딥러닝 기반 이중 등급 분류 모델을 활용한 선정적 광고 탐지 및 정화 시스템

이제희[O], 박상근

경희대학교 소프트웨어융합학과

juventa23@khu.ac.kr, sk.park@khu.ac.kr

Real-Time Filtering System for Sexually Suggestive Ads Using a Deep Learning-Based Dual-Level Classification Model

Jehui Lee[O], Sangkeun Park

Department of Software Convergence, Kyung Hee University

요 약

온라인 환경에서 사용자는 의도치 않게 선정적 광고에 노출되는 경우가 많다. 기존의 선정성 이미지 탐지 연구는 신체 노출에만 초점을 맞추거나 모호한 기준으로 성적 암시를 판단하는 한계를 지닌다. 본 연구는 이러한 한계를 극복하기 위해 한국 방송통신심의위원회의 심의 기준(SafeNet)을 바탕으로 노출 등급과 성적 암시 등급을 결합한 이중 등급 체계를 도입한다. 이를 적용한 선정성 탐지 모델은 노출 등급 F1-score 0.92, 성적 암시 등급 F1-score 0.87이라는 높은 성능을 달성했으며, 실시간 광고 차단이 가능한 크롬 플러그인으로 구현하여 그 활용성을 입증한다. 본 연구는 객관적인 심의 기준을 통해 신체 노출이 적더라도 성적 암시가 강한 이미지를 정밀하게 탐지함으로써 사용자에게 보다 안전하고 쾌적한 브라우징 환경을 제공할 수 있음을 보여준다.

주의 해당 논문에는 선정적인 표현이 포함되어 있습니다

1. 서 론*

인터넷과 모바일의 발달로 누구나 다양한 디지털 콘텐츠에 손쉽게 접근할 수 있게 되었다. 특히 아동 및 청소년들이 정보 습득, 여가 활동 등 다양한 목적으로 디지털 콘텐츠를 탐색하는 과정에서 의도치 않게 폭력적·선정적인 부정적 콘텐츠를 접하고 있어 사회적 문제가 되고 있다[1, 2]. 이와 같이 원치 않는 부정적인 콘텐츠 소비는 단순한 자극을 넘어 개인의 정서와 사회적 태도에 부정적인 영향을 미칠 위험이 있다[3].

특히, 사람들이 자주 접속하는 인터넷 뉴스 사이트에서 아동 및 청소년을 위한 접근 제한 장치 없이 선정적인 광고가 버젓이 게재되는 경우가 많다[4]. 이렇게 명확한 연령 제한 표시나 접근 제한 없이 누구나 접근 가능한 형태로 나타나는 선정적인 광고는 방송통신심의위원회의 청소년유해정보 표시 의무 규정[5]을 위반하는 사례로 볼 수 있다.

이러한 문제를 해결하기 위해 머신러닝 기반 선정적 이미지 탐지 연구가 활발히 진행되고 있다. 이미지에서 신체의 노출 정도를 기준으로 선정성을 정의하고 이를 탐지한 연구들이 수행되었다[6, 7]. 이러한 접근은 단순하면서도 적용이 용이하지만, 신체적 노출이 적으면서도 성적 암시가 강한 이미지를 탐지하지 못한다는 한계가 있다. 선정성을 신체 노출 수위와 콘텐츠에서 드러나는 성적 암시 정도를 함께 고려하여 선정성을 판단한 연구도 존재한다[8, 9, 10]. 하지만 성적 암시에 대한 정의와 기준이 연구자에 따라 주관적인 기준으로 정의되었으며, 선정성은 국가·문화적 맥락에 따라 해석이 달라질 수 있어 다른 문화권에 적용하기는 어렵다는 한계가 있다.

본 연구에서는 규정된 한국 심의 기준을 기반으로 노출 수준과 성적 암시 수준을 동시에 고려한 이미지 선정성 탐지 모델을 개발한다. 나아가 해당 아이디어의 활용성을 확인하기 위해, 크롬 웹 브라우저 확장 프로그램을 개발해서 웹 브라우저를 통한 디지털 콘텐츠 탐색 시 원치 않는 선정적 광고 이미지가 나오면 이를 탐지하고 숨겨서 보다 쾌적한 브라우징 환경을 제공할 수 있는 실시간 선정적 광고 정화 시스템을 제안한다.

