ISSN(Print) : 1976-5304, ISSN(Online) : 2713-931X
https://kdfs.jams.or.kr, DOI : 10.22798/kdfs.2022.16.1.118

디지털성범죄 예방을 위한 RNN 기반 취약군 선별 체계 연구

유 혜 진[*], 박 아 현[***], 박 윤 지[*], 이 지 원[*], 주 다 빈[*], 최 혜 인[****], 정 두 원[**]
동국대학교 경찰사법대학 경찰행정학부 (학부생)[*], (조교수)[**], 동국대학교 일반대학원 경찰행정학과 (대학원생)[***], 한국산업기술보호협회(연구원)[****]

A Study on Screening System for Vulnerable Users with RNN for Preventing Digital Sexual Crime

Hyejin Ryu[*], Ah-Hyun Park[***], Yunji Park[*], Jiwon Lee[*], Dabin Joo[*], Hyein Choi[****], Doowon Jeong[**]
**College of Police and Criminal Justice, Dongguk University (Undergraduate Student)[*], (Assistant Professor)[**],
Department of Police Administration, Dongguk University (Graduate Student)[***],
The Korean Association for Industrial Technology Security(Researcher)[****]**

요 약

정보통신 기술의 발달로 'n번방 사건'과 같은 새로운 유형의 성범죄가 성행하고 있다. 일반적으로 사이버 범죄를 예방하는 연구들은 범죄자 혹은 범죄 조직을 식별하는 프로파일링 기법을 제시하였다. 그러나 익명성을 보장하는 해외 SNS 및 메신저를 통해 피해자의 심리상태를 교묘하게 이용하여 성착취하는 디지털성범죄의 특성으로 인해 가해자의 특징을 파악할 수 있는 데이터 자체를 확보하기 어렵다. 또한, 잠재적 범죄자를 선별하는 것은 인권 침해와 관련한 문제를 야기할 수 있을 뿐만 아니라, 프로파일링 과정에서의 1종 오류는 무고한 시민을 잠재적 범죄자로 오인하여 억울한 피해자를 발생시킬 수 있다.

이에 본 논문에서는 기존 연구와는 달리 디지털성범죄를 대표적인 범죄피해 이론인 일상활동이론으로 분석하여 디지털성범죄 취약군을 선별함으로써 범죄를 예방하는 프레임워크를 제시한다. 피해자들의 SNS 게시글을 학습하여 RNN 기반의 취약군 분류기를 만들고 이를 실무적으로 활용할 수 있는 방안을 소개한다. 취약군에게 성범죄에 노출되어 있다는 사실을 인지시킴으로써 범죄자의 접근으로부터 회피할 수 있도록 유도하고 범죄 발생 시 대응 방안을 안내함으로써 증거를 빠르게 수집하고 가해자를 검거하는 데 도움이 될 수 있도록 한다.

주제어 : 디지털성범죄, 범죄예방, 일상활동이론, 딥러닝, 순환신경망

ABSTRACT

With the development of information and communication technology, new types of sexual crimes are occurring such as the "Nth room case". In general, profiling techniques have been presented to identify criminals or criminal groups to prevent cybercrime. However, it is difficult to secure data that can identify the characteristics of criminals due to the nature of digital sexual crimes that cleverly exploit victims' psychology by using foreign SNS and messengers which anonymity is guaranteed. In addition, screening potential criminals can cause problems related to human rights violations, and type 1 errors in the profiling process can treat innocent citizens as potential criminals, getting the wrong person.

In this paper, we present a framework based on the routine activity theory, to detect vulnerable users, for preventing digital sexual crimes. We introduce a way to create an RNN-based vulnerable users classifier by learning the victims' SNS posts and utilize the classifier in practice. It encourages the vulnerable users to avoid criminals by recognizing that they are exposed to sexual crimes, and guides countermeasures in case of crimes so that investigators may help to quickly collect evidence and arrest the perpetrators.

Key Words : Digital Sexual Crime, Crime Prevention, Routine Activity Theory, Deep Learning, Recurrent Neural Network

※ 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1G1A1011753).
- Received 17 July 2021, Revised 18 July 2021, Accepted 31 March 2022
- 제1저자(First Author) : Hyejin Ryu (Email : hyejinpn@dgu.ac.kr)
- 교신저자(Corresponding Author) : Doowon Jeong (Email : doowon@dgu.ac.kr)

Ⅰ. 서 론

최근 정보통신 기술의 발달로 다양한 유형의 신종 범죄가 발생하고 있다. 그 중 디지털 기기를 이용하여 타인의 성적 자율권과 인격권을 침해하는 디지털성범죄는 사회적으로 큰 파장을 일으키고 있다. 특히 다크웹에서 운영되던 아동 포르노 사이트 '웰컴 투 비디오'와 텔레그램 성착취 채팅방(이하 'n번방')에서는 성착취물이 유포되고 있을 뿐만 아니라 피해자들의 약점을 빌미로 성적인 영상물을 제작하도록 교사한 것으로 알려졌다. 범죄자들은 주로 자신의 신체를 찍은 사진이나 영상을 올리거나 조건만남을 구하는 게시글을 올리는 '일탈계' 사용자들을 범죄의 대상으로 삼고 수사기관으로 사칭하거나 신상털기로 획득한 개인정보로 피해자들을 협박한 뒤, 신체 부위를 노출한 사진을 전송할 것을 요구하거나 성적 행위에 참여하도록 강요하는 등의 성 착취를 하였다[1].

