스팸 메시지에서 불법 도박 링크 자동 추출을 위한 파이프라인 설계

김건우* 천정현* 민무홍**

*성균관대학교 컴퓨터교육과 (학부생)

**성균관대학교 컴퓨터교육과 (조교수)

A Pipeline for Automatically Extracting Illegal Gambling Links from Spam Messages

Geon Woo Kim* Jung Hyun Chun* Moohong Min**

*Department of Computer Education, Sungkyunkwan University (Undergraduate student)

**Department of Computer Education, Sungkyunkwan University (Assistant professor)

요 약

불법 도박 사이트를 홍보하기 위한 스팸 메시지는 링크를 단축하거나, 특수 문자, 한글 표현 등으로 위장해 필터링을 회피하는 경우가 많다. 이러한 링크는 탐지 및 차단이 어려워 사회적 피해를 유발할 수 있다. 본 연구에서는 스팸 메시지 내 은닉된 불법 도박 사이트 링크를 자동으로 추출하고 복원하는 파이프라인을 설계하였다. 제안하는 파이프라인은 메시지 전처리, 링크 포함 여부 분류, 규칙 기반 복원, 유효성 검증 및 리다이렉션 필터링의 과정을 포함한다. 본 연구는 실제 환경에서 적용할 수 있는 불법 도박 탐지의 자동화 가능성을 제시하였으며, 향후에는 복원된 링크의 목적지를 분류할 수 있는 탐지 모델 개발로 확장될 수 있다.

I. 서론

불법 도박 사이트를 홍보하기 위한 스팸 메시지에는 사이트의 링크들이 포함되어 있는데, 이러한 링크는 대부분은 정상적인 링크처럼 위장되거나 단축 URL, 유사 문자 조합 등을 통해 탐지를 회피하려는 경향이 있다[1][2].

불법 도박 사이트는 청소년 접근 가능성, 사기 피해, 개인정보 유출 등의 문제와 직결되어 있어 체계적인 대응이 필요하다[3]. 하지만 교묘하게 숨겨져 있는 링크로 인해, 불법 도박 사이트의 링크를 즉각적으로 확인하여 해당 사이트 접근을 차단하기는 어려운 문제점이 있다.

본 연구는 스팸 메시지 내 숨겨진 불법 도박 사이트 링크를 효과적으로 추출하고 복원하기 위한 자동화 파이프라인을 제안한다. 제안하는 파이프라인은 원본 메시지로부터 링크 포함 여부 분류, 복원 규칙 적용, 링크 유효성 검증 및 리다이렉션 필터링의 과정을 포함한다. 이를 통해 최종적으로 현재 운영 중인 불법 도박 링크를 자동으로 탐지할 수 있으며, 해당 정보를 바탕으로 현재 운영되고 있는 불법 도박 사이트에 대한 접근을 차단하여 피해를 막는 데 기여할 수 있을 것으로 기대된다.

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과(No.RS-2024-00439139, 사이버 위기 대응 능력 및 복원력 시험평가 도구 개발)이며, 과학기술정보통신부 및 정보통신기획평가원의 메타버스융합대학원의 연구 결과로 수행되었음 (IITP-2025-RS-2023-00254129)

II. 배경

스팸 메시지는 수신자의 의사와 무관하게 발송되는 광고성 메시지로, 개인정보 유출 등의 보안 위협을 야기할 수 있다. 2024년 하반기 KISA 보고서에 따르면, 1인당 월평균 약 7.32통의 휴대전화 문자 스팸을 수신하고 있으며, 불법 도박 홍보 스팸은 키워드 변형, 특수 문자 삽입, 단축 URL 등을 사용해 필터링을 우회하고 있다[2][4].

2025년 3월 기준, 특정 대량문자사업자를 통한 월간 스팸신고는 392만 건에 달해 스팸

유통의 주요 경로로 지목되고 있다[5]. 기존 스팸 탐지는 주로 키워드 필터링, 발신번호 차단 등에 의존했으나, 변형 표현과 번호 위조 등으로 탐지율이 저하되고 있다. 이에 따라 스팸 탐지 기술 및 악성 여부를 파악하고 신속 차단하기 위한 기술적 대응도 활발히 연구되고 있다. 정부 차원의 규제 강화와 국제 협력 역시 중요한 과제로 떠오르고 있다.

III. 실험방법

3.1 데이터

본 연구에서는 한국인터넷진흥원(KISA)로부터 제공받은 스팸 문자 데이터를 사용하였다. 데이터에는 1주일 동안 KISA의 자체 분류 시스템에 따라 불법 도박, 주식 및 투자, 유흥 업소 등 23개의 유형으로 분류되어 있다. 본 연구에서는 2025년 3월 12일부터 2025년 4월 2일까지 4주치의 데이터를 활용하였다.