2. 관련 연구

2.1 신체 노출 수위에 기반한 선정성 판단 연구

이미지에 나타난 신체 노출을 기준으로 머신러닝을 활용해 선정성을 판단하는 연구가 활발히 진행되었으며, 피부 노출 비율을 임계 값으로 활용하거나[6], 성기·가슴·엉덩이 등 특정 신체 부위를 탐지·분류하는 방식[7] 등이 제안되었다. 그러나 이러한 연구들은 공통적으로 나체나 특정 신체 부위의 노출을 주요 기준으로 삼기 때문에, 신체 노출 없이 자세·구도 등을 통해 성적 의미를 암시하는 이미지까지 포착하기 어렵다. 또한 대부분 이진 분류 체계를 사용해 '선정적/비선정적'만 구분하므로, 중간 수준의 성적 암시나 경계가 모호한 이미지에 대한 세밀한 판단이 제한된다는 한계가 있다.

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)

2.2 신체 노출과 성적 암시에 기반한 선정성 판단 연구

노출 중심 선정성 판단의 한계를 보완하기 위해 최근에는 선정성의 개념을 확장하여 판단에 접근하는 연구들이 제안되고 있다. Wu and Xie[8]는 이미지의 선정성을 판단할 때 성기·가슴 노출뿐만 아니라 성행위 자세 등 성적 암시까지 고려해야 함을 인지하고, 이를 기반으로 선정성을 탐지할 수 있는 딥러닝 모델을 제안하였다. Rautela et al. [9]은 포르노그래피를 성기나 성행위가 시각적으로 묘사되어 성적 자극을 유발하는 영상으로 정의하고, 노출만으로는 포르노를 판별할 수 없다고 지적했으며, He et al. [10]은 이미지의 선정성을 보다 세분화하기 위해 'Clean', 'Porn-indicative', 'Pornography'의 3단계 분류 체계를 도입하였다. 그러나 이러한 연구들은 국가·문화·종교마다 '포르노그래피'의 기준이 상이하다는 문화적 맥락 측면의 한계를 지닌다.

이에 본 연구에서는 노출 여부에만 의존하지 않고, 한국의 규정화된 심의 기준을 기반으로 광고 이미지의 선정성을 판단하는 모델을 개발하고, 이를 바탕으로 웹 브라우징 중 나타나는 선정적인 광고 이미지를 실시간으로 제거하는 시스템을 제안한다.

3. 이중 등급 체계를 활용한 선정적 광고 필터링 시스템

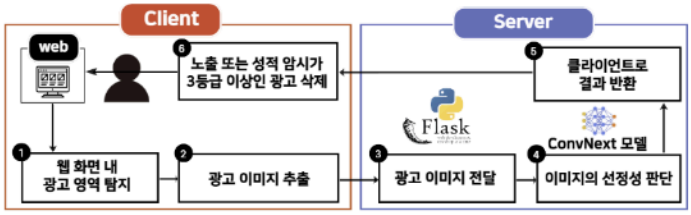


그림 1. 시스템 구조

본 연구에서는 신체 노출과 성적 암시를 함께 고려한 선정적 이미지 판단 모델을 구축하고, 이를 기반으로 웹 브라우징 중 선정적 광고를 실시간 탐지·차단하는 크롬 확장 프로그램을 개발하였다. 시스템의 전체 구조는 [그림 1]과 같다.

3.1 이미지의 선정성 분류 모델 개발

[표 1] SafeNet의 등급 기준을 참고하여 새로 만든 이중 등급 체계

등급	노출 등급	성적 암시 등급
0	인물 요소 없이 텍스트·그래픽 요소로만 구성된 이미지	
1	착의, 직접 노출 없음	성적 행위·암시 없음
2	가슴 및 둔부 일부 노출 성기·음모·항문 비노출	착의 상태에서 깊은 입맞춤 가슴·둔부 등 신체 접촉
3	전신 또는 둔부/가슴 노출 성기·음모·항문 비노출	착의 상태의 성행위/자위 성적 의도의 구도·분위기

본 연구에서는 이미지 선정성 분류 모델 학습을 위해 Roboflow 플랫폼[1]의 공개 데이터셋과 NSFW Data Scraper 레포지토리[2]에서 제공하는 이미지 데이터를 수집했다. Roboflow에서는 이미지 내 애니메이션 캐릭터 얼굴 검출 모델 학습용 공개 데이터를 다운로드하였으며, NSFW Data Scraper로 선정적 애니메이션 이미지 데이터를 추가로 확보했다. 최종적으로 두 데이터셋을 통합하여 총 1,212장으로 구성된 데이터셋을 구축하고 모델의 학습 및 평가에 활용했다.

