JDFR Journal of Data Forensics Research
데이터포렌식연구, 1권 1호, 2024년 12월, 29-46면
Journal of Data Forensics Research, Volume 1, Issue 1, December 2024, Pages 29-46.
https://doi.org/10.12972/JDFR.2024.1.1.3
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

텍스트 및 이미지 데이터를 활용한 멀티모달 온라인 성매매 홍보 불법 콘텐츠 분류 및 신고 자동화 시스템 개발

박민회[1], 박수민[2], 윤성범[3], 최호식[4], 김준철[5], 송경우[6]

[1]연세대학교 데이터사이언스연구소 연구원
[2]연세대학교 통계데이터사이언스학과 석사과정
[3]서울연구원 AI 빅데이터랩 연구원
[4]서울시립대학교 도시빅데이터융합학과 교수
[5]서울연구원 AI 빅데이터랩 연구위원
[6]연세대학교 응용통계학과, 통계데이터사이언스학과 교수

교신저자: 최호식, choi.hosik@uos.ac.kr; 김준철, kjc@si.re.kr; 송경우, kyungwoo.song@yonsei.ac.kr

요약

온라인 성매매 홍보 불법컨텐츠는 인터넷과 SNS 사용이 보편화되면서 심각한 사회 문제로 대두되고 있다. 특히 아동과 청소년을 대상으로 한 성범죄는 피해가 광범위하게 확산될 수 있어 즉각적이고 효율적인 대응이 요구된다. 본 연구는 X API를 통해 수집된 텍스트와 이미지 데이터를 활용하여 성범죄 징후를 탐지하고, 신고 과정을 자동화하는 멀티모달 성범죄 신고 자동화 시스템을 개발하는 것을 목표로 한다. 개발된 시스템은 멀티모달 대형 언어 모델(Multimodal Large Language Model, MLLM)을 기반으로 텍스트와 이미지 데이터를 통합 분석하여 성범죄의 징후를 파악하고, 자동으로 신고서 초안을 생성 기능을 포함한다. 신고서 초안은 사람의 검토를 거쳐 최종 신고서로 제출되어 신속성과 정확성을 동시에 확보하였다. 실험을 통해 개발된 시스템은 텍스트와 이미지 데이터의 결합을 통해 멀티모달 대형 언어 모델을 활용하여 성범죄 분류와 신고 절차에서 이미지를 개별적으로 학습한 모델 대비 30% 이상의 효율성을 증대시켰다. 본 연구는 성범죄 예방 및 피해자 보호를 위한 실질적 지원 도구로서 온라인 성매매 홍보 불법컨텐츠 대응의 새로운 방향을 제시할 수 있을 것으로 기대한다.

주제어

성범죄 자동 신고 시스템, 멀티모달 인공지능, 이미지 및 텍스트 분석, SNS콘텐츠 검출, 온라인 성매매 홍보 불법컨텐츠 예방

Open Access

Received: November 27, 2024
Revised: December 22, 2024
Accepted: December 24, 2024
Published: December 31, 2024

Original Article

Development of a Multimodal System for Classifying and Automatically Reporting Illegal Online Prostitution Promotional Content Using Text and Image Data

Minhoi Park[1], Sumin Park[2], Sungbum Yun[3], Hosik Choi[4], Junchul Kim[5], Kyungwoo Song[6]

[1]Researcher, Yonsei University, Republic of Korea
[2]Master's Student, Yonsei University, Republic of Korea
[3]Researcher, AI & Big Data Research Division, The Seoul Institute, Republic of Korea
[4]Professor, Department of Artificial Intelligence, University of Seoul, Republic of Korea
[5]Research Fellow, AI & Big Data Research Division, The Seoul Institute, Republic of Korea
[6]Professor, Department of Applied Statistics, Department of Statistics and Data Sceiece, Yonsei University, Republic of Korea

Corresponding Author: Hosik Choi, choi.hosik@uos.ac.kr; Junchul Kim, kjc@si.re.kr; Kyungwoo Song, kyungwoo.song@yonsei.ac.kr

ABSTRACT

Online prostitution advertisements and illegal content have emerged as serious social issues with the widespread use of the internet and social media platforms. Particularly, crimes targeting children and adolescents demand immediate and efficient responses due to their potentially far-reaching impacts. This study aims to develop an automated multimodal system for reporting sexual crimes, leveraging text and image data collected via the X API. The proposed system integrates a Multimodal Large Language Model (MLLM) to analyze text and image data comprehensively, enabling the detection of sexual crime indicators and the automatic generation of preliminary reports. These draft reports are reviewed by humans before being finalized and submitted, ensuring both speed and accuracy in the reporting process. Experimental results demonstrate that the developed system, by combining text and image data using MLLM, improves efficiency by over 30% compared to models trained on image data alone. This study contributes to the prevention of sexual crimes and the protection of victims, offering a practical tool and a novel approach for addressing illegal online content promoting prostitution.

KEYWORDS

Automated Sex Crime Reporting System, Multimodal Artificial Intelligence, Image and Text Analysis, Social Media Content Detection, Prevention of Illegal Online Prostitution Content

1. 서론

인터넷과 소셜 네트워크 서비스(SNS)의 급속한 확산은 정보와 소통의 접근성을 크게 향상시켰지만, 동시에 온라인 성매매 홍보 불법컨텐츠와 같은 새로운 형태의 범죄가 증가하는 문제를 초래하였다. 특히 아동과 청소년을 대상으로 하는 온라인 성매매 홍보 불법컨텐츠는 그 피해의 심각성과 파급력이 커, 사회적 안전망 구축이 절실히 요구되고 있다. 온라인 성매매 홍보 불법컨텐츠는 주로 SNS, 메신저 앱, 온라인 게임 등 다양한 디지털 플랫폼에서 이루어지며, 가해자들은 피해자와의 신뢰를 형성한 후 심리적 압박이나 협박을 통해 성적 이미지를 요구하거나 이를 유포하는 방식으로 범죄가 발생한다[1]. 기존의 성범죄와 달리, 온라인 성매매 홍보 불법컨텐츠는 익명성을 통해 가해자가 쉽게 신원을 숨길 수 있고, 피해가 빠르게 확산되며 지속적으로 이루어질 가능성이 높아 이를 사전에 탐지하고 예방하는 것이 어려운 실정이다[2].

최근 발생한 N번방 사건은 온라인 성매매 홍보 불법컨텐츠의 심각성을 여실히 보여주는 사례로, 아동과 청소년을 대상으로 하는 성적 이미지가 강압적으로 생산되고 SNS와 메신저 플랫폼을 통해 유포된 사건이다[1]. 이 사건은 온라인 플랫폼에서 발생하는 온라인 성매매 홍보 불법컨텐츠의 파괴력과 범죄 확산의 심각성을 강조하며, 온라인 성매매 홍보 불법컨텐츠에 대한 사전 예방과 조기 탐지의 필요성을 사회적으로 인식시키는 계기가 되었다. 특히 피해자가 자발적으로 신고하지 않는 한 이러한 범죄는 지속되기 쉽고, 기존의 사후 대응 방식만으로는 피해를 줄이는 데 한계가 있다는 점이 드러났다[3].