3.2 데이터 전처리

원본 데이터는 단일 CSV 파일 내부에 다수의 메타 정보와 주석이 포함된 구조로 되어 있어, 불법 도박에 해당하는 메시지 내용만을 추출하는 과정을 수행하였다. [(KISA:SOL)]과 [(KISA:EOL)]로 구분되는 블록 중 [(KISA)]불법 도박이라는 태그가 포함된 메시지 블록만을 필터링하였다. 이후 각 블록 내에서 마지막 두 개의 [(KISA)] 사이에 위치한 문자열을 실제 사용자에게 전송된 스팸 메시지로 간주하고, 이를 텍스트 형태로 정리하였다.

3.3 링크 포함 여부 분류

후속 처리에서 불필요한 계산을 줄이기 위해, 메시지 내에 링크 포함 여부를 분류하는 이진 분류 모델을 적용하였다. 본 연구에서는 한국어에 특화된 사전학습 언어모델인 KcELECTRA-base를 기반으로, 약 2만 건의 스팸 메시지를 GPT-4o 및 Gemini-2.5를 활용한 라벨링을 통해 학습 데이터를 구축하였다. 모델은 Hugging Face Transformers 라이브러리를 기반으로 PyTorch 환경에서 학습되었으며, Google Colab의 A100 GPU 환경에서 10 epoch 로 학습되었다.

학습된 모델은 최종적으로 F1-score 0.9835, Accuracy 97.8%의 성능을 달성하였다.

3.4 규칙 기반 링크 복원

스팸 메시지 내 불법 도박 사이트의 링크는 사용자나 필터링 시스템의 탐지를 피하기 위해 다양한 방식으로 위장되어 있다. 공백 삽입, 특수 문자 치환, 원문자(예: ⓐ, ⓑ), 한글 기반의 .com 우회 표현(예: '쩜 컴', '쩜 꺼 엄') 등이 존재한다. 이러한 표현은 기존 URL 정규식을 활용한 탐지 방식으로는 복원이 어려워 본 연구에서는 규칙 기반 링크 복원 함수를 설계하였다.

1. 특수 문자 치환 : 원 문자(예: ⓐ, ⓑ), 원숫자(예: ①, ②)를 알파벳 및 숫자로 치환.
2. 유사 문자 치환 : Ø, ß 와 같은 문자들을 알파벳에 대응 되는 것으로 치환
3. TLD 우회 표현 복원 : ‘쩜꺼엄’, ‘쩜컴’ 등 한글로 우회된 TLD 표현을 복원
4. 공백 제거 및 도메인 조립 : ‘h t t p’ 와 같이 공백이 포함된 도메인을 결합해 정규화
5. 짧은 링크 탐지 : “bit.ly, vo.la, t.ly 등” 도메인 패턴을 활용하여 단축 URL 복원
6. 일반 도메인 정합성 검사 : http/https가 없더라도, 도메인 패턴이 존재하면 링크 조립
7. 도메인 키워드 기반 후보 추정 : ‘bet’, ‘casino’, ‘slot’ 등 도박 관련 키워드가 포함되어 있을 경우 ‘.com’을 붙여 링크로 추정

위 과정은 Python 기반으로 구현되었으며, 복원이 실패한 경우는 공백으로 처리하고, 이후 유효성 검증 단계에서 제외한다. 규칙 기반 복원 방식은 특히 한국어 환경에서 우회 표현이 다양하게 나타나는 특성을 고려하여 설계되었으며, 기계 학습 없이도 높은 정확도로 위장된 링크를 원래 형태로 재구성할 수 있다.

3.5 링크 유효성 검증

복원된 링크 중 실제로 접속이 불가능하거나, 도박 사이트와는 무관한 서비스(예: 카카오톡 오픈채팅, 구글 폼)로 리다이렉트되는 경우도 존재한다.

따라서, 복원된 링크에 접속 여부를 확인하고, 유효하지 않거나 불필요한 목적지로 리다이렉트되는 경우를 제외하는 후처리 과정을 설계하였다. 복원된 링크 중 중복되는 URL은 제거하고 고유한 링크만 남겨, 유효성 검증은 고유한 링크 단위로만 수행하였다.

본 유효성 검증 프로세스는 다음과 같은 단계로 구성된다.

