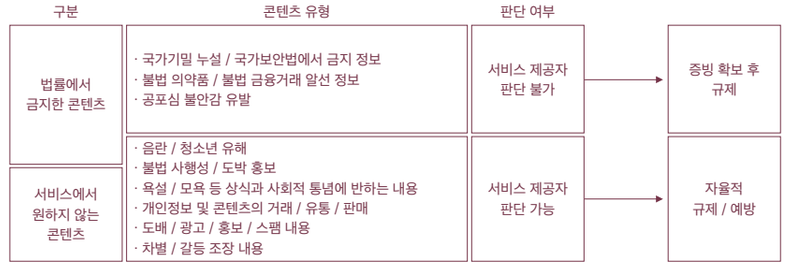
카카오,
유해 콘텐츠 차단을 위해
AI 방패를 펼치다

글 | 최은필 reyna.c@kakaocorp.com IT 업계의 에반젤리스트(evangelist)를 자처하며 빠르게 변화하는 시대에 그 누구보다 부지런히 움직이며, 좋은 것들을 나누며 살고 있습니다. 재즈 힙합(Jazz Hip-hop)을 즐겨 들으며, 언젠가 추리소설을 쓰겠다는 원대하지만 소박한 꿈을 가지고 있습니다. 전자공학을 전공했지만 실험보다는 글쓰기를 더 좋아하며, 오랜 시간 '전략, 기술, 기획'이라는 단어들과 함께하였습니다. 전 세계 수많은 사람들의 마음을 움직여보고 싶다는 커다란 꿈을 가지고, 하루하루 즐겁고 재미있게 살고 있습니다.

이용자 보호를 위한 유해 콘텐츠 관리

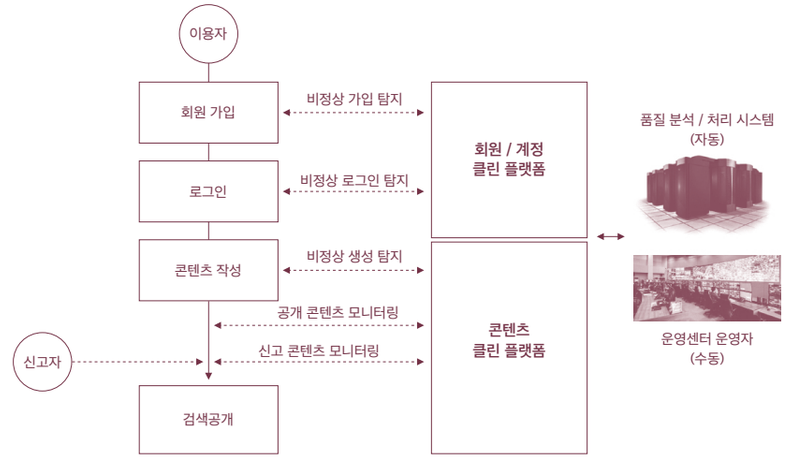
그리스 신화에 나오는 신들의 왕인 제우스는 딸 아테네에게 '아이기스(Aegis)'라는 방패를 선물한다. 이 방패는 대장간의 신 헤파이스토스가 만든 것으로 강력한 위력을 가지고 있어 벼락을 맞아도 부서지지 않으며 방패를 흔들면 무시무시한 폭풍이 일어나는 강력한 무기이다. 우리에게는 동해에 배치된 꿈의 전함이라고 불리우는 '이지스(아이기스를 영어식으로 읽음)'를 통해서 널리 알려진 용어이다.*1

[그림 1] 카카오의 서비스 제공자의 자율규제 기준과 방식



방패라는 것은 외부의 공격으로부터 지키기 위한 도구의 하나로, 카카오에서는 세상의 모든 유해 콘텐츠로부터 사용자들을 지키기 위한 방패로 유해 콘텐츠 필터링 기술을 적용하고 있다. 포털 서비스나 게시판을 제공하는 회사들은 법률에 따라 자체적으로 콘텐츠를 깨끗하게 정화해야 하는 의무를 가진다. 이것을 '서비스 제공자의 자율규제 의무'라고 하며, 단순히 법률에 의한 규제뿐만이 아닌 청소년과 어린아이들이 안전하게 인터넷 콘텐츠를 사용할 수 있는 환경을 만들어야 할 의무라고 할 수 있을 것이다. 자율 규제는 서비스 제공자의 판단 가능 여부에 따라 [그림 1]과 같은 절차로 진행된다. 서비스 제공자들이 콘텐츠 법률 위반 여부를 판단할 수 없는 경우, 법률 위반에 대한 공신력 있는 증빙 자료 확보 후 규제가 가능하다. 하지만 서비스 제공자가 자체적으로 판단 가능한 콘텐츠의 경우 자율적으로 규제하고 예방하는 의무를 가진다.

