AD Finder : 네이버 광고성 블로그 리뷰 탐지 프로그램

조민서 ○ 박상근
경희대학교 식품영양학과, 경희대학교 소프트웨어융합학과
choms0209@khu.ac.kr, sk.park@khu.ac.kr

AD Finder: Naver advertising blog review detection program

Minseo Jo○ Sangkuen Park
Department of Food and Nutrition, Kyung Hee University
Department of Software Convergence, Kyung Hee University

요 약

최근 바이럴 마케팅 광고 기법이 성행함에 따라 광고주들은 유명 인플루언서 대신 일반인을 동원하여 이용 후기를 거짓으로 작성하거나 과장,확대하여 작성하게 하여 광고임을 숨기려고 하고있다. 이로 인해 소비자들은 광고를 목적으로 하는 정보에 대한 불신감이 커져가고 있다. 따라서 이러한 '바이럴 마케팅', '뒷광고'로부터 정보를 걸러주는 도구가 필요한 환경이 되었다. 본 연구는 머신러닝 기술을 통해 광고 성 블로그 리뷰를 예측, 분류하여 사용자에게 광고 표시 기능을 제공한다. 이는 과장된 리뷰에 쉽게 속지 않도록 해줄 수 있을 뿐만 아니라 협찬 문구를 표시하지 않은 불법 블로그 리뷰까지 탐지하여 향후 건전한 소비 문화 조성에 기여할 것이다.

1. 서 론

최근 저렴하고 광고 효과가 큰 바이럴 마케팅이 성행함에 따라 많은 광고 업체들은 SNS을 통해 순수하게 정보를 제공하는 척을 하며 홍보효과를 극대화 하려는 방식을 사용하고 있다.[1] 하지만, 바이럴 마케팅의 경우 홍보에만 치중하다 보니 이용 후기를 거짓으로 작성하거나 과장, 확대하는 경우가 많다.

이로 인해 소비자들은 바이럴 마케팅에 대한 인식이 부정적으로 변하고 광고를 목적으로 하는 정보에 대해서 불신감이 커져가고 있다.

2년여전 유튜브의 유명 유튜버들 사이에서 논란이 되었던 뒷광고 사건으로 인해 공정위에서는 경제적 대가를 받고 특정 상품에 대한 후기를 작성하는 경우 광고 표시 문구를 작성하도록 표시광고법을 시행하였다. 하지만 이러한 법적제제에도 불구하고 후기 작성자들은 표시 문구를 제대로 표기하지 않고 있다. 또한 광고주들은 유명 인플루언서 대신 일반인들에게 대가를 제공하면서 광고 콘텐츠를 게시하게하여 광고임을 숨기려고 하고 있다.[2]

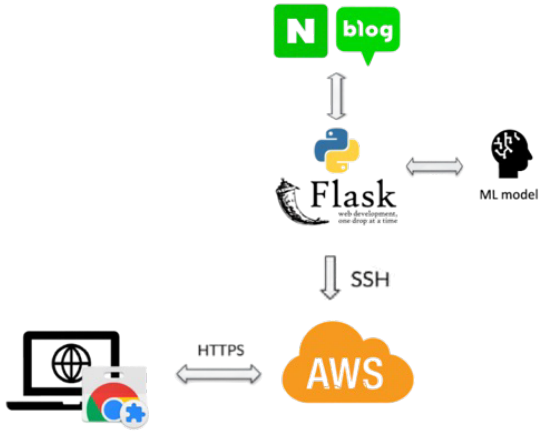
이러한 문제로 인해 소비자들은 '내돈내산' '솔직 후기'에 대한 포스팅 리뷰를 더 찾아보고 싶어하는 심리가 자연스럽게 생겨나게 되었다.