사이버 공간에서 이루어지는 성 착취 범행이 밝혀지면서 디지털성범죄를 예방할 수 있는 정책 마련에 대한 사회적 요구가 급증하였다. 경찰청에서는 이와 같은 성착취 범죄를 사전에 방지하기 위해 「디지털성범죄 특별수사본부」를 신설하고 국회에서는 일명 'n번방 방지법'이라고 칭해지는 법안을 개정하는 등 많은 정책적 노력을 시도하고 있다. 이와 함께 사이버 공간에서 발생하는 성범죄를 탐지하여 범죄자를 검거하거나 범죄자가 피해자에게 접근하는 것을 사전에 차단하는 기술적 방안에 대해서도 논의가 진행되고 있다. 범죄자의 특성에 기반하여 잠재적 범죄자를 식별하는 프로파일링 기법을 통해 디지털성범죄를 예방하는 방안이 제시되고 있다. 특히 범죄자 프로파일링 기법은 SNS상에서의 테러집단 식별 기법이나 네트워크 분석을 통한 그룹 탐지, 작성글의 습관을 학습하는 작성자 탐지 기법 등 이미 관련 연구가 다수 진행된 분야로 디지털성범죄 예방을 위해 해당 기법을 적용해볼 수 있다.

그러나 디지털성범죄에 기존의 프로파일링 기법들을 적용하는 것은 다음과 같은 이유로 한계점이 있다. 우선, 주로 가해자들은 가짜 계정을 이용하거나 일대일 대화를 통해 피해자에게 접근하여 온라인상에 SNS 사용 흔적을 거의 남기지 않아 프로파일링을 위한 데이터 수집 자체에 매우 제약이 있다. 또한, 잠재적 범죄자를 선별하는 것은 인권 침해와 관련하여 법적 문제의 소지가 있어 실무적으로 활용되는데 제한이 따른다. 마지막으로, 프로파일링의 1종 오류(False Positive) 발생으로 무고한 시민을 용의자로 판단할 가능성을 배제할 수 없다.

이에 본 논문에서는 연구 대상을 범죄자에 치중하였던 기존 연구의 관점에서 벗어나 디지털성범죄 취약군을 식별하고자 한다. 대표적인 범죄피해 이론인 일상활동이론에 기반하여 디지털성범죄를 분석하고 이를 통해 디지털성범죄 취약군 식별의 중요성을 확인한다. 또한, 취약군의 SNS 게시글을 크롤링하고 그 결과를 바탕으로 취약군을 판별하는 딥러닝 기반 알고리즘을 개발한다. 더불어 판별된 취약군에게는 성범죄에 노출되어 있다는 알람을 제공하여 성범죄를 예방하는 프레임워크를 제시함으로써 실무에 적용할 수 있는 방안도 제시한다. 마지막으로, 실험데이터를 확보하여 제시한 모델을 구현함으로써 프레임워크의 활용도를 검증한다. 약 4만 6천개의 실험데이터를 확보한 후 자연어 처리와 워드 임베딩, Gated Recurrent Units(GRU)로 구성된 모델을 학습하고 해당 모델의 정확도를 측정한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 일상활동이론에 기반하여 디지털성범죄 사례를 분석한다. 3장에서는 디지털성범죄 취약군 선별을 위한 프레임워크를 제시하고, 4장에서는 제시된 취약군 선별 모델에 대한 성능을 검증한다. 5장에서는 향후 연구계획과 함께 논문을 결론짓는다.

Ⅱ. 일상활동이론 기반 디지털성범죄 분석

2.1. 일상활동이론

일상활동이론은 시·공간적 요소와 1) 동기화된 범죄자(motivated offender), 2) 범행에 적합한 대상물(suitable target), 3) 적당한 경비 상황의 부재(absence of capable guardian), 이 세 가지 요소가 결합하여 범죄가 발생한다고 설명한다[2]. 동기화된 범죄자는 범행동기가 충만한 자를 의미하며 적당한 표적은 범죄적 경향을 일으킬 수 있는 사람 혹은 사물을 뜻한다. 범죄자는 타겟의 가치(value), 가시성(visibility), 접근성(accessibility), 부동성(inertia)을 고려하여 매력적인 타겟을 선정하고 그를 대상으로 범행을 실행한다. 적당한 경비 상황의 부재는 범죄를 억제할 수 있는 유능한 경비 상태가 존재하지 않는 것을 말하는데, 경비의 예로 경찰순찰, 이웃 및 동료, CCTV 시스템 등이 있다. 즉, 일상활동이론은 어떤 시간과 장소에 동기화된 범죄자와 적합한 대상물이 존재하고, 경비가 이루어지지 않는 상태가 수렴한다면, 범죄가 발생할 가능성이 증가하게 된다는 이론이다.

정보화 시대의 도래에 따라 일상활동이론에서 고려할 수 있는 공간적 요소가 오프라인에서 온라인으로 확장되었다. 일상활동이론에 기반하면 사이버 공간은 시공 초월적 특성으로 범죄자는 범행에 적합한 대상물에 대한 접근성이 높고, 익명성과 초국경성으로 인하여 경비의 부재가 발생하여 범죄가 발생하기 적합한 공간으로 판단된다[3].

〈Figure 1〉 The routine activity theory

2.2. 디지털성범죄 사례 분석

디지털성범죄는 디지털 기기를 이용하여 인간의 인격권과 성적 자기결정권을 침해하는 모든 행위를 뜻한다. 디지털성범죄는 자신 또는 타인의 성적 욕망을 유발하거나 만족시킬 목적으로 다른 사람의 신체를 촬영하거나 그 촬영물을 반포·임대·제시·제공·판매 또는 상영·전시하는 '유포형', 정보통신망을 통해 디지털 성폭력 행위를 제공할 목적으로 제조·수입·수출·유포하는 '제작형', 디지털 성폭력 행위의 게시글에 동조 등 가담하는 '참여형', 그리고 강간 촬영물을 소지·시청하는 '소비형'으로 분류할 수 있다[4]. 텔레그램 성착취 채팅방 운영자들은 일탈계를 운영하는 이용자에게 접근하여 수사기관으로 사칭한 후 개인정보를 탈취해 성착취물을 제작하고 이를 유포하는 방식으로 범죄를 저질러 '유포형' 및 '제작형' 성범죄자에 해당된다. '웰컴 투 비디오'의 경우 운영자들이 성착취물을 웹사이트에 게시하여 범죄수익을 거둔 바 '유포형' 성범죄자에 해당한다. 따라서 본 논문에서는 위와 같은 '유포형'과 '제작형' 성범죄를 연구 대상으로 선정하였다.

