SEO 점수 및 메타데이터를 활용한 악성 URL 다중 분류 모델 구축 연구

김동혁[1‡], 우에노고홍[1‡], 허은채[1], 한지수[1], 이성철[2], 이정빈[2]
[1]선문대학교 컴퓨터공학부 학부생
[2]선문대학교 컴퓨터공학부 교수
{rlaehdgur30, takaaaaaan, gjdmsco27, gkswltn38, sungchul, jungbini}@sunmoon.ac.kr

Development of a Multi-Class Malicious URL Classification Model Using SEO Scores and Metadata

Dong-hyeok Kim[1‡], Gohong Ueno[1‡], Eun-chae Heo[1], Ji-soo Han[1],
Sungchul Lee[2], Jung-Been Lee[2]
[1,2]Div. of Computer Science and Engineering, Sun Moon University

요 약

SEO 점수, Metadata, domain_age, 웹 성능 지표를 활용하여 악성 URL을 정밀하게 분류하는 다중 클래스 모델을 제안했다. LightGBM과 Optuna 최적화를 적용해 기존 대비 F1-score를 81%에서 92%로 향상시켰다. domain_age와 웹 성능 지표가 탐지 성능 향상에 핵심적으로 기여하였으며, 실시간 보안 솔루션 적용 가능성을 확인했다.

1. 연구배경

현대의 인터넷 환경에서는 피싱(Phishing), 멀웨어(Malware), SEO(Search Engine Optimization) 스팸 등 다양한 유형의 악성 URL이 지속적으로 증가하고 있으며, 이러한 위협은 빠르게 진화하고 있다 [1,2]. 단순한 문자열 기반 탐지 방식[3,4]은 주로 URL의 길이, 도메인 패턴, 특수 문자 사용 여부 등에 집중해 악성 여부를 탐지하고, 웹사이트 성능 지표나 Metadata를 고려하지 않아 탐지 정확도가 낮고 정상 웹사이트로 위장된 SEO 스팸이나 새롭게 생성된 도메인을 탐지하는데 한계가 있다. 이에 따라 사용자 보호와 신속한 대응을 위해 정밀하게 분류할 수 있는 모델이 필요하다.

따라서 본 연구는 정상, 피싱, 멀웨어와 같은 다중 클래스 분류 모델을 구축하기 위해 URL 구조 기반 feature와 함께 SEO score, domain age, Metadata 존재 여부, 웹 성능 지표를 주요 변수로 설정하였다. 이러한 데이터를 크롤링과 Google PageSpeed Insight API를 활용해 수집하고 이상치 제거를 위해 log 변환 등의 전처리를 진행하였다. 이후 LightGBM 기반의 앙상블 모델을 적용하여 보다 정밀하게 악성 URL 분류를 수행하였다.

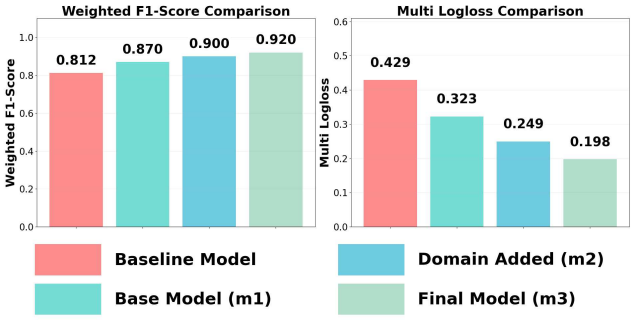
2. SEO 및 메타데이터 기반 URL 분류시스템

본 논문에서는 웹페이지의 <meta>, <og: title>, <title> 태그 존재 여부, 웹 성능 지표(예: PageSpeed API), domain_age 등을 기반으로 머신러닝 모델을 학습시켜 정밀한 다중 클래스 분류를 수행하였다.

Python의 requests, dotenv 라이브러리를 활용해 총 10만건의 URL에 대하여 PageSpeed API에서 웹 성능 데이터를 병렬로 수집하였고, 모델 학습의 안정성을 확보하기 위해 결측치 제거, 로그 변환, 이상치 필터링을 통해 데이터를 정제하였다. 분류 모델은 LightGBM 기반의 앙상블 모델을 사용하였으며, 하이퍼파라미터 최적화는 Optuna로 수행하였다. 웹페이지의 <head> 영역에 포함된 Metadata 존재 여부로 이진 변수(예: has_title)를 생성하고, 도메인 신뢰도를 반영하기 위해 domain_age를 주요 변수로 추가하였다.

기존 연구 모델은 URL 문자열 기반 특징(길이, 특수문자 빈도, 서브도메인 구조 등)을 XGBoost 분류기에 적용하여 악성 여부를 판별하는 방식이었다 [3]. 이 모델과 성능을 비교 평가한 결과, 기존 모델의 F1-score는 81%, 본 연구의 기본 모델(m1)에서의 F1-score는 87%이었다. domain_age를 추가한 모델(m2)의 F1-score는 90%, 로그 변환 및 하이퍼파라미터 튜닝을 적용한 최종 모델(m3)의 F1-score는 92%까지 향상되었다. 이는 domain_age 변수가 피싱·악성 URL의 짧은 주기를 효과적으로 포착했기 때문이며, 로그 변환된 웹 성능 지표들이 극단값의 영향을 완화하면서 모델의 안정성을 높인 결과로 해석된다.