하지만 대부분의 블로그 포스팅 리뷰에는 협찬 문구가 글의 마지막에 위치해있어 소비자가 포스팅을 끝까지 읽어 직접 확인하지 않는 이상 광고를 식별하기 어렵다.[3]

해당 문제를 해결하기 위해 블로그 포스팅 리뷰 광고 의심 여부를 표시하는 기능을 도입하여, 소비자들이 믿을 수 있는 정보에 더 쉽게 접근하고 현명한 소비를 할 수 있도록 하는 프로그램을 개발하고자 한다.

2. 연구 설계

본 연구는 광고 블로그 리뷰 모델 개발과 API 서버 구축, 배포 단계로 구성되어 있다.
전체적인 연구 진행 구조는 [그림 1]과 같다.



[그림 1] 네이버 광고성 블로그 리뷰 탐지 서비스 구조도

2.1 데이터 수집

네이버 블로그 리뷰 데이터를 수집하기 위해 주요 검색 키워드를 선정하여 총 530개의 블로그 포스팅 데이터를 수집하였다. 블로그 리뷰 특징 정보를 수집하기 위한 웹 크롤링에서는 Python 라이브러리 BeautifulSoup[1]가 사용된다.

BeautifulSoup을 활용해서 블로그의 포스트 제목, 포스트 URL, 포스트 작성날짜, 포스트 내용, 포스트 내 이미지 개수, 포스트 내 비디오 포함 여부, 댓글 수 총 7개의 정보를 스크래핑하였다.

2.1.2 블로그 특징 추출

블로그 포스팅 내의 구조는 일반인 블로그와 상업적 목적을 갖는 블로그에 따라 차이가 있을 것이라고 판단하였다.[4] 따라서 광고 블로그를 판독하기 위하여 광고 포스팅만의 특정 패턴을 나타낼 수 있는 총 8가지 예측 변수를 설정하였다.

추출된 8가지 변수는 <표 1>과 같다.

<표 1> 예측 변수

변수명	변수 설명
Post Count	카테고리 내 게시글 수
Posted A week ago	최근 일주일 내 게시글 작성 여부(1,0)
Post length	포스트 내용 길이
Sponsored Word	포스트 내용 내 협찬문구 포함 여부(1,0)
Keyword('내돈내산')	포스트 내용 내 '내돈내산' 키워드 포함 여부(1,0)
Image Count	포스트 내용 내 이미지 개수
Video or not	포스트 내용 내 비디오 콘텐츠 포함 여부(1,0)
Review Count	포스트 댓글 개수

2.1.3. 데이터 레이블링

광고와 비광고를 분류해주기 위한 모델 학습 데이터셋을 구축하기 위해 직접 레이블링 과정을 수행하였다. 국내에서는 공정거래위원회의 추천 ·보증 등에 관한 표시·광고 심사지침에 의해 경제적 대가를 받고 특정 상품에 대한 후기를 작성하는 경우 표시광고법에 따라 [그림 2]와 같이 표시 문구를 작성해야한다.[5]

따라서 블로그 포스팅 내 광고 표시법의 권장 문구의 여부에 따라 광고 리뷰 판정과 레이블링 작업을 수행하였다.

[그림 2] 추천·보증 등에 관한 표시·광고 심사지침 실제 이행 예시

2.1.4 데이터 전처리

수집된 데이터 전처리를 위해 결측치, 이상치, 그리고 정규화 작업을 진행하였다. 구축된 데이터 셋은 협찬 문구가 있으면 모두 광고성 데이터로 레이블링 해주었기 때문에 Sponsored Word 변수의 값이 1이면 레이블 값은 1이 될 수 밖에 없다. 이렇게 되면 모델은 Sponsored Word 변수에만 지나치게 의존하게 되어 광고이지만 협찬문구를 표기하지 않은 불법성 블로그 리뷰를 잡아내지 못한다는 문제가 생긴다.