'n번방' 사건과 '웰컴 투 비디오' 사건을 일상활동이론에 기반한 결과는 [표 1]과 같다. 'n번방'의 경우 SNS 상에서 접근이 쉬운 일탈계 사용자들이 타겟이 되었다. 특히, 트위터에서는 신체 특정부위를 촬영한 사진이나 영상을 공유하는 게시물을 제한하는 규정이 없어 피해자들은 별도의 필터링 없이 범죄자들에게 노출되었으며 익명성 강한 텔레그램의 특성으로 인해 유통 단계에서도 적절한 감시가 이루어지지 못했다. '웰컴 투 비디오' 사건에서는 고수익을 거둘 수 있는 아동들이 타겟이 되었고 익명성이 보장되는 토르 네트워크 환경에서 성착취물이 거래 및 유포되었다.

〈Table 1〉 Case analysis based on the routine activity theory

사례	요소	설명
n번방	동기화된 범죄자	성범죄자
	범행에 적합한 대상물	알탈계, 섹트를 운영한 이용자
	적당한 경비 상황의 부재	SNS 이용 가이드라인 및 익명성으로 인한 이용자 관리의 불가
	공간	트위터, 텔레그램
웰컴 투 비디오	동기화된 범죄자	성범죄자
	범행에 적합한 대상물	아동, 미성년자
	적당한 경비 상황의 부재	익명성으로 인해 이용자 관리의 불가
	공간	다크웹

2.3. 취약군 선별의 필요성

일상활동이론에 기반하여 주요 디지털성범죄 사례를 확인한 결과, '적당한 경비 상황의 부재'가 범죄 발생에 직접적으로 영향을 준 것임을 알 수 있다. 이에 감시의 강화는 범죄 예방을 위한 일차적인 해결책으로 고려될 수는 있으나, 익명성을 기반으로 하는 텔레그램과 토르 네트워크의 특수성 및 프라이버시 침해 문제로 인해 현실적으로 어려움이 따른다.

한편, 디지털성범죄를 예방하기 위해 '동기화된 범죄자'를 사전에 검거하는 방법을 활용해볼 수 있겠다. 실제로 일반 사이버범죄 예방을 목적으로 진행된 연구 중 다수가 범죄자를 사전에 탐지하는 데 초점을 맞추고 있다. 특히 SNS 상에서 테러 단체를 탐지하는 여러 연구가 진행되었는데 트위터 상에서 텍스트 내용에 대한 정보를 사용하지 않고 Social graph feature를 이용하여 테러단체 추종자를 탐지하는 연구[5], Centrality measure와 fuzzy clustering을 통해 실시간으로 메시지를 분석하여 사용자의 영향을 추출하여 테러를 조직하는 지도자와 추종자를 구분하는 연구[6], 트위터 상에서 머신러닝 기법을 활용하여 ISIS 채용 목적의 게시글을 탐지하는 연구[7] 등이 발표되었다. 국내에서는 소셜네트워크 분석을 활용하여 사이버금융범죄 조직을 탐지하는 기법[8], 실무 관점에서의 사이버성범죄 가해자 검거를 위한 위장수사시스템[9] 등이 연구되었다.

이러한 접근 방식들은 디지털성범죄 예방의 관점에서 잠재적 범죄자를 기술적으로 정확하게 탐지하는 경우 큰 효과를 거둘 수 있다. 그러나 부정확한 탐지 즉, 1종 오류가 발생하였을 때, 잠재적 범죄자로 분류된 서비스 이용자는 표현의 자유와 사생활의 자유를 침해받는 것은 물론, 이용자의 특정 게시물을 근거로 위법한 개인정보 수집이 발생할 수도 있다. 결과적으로 이는 수사기관의 신뢰도를 저하시키는 결과를 야기하게 될 것이다. 또한 앞선 사례분석에서와 같이 가해자들은 익명성이 보장되는 서비스를 이용하고, 가짜 계정을 이용하거나 일대일 대화 기능만을 사용하는 등 사이버 공간에서 자신들의 흔적을 최소화하고자 노력하므로 범죄자의 특징을 파악할 수 있는 데이터 수집 자체가 매우 어렵다.

따라서 '범행에 적합한 대상물'을 감소시키는 접근이 대안으로 제시될 수 있다. 디지털성범죄 취약군을 식별하여 성범죄에 노출되어 있다는 것을 인지시킴으로써 취약군의 가치나 가시성, 접근성, 부동성을 변화시켜 범죄자의 입장에서 덜 매력적인 타겟으로 전환시키는 방안이다. 범죄 고위험군에게 범죄의 타겟이 될 수도 있다는 알람을 제공하여 범죄 발생 위험에 대해 환기시킴으로써 궁극적으로 취약군의 행동변경을 이끌어낼 수 있다[10]. 또한, 범죄자가 접근하거나 범죄가 발생했을 때의 대응 방안도 함께 공지하여 피해를 최소화하고 효율적인 증거 확보를 통해 신속히 가해자를 검거하는 데 도움이 될 수 있다.

특히, 취약군 선별은 잠재적 범죄자 탐지에 비해 기술 구현 측면에서도 유리하다. 취약군들은 주로 온라인 상에서 다수의 글과 사진, 영상을 공유하므로 학습데이터의 수집이 용이하다. 또한, 본 연구에서 제시하는 프레임워크는 취약군으로 판별된 사용자들에게 알람을 주는 목적으로 구현되므로 잠재적 범죄자 식별 시 발생할 수 있는 인권 침해 문제 혹은 1종 오류로 인한 억울한 피해자 발생의 문제로부터 상대적으로 자유롭다는 장점도 있다.