1.1. 온라인 성매매 홍보 불법컨텐츠에 대한 기존 대응 방식의 한계

온라인 성매매 홍보 불법컨텐츠에 대한 기존의 대응 방식은 주로 사건 발생 후 사후 처벌과 법적 제재에 중점을 두고 있다. 그러나 이러한 방식은 몇 가지 중요한 한계를 가지고 있다. 첫째, 피해자가 스스로 신고하지 않는 한 성범죄가 발생하는 동안 이를 감지하거나 예방하기가 어렵다는 점이다. 피해자는 주로 협박이나 심리적 압박으로 인해 신고를 주저하게 되며, 이로 인해 범죄가 오랜 기간 동안 지속되는 경우가 많다[4]. 둘째, 온라인 플랫폼에서 활동하는 가해자들은 익명성을 악용하여 추적을 회피하고, 새로운 계정을 통해 반복적으로 범죄를 저지르기 때문에 기존의 수사 방식만으로는 즉각적이고 실질적인 대응이 어렵다[5]. 셋째, 대부분의 대응 방식은 사건이 발생한 이후의 처벌에 중점을 두고 있어, 실제로 피해를 사전에 예방하거나 피해 규모를 줄이는 데 한계가 있다. 이러한 사후 대응 중심의 구조는 이미 발생한 피해를 관리할 수밖에 없기 때문에, 온라인 성매매 홍보 불법컨텐츠를 조기에 인지하고 피해자를 보호하는 데 있어 한계를 보인다[6].

1.2. 신고 자동화 시스템의 필요성

온라인 성매매 홍보 불법컨텐츠를 효과적으로 대응하기 위해서는 기존의 사후 대응을 보완할 수 있는, 사전 탐지 및 조기 대응 시스템이 필요하다. 특히 AI 기반의 신고 자동화 시스템은 성범죄 징후를 탐지하고 신고서를 자동화하고, 이메일과 문서를 자동으로 생성하여 범죄 발생 시 신속히 대응할 수 있는 솔루션이다. 이러한 시스템은 온라인 성매매 홍보 불법컨텐츠의 특징을 반영하여 데이터 분석 및 자동화된 문서 생성 기능을 통해 신고 과정의 효율성을 높이며, 피해자가 직접 신고하기 어려운 상황에서도 위험 상황을 인지하고 조치를 취할 수 있도록 돕는다[7].

본 연구는 텍스트와 이미지 데이터를 활용하여 온라인 성매매 홍보 불법컨텐츠를 분류하고,

신고를 자동화하며, 신고에 필요한 e-mail 및 문서 생성 과정을 자동화하는 AI 기반 멀티모달 성범죄 신고 자동화 시스템을 개발하고자 한다. 성범죄의 징후를 텍스트 및 이미지 데이터에서 동시에 탐지하는 멀티모달 대형 언어 모델(MLLM)을 사용하여, 단일 형태의 데이터보다 더 정확하게 성범죄를 분류체계에 따라 분류할 수 있도록 한다. 예를 들어, 특정 게시글과 성적 이미지가 함께 나타나는 경우, MLLM은 텍스트와 이미지의 의미적 관계를 파악하여 성범죄 분류체계에 따라 분류를 진행할 수 있다[8].

1.3. 연구 목표와 접근 방식

본 연구는 X API를 통해 수집된 텍스트와 이미지 데이터를 기반으로, 성범죄를 분류하고 신고를 자동화하고, 신고를 위한 이메일과 신고서를 생성하는 멀티모달 신고 자동화 시스템을 개발하는 것을 목표로 한다. 텍스트와 이미지 데이터를 통합 분석할 수 있는 MLLM을 활용하여 성범죄 징후를 분류하고, 자동으로 신고서를 작성 및 이메일을 생성하는 시스템을 설계하였다. 이 시스템은 SNS를 통해 자동으로 신고하고, 최종 단계에서 사람이 신고서를 검토하고 승인하도록 설계되었다[7].

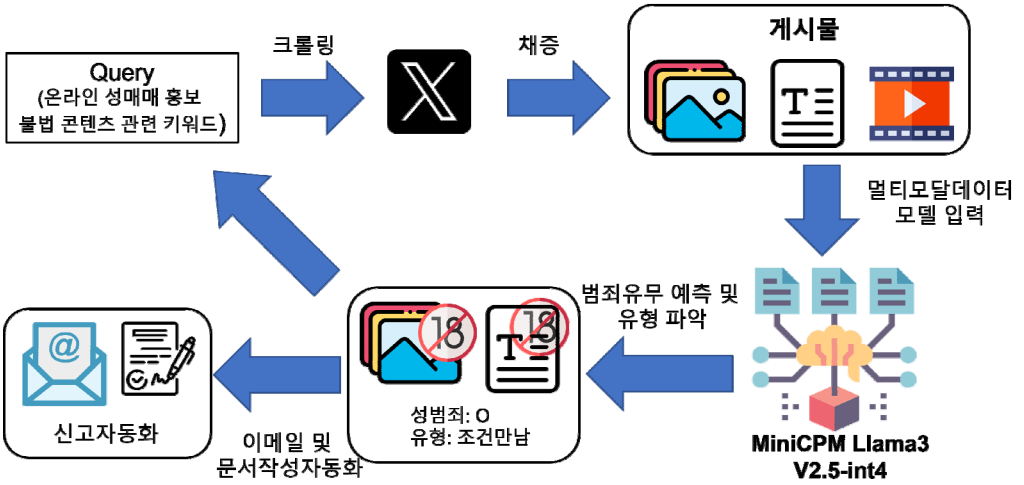
본 연구에서 제안하는 성범죄 신고 자동화 시스템의 주요 목표는 성범죄의 효율적 탐지, 신고 자동화 및 신고 이메일 및 문서 자동생성을 통한 효율화 이다.

첫째, 성범죄 효율적 탐지를 위해 텍스트와 이미지 데이터를 기반으로 성범죄와 관련된 징후를 통합 분석하여 정확하게 탐지하는 기능을 구현한다. MLLM을 사용하여 텍스트와 이미지 간의 복합적 관계를 분석함으로써 성범죄 가능성을 효과적으로 판단하고, 이를 통해 조기 대응을 가능하게 한다[9].

둘째, 신고 자동화 및 신고 이메일 및 문서 자동생성을 통한 효율화를 위해 성범죄 징후가 감지되면 자동으로 신고서 초안을 작성하고, 또한 신고를 위한 이메일 초안을 작성하여 최종 승인 단계로 전달하여 관련 법 집행 기관에 제출할 수 있는 구조를 설계하였다. 이러한 신고 자동화와 이메일 및 문서 생성 자동화를 통해 전체 신고 과정의 효율성을 증대시키며, 신고서를 사람이 검토할 수 있도록 하여 신뢰성을 확보한다[7].

1.4. 기대 효과 및 연구의 기여

본 연구에서 제안하는 AI 기반의 멀티모달 성범죄 신고 자동화 시스템은 온라인 성매매 홍보 불법컨텐츠에 대한 선제적 대응 방안을 제시하며, 피해자가 직접 신고하지 않아도 자동화된 방식으로 성범죄의 징후를 탐지하고 필요한 조치를 취할 수 있도록 한다. 본 시스템은 성범죄 예방 및 피해자 보호를 위한 실질적인 지원 도구로서, 온라인 성매매 홍보 불법컨텐츠 대응에 있어 새로운 방향을 제시할 수 있을 것으로 기대된다. 이를 통해, 사회적으로 중요한 아동 및 청소년의 디지털 환경 안전을 강화하고, 더 나아가 온라인 성매매 홍보 불법컨텐츠의 피해를 최소화하는 데 기여할 수 있을 것이다.