따라서 모델이 한 가지 변수에 지나치게 의존하게 되는 문제를 해결하고자 데이터 증강기법을 활용하였다. 광고 데이터 (label =1)의 Sponsored Word 변수 값을 모두 0에서 1로 변경한 추가데이터셋을 기존의 데이터셋에 결합해주었다. 이를 통해 기존보다 277개의 추가된 총 777개의 데이터셋을 구축하였다.

2.1.5 EDA 및 통계 분석

탐색적 데이터 분석 (Exploratory Data Analysis, EDA)은 데이터의 패턴, 관계, 분포를 파악하는데 중요한 역할을 한다.[2] 이 과정에서 통계적 가설 검정을 함께 진행하여 데이터로부터 얻은 결과가 우연히 일어난 것인지 아니면 통계적으로 유의미한 지를 평가할 수 있다.

각 변수 별로 유의수준 0.05 하에서 T-test, 카이제곱검정을 진행하였다.

그 결과, 포스트 길이 (Post length), 이미지 개수(Image Count), 비디오 포함 여부(Video or not) 변수에서 집단간의 차이가 유의미한 것을 확인할 수 있었다.

2.1.6 예측 모델 생성 및 선정

블로그 포스트의 광고성, 비광고성 리뷰 예측을 위해 본 연구에서는 머신러닝의 지도학습 방법을 사용하였다. 예측 모델 중 선행 연구를 바탕으로 예측 정확도가 높은 의사 결정 나무 기반의 앙상블, 로지스틱 회귀, SVM의 총 5가지 종류의 모델을 선정하였다.

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[2] https://www.ibm.com/topics/exploratory-data-analysis

3. 예측 모델 평가 및 선택

3.1 예측 모형 평가

예측 모형의 성과 검증 지표는 Scikit-learn 의 GdSearchCV 라이브러리[6]를 통한 교차 검증(cross-validation)을 사용하였으며 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score, AUC 을 측정하여 검증을 수행하였다. 모든 모델의 성능 결과가 비슷하게 비슷하게 나온 것을 확인할 수 있었고, 이 중 가장 높은 XGBoost 를 최종 모델로 선정하였다. 검증 결과는 <표 2>와 같다.

<표 2> 모델 성능 비교

	XGBoost	LightGBM	CatBoost	Logistic Regression	SVM
Accuracy	0.79	0.79	0.78	0.78	0.78
Recall	0.9	0.87	0.87	0.89	0.89
Precision	0.8	0.82	0.8	0.79	0.79
F1 score	0.85	0.84	0.83	0.83	0.84
AUC	0.86	0.85	0.87	0.86	0.86

4. 크롬 익스텐션 어플리케이션 개발

4.1 블로그 리뷰 표시 기능 서버 구축

사용자가 네이버에 특정 키워드를 검색하였을 때 네이버 view 블로그 창에서 광고 표시 기능을 제공하기 위해 Flask API[3] 서버를 구축하였다. 사용자가 특정 키워드를 검색하게 되면 해당 페이지의 HTML source 데이터를 서버로 전송하여 서버는 HTML 파일을 크롤링하게 되어 블로그 특징 정보 데이터를 축적하게 된다. 그리고 이를 개발한 분류 모델에게 전달해주고 모델은 이에 대해 광고인지 아닌지 예측을 하게 된다. 그리고 그에 대한 결과값을 다시 클라이언트로 전송하여 광고 블로그의 표시 기능을 제공하게 된다.