Ⅲ. 디지털성범죄 취약군 선별을 위한 프레임워크 제안

본 논문에서 제안하는 디지털성범죄 취약군 선별을 위한 프레임워크는 [그림 2]와 같다. 프레임워크는 취약군 모니터링 시스템(Digital Sexual Crime Monitoring System)과 데이터 수집 모듈(Data Collection Module), 분류기 생성 모듈(Classifier Generation Module), 분류기 업데이트 모듈(Classifier Update Module)로 구성되어 있다.

3.1. 데이터 수집 모듈

디지털성범죄가 일어났을 경우 피해자가 온라인 상에 게시한 콘텐츠를 수집한다. 트위터, 페이스북을 비롯한 주요 SNS에는 특정 주제를 쉽게 찾을 수 있도록 해시태그 기능을 제공하고 있으므로 이를 실제 작성 내용과 함께 획득한다. 콘텐츠에는 텍스트, 사진, 영상, 음성 등이 있으며 SNS 서비스의 특성에 따라 형태가 다르므로 이를 고려한 저장 체계, 즉 데이터베이스 스키마를 구축한다.

3.2. 분류기 생성 모듈

취약군 선별을 위한 최초의 분류기를 생성한다. 기존의 '유포형' 및 '제작형' 성범죄 피해자들의 콘텐츠와 일반 게시글들의 콘텐츠들로 데이터 세트를 구축한다. 분류기는 모니터링 시스템에서 SNS의 게시글을 분석하여 게시자가 디지털성범죄 취약군에 해당되는지 판단하는 목적으로 개발된다. 이진 분류 문제에 해당되므로 레이블은 취약군과 non-취약군으로 설정된다. 또한 콘텐츠 내의 연속된 데이터들을 분석 후 하나의 결론을 도출하기 때문에 Many-to-one 문제에 해당된다. 이러한 특성을 반영하여 순환신경망(Recurrent Neural Network, RNN)을 이용하여 분류기를 생성한다[11]. 순환신경망은 딥러닝 모델 중 하나로 순차성을 기억할 수 있는 메모리 셀을 바탕으로 연속된 데이터를 다룰 수 있다. 또한, 입출력 데이터를 시퀀스의 길이와 관계없이 나타낼 수 있다. SNS 게시글들은 텍스트들이 순차적으로 연속된 형태이므로 순환신경망 적용에 적합하다.

3.3. 분류기 업데이트

데이터 수집 모듈에서 수집된 새로운 피해자의 콘텐츠를 반영하여 기존 분류기를 업데이트한다. 새로 발생한 피해자의 트윗을 크롤링 할 때, 해시태그 부분을 따로 추출해서 기존의 해시태그 리스트에 없는 해시태그가 있으면 해당 해시태그를 추가하는 방법으로 해시태그 리스트를 업데이트한다. 이외에 수사관의 판단 하에 범죄자들의 타겟이 될 것으로 예상되는 해시태그도 추가함으로써 추가 콘텐츠를 수집한다. 이후 피해자나 취약군이 작성했던 글이나 사진, 영상 등이 기존 데이터 세트에 동일하게 존재하는지 체크하고 해당 콘텐츠를 학습 데이터 세트에 추가한다. 기존에 개발되었던 분류기를 전이학습하여 모델의 정확도를 향상시킨다. 해시태그 리스트와 모델이 업데이트되면 취약군 모니터링 시스템에 이를 전달한다.

3.4. 취약군 모니터링 시스템

온라인 상에서 지속적으로 디지털성범죄 취약군을 탐지한다. 데이터 수집 모델 및 분류기 업데이트 모듈에서 생성된 해시태그 리스트를 기반으로 게시글들을 일차적으로 필터링한 뒤 크롤링한다. 크롤링된 데이터를 생성된 분류기를 이용하여 취약군 여부를 판단한다. 취약군으로 판단되는 경우, 이용자에게 성범죄에 노출되어 있다는 사실을 인지시키고 범죄자가 접근하거나 범죄가 발생했을 때의 대응 방안을 사전에 전달한다.

Ⅳ. 취약군 선별 모델 구현

본 장에서는 앞서 제안한 프레임워크를 기반으로 취약군 선별 모델을 구현하고 검증한다. 실험을 위해 일상활동이론과 'n번방' 사건을 분석하여 범죄자 관점에서의 취약군의 기준을 설정한다. 설정된 기준에 적합한 트위터 글들을 크롤링하고 데이터 세트를 구축한 뒤 분류기를 구현한다.

4.1. 데이터 수집

'n번방' 사건의 가해자들은 주로 트위터에 선정적인 게시물을 올린 미성년자를 타겟으로 설정한 후 수사기관을 사칭하여 신상정보를 얻어내었다. 범죄자들이 미성년자를 범행 대상으로 설정한 이유는 미성년자는 성인에 비해 판단능력이 부족하고 사법제도와 절차에 관해 낮은 이해도를 가지고 있으므로 협박을 받았을 때 성인에 비해 적절한 대응을 하지 못할 것이라고 판단하였기 때문이다. 따라서 연구팀은 취약군의 기준을 선정적인 내용과 함께 개인정보가 포함된 게시물을 올린 사람과 선정적인 게시물을 올리는 미성년자로 설정하였다.