<Figure 1> Workflow for Sexual Crime Classification and Automated Reporting System

2. 관련 연구 및 배경지식

2.1. 관련 연구

온라인 성매매 홍보 불법컨텐츠 탐지에 대한 연구는 주로 SNS와 메신저 앱에서 이루어지는 텍스트 데이터 분석에 중점을 두고 있으며, 특정 패턴이나 언어적 특징을 통해 범죄 가능성을 감지하려는 접근이 주를 이룬다. Gomez et al. [10]은 SNS에서 특정 표현과 언어 패턴을 분석하여 성범죄 가능성을 탐지하는 방식을 제안하며, 주로 텍스트 기반 데이터만을 활용하는 단일 모달 접근법을 사용하고 있다. 그러나 이러한 단일 모달 접근법은 이미지가 포함된 경우의 성범죄 징후를 탐지하는 데 한계가 있다. 예를 들어, 특정 단어가 없는 이미지 단독의 경우에도 성범죄 징후가 포함될 수 있으며, 이를 보완하기 위해 본 연구는 텍스트와 이미지를 결합하여 분석하는 멀티모달 모델을 도입하여 성범죄 탐지의 정확도를 높이고자 한다.

멀티모달 접근법은 특히 성범죄와 같은 복합적 상황에서 정확도를 향상시키는 데 효과적이다. Baltrušaitis et al. [5]은 텍스트와 이미지 데이터를 동시에 분석하여 불법 콘텐츠를 탐지하는 방안을 제안하며, 이는 이미지와 텍스트 간의 의미적 연관성을 통합 모델로 다루어 단일 데이터 소스보다 정밀한 탐지가 가능함을 보여준다. 본 연구는 최신 멀티모달 대형 언어 모델(MLLM)을 활용하여 텍스트와 이미지 간 의미적 관계를 통합 분석함으로써 성범죄와 관련된 복합적 징후를 탐지하고자 한다. 이러한 접근은 성범죄와 같은 복합적 범죄의 징후를 포착하는 데 있어 더 높은 정확성을 보장할 수 있다.

또한, 본 연구는 자동화된 신고서 생성과 이메일 발송 기능을 추가하여 탐지된 성범죄 징후에 대해 실질적인 대응 방안을 제공한다. 기존의 대부분 연구는 성범죄 탐지를 목표로 하지만, 탐지된 결과를 자동화된 신고 및 문서 발송 시스템으로 연결하는 사례는 부족하다. Hosseini et al. [11]의 연구는 악성 댓글 탐지 시스템의 한계와 이를 우회하는 방법을 논의하고, 신속하고 안정적인 자동화 신고시스템의 필요성을 강조한다. 본 연구는 신고 자동화와 자동화된 신고서 초안을 생성하고, 이를 이메일을 통해 관련 기관에 전송하는 절차를 구현하여, 탐지와 즉각적 대응을 원활히 연결하는 데 중점을 두고 있다.

본 연구는 기존 텍스트 또는 이미지 기반 탐지 모델의 한계를 보완하고, 멀티모달 분석을 통해 성범죄 징후를 정밀하게 탐지할 수 있는 시스템을 개발한다. 또한 자동화된 신고 절차와 문

서 생성 기능을 통해 온라인 성매매 홍보 불법컨텐츠에 대한 실질적이고 신뢰할 수 있는 대응 방안을 제시하고자 하며, 이는 성범죄 대응에서의 실용적 기여를 목표로 한다.

2.2. 배경지식

2.2.1 성범죄 분류 체계

성범죄 분류 체계는 성매매 관련 범죄의 유형을 기준으로 세분화 되어 있으며, 각 유형은 범죄의 장소나 서비스의 형태에 따라 분류된다. 아래는 해당 분류체계를 구조적으로 정리하여 설명한 것이다.

<Table 1> Sexual Crime Classification System

분류		내용
출장형		특정 장소로 출장 서비스를 제공하는 형태의 성범죄로 출장샵, 애인대행, 섹파매칭 등이 있으며, 주로 온라인 플랫폼에서 고객과 연결되어 원하는 장소로 파견 서비스를 제공.
업소형	주점형	단란주점, 유흥주점 등에서 성매매 서비스를 제공하는 형태로 영업 형태에 따라 룸 형태 또는 풀형대로 제공되어 룸싸롱, 풀싸롱이 있음. 또한 서비스의 내용과 의상에 따라 분류되며, 특정의상(셔츠, 레깅스 등)을 입은 종원원이 응대하는 방식에 따라 셔츠룸, 레깅스룸이 있으며, 맥양집이라 불리는 특정 주점 형태로 성매매 서비스가 제공되는 곳을 주점형을 분류 할 수 있음.
	위장형	표면적으로 일반적인 오피스텔, 마사지샵, 안마시설 등으로 운영되나 실제로는 성매매가 이루어지는 장소로서 오피스텔, 마사지, 안마, 휴게텔, 키스방 등이 포함됨. 이러한 시설은 성매매를 제공하지만, 외부적으로 성매매와 무관한 서비스인 것처럼 가장하는 형태에 해당함.
조건만남		주로 1:1로 진행되는 성매매 형태로, 특정 조건(예: 금전적 대가)을 합의하여 만남이 이루어지는 방식. 주로 온라인 커뮤니티나 모바일 앱을 통해 연결되며, 일회성 성매매나 지속적인 성적 만남이 포함됨.
성매매 알선 포털사이트		성매매 업소에 대한 정보를 제공하는 온라인 포털로, 지역별 및 업종별 성매매업소 정보와 이용후기를 제공함. 이러한 포털사이트에서는 성매매 후기 및 평점 등을 기재하여 이용자들이 특정 업소의 정보를 확인하고 선택 할 수 있도록 함.
기타		위에 명시된 유형 외에도 성매매 관련 구인 및 이용 후기 제공 등의 방식으로 성매매를 조장하는 형태로서 성매매 구인광고, 성매매업소 이용후기, 보도형태의 성매매 소개 등이 해당.

2.2.2 멀티모달 모델

멀티모달 모델은 서로 다른 형태의 데이터를 결합하여 분석하는 방식으로, 단일 모달 모델보다 더 정확한 예측과 풍부한 정보를 제공할 수 있다. 예를 들어, 텍스트 분석으로만 감지하기 어려운 성범죄 징후를 이미지 분석과 결합하면 더 명확하게 탐지할 수 있다. 특히, 멀티모달 모델은 감정 분석, 위험 탐지, 범죄 예측 등 복잡한 의미 해석이 필요한 상황에서 중요한 역할을 한다[5,13].

멀티모달 모델은 다음과 같은 구성 요소를 포함하여 서로 다른 데이터 유형을 통합하고 분석한다.

<Table 2> Multimodal Components

구성요소	
모달리티 간 임베딩	텍스트, 이미지 등의 데이터를 동일한 표현 공간으로 변환하여 서로 다른 모달리티 간의 의미적 유사성을 파악함. 예를 들어, BERT와 같은 텍스트 임베딩 모델과 ResNet과 같은 이미지 임베딩 모델을 결합하여 일관된 표현 공간에서 데이터를 비교할 수 있음 [15][16].
멀티헤드 어텐션기법	Transformer 기반의 멀티헤드 어텐션은 텍스트와 이미지 간의 상호작용을 효과적으로 이해하게 해주는 핵심 기술로서, 이를 통해 모델은 텍스트와 이미지가 서로 어떠한 연관성을 가지고 있는지, 각 모달리티 간의 중요한 부분이 무엇인지를 학습할 수 있음 [16].
데이터 정렬과 통합	여러 모달리티 간 데이터를 효과적으로 연결하고 통합하여, 모델이 통합된 정보를 바탕으로 더 나은 판단을 내리도록 하며, 이를 통해 텍스트와 이미지가 함께 나타나는 경우 성범죄와 같은 복합적 사건의 가능성을 높게 평가할 수 있음 [18].

