오염 공격에 대한 SEIR 전염병 모델을 활용한 AI 스크래퍼 악용 시나리오 분석

권순홍[1], 손우영[2], 이종혁[3]
[1]세종대학교 정보보호학과 & 지능형드론 융합전공 박사과정
[2]세종대학교 프로토콜공학연구실 학부과정
[3]세종대학교 정보보호학과 & 지능형드론 융합전공 교수
soonhong@pel.sejong.ac.kr, wooyoung@pel.sejong.ac.kr, jonghyouk@sejong.ac.kr

Analysis of AI Scrapers Exploitation Scenarios via the SEIR Epidemic Model for Poisoning Attacks

Soonhong Kwon[1], Wooyoung Son[2], Jong-Hyouk Lee[3]
[1]Dept. of Computer and Information Security & Convergence Engineering for Intelligent Drone, Sejong University
[2]Protocol Engineering Lab., Sejong University
[3]Dept. of Computer and Information Security & Convergence Engineering for Intelligent Drone, Sejong University

요 약

AI(Artificial Intelligence) 시대가 도래됨에 따라 범국가적인 형태로 AI 산업 시장에서 기술 패권을 선도하고자 하는 노력이 이루어지고 있다. 해당 분야에서 가장 중요시 되고 있는 것은 데이터이다. 양질의 데이터를 보유하고 있다는 것은 AI 기술을 활용하여 양질의 데이터를 생성할 수 있음을 의미한다. 이에 따라 AI 기술에 기반을 두고 있는 기업들은 자동으로 데이터를 크롤링하기 위하여 AI 스크래퍼를 인터넷 상에 배포하고 있다. 하지만, AI 스크래퍼를 통해 데이터를 크롤링할 시 발생하는 트래픽으로 인하여 시스템 중단이 초래될 수 있을 뿐만 아니라 검증되지 않은 데이터가 크롤링되어 학습이 이루어지는 경우, 생성형 AI 모델 붕괴를 초래함으로써 성능 저하로 이어질 수 있다. 이에 본 논문에서는 인터넷 상에 무분별하게 배포되어 있는 AI 스크래퍼가 오염 공격에 노출되었을 경우, 미칠 수 있는 파급력을 SEIR 전염병 모델을 통해 정량적으로 분석하며, 이를 통해 대규모 시스템에서 단일 감염으로 인한 확산이 큰 파급효과를 불러일으킬 수 있음을 보인다.

1. 서론

4차 산업혁명 이후, 빅데이터 시대를 거쳐 현재 우리는 AI(Artificial Intelligence) 중심의 시대에 살고 있다. AI 기술은 어느 순간 등장한 기술이 아닌 지속적인 발전을 이루어 왔으며, 이는 앨런 튜링의 튜링 테스트에 의해 처음 정립된 개념에서 비롯되었다. 이후, 수십년의 발전에 거쳐 1980년대에는 사람이 설정한 규칙에 의거하여 자동으로 판정을 내리는 시스템이 도입되면서 2차 AI 붐을 맞이하게 되었으며, 2010년대에는 딥러닝의 성장, 2020년대에는 생성형 AI의 혁명으로 3차 AI 붐이 여전히 진행되고 있는 상황이다 [1].

이에 따라 범국가적인 형태로 AI 산업 시장에서 기술 패권을 선도하고자 하는 노력이 이어지고 있다. 해당 분야에서 가장 중요시 되고 있는 것은 데이터이다. 양질의 데이터를 보유하고 있다는 것은 AI 기술을 활용하여 양질의 데이터를 생성할 수 있음을 의미하기 때문이다. 이에 따라 AI 기술에 기반을 두고 있는 기업들은 자동으로 데이터를 크롤링하기 위하여 AI 스크래퍼를 인터넷 상에 배포하고 있다.

하지만, 이와 같은 상황은 여러 가지 문제를 야기하고 있다. 첫 번째로는 AI 스크래퍼를 통해 데이터를 크롤링할 시 발생하는 트래픽으로 인하여 시스템 중단을 야기할 수 있다. 하나의 예로 생성형 AI를 통해 이미지를 생성하는 서비스를 제공하는 Midjourney에서는 타사 AI 스크래퍼의 방대한 데이터 크롤링으로 인해 시스템 중단이 발생하여 AI 스크래퍼 차단 정책을 수립한 경우가 있다 [2]. 두 번째로는 AI 스크래퍼를 통해 수집되는 데이터가 생성형 AI를 통해 생성된 검증되지 않은 데이터인 경우, 해당 데이터를 기반으로 한 학습이 이루어졌을 때, 양질의 데이터를 생성할 수 없다는 문제가 있다. 해당 문제는 생성형 AI의 모델 붕괴를 초래하며, 이로 인해 성능이 점차 저하될 수 있다 [3]. 현재 직면하고 있는 문제들에 대해 지속적으로 우려의 목소리가 높아지고 있으나, 빅테크 기업에서는 AI slop과 같은 데이터도 모델의 성능이 좋아지는 경우, 해결될 수 있는 문제로 고려하고 있다.