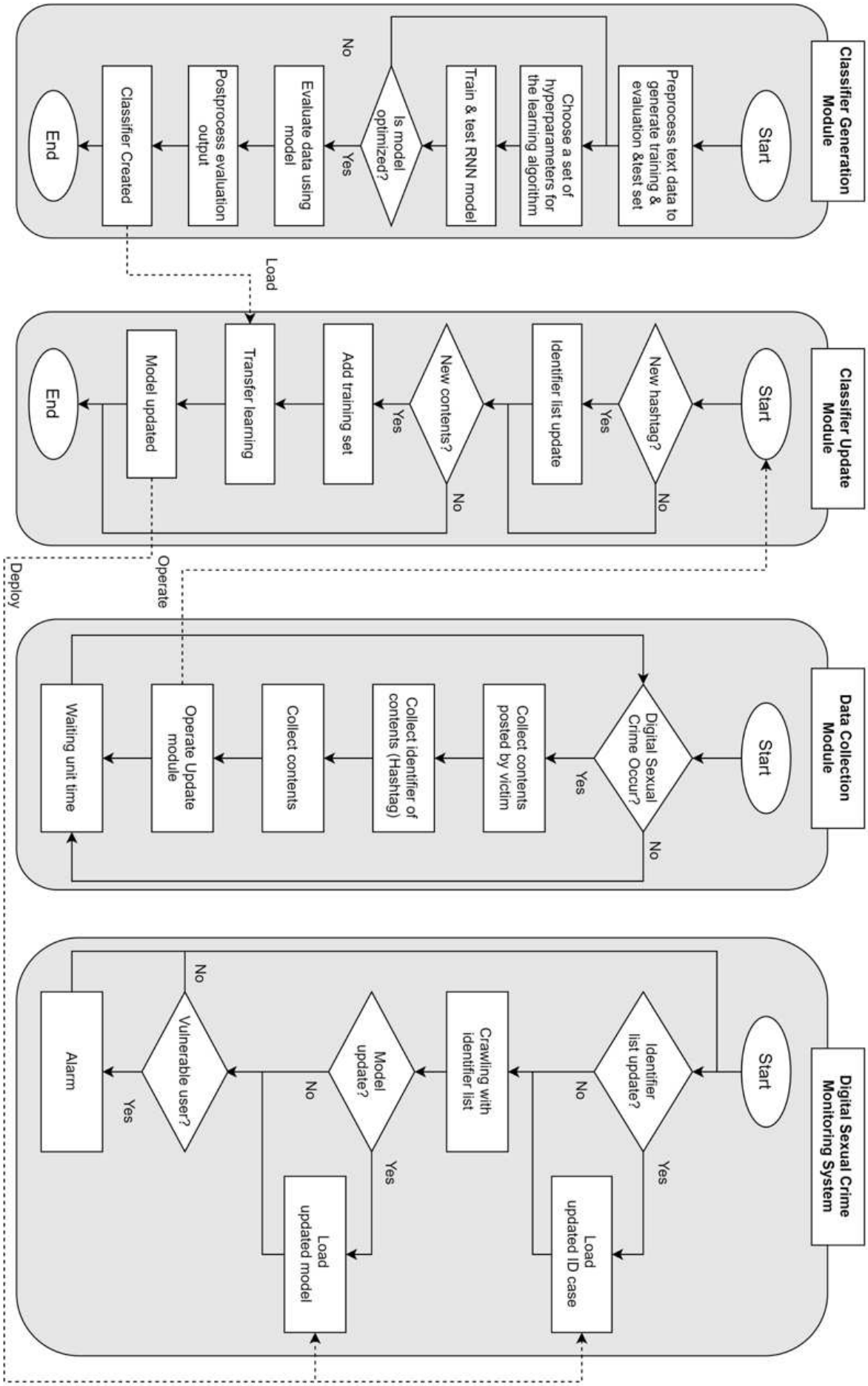
키워드 선정을 위해 트위터 게시글들을 분석한 결과 '일탈', '섹트', '오프', '조건만남', '맘눌뎀'과 같은 해시태그를 포함한 글들이 본 연구에서 설정한 취약군 기준에 부합한 내용을 담고 있음을 확인하였다. 해시태그를 포함하는 트윗들은 트위터에서 제공하는 API 플랫폼을 활용하여 크롤링하였다. 그 결과, 총 45,991개의 콘텐츠가 수집되었다. 앞서 제시한 5개의 키워드를 이용하면 상당 수의 취약군은 탐지할 수 있지만 모든 취약군을 탐지하기에는 한계가 있다. 따라서 5개의 키워드로 데이터 수집을 시작한 후 크롤링 하는 과정에서 취약군의 트윗에 포함된 해시태그를 따로 수집한 뒤, 수사관의 판단 하에 모델을 업데이트 하는 과정에서 조사 대상 해시태그 리스트를 추가하여 키워드를 확장하여 충분한 트윗을 확보할 수 있도록 하였다.

4.2. 데이터 세트 구축

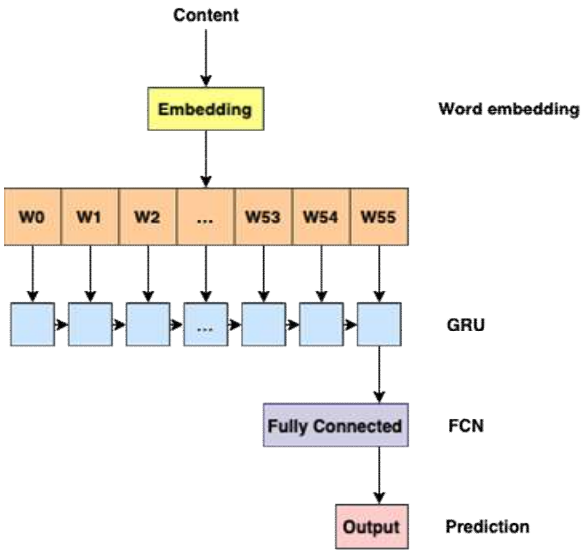
앞서 설정한 취약군 기준에 부합한 키워드를 포함한 게시물의 텍스트 데이터를 수집한 다음, 학습 데이터의 품질을 향상시키기 위해 중복 제거와 레이블링을 진행하였다. 레이블링은 수집된 데이터들을 연구팀이 직접 확인하여 구분하였으며 델파이 기법을 적용하여 구체적인 취약군 구분 방안을 설정하였다. 선정적 해시태그를 포함하였더라도 ① 광고성 게시글, ② 성적 취향과 관련된 글, ③ 오프라인 만남을 구하는 글이지만 특정 지역 또는 개인정보를 밝히지 않은 글, ④ 해당 키워드를 해시태그로 포함하고 있지만 일상적인 내용을 담은 글, ⑤ 리트윗을 한 글, ⑥ 게시물에 내용은 없고 해당 키워드를 포함한 해시태그만 가득한 글, ⑦ 오프라인 만남 후기와 같은 내용을 포함한다면 취약군이 아닌 것으로 레이블링하였다. 그 결과 수집한 45,991개 중 23,291개의 데이터가 취약군으로 레이블링되었다.