<Table 3> Multimodal Model

멀티모달 모델	
CLIP (Contrastive Language-Image Pretraining)	OpenAI의 CLIP 모델은 이미지와 텍스트 쌍을 대규모로 학습하여 이미지 내 객체와 텍스트 간의 관계를 이해하는 데 탁월한 성능을 보임. CLIP은 성범죄 탐지에서도 텍스트와 이미지의 의미적 연관성을 통해 위험성을 판단하는 데 활용될 수 있음 [4].
VisualBERT	BERT 기반 모델인 VisualBERT는 텍스트와 이미지 간의 상관관계를 학습하기 위해 개발된 모델로, 이미지 캡션 생성 및 이미지-텍스트 매칭 등의 작업에서 우수한 성능을 보임. 이러한 모델은 성범죄 신고 시스템에서 멀티모달 데이터의 의미를 해석하고 적절히 반응할 수 있도록 하는 데 유용함[18].
UNITER	Unified Vision and Text Representation(UNITER)모델은 텍스트와 이미지 간 관계를 보다 정밀하게 학습하여, 복합적인 상황에서 텍스트와 이미지 간의 관계를 명확하게 파악할 수 있으며, 이는 성범죄와 같이 여러 모달리티에서 징후가 나타나는 경우 탐지 성능을 크게 향상시킬 수 있음 [19].
Mini-CPM-Llama3-V2.5	멀티모달 대형 언어 모델로, 텍스트와 이미지의 복합적 데이터를 통합 분석할 수 있도록 설계된 모델로, 성범죄 탐지와 같은 복잡한 시나리오에서 텍스트와 이미지의 상호작용을 이해하는 데 특히 유용함. 이 모델은 효율성과 확장성을 겸비한 경량화된 구조를 채택하여 대규모 데이터 처리와 다양한 응용 가능성을 제공함. 텍스트와 이미지 데이터를 동시에 분석하고, 텍스트와 이미지 간의 관계를 모델링하여 성범죄 징후와 같은 복합적 사건을 보다 정확하게 탐지할 수 있음. 이를 통해 성범죄 탐지 자동화 시스템에서 신속하고 신뢰성 높은 분석 지원 가능[20].

성범죄 탐지에서는 텍스트와 이미지가 동시에 나타나는 상황이 많기 때문에, 멀티모달 모델이 특히 효과적이다. 텍스트 기반의 위험 대화와 이미지 기반의 시각적 성적 암시가 함께 나타나는 경우 이를 결합하여 분석함으로써 탐지 정확도를 크게 높일 수 있다. 예를 들어, 성적 암시를 포함하는 이미지를 분석하고, 해당 이미지와 관련된 텍스트 대화에서 위험성을 평가함으로써 멀티모달 모델은 복합적인 성범죄 상황을 효과적으로 탐지할 수 있다 [17].

2.2.3 자동화시스템의 워크플로우

자동화 시스템의 워크플로우는 특정 조건이 충족되었을 때 자동으로 작업을 수행하도록 설계된 구조다. 성범죄 신고 자동화 시스템에서는 성범죄 징후가 탐지되면 시스템이 자동으로 문

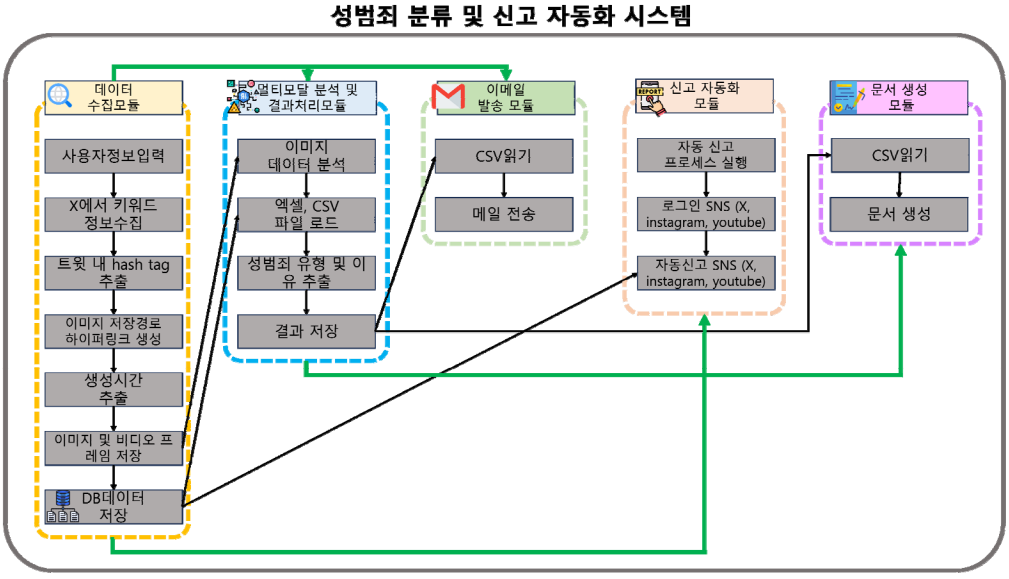
서를 작성하고, 이메일을 통해 관련 기관에 이를 전송하는 절차가 포함된다. Johnson & Zweig [21]은 자동화된 워크플로우 시스템이 조직의 효율성을 증대시킬 수 있다고 주장하며, 신뢰성 있는 데이터 흐름을 보장할 수 있는 시스템의 중요성을 강조한다. 이러한 자동화된 워크플로우는 피해자가 직접 신고하지 않더라도 위험 상황을 자동으로 인지하고 신속히 신고할 수 있는 환경을 제공한다.

2.2.4 이메일 및 문서 자동생성 이메일 전송

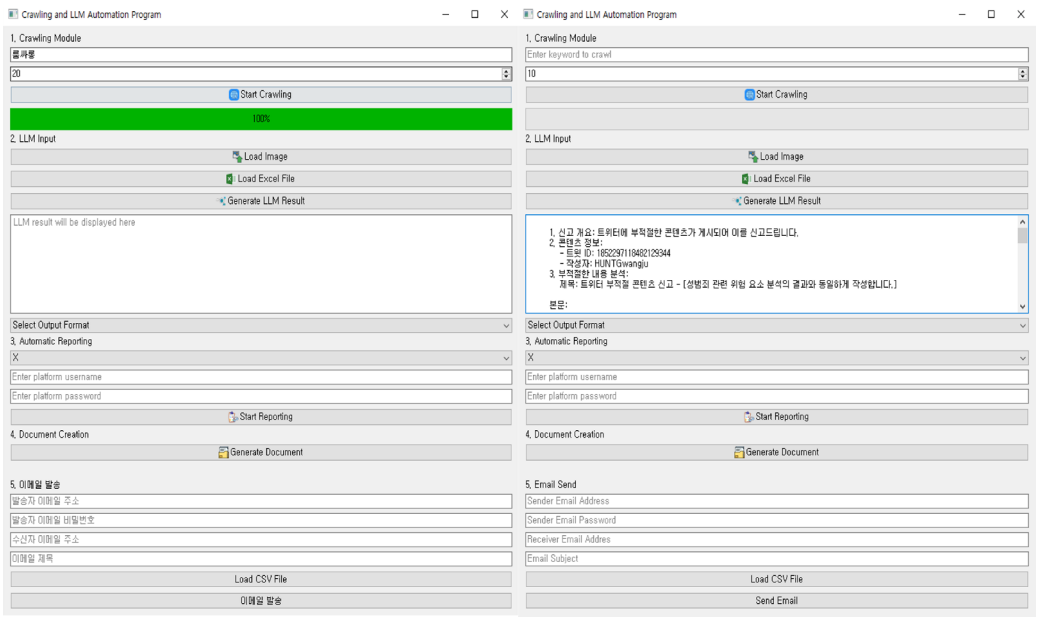
신고서 생성과 이메일 전송은 자동화된 성범죄 탐지 시스템의 핵심 요소로, 탐지된 성범죄 징후를 바탕으로 신고서를 자동 생성하고 상황에 맞는 데이터를 입력하여 관련 기관에 신속히 발송하는 과정을 포함한다. Casey [7]는 디지털 증거를 기반으로 자동화된 보고서 생성을 통해 수사 과정의 효율성을 높일 수 있음을 제시하며, 이러한 자동화 시스템이 신속한 대응을 가능하게 함을 강조한다. Nguyen et al. [22]은 최신 자동화 기술을 사용하여 신고서와 같은 중요한 문서의 작성과 전송을 효과적으로 구현할 수 있음을 보였다. 자동화된 이메일 전송 기술은 SMTP 또는 API를 통해 구현되며, 신고서가 빠르고 안전하게 전달되도록 설계된다.