이에 본 논문에서는 인터넷 상에 무분별하게 배포되어 있는 AI 스크래퍼가 오염 공격에 노출되었을 경우의 파급력을 분석하고자 SEIR(Susceptible-Exposed-Infected-Recovered) 전염병 모델을 통한 정량적 분석을 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 SEIR 전염병 모델에 대해 설명하며, 이를 기반으로 한 AI 스크랩퍼 오염 공격의 전파 구조를 제시한다. 3장에서는 SEIR 전염병 모델을 기반으로 AI 스크래퍼가 오염 공격에 노출되었을 경우의 시나리오를 제시함으로써 AI 스크래퍼 감염 확산에 대한 정량적 분석을 수행한다. 4장에서는 본 논문의 결론을 맺는다.

2. SEIR 전염병 모델 기반 AI 스크래퍼 오염 공격 전파 구조

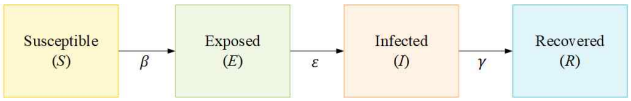
전염병 모델은 용어 자체에서 유추할 수 있듯이 질병 확산 모델링에서 활용되고 있는 모델이다. 감염병 모델링의 이점으로는 적은 양의 데이터로 모형을 구성하기 용이하고, 예측 가능한 시나리오에 대해 다양하게 제시할 수 있다는 점이다. 전염병 모델은 Susceptible-Infectious 구조에서부터 Susceptible-Exposed-Infectious-Recovered-Susceptible 구조까지 다양한 구획 모델로 표현될 수 있다 [4].

다양한 구획 모델 중에서 본 논문에서는 AI 스크래퍼가 오염 공격에 노출되었을 경우를 묘사하기 위해 SEIR 모델을 활용하여 표현하고자 하였다. AI 스크래퍼는 자동화된 시스템으로 구성되어 있으며, 하나의 스크래퍼가 감염되어 AI slop 데이터를 생성하는 경우, 해당 데이터가 다른 AI 스크래퍼에 전이되어 영향을 미칠 수 있기 때문이다. SEIR 전염병 모델 기반 AI 스크래퍼 오염 공격 전파 구조에 대해 설명하기 전, 기존 의학적 의미를 기반으로 동작하는 SEIR 모델의 각 상태에 대해 AI 스크래퍼 오염 확산을 기준으로 <표 1>과 같이 정의하였다.

<표 1> SEIR 전염병 모델의 각 상태 정보

상태	의학적 의미	AI 스크래퍼 기준 의미
S	감염된 상태는 아니지만, 감염될 수 있는 상태	보안 인증을 수행하지 않는 AI 스크래퍼 시스템
E	감염원에 노출되었지만, 감염이 되지 않은 상태	프롬프트 인젝션 공격을 당했으나, 아직 실행이 되지 않은 상태
I	감염이 이루어져 타인에게 전파될 수 있는 상태	오염된 프롬프트를 실행함으로써 악성 응답을 생성하고 전파 가능한 상태
R	감염에서 회복되어 전염성이 없는 상태	오염을 인지하고, 사후 처리를 통해 더 이상 감염되지 않은 상태

<표 1>을 기반으로 도식화하여 보면 SEIR 전염병 모델 기반 AI 스크래퍼 오염 공격 전파 구조는 (그림 1)과 같은 구조를 가진다.



(그림 1) SEIR 모델 플로우 차트

SEIR은 연립 미분방정식을 기반으로 설명이 될 수 있으며, β는 감염율을 의미한다. 감염된 노드를 I라고 했을 때, I를 통해 도출될 수 있는 S에 대한 오염 공격 전파될 확률을 의미한다. 이에 따라 수식 (1)에서 설명되고 있는 βSI는 단위 시간 당 노출 상태인 E 상태로 전이되는 스크래퍼의 수를 의미한다.