4.3. 딥러닝 기반 분류기

'3.2. 분류기 생성 모델'에서 설명했듯이 취약군 분류는 Many-to-one 이진 분류 문제에 해당하므로 RNN을 이용한다. RNN 계열 중 기존 Long Short-Term Memory (LSTM) 모델을 변형하여 연산량을 감소시킨 Gated Recurrent Units (GRU) 모델을 사용한다[12]. 크롤링된 트윗을 워드 임베딩을 통해 100차원 벡터로 변환하고 이를 GRU에 입력한다. 입력받은 벡터에 대해 학습하고 이를 완전 연결 계층(Fully Connected Layer, FC Layer)을 통과시킨 후 시그모이드(Sigmoid) 활성화 함수를 이용하여 어느 범주에 속하는지 최종 출력한다. 분류기 모델을 도식화한 것은 [그림 3]과 같다.



〈Figure 2〉 Digital sexual crime vulnerable group discrimination framework

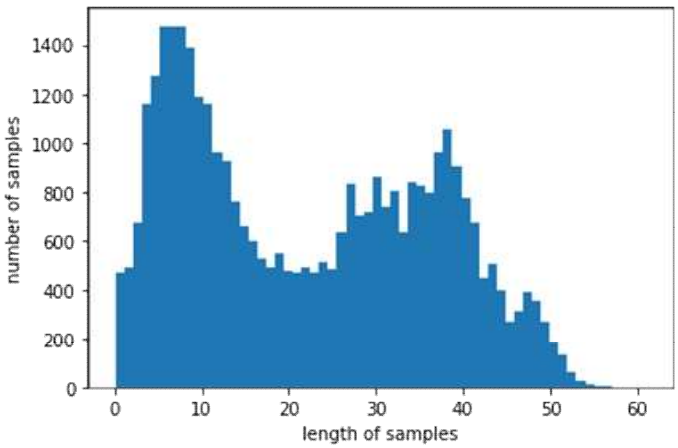


〈Figure 3〉 Classifier model

4.4. 실험 결과 및 분석

분류 모델 학습은 64GB RAM 용량 및 Nvidia의 RTX 2080Ti 그래픽카드 2대가 장착된 컴퓨터 환경에서 진행되었다. 모델 생성 및 학습은 Tensorflow를 이용하였다. 테스트 세트를 분류 모델에 입력할 수 있도록 토큰화, 패딩, 임베딩 등을 수행하였다. 토큰화는 Mecab 기반 한국어 형태소 분석기를 사용하였다.

모델 입력 시 학습데이터의 벡터 길이를 일정하게 설정하기 위하여 데이터 길이의 분포를 분석하였다. 평균 길이는 22.20이었으나 최대 길이는 61이었으며 〔그림 4〕와 같이 두 개의 가우시안이 혼합된 분포를 이루고 있다. 벡터 길이를 55로 설정하였을 때, 데이터의 99.97%가 온전하게 보존되므로 모든 학습 데이터의 길이를 55로 고정하고 패딩을 수행했다.