3. 시스템 설계 및 구현

본 연구의 성범죄 자동 신고 시스템은 다양한 온라인 데이터에서 성범죄 관련 텍스트 및 이미지 데이터를 수집하고, 이 데이터를 통해 성범죄 가능성을 평가하여 자동으로 신고 문서를 생성하고 전송하는 일련의 자동화된 과정으로 구성된다. 이 시스템은 데이터 수집 모듈, 멀티모달 분석 모듈, 신고 자동화 모듈, 문서 생성 모듈, 이메일 발송 모듈로 구성되며, 각 모듈의 역할을 아래에 상세히 설명하였다.



<Figure 2> Architecture of Sexual Crime Classification and Automated Reporting System Green arrows show the relationship flow between modules, while black arrows show the relationship flow between functions.

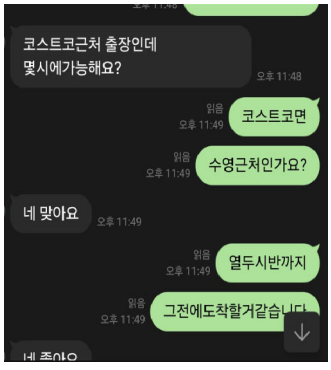


<Figure 3> System GUI

3.1. 데이터 수집 모듈

데이터 수집 및 전처리: 다양한 온라인 플랫폼 중 본 연구에서는 X의 유료 API를 통해 텍스트와 이미지 데이터를 수집하였다. 수집된 데이터는 분석에 적합한 형식으로 전처리되며, 텍스트와 이미지 데이터 각각에 대해 정규화, 표준화 작업이 이루어진다. 특히 텍스트 데이터는 불필요한 필요 정보들로만 컬럼에 맞게 데이터를 저장하며, 이미지 데이터는 분석에 적합한 해상도로 변환된다. 데이터 수집 모듈은 Figure 3의 1. Crawling Module로서 수집을 원하는 키워드를 입력하고, 한번에 수집 하고자하는 트윗의 수를 입력한 다음 Start Crawling을 클릭하면 데이터들이 지정된 경로에 이미지 데이터와 엑셀 데이터로 입력이 된다. Figure 3 좌측 이미지에서 보이는 것과 같이 프로세스가 진행되는 상태를 확인 할 수 있도록 구현하였다.

<Table 4> Example of Collected Data

이미지	텍스트
	1. Text : #부산게이 #부산ㄱㅇ #부산호기심 #부산섹트 #부산msg #울산게이 #창원게이 #김해게이 #진해게이 부산게이 마사지 방문.출장가능합니다! 라인 vmmooo 카톡 zc96 https://t.co/dPCKtZitLD 2. Hashtags : #부산게이 #부산ㄱㅇ #부산호기심 #부산섹트 #부산msg #울산게이 #창원게이 #김해게이 #진해게이

모듈을 통해 수집한 데이터는 이미지1,500장, 텍스트 1,500건을 수집하였으며, 각각의 데이터는 고유값(아이디)을 기준으로 이미지데이터는 파일명을 정하여 저장되었으며, 텍트스 데이터는 테이블데이터로 저장하였다.

<Figure 4> Tabular data example

3.2. 멀티모달 분석 모듈

텍스트 및 이미지 결합 분석: 멀티모달 분석 모듈은 텍스트와 이미지 데이터를 통합하여 성범죄 분류한다. 최신 멀티모달 모델인 Mini-CPM-Llama3-V2.5 기반으로 설계되어, 텍스트와 이미지 간의 상관관계를 학습하고 지정된 프롬프트에 따라 이미지 데이터를 분류하고 이메일과 문서작성에 필요한 결과를 출력한다.

성범죄 분류 : 텍스트 분석에서는 특정 키워드와 문맥을 분석하고, 이미지 분석에서는 컴퓨터 비전 기술을 통해 성적 암시를 감지한다. 텍스트와 이미지를 분석하여 성범죄 분류체계에 따른 범죄유형으로 분류하고 각 게시물의 신고를 위한 이메일, 문서작성을 위한 신고 글을 생성한다.

Figure 3의 2. LLM Input 영역에서 필요한 이미지와 DB데이터를 각각 업로드하면 해당 이미지와 DB를 분석하여 작성된 내용은 Figure 3의 우측의 이미지와 같이 “LLM result will be displayed here”에 출력되고, CSV파일로도 저장된다. CSV파일은 기본 DB에 모델이 생성한 이메일 작성과 문서작성을 위한 글을 새로운 컬럼을 생성하여 각 데이터에 대하여 저장된다.

3.3. 신고 자동화 모듈

데이터 수집 모듈을 통하여 수집된 성범죄 관련 URL은 Figure3의 3. Automatic Reporting을 통하여 SNS의 플랫폼을 선택한 뒤 플랫폼의 ID, Password를 입력하여 신고 자동화 프로그램 모듈을 실행하게 된다. 해당 모듈은 셀레니움기반으로 headless 상태로 백그라운드에서 프로그램이 실행되며, 업로드 된 엑셀 데이터에서 URL 컬럼 첫 번째 행부터 마지막까지 탐색하여 URL을 통해 게시글들을 신고하고, 신고완료 혹은 에러상황에 대한 부분은 터미널 창을 통해 확인이 가능하다.

3.4. 이메일 발송 모듈

Figure 3의 5.이메일발송의 영역에서 진행된다. MLLM을 통해 생성되어 기존의 DB에 저장된 이메일 버전의 초안을 불러와 이메일로 발송하게 된다. 이 모듈에서는 3.4의 문서생성 모듈과 같이 Load CSV File을 통해 CSV파일을 불러온다. 기본적으로 메일발송에 필요한 발송자 메일주소, 비밀번호, 수신자 메일주소를 작성하도록 하였으며, 이메일 제목을 넣어 메일을 발송할 수 있도록 하였다. SMTP 또는 전용 API를 통해 메일이 발송되며, 이때 업로드된 CSV파일도 첨부되어 발송 할 수 있다.

<Table 5> Email Sending Module Results 1

제목: 트위터 부적절 콘텐츠 신고 - 업소형의 조건만남
신고 개요:
트위터에 부적절한 콘텐츠가 게시되어 이를 신고드립니다.
콘텐츠 정보:
- 트윗 ID: 1832029672483455270
- 작성자: nanbam697036
- 게시 날짜: 2024-09-06 12:14:46
부적절한 내용 분석:
해당 트윗은 유흥 포털을 홍보하고, 전국의 유흥(마사지, 오피, 풀사롱, 휴게텔, 북창동시스템) 정보를 제공하며, 프로필부터 부킹까지 한곳에서 확인할 수 있다는 것을 밝히고 있다. 해시태그도 유흥 관련 내용을 포함하고 있어, 성범죄 분류체계 중 '업소형'의 '조건만남'으로 분류할 수 있다.
법적 위반 가능성:
해당 콘텐츠는 트위터 이용약관 및 관련 법규에 위반될 수 있습니다.
요청 사항:
위의 분석 결과를 바탕으로 해당 콘텐츠의 즉각적인 삭제 및 추가 조치를 요청드립니다.