모든 이미지는 방송통신심의위원회의 SafeNet 등급[3] 기준을 참고해 노출 등급과 성적 암시 등급을 각각 0-3등급으로 연구자가 직접 한 장씩 검토하여 수동으로 라벨링하였다. 이미지에 인물 요소가 전혀 없는 경우에는 해당 이미지에 0등급을 부여했다 [표1].

[노출 등급(1-3) x 성적 암시 등급(1-3)]은 총 9개의 조합이 나오며, 이미지에 인물이 전혀 포함되지 않은 0등급을 합치면 총 10개의 클래스 조합이 만들어진다. 전체 1,212장의 이미지는 이 10개 조합 중 하나의 조합이 부여되는 것이다. 클래스 조합 분포를 분석한 결과, 10개 클래스 조합의 평균 이미지 개수는 121.2장(표준편차 149.3)으로 클래스 조합간 불균형이 크게 나타났다. 이를 해소하기 위해 데이터가 많은 조합은 하향 샘플링하고, 데이터가 적은 조합은 Albumentations[4] 기반 이미지 증강으로 보완하였다. 그 결과 각 조합별로 200장씩 균등하게 데이터셋을 구성해서 최종 2,000장의 학습용 데이터를 구축하였다.

본 연구에서는 Liu et al. [11]이 제안한 멀티태스크 분류 모델을 파인튜닝했다. 전체 데이터의 10%를 테스트셋으로 분리하고, 나머지 90%에 대해 Stratified K-Fold 교차검증(K=5)을 수행하였다. 모든 Fold에서 유사한 F1-score가 나타나 데이터 분할에 관계없이 안정적으로 학습되었음을 확인하였다. 이 중 검증 손실이 가장 낮은 모델을 최종 모델로 확정하였으며, 독립된 테스트셋을 활용한 최종 평가 결과 [표 2]와 같이 Macro 평균 기준으로 높은 분류 성능을 달성하였다.

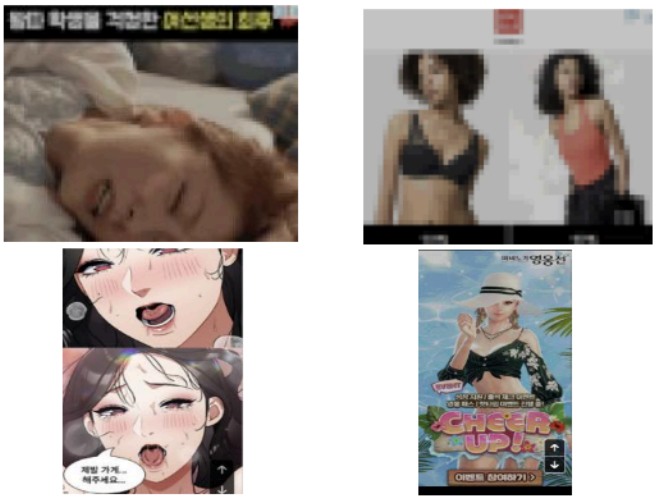
[표 2] 노출 등급 및 성행위 등급 분류 성능

구분	Precision	Recall	F1-score
노출 등급	0.92	0.92	0.92
성행위 등급	0.88	0.87	0.87

노출 등급과 성적 암시 등급이라는 이중 등급 체계를 적용하여 모델을 학습함으로써, 노출이 없더라도 선정적인 이미지(낮은 노출 등급 + 높은 성적 암시 등급)와 노출이 있더라도 선정적이지는 않은 이미지(높은 노출 등급 + 낮은 성적 암시)를 분류할 수 있게 되었다. 이를 활용해, 노출 등급이 '전신 또는 주요 부위 노출(3등급)'이거나, 성적 암시 등급이 '암시 또는 행위 표현(3등급)'일 경우 해당 이미지를 선정적인 광고로 판단하기로 했다.

---

[1] https://universe.roboflow.com/
[2] https://github.com/alex000kim/nsfw_data_scraper
[3] https://www.safenet.ne.kr/dstandard.do
[4] https://pypi.org/project/albumentations/



(a) (b)

그림 3. 선정성 이미지 분류 예시

[그림 3]은 해당 모델의 선정성 분류 결과 예시이다. (a)는 노출이 적더라도 성적 암시가 강해서 모델이 선정적이라고 판단했으며, (b)는 노출 등급은 높지만 성적 암시가 약해서 비선정적 이미지(예: 수영복 광고)라고 판단했다.