1. HTTP 응답 여부 확인 : Python requests 라이브러리를 활용하여 각 복원된 링크에 접속을 시도하고, 응답 상태 코드를 확인한다.
2. 리다이렉션 대상 필터링 : 최종 리다이렉션된 URL이 'open.kakao.com', 'docs.google.com' 등과 같은 정상적인 서비스 플랫폼일 경우, 해당 링크는 제외한다.
3. 본문 내 404 페이지 탐지 : HTTP 응답이 성공하더라도, 본문 내에 '404', 'page not found', '존재하지 않음' 등이 포함된 경우, 해당 링크는 무효한 것으로 간주하고 제외한다.
4. 예외 처리 및 유지 : 접속 중 예외가 발생한 경우(예: 연결 차단, 리셋 등)는 로그로 기록하되, 향후 재검토를 위해 복원된 링크는 유지하고 예외 목록에 따로 저장한다.

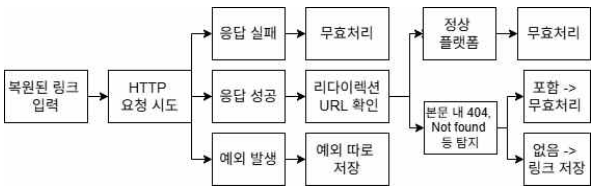

그림 1 링크 유효성 검증 파이프라인 구조

IV. 결과 분석

해당 파이프라인을 거쳐, 2025년 3월 27일 ~ 2025년 4월 2일의 스팸메시지 데이터셋을 추출/복원한 결과는 다음과 같다.

전체 메시지	16,316건
불법 도박 메시지	9,329건 (57.18%)
링크 포함된 메시지	7,088건 (43.44%)
정상 작동하는 링크	798개
예외처리된 링크	262개

표 1 스팸 메시지 내 불법도박 링크 추출/복원 결과

전체 메시지 중 '불법 도박'으로 분류되어 있는 메시지는 9,329건(57.18%)으로 이 중 학습된 모델로 링크 포함 여부를 분류하였을 때, 링크가 포함된 메시지는 7,088건(43.44%)이었으며, 불법 도박 메시지 중에서는 75.98%에 해당한다.

파이프라인에 따라 링크를 추출하였을 때 1,060개의 링크가 추출되었으며, 그중 정상 작동하는 링크는 798개이고, 복원된 링크 중 75.2%에 해당한다. 예외처리된 링크는 262개로 전체 링크 중 24.7%에 해당한다.

V. 결론 및 향후 과제

본 연구는 스팸 메시지 내 숨겨진 불법 사이트 링크를 자동으로 추출하고 복원하는 파이프라인을 제안하였다. 링크 포함 여부 분류, 규칙 기반 복원, 유효성 검증 및 리다이렉션 필터링을 단계적으로 수행함으로써 접속 가능한 위험 링크를 효과적으로 선별할 수 있었다.

불법 도박 홍보의 경우 메시지뿐만 아니라 SNS, 이메일 등 다양한 수단을 활용한다. 해당 과정을 통해 추출된 링크에 대한 접근을 차단하는 과정을 거친다면 다른 수단을 통해 접한 사이트에 대해서도 접근을 막을 수 있다.

다만, 불법 도박 메시지로부터 복원된 링크가 정상적으로 작동하더라도 실제로 불법 도박 사이트인지까지는 판단하고 있지 않다는 한계가 존재한다. 따라서 향후 과제로 최종 목적지의 불법 도박 사이트 여부를 분류할 수 있는 탐지 모델을 개발 및 적용하여 해당 링크에 대한 실시간 차단이나 즉각적인 조치를 가능하게 하는 연구로 확장할 필요가 있다.

[참고문헌]

[1] 정혜영, 김민철, "불법 온라인 도박 사이트 홍보 스팸문자의 언어적 특징 분석," 언어와 정보학, 제26권, 제3호, pp.17-24, 2022.

[2] 최현석, 박은영, "스팸 SMS URL의 악성 여부 판별을 위한 머신러닝 기반 탐지 기법," 한국정보과학회 학술대회 논문집, 제48권, 제2호, pp.337-344, 2021.

[3] 안재현, 김민석, "스팸 문자의 사회적 피해와 대응 기술 현황," 사이버보안학회지, 제5권, 제1호, pp.78-88, 2023.

[4] 한국인터넷진흥원(KISA). (2025). 2024년 하반기 스팸 유통현황 보고서.

[5] 한국인터넷진흥원(KISA), "2025년 3월 대량문자사업자별 스팸신고 현황," 불법스팸대응센터, 2025.04.14.