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

수식 (2)는 E에 대한 수식으로 <표 1>에서 정의한 바와 같이 프롬프트 인젝션 공격을 당했으나, 아직 실행이 되지 않은 상태를 의미한다. (그림 1)을 통해 확인할 수 있듯이 E에서 I로 상태가 전환되는 상황에서 ε은 감염 전이율로써, 이미 오염되어 있는 프롬프트가 실행됨으로써 악성 응답을 생성할 확률을 의미한다.

$$\frac{dE}{dt} = \beta SI - \varepsilon E \tag{2}$$

수식 (3)은 I에 대한 수식으로 <표 1>에서 정의한 바와 같이 오염된 프롬프트를 실행함으로써 악성 응답을 생성하고 전파 가능한 상태를 의미한다. (그림 1)을 통해 확인할 수 있듯이 I에서 R로 상태가 전환되는 상황에서 유추할 수 있듯이 γ는 회복률을 의미한다. γ는 I에서 R로 전환되는 비율, 즉 보안 조치에 의해 회복된 확률을 의미한다.

$$\frac{dI}{dt} = \varepsilon E - \gamma I \tag{3}$$

수식 (4)는 R에 대한 수식으로 <표 1>에서 정의한 바와 같이 오염을 인지하고, 사후 처리를 통해 더 이상 감염이 되지 않은 상태를 의미한다. 즉, 이는 감염이 되었지만, 보안 조치를 통해 더 이상 감염이 이루어지지 않은 상태가 되고, 이를 기반으로

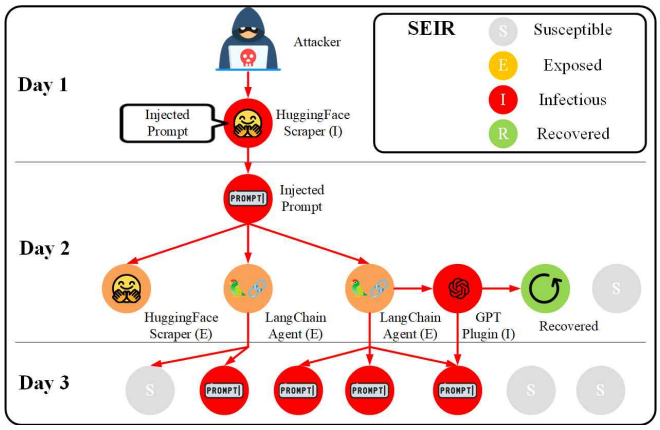
다른 AI 스크래퍼에 감염을 확산시키지 않은 스크래퍼를 의미한다.

$$\frac{dR}{dt} = \gamma I \quad (4)$$

이처럼 SEIR 전염병 모델 기반 AI 스크래퍼 오염 공격 전파 구조를 통해 확인할 수 있듯이 AI 스크래퍼는 총 4가지의 상태를 가질 수 있음을 확인할 수 있으며, 이는 시간의 흐름에 따라 순차적으로 변화함을 알 수 있다.

본 논문에서 제시한 모델을 통해 인증되지 않은 AI 스크래퍼에 오염된 프롬프트가 전달되고 해당 AI 스크래퍼에서 이를 실행함으로써 모델 학습에 오염된 데이터가 활용됨을 알 수 있다. 이는 AI slop 데이터를 생성할 수 있도록 하며, 이렇게 생성된 데이터는 다른 AI 스크래퍼에 의해 수집되어 전파될 수 있음을 보이는데 의의가 있다.

3. SEIR 전염병 모델 기반 오염 공격 시나리오별 AI 스크래퍼 감염 확산 분석



(그림 2) 오염 공격에 노출된 AI 스크래퍼 감염 확산

앞서, 2장에서는 SEIR 전염병 모델을 기반으로 하여 AI 스크래퍼가 오염 공격에 노출되었을 시, 전파 구조에 대해 살펴보았다. 본 장에서는 앞서 정의한 SEIR 모델을 기반으로 하여 AI 스크래퍼가 오염 공격에 노출된 시나리오를 구성하고 이로 인한 파급효과에 설명하기 위한 정량적인 평가를 수행한다.

정량적 평가를 수행함에 앞서, 오염 공격에 노출된 AI 스크래퍼에 의해 시간이 지남에 따라 감염이 확산되는 시나리오를 (그림 2)와 같이 표현하였다. 이를 시간에 따라 감염 상태, 노출 사태, 감염 전파 상태, 회복 상태로 요약하면 <표 2>와 같이 정리할 수 있다.