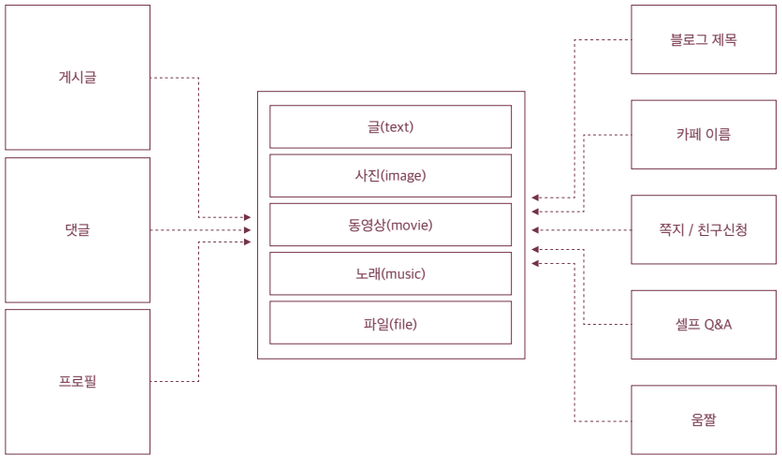
[그림 2] 카카오의 콘텐츠 품질관리 라이프사이클(life cycle)



*1 참고 | http://100.daum.net/encyclopedia/view/73XXXXXXKS912

현재 유해 콘텐츠는 대부분 신고에 의해서 처리되지만, 신고되지 않는 항목들은 실시간 모니터링을 통해 필터링하여 노출되지 않게 처리하는 방법이 있다([그림 2]). 하지만 콘텐츠의 이상 여부를 시스템을 통해서 모니터링하더라도 결국 사람이 직접 콘텐츠를 보고 판단해야 하는 시점이 오게 된다. 지금까지는 육안을 통해서 사람이 콘텐츠 이상 여부를 판단하였지만, 최근에는 머신러닝 기술을 활용하여 콘텐츠의 유해성을 판단하고 걸러내는 작업이 진행되고 있다. 현재 국내외 다양한 콘텐츠 서비스 제공 회사들이 스마트한 유해 콘텐츠 차단을 위해 AI 기술을 도입하고 있으며, 카카오 역시 끊임없이 진화하고 있는 유해 콘텐츠로부터 사용자들을 보호하고 더욱 안전한 인터넷 환경을 만들기 위해 내부적으로 지속적인 연구개발과 투자를 강화하고 있다. 현재 카카오에서는 이미지 외에 글, 동영상, 음악, 파일 등 다양한 콘텐츠의 유해성을 관리하고 있으며, [그림 3]과 같은 항목들을 집중 모니터링하고 있다.

[그림 3] 카카오 자율규제 관리 대상 공개 콘텐츠 종류



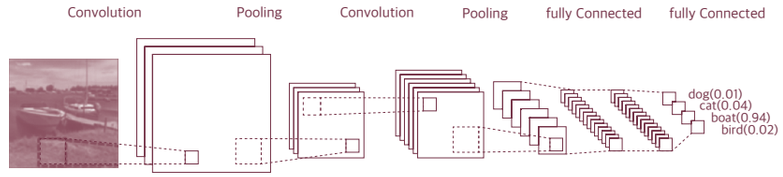
유해 콘텐츠 차단을 위한 AI 기술은 도입 초창기보다는 완전 자동화 상태로 발전하고 있으며, 유해 콘텐츠 관리 기준은 국가 및 회사, 서비스마다 개별 기준을 적용하고 있기 때문에 공통적인 기준보다는 회사&국가별 다른 기준을 가지고 있다.

카카오는 상대적으로 타이트한 규정을 적용하고 있으며, 향후에는 카카오 TV 및 티스토리 등 주요 서비스들을 중심으로 확대 전개할 예정이다. 특히 카카오 TV와 같은 서비스는 라이브 방송이 되는 경우가 많은데, 실제 운영자들이 감시를 할 경우 스폿성으로 노출되는 유해 콘텐츠를 직접 확인하기 어려운 상황들이 발생한다. 그래서 AI 기술을 적용하여 사용자들에게 유해 콘텐츠 노출을 최대한 방지하기 위해서 첨단 기술을 도입하게 되었다.

유해 콘텐츠 필터링 기술의 원리와 현황

유해 콘텐츠를 필터링하는 기술의 원리를 간단하게 살펴보자. 카카오의 유해 콘텐츠 필터링 기술은 텐서플로를 기반으로 하고 있다. 예를 들어 하나의 이미지를 숫자로 바꾸고, 이미지를 하나의 행렬로 표현한다. 텐서플로 모델의 규칙에 따라서 행렬 데이터를 잘게 나누고 벌집처럼 나누어서, 적용 알고리즘에 따라 특징을 분석하게 된다. 특히 이미지의 경우 카카오는 합성곱 신경망(convolutional neural network, CNN) 모델이라는 학습 알고리즘을 사용하고, 알고리즘 기술은 각 회사들마다 다른 특징 기술을 가지고 있다.