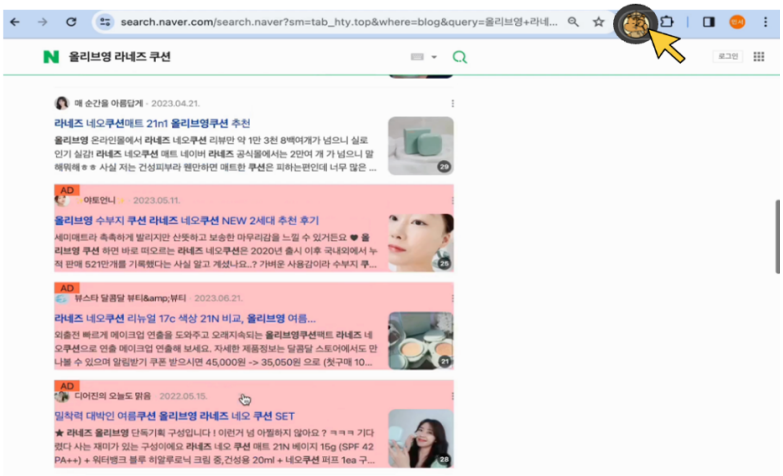
4.2 크롬 익스텐션 애플리케이션 개발

크롬 익스텐션 어플리케이션 형태로 서비스를 제공하기 위하여 크롬에서 제공하는 확장 프로그램 tool[7]을 활용하여 가이드라인에 맞춰 프로그램을 설계하였다. 프로그램 실행동안 발생하는 이벤트를 처리하는 이벤트 핸들러를 Background.js 에 정의하여 사용자가 익스텐션 아이콘을 클릭하였을 때 해당 프로그램이 실행될 수 있도록 하였다.

4.3 AD Finder : 광고성 블로그 포스팅 표시 서비스

AD Finder 어플리케이션은 네이버 블로그 View 창에서 원하는 키워드 검색 후 상단의 크롬 탭에 나타나는 플러그인 아이콘만 누르면 블로그 표시 기능을 제공한다.
개발된 AD Finder 익스텐션 어플리케이션의 UI 는 [그림 3]과 같다.

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



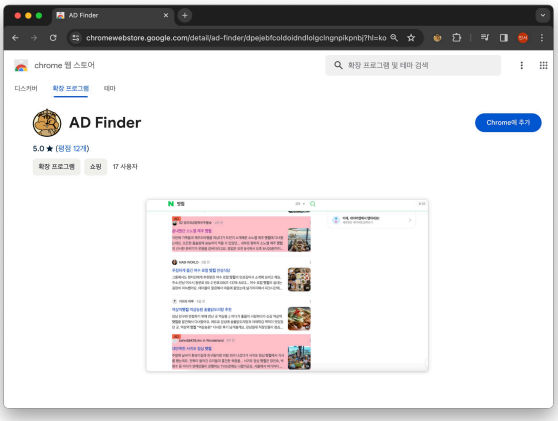
[그림 3] AD Finder 크롬 익스텐션 애플리케이션 UI

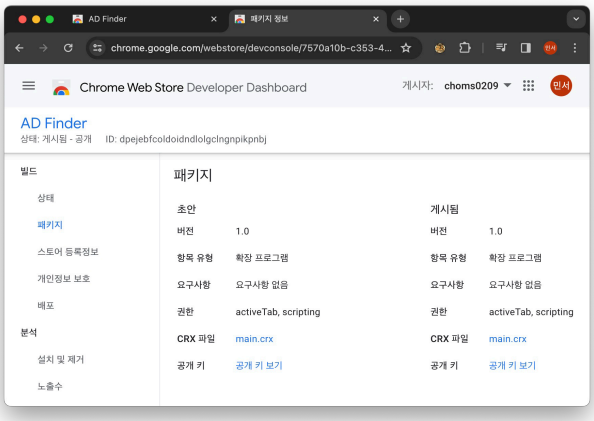
5. 애플리케이션 배포 및 결과

5.1 크롬 익스텐션 애플리케이션 배포

개발한 서비스를 운영하기 위해 AWS EC2 서비스를 이용하여 애플리케이션 배포 과정을 수행하였다. AWS EC2 에 Flask API 서버를 구동 시키고[8] 도메인을 구매하여 SSL 인증서를 받아 클라이언트와 서버가 HTTPS 프로토콜로 통신할 수 있게 구성하였다.