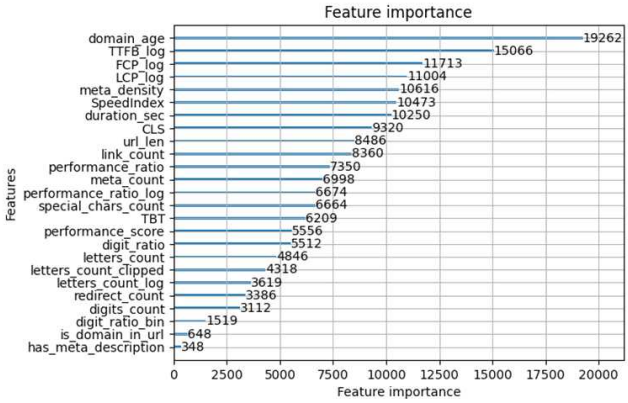
또한 그림 1과 같이 기존 연구 대비 본 연구는 multi-logloss 0.2329 만큼 감소시켰는데, 이는 모델의 예측 확률 분포가 더욱 정밀해졌음을 의미한다.



(그림 1) 모델 성능 비교 결과

3. 악성 URL분류 모델 결과 해석

그림 2에서 보는 바와 같이 도메인 연령과 SEO 점수 기반 변수는 기존 문자열 기반 탐지 방식 대비 높은 분류 성능을 보였으며, domain_age는 웹사이트의 신뢰성, 권위, SEO 및 안정성에 크게 기여하여 모델 예측에 가장 중요한 특성임이 확인되었다.



(그림 2) 악성 URL 분류 모델의 특징 중요도

이는 피싱·악성 URL이 일반적으로 짧은 기간 동안 운영되며, 이에 따라 도메인 생성 연도가 최근일수록 악성일 가능성이 높다는 것을 의미한다. TTFB_log, FCP_log, LCP_log와 같은 웹 페이지 로딩 속도 지표는 웹사이트의 사용자 경험, SEO, 최적화 수준을 반영하며, 악성 URL이 정상 URL 대비 로딩 속도가 비정상적으로 느리거나 특정 패턴을 보이는 특성을 포착해 높은 예측 기여도를 보였다.

4. 결론

기존 악성 URL 탐지 기술은 문자열 기반 분석에 의존해 왔으나, 본 연구에서는 SEO 점수, Metadata, domain_age 등 외부 API 기반 고차원 정보를 활용한 악성 URL 다중 클래스 분류 모델을 구축하여 차별화된 분류 성능을 입증하였다. 또한 단순한 특성 분석을 넘어, 실제 웹페이지의 성능 지표를 반영하여 실서비스에서 활용할 수 있는 실용성을 확보하였다. 최종적으로 F1-score 0.92의 성능을 보이는 악성 URL 분류 모델을 구축하였다.

이 시스템은 URL 필터링, 기업용 웹 보안 솔루션, 브라우저 확장 프로그램 등 다양한 보안 환경에 적용할 수 있으며, 특히 실시간 URL 분석이 필요한 백신 엔진이나 악성 코드 탐지 서비스에서 높은 활용도를 기대할 수 있다.

향후 연구에서는 URL 구조의 의미론적 분석, 텍스트 기반 피싱 요소 추출 등 심층 특성 엔지니어링을 통해 성능을 더욱 향상 시킬 예정이다. 또한 오분류 사례를 바탕으로 설명할 수 있는 AI 기법을 도입하는 방향도 고려하고 있다.

감사의 글

이 논문은 2025년도 과학기술정통신부 및 정보통신기획평가원의 SW중심대학지원사업(2024-0-00023) 및 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2023-00243114).

참고문헌

[1] 한 채림, 윤수현, 한명진, 이일구. “머신러닝 기반 악성 URL 탐지 기법”. 한국정보보호학회. 32. 3. 555-566. 2022.

[2] M. Shimamura, S. Matsugaya, K. Sakai, K. Takeshige and M. Hashimoto, “An Analysis of the Relationship Between Black-Hat SEO Malware Families Leveraging Information from Redirected Fake E-Commerce Scam Sites.”, IEEE Conference on Dependable and Secure Computing (DSC 2024). Tokyo, Japan. 06-08 November 2024.

[3] 김영준, 이재우. “URL 주요특징을 고려한 악성 URL 머신러닝 탐지모델 개발.” 한국정보통신학회논문지, 26. 12. 1786-1793. 2022.

[4] 장민해, 송재주, 김명수., “A Study on the Detection Method for Malicious URLs Based on a Number of Search Results Matching the Internet Search Engines Combining the Machine Learning.”, Journal of electrical engineering & technology, 17. 1. 617-626. 2022.

[5] Google Developers, “PageSpeed Insights API”, https://developers.google.com/speed/docs/insights/v5/about