3.2 실시간 광고 정화 시스템 구현

본 연구에서 개발한 선정성 탐지 모델의 활용성을 확인하기 위해, 사용자가 웹페이지를 탐색하는 과정에서 노출되는 선정적인 광고 이미지를 실시간으로 탐지하고 제거하는 크롬 확장 프로그램[5]을 개발했다. 클라이언트는 웹페이지의 DOM을 실시간으로 스캔하고, 광고로 추정되는 요소를 탐지한다. 이후, 현재 화면을 캡쳐하고 광고 영역만 분리하여 서버로 전송한다. 서버에서는 선정성 분류 모델로 광고 이미지의 노출 등급과 성적 암시 등급을 예측하고, 선정적 광고로 판단되면 보이지 않도록 해당 광고를 DOM에서 제거한다.

[그림 4]는 제안한 시스템의 실제 동작 장면을 보여준다. 좌측 화면은 광고 제거 전 상태로, 붉은 테두리로 표시된 영역에서 선정적 광고가 노출된 것을 확인할 수 있다. 해당 광고는 노출 등급 2(부분 노출), 성적 암시 등급 3(성행위 암시)으로 분류되어 시스템에 의해 자동 제거되었다. 우측 화면은 동일 페이지에서 해당 광고가 제거된 이후의 상태로, 제안한 시스템이 정상적으로 작동함을 확인할 수 있다.

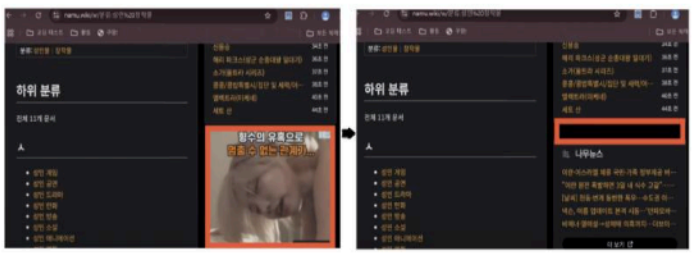


그림 4. 선정적 광고 탐지 및 제거 결과 화면 예 (나무위키 페이지)

[5] https://youtu.be/AUoglYzxsR8?si=ISrFTVGNfsPaKDLL

4. 결론 및 향후연구

본 연구에서는 웹 환경에서 노출되는 선정적 광고 이미지를 실시간으로 탐지하고 제거할 수 있는 크롬 확장 프로그램 기반 시스템을 제안하였다. 한국 방송통신심의위원회의 SafeNet 등급 기준 기반의 이중 등급 체계를 적용함으로써, 단순 노출 여부에 의존하지 않고 노출이 없더라도 성행위를 암시하는 이미지를 효과적으로 탐지했다. 본 연구는 기존 선정성 탐지 연구와 명확한 차별점을 지닌다. 또한, 실제 사용자 피드백을 통해 제안한 모델이 실사용 환경에서도 효과적으로 작동함을 확인하였다.

7. 참고문헌

[1] 한국청소년정책연구원. 청소년 미디어 이용 실태 및 대상별 정책대응방안 연구Ⅱ : 10대 청소년, 2021.

[2] 초록우산 어린이재단. 인터넷 뉴스를 보던 아동이 깜짝 놀란 이유는? 아동 온라인 유해환경의 실태. 2023.

[3] Burnay et al. Effects of violent and nonviolent sexualized media on aggression-related thoughts, feelings, attitudes, and behaviors: A meta-analytic review. Aggressive Behavior. 48. 1. 111-136. 2022.

[4] 한국소비자원. 소셜 네트워크 서비스(SNS) 부당광고 실태조사. 2021.

[5] 방송통신위원회. 청소년 유해매체물의 표시방법 고시. 고시 제2013-21호. 2013.

[6] Gajula et al. A Machine Learning Based Adult Content Detection Using Support Vector Machine, Proceedings of the 14th INDIACom: 2020 7th International Conference on “Computing for Sustainable Global Development”. Vol. 14, No. 1, pp. 1-8, 2020.

[7] Zhang et al. Sensitive Information Detection Based on Deep Learning Models, Applied Sciences. Vol. 14, No. 17, pp. 7541-7560, 2024.

[8] Wu and Xie. Fine-Grained Pornographic Image Recognition with Multi-Instance Learning, Computer Systems Science and Engineering, Vol. 47, No. 1, pp. 300-315, 2023.

[9] Rautela et al. Obscenity Detection Transformer for Detecting Inappropriate Contents from Videos, Multimedia Tools and Applications, Vol. 83, No. 10799-10814, pp. 10799-10814, 2024

[10] He et al. Sensitive Image Classification by Vision Transformers, Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1704-1711, 2024.

[11] Liu et al. A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. 11976-11986.