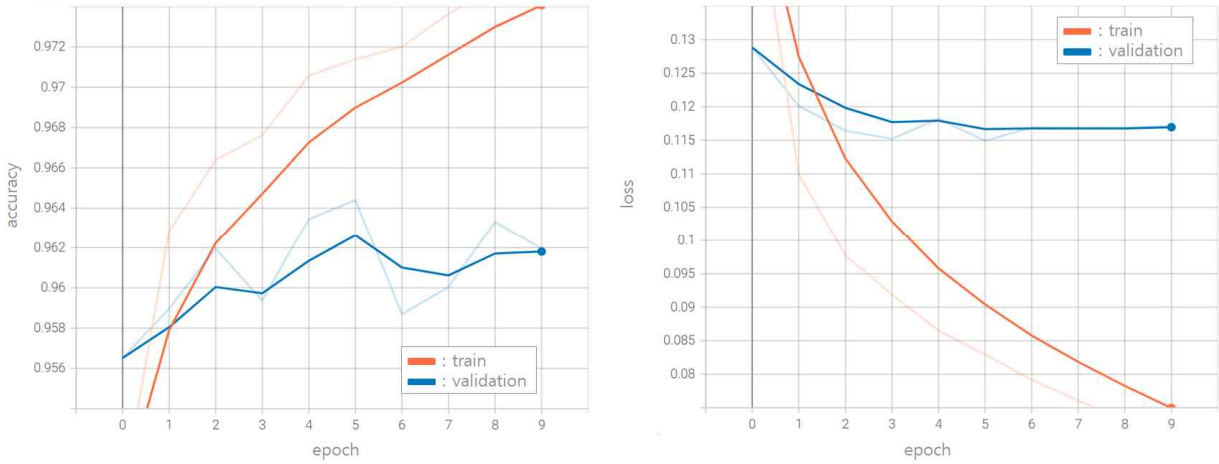


〈Figure 4〉 Word length distribution of dataset

학습 시 배치사이즈는 60으로 설정하였으며 학습 데이터 세트의 20%를 검증 데이터 세트로 설정하였다. 과적합을 회피하도록 조기종료(early stopping)를 validation loss 기준 epoch 4회로 설정하였다. 〔그림 5〕는 각 반복 횟수에 따른 Accuracy의 변화를 그래프로 나타낸 것이다. 〔그림 5〕 그래프에 있는 옅은 선은 실제 값을 나타내는 선이고, 짙은 선은 실제 값을 보기 쉽게 smooting 한 값이다. 조기종료로 인하여 epoch 9회에서 학습이 종료되었으며 6번째 학습 시 검증 데이터 세트에 대하여 가장 좋은 성능을 보였다. 테스트 세트에 대한 Accuracy는 96.52%로 측정되었다. 추가적으로 해당 모델의 정확도를 확인하기 위하여 Precision, Recall, F1-Score를 측정한 결과 Precision은 93.22%, Recall은 87.62%, F1-Score는 90.33%로 측정되었다. Recall 값이 87.62%로 상대적으로 낮은 수치를 보였는데 이는 취약군의 게시글 중 non-취약군 분류 기준으로 활용한 광고성 단어나 성적 취향 관련 단어가 포함된 게시글들이 False Negative를 발생시켜 나타난 현상으로 보인다.

임의의 글이 취약군이 작성한 글인지를 판단할 수 있는 프로그램을 학습된 모델을 기반으로 개발하였다. 해

당 프로그램은 분류기의 예측값이 임계치 0.5를 초과할 경우 취약군이 작성한 글로 판단하고, 그 확률을 함께 출력한다. 임계치는 1종 오류와 2종 오류 조절을 위해 변경 가능한 수치이다. [표 2]는 개발된 취약점 탐지 프로그램을 테스트한 결과이다.



〈Figure 5〉 Accuracy and loss trends according to the epoch

〈Table 2〉 The result of vulnerable group detection

테스트 데이터	판별 결과
출퇴근이 전부인 외로운 20대 직장입니다 ㅠㅠ	취약군 79.19%
안녕하세용 수원사는 25살 직딩이구요 즐기실 분 라인	취약군 96.01%
자영 판매해요 본인 영상 많구요 인증 원하시면 해드려요	취약군 91.84%
고딩 자영 판매합니다 문의는 라인으로 연락주세요	취약군 86.83%
오늘 조건 구해요 만나실 분 라인 개인 예약 환영	취약군 77.46%
자습할 거 많지만 너무 심심해	non-취약군 99.23%
맞팔해주세요	non-취약군 99.80%
잘생긴 중고딩 안경남에게 진심인 편	non-취약군 98.60%
나랑 여름방학 내내 놀러갈 사람	non-취약군 99.63%
이제 밤 9시에 통화를 못 한다는 사실이 이별을 깨닫게 만든다	non-취약군 99.64%

4.5. 논의

구현한 취약군 분류기는 실제 피해자의 데이터를 이용하여 학습한 것이 아니기에 곧바로 실무에 활용하기에는 제한이 있다. 그러나 동일한 해시태그를 사용한 선정성 높은 글들 중 취약군이 작성한 글을 높은 성능으로 분류할 수 있다는 것을 증명하였다. 이는 제안한 프레임워크에 실데이터를 적용할 경우 디지털성범죄 예방에 높은 기여를 할 가능성을 보여준다.

실험 결과를 분석하는 과정에서 취약군이 작성한 것으로 판단할 수 있는 게시글 중 해시태그에 데이터 구축 시 사용한 5개의 대표 키워드 외의 키워드가 발견되기도 하였다. 또한, 게시글에 해시태그 자체를 포함하지 않는 경우도 발견되었다. 이는 모델의 미탐에 해당되나, 해당 게시글들을 작성한 취약군 계정들이 등록한 타 트윗들은 취약군으로 분류되어 결론적으로 취약군 계정은 탐지 가능하였다. 따라서, 실제 피해자가 작성한 데이터로 모델을 구축할 때에는 단일 트윗 단위가 아닌 계정 단위의 취약군 탐지가 가능할 것으로 보인다.

본 연구에서는 텍스트 데이터만을 활용하였으나 크롤링된 트윗 중 일부는 텍스트 데이터 없이 사진과 영상으로만 게시된 경우도 있었다. 따라서 향후 멀티미디어 데이터에 대한 취약군 분류 방안 연구도 필요해 보인다. 다만, 선행 연구들[13, 14]에서 언급되었듯이 음란물 조사 시 선정적인 데이터를 직접 확인해야만 하는 조사자의 정신적 건강의 저해 가능성이 있으므로 레이블링 작업 시 사람의 개입을 최소화할 수 있는 방안에 대한 연구가 선행되어야 할 것이다.

Ⅴ. 결 론

본 논문은 대표적인 범죄피해 이론인 일상활동이론을 최근 사회적 공분을 일으키고 있는 디지털성범죄에 적용함으로써 범죄자 식별에 초점을 맞추는 기존 연구들과는 달리 취약군을 탐지하여 범죄를 예방하는 새로운 관점을 제시하였다. 또한, 실질적으로 취약군을 모니터링할 수 있는 딥러닝 기반의 디지털성범죄 취약군 선별 체계를 제시하였다. 취약군 모니터링 시스템, 데이터 수집 모듈, 분류기 생성 모듈, 분류기 업데이트 모듈로 구성된 프레임워크를 제안하고 이를 기반으로 취약군 선별 모델을 구현하였다. 데이터세트를 구축해 해당 모델의 효과를 검증한 결과, 테스트 세트에 대한 정확도는 96.52%로 측정되었다. 선별된 취약군에게 범죄의 타겟이 될 수 있다는 위험 및 성범죄 피해 발생 시의 대처방법에 대한 알람을 제공함으로써, 디지털 성범죄를 사전에 예방하고 피해 발생 시에도 피해를 최소화할 수 있을 것으로 기대된다.

