는 데 도움이 될 수 있다.

본 연구를 수행하기 위해서, 실제로 가짜 리뷰어를 모집하는 오픈채팅방에 접속하여 가짜 리뷰가 작성되는 제품 리스트를 확보하고, 이를 기반으로 가짜 리뷰의 특징 등을 선별했다. 하지만 여전히 가짜 리뷰라고 가정한 계정이 100% 가짜 리뷰 계정이라고 결정지을 수 없다는 한계가 존재한다. 이에 따라 생존 편향(Survivorship bias) 문제가 있을 수 있다. 향후 연구에서는 가짜 리뷰 계정과 일반 계정을 보다 정확하게 분류할 수 있는 기준을 마련하고, 적절한 K 값 설정을 통한 기존 KNN 알고리즘의 성능 향상 및 앙상블 등 다양한 기법을 활용해 모델의 정확성과 서비스의 활용성을 높이고자 한다.

Acknowledgement

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2023년도 SW중심대학사업의 결과로 수행되었음" (2023-0-00042)

참고 문헌

[1] 김세라. 온라인 쇼핑 이용후기 '실제 구매에 큰 영향' 미쳐. 소비자 경제, https://www.dailycnc.com/news/articleView.html?idxno=209683, 2023.

[2] Young Eun Huh, Joachim Vosgerau, Carey K. Morewedge. Social Defaults: Observed Choices Become Choice Defaults. Journal of Consumer Research, 41, 3, 746-760, 2014.

[3] Sherry He, Brett Hollenbeck, Davide Proserpio. The Market for Fake Reviews. Marketing Science, 41, 5, 896-921, 2022.
[4] Saleh Nagi Alsubari, Sachin N. Deshmukh, Ahmed Abdullah Alqarni, Nizar Alsharif, Theyazn H. H. Aldhyani, Fawaz Waselallah Alsaade, Osamah I. Khalaf. Data Analytics for the Identification of Fake Reviews Using Supervised Learning, Computers, Materials & Continua, 70, 2, 3189-3204, 2022.
[5] Rami Mohawesh, Shuxiang Xu, Son N. Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, Sumbal Maqsood. Fake Reviews Detection: A Survey, IEEE Access, 9, 65771-65802, 2021.
[6] 강지우, 김동욱, 송이현, 이석범, 이범진, 정윤경. 음식점 가짜 리뷰 판별을 위한 기계학습 방법 비교. 한국정보과학회 학술발표논문집, 1980-1982, 2017.
[7] N. A. Patel & R. Patel. A Survey on Fake Review Detection using Machine Learning Techniques, 2018 4th International Conference on Computing Communication and Automation (ICCCA), 1-6, 2018.
[8] 이민철 & 윤현식. 머신러닝을 활용한 가짜리뷰 탐지 연구: 사용자 행동 분석을 중심으로. 지식경영연구, 21, 3, 177-195, 2020.