[그림 4] CNN 모델 원리*2



모든 머신러닝 기술이 그러하듯이 유해 콘텐츠 필터링 기술을 고도화하기 위해서는 다양한 형태의 샘플링 데이터가 필요하고 학습을 시켜야 한다. 예를 들어 명화나 조각상 같은 이미지들을 디테일하게 분석하고 정확하게 판별하기 위해서는 오랜 시간 학습이 필요하다. 최근 들어서 콘텐츠 유해 필터링 기술이 급속도로 발전하고 있다. 실제 이 기술들은 1970년대에 최초 공개되었으나 최근 클라우드, GPU 등 하드웨어 성능이 좋아지면서 필터링 기술 역시 고도화되었으며, 특히 텐서플로 기술은 유해 콘텐츠 필터링에 큰 기여를 했다고 말할 수 있다.

유해 콘텐츠 필터링 기술의 미래

실제 유해 콘텐츠에 노출되면 나타나는 후유증은 상당히 위험하다. 특히 재발을 방지하는 것이 우선되어야 하는데, 이를 위해서는 단순히 기술 고도화를 통한 1차적 방지뿐만이 아니라, 재발이 발생하지 않도록 계도가 필요하다. 아무리 기술이 고도화되더라도 기술과 사람은 공존해야 하고, 사람의 실수를 기계가 보완하는 형태가 되어야 한다.

특히 유해 콘텐츠를 필터링하는 기술은 무엇보다 유해성 판별의 자동화에 초점을 두어야 할 것이다. 현재는 1차 유해 콘텐츠를 필터링하는 작업은 기술에 의존하며, 사람의 육안으로 한 번 더 점검하면서 콘텐츠를 구분하고 있다. 하지만 사람의 육안으로 유해 콘텐츠를 걸러내기에는 한계가 있기 때문에, 향후에는 신속 정확하게 유해 콘텐츠를 걸러내야 할 필요가 있다. 그리고 이것은 사용자들에게 더 좋은 콘텐츠를 제공하고자 하는 카카오의 책임과 의무 사항이라고 볼 수 있을 것이다. 향후에는 지금보다 더 다양한 글로벌 콘텐츠들이 유입될 수 있다. 음란/유해 콘텐츠 수준은 갈수록 심화되고, 더욱 자극적인 영상들이 생성되고 있다. 그리고 인터넷 환경이 좋아지면서 유해 콘텐츠를 노출할 수 있는 사용자들 장비 역시 성능이 더 좋아지면서, 더욱 고도화된 방식으로 유해 콘텐츠들을 의도적으로 노출하고 있다고 한다. 실제 카카오에서 유해 콘텐츠 필터링 기술을 개발하고 콘텐츠 품질을 관리하며 이용자 보호를 위해 업무를 하고 있는 담당 크루 분들에 의하면, 본인들 역시 업무와 관련된 유해 콘텐츠 상황을 직접 확인했고 그 당시 충격이 상당했다고 한다.

*2 참고 | https://cdn-images-1.medium.com/max/1600/1*N4h15gwbWNmtrRhszM9EJg.png

깨끗한 인터넷 환경을 위한 카카오의 노력

국내에서 유해 콘텐츠와 관련된 논란이 시작된 것은 2007년 야후코리아의 UCC 서비스 때문이었다. 1분 정도의 유해 동영상이 6시간 동안 서비스에 노출되었고, 결국 해당 사건 이후 서비스 중단 및 관련 법률, 정부 규제 정책들이 쏟아져 나오게 되었다. 이 사건은 서비스가 중단될 정도로 파급력이 큰 사건이었다.

현재 카카오는 구글의 텐서플로 기술을 기반으로 유해 콘텐츠를 필터링하고 있으며, 향후에는 자체 모델을 설계하여 카카오 서비스에 최적화된 방어막을 만들 수 있는 방안을 구성하려고 한다. 현재 이미지, 텍스트, 사용자 정보 등을 개별적으로 모니터링하고 있지만, 향후에는 이 모든 항목을 통합적으로 모니터링하여 유해 콘텐츠를 방지하는 쪽으로 구상하고 있다. 앞에서 언급한 방패 '아이기스'처럼 모든 유해 콘텐츠로부터 사용자를 보호하고 더욱 나은 세상을 구상하기 위해 카카오는 끊임없이 노력하고 나아가기 위해 노력하는 것이다.

각 회사들의 유해 콘텐츠 관련 방침은 사실 서비스를 활성화하고 성장만을 바라보는 회사에서는 쉽게 간과할 수 있는 부분이다. 앞으로 기술이 고도화되고 사회가 더욱 복잡해지면, 콘텐츠 유해성에 대한 논란은 더욱 가중될 것으로 예상된다. 향후 기술은 지속적으로 발달할 것이고, 그에 따라 사회적 이슈 역시 더욱 복잡하고 다양한 형태로 우리를 찾아올 것으로 생각된다. 향후 카카오i의 클린 엔진, 세이프 엔진을 통해서 외부 기관에서도 사용할 수 있도록 파이프라인 역할도 생각하고 있다. 카카오 역시 콘텐츠를 제공하는 서비스 제공자 입장에서 사회적 책임을 다하고, AI 기술을 통해 더 좋은 세상을 만들기 위해 노력하고 있다.

〈카카오, 유해 콘텐츠 차단을 위해 AI 방패를 펼치다〉 브런치로 연결되는 QR 코드입니다.