그 후 구글 웹스토어에 애플리케이션 제출 하여 심사기간을 거쳐 최종적으로 다른 호스트 서버에서도 애플리케이션을 다운받아 이용할 수 있게 하였다.[4]





[그림 4,5] Chrome Webstore 에 출시된 AD Finder

[4] https://chromewebstore.google.com/detail/ad-finder/dpejebfcoldoidndlolgclngnpikpnbj?hl=ko

5.2 사용자 반응 결과 확인

구글 크롬 웹스토어의 사용자 리뷰를 통해 실제 해당 서비스를 이용한 사용자들의 의견을 수렴하였다.[5] '클릭 전에 미리 광고 글을 거를 수 있었다.', '진짜 후기만 볼 수 있어 정보 수집이 효율적이다.' 등의 긍정적인 사용자 반응을 확인할 수 있었다.

6. 결론 및 향후 과제

본 연구를 통해 개발한 AD Finder 는 사용자에게 홍보에만 치중하는 광고성 리뷰의 노출을 줄이고 정확한 이용 후기 정보만을 제공하는데 도움을 줄 것이다.

AD Finder 애플리케이션 사용자는 해당 포스트의 URL 에 직접 접속하지 않고도 클릭 한번으로 광고 여부를 판단할 수 있다. 이로써 사용자는 다양한 제약사항에 구애 받지 않고 편리하게 서비스를 이용할 수 있을 것이다.

또한 머신러닝 모델이 협찬 문구 여부 변수에 의존적으로 학습하지 않도록 데이터 증강기법을 적용해주었기 때문에 해당 모델은 협찬 문구가 없는 불법 광고 리뷰까지 탐지할 수 있다. 이는 공정거래위원회가 불법 광고 사례를 탐지하는데 도움을 줄 것이다.

해당 서비스를 더 많은 사용자들에게 확장 시키기 위해 광고 표시 기능 뿐만 아니라 광고 블로그 리뷰는 제외하고 볼 수 있는 기능을 추가하여 필터링 효과를 제공하고자 한다. 또한 네이버 블로그 이외의 데이터를 활용하여 다양한 플랫폼에서 해당 서비스를 이용할 수 있도록 발전시켜볼 예정이다.

[5] https://chromewebstore.google.com/detail/ad-finder/dpejebfcoldoidndlolgclngnpikpnbj/reviews?hl=ko

참고문헌

[1] 강미영. 인플루언서의 뒷광고에 대한 처벌 및 제재 가능성. 법학연구, 2021, 65: 115-146.

[2]정진호, "SNS 에 슬쩍 '뒷광고'…공정위, 9 개월간 1 만 7020 건 적발", 중앙일보, 2022.02.03. https://www.joongang.co.kr/article/25045134#home

[3]서병곤, "SNS 상 '뒷광고' 1 만 7020 건 적발...인스타 56% 차지", 이투데이, 2022.02.02. https://www.etoday.co.kr/news/view/2101782

[4] 이민철, 윤현식. 머신러닝을 활용한 가짜리뷰 탐지 연구: 사용자 행동 분석을 중심으로. 지식경영연구, 2020, 21.3: 177-195.

[5] 공정거래위원회, "추천·보증 등에 관한 표시·광고 심사지침 가이드라인 배포". https://developer.chrome.com/docs/extensions/get-started?hl=ko

[6]Scikit-learn, "GridsearchCV". https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[7]Chrome Extension Guidelines. https://developer.chrome.com/docs/extensions/get-started?hl=ko

[8] AWS EC2 Deploy WebApp Guidelines. https://aws.amazon.com/ko/getting-started/guides/deploy-webapp-ec2/module-one/

[9] 이기성; 이종찬. 머신러닝을 이용한 의료 및 광고 블로그 분류. *한국산학기술학회 논문지*, 2018, 19.11: 730-737.

[10] 노영주, et al. 빅데이터 분석을 활용한 가짜 리뷰 필터링 시스템 ADDAVICHI. 한국인터넷방송통신학회 논문지, 2019, 19.6: 1-8.