<Table 6> Email Sending Module Results 2

제목: 트위터 부적절 콘텐츠 신고 - 출장형
본문:
1. 신고 개요:
트위터에 부적절한 콘텐츠가 게시되어 이를 신고드립니다.
2. 콘텐츠 정보:
- 트윗 ID: 1829446603746513026
- 작성자: lingdingdi33117
- 게시 날짜: 2024-08-30 09:10:34
3. 부적절한 내용 분석:
이미지와 엑셀 데이터를 분석한 결과, 트윗의 내용은 '광주떡집추천'과 관련된 여행을 포함하고 있다. 해시태그 #요리하기는 요리를 추천하는 것을 의미한다. 이 콘텐츠는 출장형 성범죄에 해당한다. 이유는 트윗의 내용에서 광주 떡집이 추천되며, 이는 출장샵(출장안마, 애인대행, 섹파매칭)과 유사한 의도와 목적을 가질 수 있다는 것을 보여준다.
4. 법적 위반 가능성:
이러한 콘텐츠는 관련 법규나 트위터 이용약관 위반 사항을 일으킬 수 있다.
5. 요청 사항:
위의 분석 결과를 바탕으로 해당 콘텐츠의 즉각적인 삭제 및 추가 조치를 요청드립니다.

3.5. 문서 생성 모듈

Figure 3의 4. Document Creation은 Word 문서를 생성하는 기능이다. 이 과정에서 Figure 3의 2. LLM Input에서 MLLM으로 분석한 이미지와 텍스트 데이터를 활용한다. 특정 포맷으로 저장된 신고 관련 문서 초안(CSV 파일)을 Load CSV File로 불러온 뒤, Generate Document를 클릭하면 문서가 생성된다. 이 모듈은 사전에 정의된 템플릿을 사용하여 신고를 위한 ID, Username, 게시시간 등을 포함하고 있다. 자동화된 문서 생성으로 신고초안을 작성하여 사용자의 문서작성에 대한 신속하고 효율적인 보조 역할을 가능하게 한다.

<Table 7> Document Creation result 1

제목: 불법/유해 콘텐츠 신고 보고서-[업소형 중 주점형]

1. 개요
   - 신고 일시: 2024년 09월 09일 14시 06분
   - 신고 대상: 소셜 미디어 게시물 (플랫폼: 트위터)
   - 신고 사유: 상무지구 하이 퍼블릭 룸의 운영 및 관련 정보 공유
2. 콘텐츠 세부 정보
   - 게시물 ID: 1832734593818226786
   - 게시자: OndovG40593
   - 게시 일시: 2024-09-08 10:55:52
   - 콘텐츠 유형: photo
3. 이미지 분석 결과
   - 상무지구 하이 퍼블릭 룸 내의 여자들이 모인 사진을 볼 수 있다. 사진은 공중전화실과 비슷한 환경에서 찍어졌다.
4. 성범죄 관련 분석
   - 해시태그와 트윗 내용에서 "하이퍼블릭"과 "룸카페"와 같은 단어들이 사용되어, 이러한 용어가 일반적으로 성매매와 관련된 용어로 인식된다. 또한, "DM&카톡문의:Rid2040"라는 문구가 exists, 이는 서비스 제공자의 문의 정보를 제공하는 것으로, 이는 성매매 알선 포털 사이트에 해당하는 유형으로 간주될 수 있다.
5. 법적 위반 사항
   - 트위터의 사용 약관에 따라, 불법적인 콘텐츠나 성범죄와 관련된 콘텐츠는 허용되지 않는다. 이 콘텐츠는 조건만남과 관련된 알선 포털 사이트의 유형에 해당하여, 관련 법률을 위반할 수 있다.
6. 권고 사항
   - 트위터가 관련 콘텐츠를 즉각적으로 삭제하고, 추가 조사를 통해 조치가 필요한지 확인해야 한다.
7. 결론
   - 이 콘텐츠는 조건만남과 관련된 알선 포털 사이트의 유형에 해당하여, 법적 위반 사항이 있다. 트위터는 관련 콘텐츠를 즉각적으로 삭제하고, 추가 조치를 취해야 한다.

<Table 8> Document Creation result 2

제목: 불법/유해 콘텐츠 신고 보고서-[조건만남]

1. 개요
   - 신고 일시: 2024년 09월 09일 15시 17분
   - 신고 대상: 소셜 미디어 게시물 (플랫폼: 트위터)
   - 신고 사유: 이미지와 트윗 내용에서 성범죄 관련 콘텐츠가 포함되어 exists.
2. 콘텐츠 세부 정보
   - 게시물 ID: 1832747156954100217
   - 게시자: ACurits53177
   - 게시 일시: 2024-09-08 11:45:47
   - 콘텐츠 유형: photo
3. 이미지 분석 결과
   - 이미지의 일부는 성적 내용을 암시하는 요소를 포함하고 있다.
4. 성범죄 관련 분석
   - 트윗 내용과 해시태그는 조건만남과 성매매 알선 포털 사이트와 유사한 성범죄의 경로를 암시한다.
   - 해시태그: #군산 #조건 #만남 #지인능욕 #능욕 #맘눌뎀 #야노 #오프 #야썰
5. 법적 위반 사항
   - 이미지와 트윗 내용은 조건만남과 성매매 알선 포털 사이트에 해당하는 법률 위반 사항이 될 수 있다.
6. 권고 사항
   - 해당 콘텐츠는 조건만남과 성매매 알선 포털 사이트와 관련된 법률 위반 사항이 될 수 있으므로 즉각적인 삭제와 추가 조사가 필요하다.
7. 결론
   - 이 리포트는 공식 문서이므로 객관적이고 전문적인 언어를 사용하여 작성하였으며, 불필요한 개인적 의견이나 추측은 배제하고 관찰된 사실에 근거하여 작성하였다.

4. 시스템 개발

스템 개발은 크게 데이터 수집 및 전처리, 텍스트 및 이미지 통합 분석, 자동 신고, 이메일 및 문서 생성, 시스템 최적화와 성능 평가의 다섯 가지 단계로 이루어져 있으며, 각 과정에서의 구체적인 구현 방식은 다음과 같다.

4.1 데이터 수집 및 전처리

성범죄 자동 신고 시스템의 성능은 학습에 사용된 데이터의 품질에 크게 의존한다. 이에 따라 텍스트와 이미지 데이터를 포함한 고품질 학습 데이터를 구축하기 위해 X API를 활용한 크롤링 모듈을 사용하여 데이터를 수집하였으며, 이 데이터를 멀티모달 모델의 입력 형식에 맞추어 전처리하였다.

본 연구를 위하여 수집한 데이터는 이미지 데이터는 약1,500개, 텍스트 데이터 1,500개가 매칭될 수 있도록 수집하였다. 이미지와 텍스트 데이터는 고유 ID를 통하여 매칭될 수 있도록 crawling module에서 저장하도록 하였다.