<표 2> 시나리오에 따른 SEIR 상태 요약

Day	감염 상태	노출 상태	감염 전파 대상	회복 상태
1	HuggingFace Scraper	-	Langchain Agents / GPT Plugin	-
2	HuggingFace Scraper / GPT Plugin	HuggingFace Scraper / LangChain Agents	Susceptible 노드 3개 (Day 3)	-
3	Prompt 노드 4개	-	Susceptible 노드 다수 (이어짐)	GPT Plugin

<표 2>의 내용을 기반으로 앞서 2장에서 설명한 주요 파라미터인 β, ε, γ를 활용하여 일반적인 환경, 고위험 환경, 보안 조치 적용 환경 시나리오에 따른 시뮬레이션을 수행한다. 이를 위해 주요 파라미터를 기반으로 한 시뮬레이션 시나리오를 <표 3>과 같이 정리하였다.

<표 3> 시나리오 케이스 별 주요 파라미터 설정

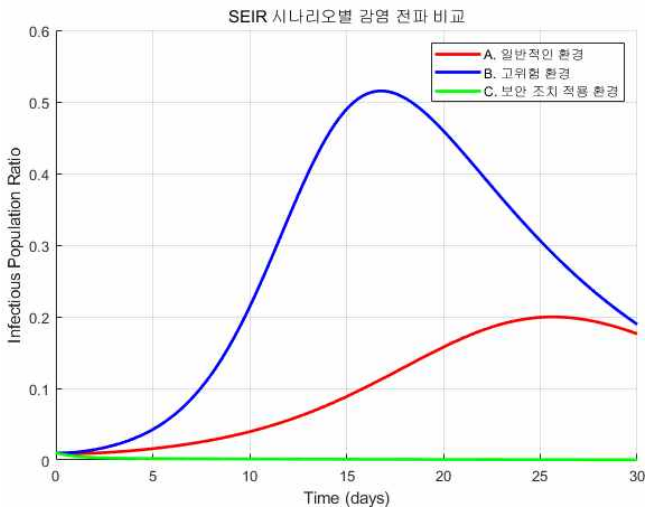
Case	β	ε	γ
일반적인 환경	0.6	0.4	0.2
고위험 환경	0.8	0.6	0.1
보안 조치 적용 환경	0.5	0.3	0.7

일반적인 환경은 우리가 일반적으로 알고 있는 환경으로 감염과 회복이 혼합적으로 일어나는 환경을 의미하며, 고위험 환경은 감염률이 높은 동시에 감염 전이율 역시 높은 환경, 보안 조치 적용 환경은 감염률과 감염 전이율이 회복률에 비해 상대적으로 낮은 환경을 의미한다.

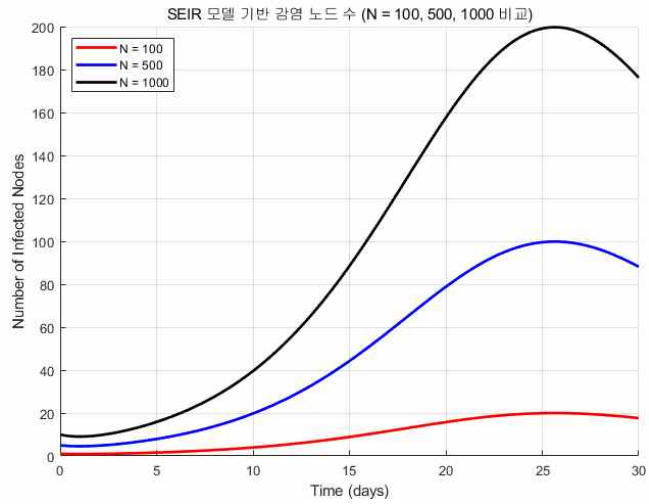
이에 따라 <표 3>의 데이터를 기반으로 하여 수식 (1), 수식 (2), 수식 (3), 수식 (4)에 대입하면 (그림 3)과 같은 결과를 확인할 수 있다. (그림 3)을 통해 확인할 수 있듯이 고위험 환경의 경우, 16일부터 최대 감염률이 50% 이상 도달한 것을 확인할 수 있으며, 일반적인 환경에서는 25일 정도에 최대 감염률 20%를 달성하고, 평균적인 감염 확산이 이루어지는 것을 보여준다. 보안 조치가 적용된 환경의 경우에는 회복률이 높음에 따라 감염 확산이 거의 일어나지 않는 것을 확인할 수 있다.