한편, 데이터 수집 시 실제 피해자의 콘텐츠가 아닌 연구팀이 취약군으로 설정한 데이터로 테스트가 진행된 점, 멀티미디어 콘텐츠는 제외하고 텍스트에만 한정하여 실험을 수행하였다는 점에서 한계를 갖는다. 다만, 선정적인 데이터들로부터 취약군이 작성한 글들을 높은 정확도로 구분할 수 있다는 점에서 제안한 프레임워크에 실데이터를 학습시킨다면 실무적으로도 높은 활용도를 보일 것으로 예상된다.

향후 연구에서는 멀티미디어 데이터를 함께 사용하여 피해 위험성을 다방면으로 고려하고, 트위터 이외의 SNS도 함께 고려하여 각 SNS의 특성에 따른 이용자의 취약성 분석을 진행할 것이다. 또한, 디지털성범죄 수사관의 정신적 스트레스를 경감시킬 수 있도록 멀티미디어 데이터 레이블링을 자동화하는 방안에 대한 연구도 진행할 계획이다.

참 고 문 헌 (References)

[1] H. J. Kim, "A study on the actual conditions of digital sex crime policies in major countries and issues and direction of the sex crime policy in Korea: case studies of the U.S., Australia, Japan, and Germany" , Digital Convergence Research,18(8), pp.86, 2020.

[2] L. E. Cohen and M. Felson, "A routine activity approach," American sociological review, pp.588-608, 1979.

[3] Y. H. Lee, D. W. Kim, and Y. J. Yoo, "A Study on Routine Activity and Cyber-Crime Victimization in Cyberspace," Korean Journal of Public Safety and Criminal Justice Vol.20, No. 3, pp. 214-240, 2011.

[4] D. Jang and S. O. Kim, "A study on the punishment and control of online sexual violence crime," Seoul: Korean Institute of Criminology, 2018.

[5] M. Petrovskiy and M. Chikunov, "Online extremism discovering through social network structure analysis," in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT),* pp.243-249, 2019.

[6] C. Sánchez-Rebollo, C. Puente, C. Piriz, J. P. Fuentes, and J. Jarauta, "Detection of jihadism in social networks using big data techniques supported by graphs and fuzzy clustering," *Complexity*, 2019.

[7] M. Nouh, J. R. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on twitter," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp.98-103, 2019

[8] H. C. Kim and J. W. Yoon, "A Case of Cyber Financial Crime Investigation Through Social Network Analysis (2-Mode Concepts," *Journal of Digital Forensics,* Vol.14, No.4, pp.449-465, 2020.

[9] Y. H. Chae and H. K. Kim, "A Study on Undercover Investigation System Introducation Plan for Digital-sexual crime," *Journal of Digital Forensics,* Vol.14, No.4, pp.436-448, 2020.

[10] Y. S. Kim, "A Review of Risk Interpretation Model on the Relationship of Fear of Crime and Self-Protection," Korean Journal of Public Safety and Criminal Justice, Vol.25, No.1, pp.63-91, 2016.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling,". arXiv preprint arXiv:1412.3555. 2014.

[13] P. Q. Brady, "Crimes against caring: Exploring the risk of secondary traumatic stress, burnout, and compassion satisfaction among child exploitation investigators," Journal of Police and Criminal Psychology, Vol.32, No.4, pp.305-318, 2017.

[14] G. W. Burruss, T. J. Holt, and A. Wall-Parker, "The hazards of investigating internet crimes against children: Digital evidence handlers' experiences with vicarious trauma and coping behaviors," American Journal of Criminal Justice, Vol.43, No.3, pp.433-447, 2018.

저 자 소 개

유 혜 진 (Hyejin Ryu)
준회원

2019년 3월~현재 : 동국대학교 경찰사법대학 경찰행정학부
2022년 3월~현재 : 동국대학교 일반대학원 경찰행정학과 학석사과정
관심분야 : 디지털 포렌식, 딥러닝, 블록체인 등

박 아 현 (Ah-Hyun Park)
준회원

2022년 2월 : 동국대학교 경찰사법대학 경찰행정학부 졸업
2022년 3월~현재 : 동국대학교 일반대학원 경찰행정학과 석사과정
관심분야 : 디지털 포렌식, 딥러닝, 암호화폐 등

박 윤 지 (Yunji Park)
준회원

2019년 3월~현재 : 동국대학교 경찰사법대학 경찰행정학부
2022년 3월~현재 : 동국대학교 일반대학원 경찰행정학과 학석사과정
관심분야 : 디지털 포렌식, 모바일 포렌식, 정보보호 등

이 지 원 (Jiwon Lee)
준회원

2019년 3월~현재 : 동국대학교 경찰사법대학 경찰행정학부
2022년 3월~현재 : 동국대학교 일반대학원 경찰행정학과 학석사과정
관심분야 : 디지털 포렌식, 모바일 포렌식, 딥러닝, 침해사고대응 등

주 다 빈 (Dabin Joo)
준회원

2018년 3월~현재 : 동국대학교 경찰사법대학 경찰행정학부
2022년 3월~현재 : 동국대학교 일반대학원 경찰행정학과 학석사과정
관심분야 : 디지털 포렌식, 정보보호 등

최 혜 인 (Hyein Choi)
준회원

2021년 2월 : 동국대학교 경찰사법대학 경찰행정학부 졸업
2021년 4월~현재 : 한국산업기술보호협회
관심분야 : 사이버 보안, 산업보안컨설팅 등

정 두 원 (Doowon Jeong)
정회원

2019년 2월 : 고려대학교 정보보호대학원 공학박사
2020년 9월~현재 : 동국대학교 경찰사법대학 조교수
2022년 1월~현재 : 동국대학교 융합안전학술원 사이버안전연구센터 센터장
관심분야 : 디지털 포렌식, 정보보호 등