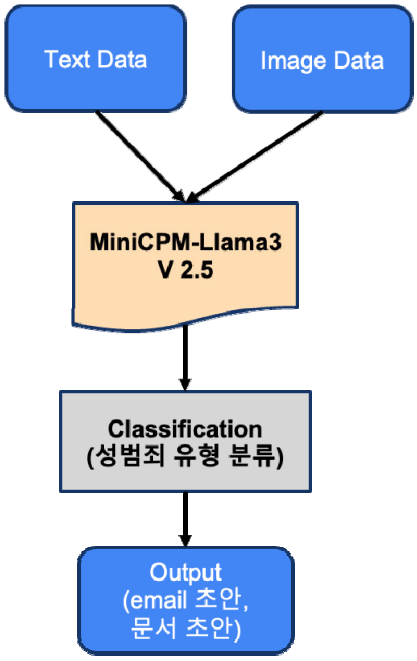
텍스트 데이터 수집 및 전처리는 X API를 활용한 크롤링 모듈을 통해 X에서 성범죄와 관련된 특정 키워드를 포함하는 텍스트 데이터를 수집한다. 텍스트 데이터는 성범죄와 관련된 특정 표현, 즉 성범죄 분류체계의 분류에 해당하는 키워드의 텍스트가 포함되도록 구성한다. 수집된 데이터는 학습에 성범죄 분류 및 신고자동화에 활용할 수 있도록 ID, 작성자, 생성일자, 게시글과 같은 정보가 포함되며, 별도의 Hash TAG 컬럼으로 게시글 내 포함된 hash tag를 구분하여 데이터를 저장 한다.

이미지 데이터 수집 및 전처리는 이미지 데이터는 텍스트 데이터와 같이 X API를 통해 수집하며, 특정 키워드를 통해 수집되는 이미지, 동영상의 첫프레임을 저장하여 성범죄와 관련된 의심스러운 이미지나 불법 콘텐츠가 포함된 데이터를 수집한다. 각 이미지는 성범죄 분류 체계에 따른 키워드로 파일별 라벨링이 된다. 이미지 데이터는 이미 자체로서 성범죄와의 연관성이 있을 수 있으며, 또한 성범죄에 활용하기 위한 문구들이 이미지화 되어 있는 경우가 존재한다. 즉 이미지 내 성적 콘텐츠가 아닌 텍스트 요소에도 집중할 수 있도록 한다.

멀티모달 입력 형식은 텍스트와 이미지 데이터는 멀티모달 입력 형식에 맞춰 통합되어 모델에 입력될 수 있도록 구성한다. 예를 들어, 텍스트와 이미지가 동시 입력값으로 주어 모델이 텍스트와 이미지 간의 상호 관계를 파악하여 성범죄 분류를 보다 정확하게 수행할 수 있도록 한다.

4.2. 텍스트 및 이미지 통합 분석

성범죄 신고 자동화 시스템의 탐지 모델은 텍스트와 이미지 데이터를 동시에 분석하는 멀티모달 대형 언어 모델(MLLM)을 기반으로 구축한다.



<Figure 5> Visualization of Text and Image Integration for Analysis

MiniCPM-Llama3-V 2.5 모델을 활용하여, 본 연구에서는 경량화된 버전인 int4로 양자화된 모델을 사용하였다. 텍스트와 이미지 데이터 간의 상호 관계를 통합적으로 분석하는 구조를 채택한다. 텍스트와 이미지가 동시 입력값으로 주어질 경우 모델이 텍스트와 이미지 간의 관계를 파악하여 성범죄 분류를 보다 정확하게 탐지할 수 있도록 한다.

이미지 내 포함된 텍스트를 분석하기 위해 OCR(광학 문자 인식) 기능을 통합하여 사용한다. 이를 통해 이미지에서 텍스트를 정확히 추출하고, 성범죄와 관련된 문구나 대화 내용을 파악한다. 예를 들어, 성적 이미지를 포함한 텍스트가 이미지에 포함된 경우 OCR 기능을 통해 텍스트를 분리하여 분석함으로써 성범죄와 관련된 구체적인 증거를 확보한다.

4.3. 신고자동화

수집된 데이터는 특정 키워드를 통하여 성범죄 징후가 감지되는 데이터들로 텍스트 데이터에 저장된 게시글의 URL을 활용한다. 이때 셀레니움 기반의 자동 신고 모듈을 활용해 수집된 데이터를 기반으로 신고과정을 진행하여 성범죄 관련 게시글을 신속하게 신고할 수 있도록 한다.

4.4. 이메일 및 문서 생성

시스템은 자동으로 신고서를 생성하고, 이를 관련 기관에 전달하는 자동 신고 및 문서 생성 프로세스를 수행한다. 시스템이 감지한 위험 상황에 대한 신고서는 자동으로 생성된다. 신고서에는 성범죄 징후의 구체적인 내용, 게시글 작성자의 ID, Username, 작성 시간, 텍스트 내용, hashtag와 같은 정보가 포함되어 글이 작성된다. 이메일과 문서의 포맷은 약간의 차이가 있지만 수집된 이미지와 텍스트 데이터가 매칭되어 그 정보를 탐색하여 MLLM을 통해 생성된 글이 CSV파일로 저장하여 이를 활용한다. 자동 생성된 신고서는 신속한 대응을 가능하게 한다. 이

를 통해 성범죄와 관련된 사건의 경위를 구체적이고 명확하게 전달할 수 있도록 구성되었다.

4.5. 시스템 최적화 및 성능평가

성범죄 자동 신고 시스템의 실효성을 높이기 위해 최적화와 성능 평가 절차를 수행하였다. 실시간 데이터 처리와 빠른 응답 속도를 위해 모델의 파라미터 수를 줄이고, 모바일 환경에서도 원활히 작동할 수 있도록 int4로 경량화된 모델을 활용하였다. 또한 양자화 기법을 적용하여 낮은 성능의 하드웨어에서도 시스템이 안정적으로 작동하도록 설계되었다.

성능 평가는 실제 시나리오에서 시스템의 탐지 정확도와 신고의 신뢰성을 평가 하였다. 모델의 성능은 각 카테고리별 정확도(accuracy)를 기준으로 평가하였으며, 이를 100%로 환산하여 아래 Table9에 요약하였다.

<Table 9> The sample table for describing editing rules

MiniCPM-Llama3 V2.5 int4	Accuracy (total)	category1 (출장형)	category2 (주점형)	category3 (위장형)	category4 (조건만남)
이미지	28%	8%	45%	27%	28%
이미지&텍스트	41%	57%	47%	38%	75%
이미지&텍스트 + CoT	62%	44%	63%	52%	73%

이를 통해 제안된 시스템은 CoT(Chain of Thought) 기법을 추가했을 때 가장 높은 정확도를 달성하였음을 확인할 수 있다.

5. 결과 및 논의

5.1. 주요결과 요약

본 연구에서 구현한 온라인 성매매 홍보 불법컨텐츠 자동 신고 시스템은 성범죄와 관련된 텍스트 및 이미지 데이터를 효과적으로 수집하고, 멀티모달 대형 언어 모델(MLLM)을 통해 성범죄 징후를 실시간으로 탐지하며, 신고 및 문서 생성 절차를 자동화하는 기능을 포함한다. 시스템의 성능을 평가한 결과, 텍스트 및 이미지 데이터의 결합 분석을 통한 성범죄 탐지 정확도가 기존의 단일 모달 분석보다 높다는 것을 확인하였다. 또한, 자동 신고 및 문서 생성 과정에서 신속성과 효율성이 확보되어 성범죄 관련 콘텐츠의 조기 탐지 및 대응에 효과적임을 입증하였다.

5.2. 성과와 한계 및 개선 방안

본 연구에서 제안한 시스템은 텍스트와 이미지 간의 상호 관계를 효과적으로 분석하고, 신고 자동화, 이메일 발송, 문서 생성 과정에서 사용자 개입을 최소화하여 효율성과 정확성을 크게 향상시켰다. 특히 멀티모달 모델을 활용한 접근은 성범죄 신고 절차의 신속성과 신뢰성을 높이는 데 기여했다. 그러나 몇 가지 한계점도 확인되었다.

첫째, 모델은 오탐률(False Positive)과 미탐률(False Negative) 문제를 완전히 해결하지 못했다. 복잡한 이미지나 문맥이 모호한 텍스트의 경우 성범죄 여부를 정확히 판단하지 못하는 사례가 발생했으며, 이는 데이터셋의 다양성과 훈련 데이터 품질 부족에서 기인한 것으로 보인다.