또한, AI 스크래퍼 수가 기하급수적으로 증가하는 현황임에 따라 AI 스크래퍼 수가 증가할수록 감

염 확산에 어떠한 영향을 미치는지에 대해 분석이 이루어져야 한다. 이에 따라 <표 2>의 주요 파라미터를 기반으로 AI 스크래퍼가 100개, 500개, 1,000개일 경우의 최대 감염 노드 수 및 발생 시점에 대해 분석을 진행하였다. 이는 (그림 3)의 그래프에 AI 스크래퍼 수를 곱함으로써 분석을 수행할 수 있다.



(그림 3) SEIR 시나리오 별 감염 전파 비교



(그림 4) SEIR 모델 기반 감염 노드 수

(그림 3)의 파라미터를 기반으로 하였음에 따라 감염률은 동일하지만, AI 스크래퍼 수가 증가할수록 감염자 수 역시 증가하는 것을 (그림 4)를 통해 확인할 수 있다. 또한, 25일이 경과한 시점에서 최대 감염자 수를 기록하는 것을 확인할 수 있으며, 100명일 때보다 500명일 때 5배의 피해를 발생시킬 수 있음을 보이고 있으며, 100명일 때보다 1,000명일 때 10배의 피해를 발생시킬 수 있음을 확인할 수 있다. 즉, 이는 대규모 시스템에서 단일 감염이 발생할 때 파급 효과가 큰 것을 확인할 수 있다. 이에 따라 데이터 수집을 위해 무분별하게 AI 스크래퍼를 인터넷 상에 배포하는 것은 빅테크 기업에서 생각하고 있는 것보다 심각한 파급 효과를 불러일으킬 수 있음에 따라 대규모 시스템 환경에서 이를 동적으로 보안할 수 있는 정책 및 기술이 모색되어야 할 필요가 있다.

4. 결론

4차 산업혁명 시대 이후, 빅데이터 시대를 거쳐 우리는 AI 중심의 시대에서 살아가고 있다. 이에 따라 범국가적으로 AI 산업 시장에서 기술 패권을 선도하고자 하는 노력이 주를 이루고 있으며, 방대한 양의 데이터를 보유하는데 집중하고 있다. 이를 위해 빅테크 기업에서는 AI 스크래퍼를 인터넷 상에 배포하여 자동으로 데이터를 크롤링하고 있다.

하지만, 이와 같은 상황은 AI 기반의 서비스를 제공하고 있는 기업의 시스템 중단을 초래하거나 모델 붕괴로 인한 성능 저하를 야기할 수 있다. 하지만, 빅테크 기업에서는 이에 대한 심각성을 느끼지 못하고 데이터 수집을 위한 AI 스크래퍼 배포에만 집중하고 있다. 이에 본 논문에서 AI 스크래퍼가 오염 공격이 노출되었을 경우에 대해 SEIR 전염병 모델을 기반으로 정량적 평가를 수행함으로써 파급력을 분석하였으며, 대규모 시스템 즉, AI 스크래퍼가 인터넷 상에 다수 배포될수록 단일 감염으로 인한 파급 효과가 클 수 있음을 보였다. 향후 연구로는 대규모 시스템 상에서 이를 동적으로 보안하여 회복률을 높일 수 있는 정책 기술에 대한 연구를 수행하고자 한다.

Acknowledgment

이 논문은 2024년도 정부(개인정보보호위원회)의 재원으로 한국인터넷진흥원의 지원을 받아 수행된 연구임 (2780000004, 블록체인 환경에서의 개인정보보호 표준개발).

참고문헌

[1] “AI의 시작과 발전 과정, 미래 전망”, [Online]. Available: https://news.skhynix.co.kr/all-around-ai-1/ [Accessed: 2025-04-03].

[2] “Midjourney accuses Stability AI employees of scraping its data, bans them from using the AI tool”, [Online]. Available: https://www.indiatoday.in/technology/news/story/midjourney-accuses-stability-ai-employees-of-scraping-its-data-bans-them-from-using-the-ai-tool-2513845-2024-03-12?utm_source=chatgpt.com [Accessed: 2025-04-03].

[3] SHUMAILOV, Ilia, et al. “AI models collapse when trained on recursively generated data”. *Nature*, 2024, 631.8022: 755-759.

[4] “Extending the Basic SIR Model”, [Online]. Available: https://towardsdatascience.com/extending-the-basic-sir-model-b6b32b833d76/?gi=14b874bac287 [Accessed: 2025-04-03].