Flow-to-Text 기반 IoT 침입 탐지를 위한 텍스트 표현 방식 비교 연구

김유담, 안석현, 박수현, 최희수, 조성제
단국대학교 소프트웨어학과, 단국대학교 인공지능융합학과, 단국대학교 컴퓨터학과
{kyd1012kr, seokhyun, parksh, cv3686, sjcho}@dankook.ac.kr

1. 서론

- IoT 환경의 네트워크 공격 증가 → 효율적인 IDS의 중요성 증대
- 기존 ML/DL 기반 IDS: 높은 성능이나 특징 설계 의존성 및 새로운 공격 패턴 일반화 한계
- 최근 PLM을 구조적 데이터에 적용하는 연구 증가 → 텍스트 표현 구조가 성능에 미치는 영향 분석은 부족

연구 목적: 세 가지 텍스트 템플릿(T1~T3)을 통해 RoBERTa 기반 침입 탐지 성능을 실험적으로 비교 분석

2. 실험 방법

(1) 데이터셋 및 입력 특징

- 데이터셋 : 실제 IoT 기기 환경의 트래픽을 반영한 IoT-23 데이터셋 활용
- 특징 추출 : Zeek conn.log에서 보안 탐지에 유효한 12개 핵심 피처 선별
  - 네트워크 기본 정보 : proto, service, conn_state
  - 양방향 통계 정보 : duration, orig_bytes/resp_bytes , orig_pkts/resp_pkts
  - 기타 특징 : missed_bytes, history, id.orig_p, id.resp_p
- 분류 레이블
  - 다중 분류(5class) : Benign, PortScan, DDoS, Okiru, C&C
  - 이진 분류(2class) : Benign vs. Malicious
- 데이터 분할: 학습 80% / 평가 20% (group 기반, 중복 없음)

(2) Flow-to-text 변환

템플릿	설명	예시
T1	자연어 서술형	A network flow was observed using the tcp protocol. The communication lasted unknown seconds...
T2	키=값 구조형	Flow summary: protocol=tcp; duration=unknown seconds; source_port=39000; ...
T3	확장 자연어형	This network session reflects communication behavior over tcp. A device initiated traffic...

(3) 실험 설정

- 기반 모델: roberta-base (BERT의 사전학습 절차를 개선한 모델로, 언어 이해 및 텍스트 분류 작업에서 널리 활용)
- 학습 파라미터
  - Learning Rate: 2×10^{-5}
  - Batch Size: 8
  - Epochs: 최대 5
  - Max Length: 256
  - Macro F1-score 기준 조기 종료 적용 (patience=2)
- 분류 레이블
  - ML: Random Forest, XGBoost
  - DL: MLP
  - IoT-23 데이터셋의 클래스 불균형 완화를 위해 각 모델의 학습 과정에서 클래스 가중치를 개별적으로 적용
- 평가 지표
  - 주 지표 : Macro F1-score — 각 클래스의 F1-score를 동일한 비중으로 평균하여 클래스 불균형 환경에서도 성능을 왜곡 없이 평가
  - 보조 지표 : 보조 지표: Accuracy, 클래스별 F1-score (다중 분류)

3. 실험 결과

- 이진 및 다중 분류 성능 비교

Model	이진 분류		다중 분류	
	Acc	F1	Acc	F1
RF	0.9383	0.8873	0.9387	***0.9471***
XGBoost	0.9439	0.8993	0.8839	0.8759
MLP	0.8667	0.7699	0.7851	0.7721
RoBERTa-T1	0.9453	0.8953	0.8793	0.9036
RoBERTa-T2 ★	***0.9495***	***0.9054***	***0.9390***	***0.9437***
RoBERTa-T3	0.9446	0.8938	0.8796	0.9040

- 이진 분류: RoBERTa-T2 Macro F1 90.54%로 최고 성능
- 다중 분류: RF(94.71%) > RoBERTa-T2(94.37%, 0.34%p 차이)
- 다중 분류에서 RoBERTa-T1, T3는 Accuracy는 낮으나 Macro F1-score 높음 → 클래스 불균형 환경에서 소수 클래스 예측 성능 우수

- 클래스별 F1-score 분석

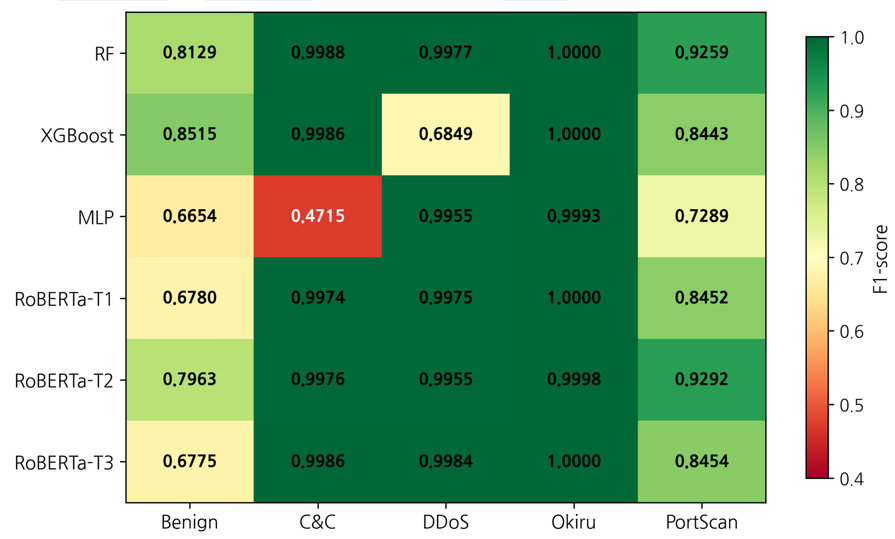

그림. 모델별 클래스 F1-score 히트맵

- Okiru: 모든 모델에서 F1-score 99.9% 이상
- XGBoost DDoS: 68.49%로 저하 → PortScan 샘플 일부가 DDoS로 오분류
- MLP C&C: 47.15%로 크게 저하 → PortScan 샘플 다수가 C&C로 오분류
- RoBERTa-T2 PortScan: 92.92%로 T1(84.52%), T3(84.54%) 대비 8.40% 높음
- RoBERTa-T2가 전 클래스에 걸쳐 가장 안정적인 성능 유지

4. 결론

- T2(키=값 구조형)가 이진/다중 분류 모두에서 RoBERTa 템플릿 중 최고 성능
- 이진 분류: 제시 모델 중 RoBERTa-T2 Macro F1 90.54%로 최고 성능
- 다중 분류: RoBERTa-T2 94.37% (RF 94.71%와 유사)
- 구조적 데이터에서는 피처명·값을 명시적으로 보존하는 표현 방식이 PLM 학습에 효과적
- 한계 및 향후 연구
  - 단일 데이터셋(IoT-23), 단일 모델(RoBERTa), 3가지 템플릿만 비교
  - 다양한 데이터셋 검증, 경량/생성형 언어모델 비교, attention 분석을 통한 해석 가능성 연구 필요

5. Acknowledgement

이 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학석사연계ICT핵심인재양성 지원을 받아 수행된 연구임(IITP-2026-RS-2023-00259867) . 또한, 본 연구는 2026년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(2024-0-00035)