둘째, 개인정보 보호 문제가 있다. 성범죄 관련 데이터를 수집하고 분석하는 과정에서 민감한 개인정보가 포함될 가능성이 높아 안전한 데이터 처리와 보안 강화가 필수적이다.

이러한 한계를 극복하기 위해 정교한 프롬프트 엔지니어링 기법을 도입하고, 모델 구조 최적화 및 하이퍼파라미터 튜닝을 통해 성능을 개선할 필요가 있다. 또한, 데이터 암호화 기술을 활용하여 민감 정보를 안전하게 관리하고, 법적 및 윤리적 기준에 맞춰 시스템을 설계함으로써 개인정보 보호 문제를 해결해야 한다.

5.3. 이론적 및 실무적 시사점

본 연구는 성범죄 탐지 및 신고의 자동화를 위한 새로운 접근 방식을 제안함으로써, 온라인 성매매 홍보 불법컨텐츠 대응 시스템의 가능성을 이론적으로 확장하였다. 멀티모달 모델을 활용한 접근 방식은 성범죄 관련 콘텐츠를 보다 정확하게 탐지하고 분류할 수 있음을 보여주었으며, 이러한 시스템이 실무적으로 다양한 상황에 적용될 수 있는 가능성을 제시하였다.

실무적으로, 본 시스템은 법 집행 기관이나 사회적 보호 기관에서 즉각적인 성범죄 대응이 가능하도록 지원할 수 있으며, 성범죄 피해 예방 및 피해자 보호에도 기여할 수 있다. 특히, 자동 신고 기능은 사용자가 직접 신고하지 못하는 상황에서도 신속히 대응할 수 있는 방안을 제공한다.

6. 결론 및 향후 연구

6.1. 연구요약

본 연구에서는 온라인 성매매 홍보 불법컨텐츠 신고 자동화 시스템을 개발하여 성범죄와 관련된 텍스트 및 이미지 데이터를 탐지하고, 신고 자동화와 이메일 및 문서 생성을 수행하는 과정을 제안하였다. 시스템은 크롤링 모듈, 멀티모달 분석 모듈, 신고 자동화 모듈, 문서 생성 모듈, 이메일 발송 모듈로 구성되며, 각 모듈이 유기적으로 연동되어 성범죄 탐지 및 대응의 효율성을 극대화한다. 이로써 본 연구는 온라인 성매매 홍보 불법컨텐츠 대응을 위한 자동화 시스템의 가능성을 실증적으로 입증하였다.

6.2. 자동 신고 시스템의 잠재적 영향

본 연구에서 제안한 자동 신고 시스템은 사회적 측면에서 중요한 영향을 미칠 수 있다. 특히, 온라인 성매매 홍보 불법컨텐츠의 조기 발견과 빠른 대응이 가능해져 피해자 보호에 중요한 역할을 할 수 있으며, 사회적 안전망 구축에도 기여할 수 있다. 자동화된 시스템을 통해 성범죄와 관련된 콘텐츠가 즉시 신고될 수 있어, 잠재적 피해를 최소화하고 범죄 예방에 기여할 수 있다. 이러한 시스템이 확산될 경우, 온라인 성매매 홍보 불법컨텐츠에 대한 경각심을 높이고 사회적 경계를 형성하는 데 기여할 것이다.

6.3. 향후 연구 방향

향후 연구에서는 본 시스템의 성능을 더욱 향상시키기 위해 다양한 형태의 성범죄 데이터를 수집하고, 데이터셋의 다양성을 높이는 방향으로 나아가야 한다. 또한, 모델의 정확도와 신뢰성을 개선하기 위해 하이퍼파라미터 최적화 및 고급 딥러닝 기법을 추가적으로 적용할 필요가

있다.

또한, 성범죄 신고 자동화 시스템의 법적, 윤리적 측면에 대한 연구도 병행되어야 한다. 온라인 성매매 홍보 불법컨텐츠와 관련된 개인정보 보호와 데이터 보안이 중요한 만큼, 시스템의 개발 및 활용에 있어서 법적 기준과 윤리적 책임을 고려하는 것이 중요하다. 이를 통해, 온라인 성매매 홍보 불법컨텐츠 대응 시스템이 사회적 신뢰를 얻고 실질적인 문제 해결에 기여할 수 있도록 지속적인 연구가 필요하다.

이와 같은 연구와 개발 방향을 통해 성범죄 자동 신고 시스템이 더욱 발전할 수 있을 것이며, 사회적 안전과 디지털 범죄 예방에 기여하는 중요한 기술로 자리잡을 가능성이 크다.

감사의 글(Acknowledgements)

본 논문은 서울연구원 (2024-PR-39. 서울 디지털 성범죄 통합대응체계 구축방안)의 지원을 받아 수행된 연구임.

참고문헌(References)

[1] National Youth Policy Institute. 2021. Research on the Status of Digital Sexual Crimes Against Children and Adolescents and Countermeasures (아동·청소년 대상 디지털 성범죄 현황 및 대응방안 연구). Sejong, Korea: National Youth Policy Institute.

[2] Marciniak R, Stefańska M. 2021. Detection and Prevention of Child Sexual Exploitation Online: An Overview of Existing Techniques. IEEE Access.

[3] Hancock JT, Thom-Santelli J, Ritchie T. 2004. Deception and Design: The Impact of Communication Technology on Lying Behavior. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[4] Radford A, Kim JW, Hallacy C, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. ICML.

[5] Baltrušaitis T, Ahuja C, Morency LP. 2018. Multimodal Machine Learning: A Survey and Taxonomy. ACM Computing Surveys.

[6] Abdullah A, Mohammad RM. 2020. Automated Detection of Online Child Exploitation Content Using AI and Machine Learning Techniques. Computers & Security.

[7] Casey E. 2011. Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet. Academic Press.

[8] Mundie DA, Taylor MS. 2018. Privacy-Preserving Data Analysis for Public Safety and Law Enforcement. IEEE Security & Privacy.

[9] Dwork C, Roth A. 2014. The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science.

[10] Gomez R, Zervas P, Sampson DG. 2019. Automatic detection of offensive language in social media using deep learning. Social Network Analysis and Mining.

[11] Hosseini H, Xiao B, Poovendran R. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[12] Schmidt A, Wiegand M. 2017. A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.

[13]Audebert N, Le Saux B, Lefevre S. 2019. Deep Learning for Classification of Multimodal Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing.

[14] Sharma A, Bhatt C, Mittal, M. 2020. Multimodal Emotion Recognition Using Cross-Modal Attention and Contextual Augmentation. Neurocomputing.

[15] Devlin J, Chang M, Lee K, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[16] Vaswani A, Shazeer N, Parma N, et al. 2017. Attention is All You Need. Advances in Neural Information Processing Systems.

[17] Jiang Z, Lin J, Wang, M. 2019. Convolutional Neural Networks for Child Exploitation Image Classification: An Empirical Study. IEEE Access.

[18] Li LH, Yatskar M, Yin D, et al. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03557

[19] Chen YC, Li L, Yu L, et al. 2020. UNITER: UNiversal Image-TExt Representation Learning. ECCV.

[20] Yao Y, Yu T, Zhang A, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800

[21] Johnson RD, Zweig D. 2021. The Impact of Automated Workflow Systems on Organizational Efficiency. Journal of Organizational Behavior.

[22] Nguyen D, Luong MT, Manning CD. 2015. Efficient Sequence Labeling with RNNs and CRFs. Proceedings of the Conference on Empirical Methods in Natural Language Processing.