Transformer 기반
HEAT 악성 URL 탐지 시스템

김예지 고가은 곽지현

원팀

table of contents

원팀

기존 블랙리스트 한계를 넘어 Transformer로 적응형 악성 URL 탐지 구현

HEAT 공격은 URL을 살짝 변형해 필터를 회피하지만,
이 프로젝트는 문맥 기반으로 이를 탐지합니다.

[문제 상황]

- 피싱, 스팸 URL 공격 급증
- 기존 블랙리스트 방식의 한계
- 변형 URL(HEAT 공격)에 대한 대응 불가

[HEAT 공격 특징]

- 무의미한 문자열 삽입
- 도메인/경로 변조
- 인코딩 교란을 통한 탐지 회피

[우리의 해결 방향]

- Transformer의 문맥 이해력 활용
- 문자 단위 인코딩으로 변형 패턴 학습
- 새로운 공격에도 적응 가능한 탐지 모델

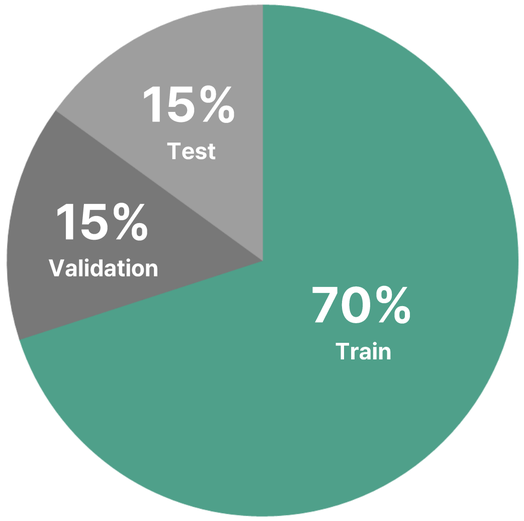
Pipeline

시스템 전체 구조

Distribution

데이터셋 구성 및 전처리

데이터셋 구성 비율 (Dataset Split)



[Train] (70%)

모델이 패턴을 학습하도록 사용하는
데이터로, 전체 70%를 차지

[Validation] (15%)

학습 중 모델의 성능을 점검하고
하이퍼파라미터를 조정하는 데 사용

[Test] (15%)

학습이 완료된 모델의 최종 성능과
일반화 능력을 평가하기 위해 사용

총 데이터 개수 : 1,782,941 / Train : 1,248,058 / Validation : 267,441 / Test : 267,442
출처 : Kaggle, Alexa Top Sites, KISA

데이터셋 샘플 (Dataset Sample)

정상 URL

	A	B	C
73	73	google.com.tw	
74	74	google.com.au	
75	75	whatsapp.com	
76	76	google.pl	
77	77	xhamster.com	
78	78	detail.tmall.com	
79	79	diply.com	
80	80	google.co.id	
81	81	adobe.com	
82	82	coccoc.com	
83	83	craigslist.org	
84	84	nicovideo.jp	
85	85	txxx.com	
86	86	dropbox.com	
87	87	amazon.de	
88	88	google.com.ar	
89	89	amazon.in	
90	90	googleusercontent.c	
91	91	google.com.pk	

Normal_URL

악성 URL

	A	B	C
82	82	http://vb.fg6e.yachts	
83	83	http://nlx.o4gs.yachts	
84	84	http://xls.o4gs.yachts	
85	85	http://byx.o4gs.yacht	
86	86	http://nly.o4gs.yachts	
87	87	http://ion.r7pm.yacht	
88	88	http://ynr.q6yd.yachts	
89	89	http://yie.q6yd.yachts	
90	90	http://ynv.q6yd.yacht	
91	91	http://yiu.q6yd.yachts	
92	92	http://yna.q6yd.yacht	
93	93	http://zib.t9zd.yachts	
94	94	http://ziu.u0tk.yachts	
95	95	http://nlz.o4gs.yachts	
96	96	http://xlv.o4gs.yachts	
97	97	http://nlw.o4gs.yacht	
98	98	http://byz.o4gs.yacht	
99	99	http://byn.o4gs.yacht	
100	100	http://hit.s8vn.yachts	

Phishing_URL

전처리 과정 (Preprocessing Steps)

중복 및 비정상 **URL** 제거 **->** 특수문자 정규화 및 소문자 변환 **->** 라벨 밸런싱 **(정상:악성 = 1:1) ->** 데이터 무작위 셔플 및 인코딩 통일

Model design

모델 설계 및 학습 설정

핵심 설정

- batch=4, grad_accum=8
- max_length=192
- learning_rate=2e-5
- dropout=0.1, weight_decay=0.01

정규화 기법

EarlyStopping, Scheduler(linear), FP16

기반 모델

DeBERTa-v3-Large

훈련 환경

Colab T4 GPU, PyTorch 2.3

최종 모델

20000

Model performance

모델 성능 향상 과정

v1

기본 학습

F1 Score: 0.904

초기 모델, underfitting

v2

Regularization + LR Scheduler

F1 Score: 0.934

수렴 안정화

v3

FP16 + EarlyStopping

F1 Score: 0.9587

최종 안정적 수렴

v3-FP16

경량화 후 성능 유지

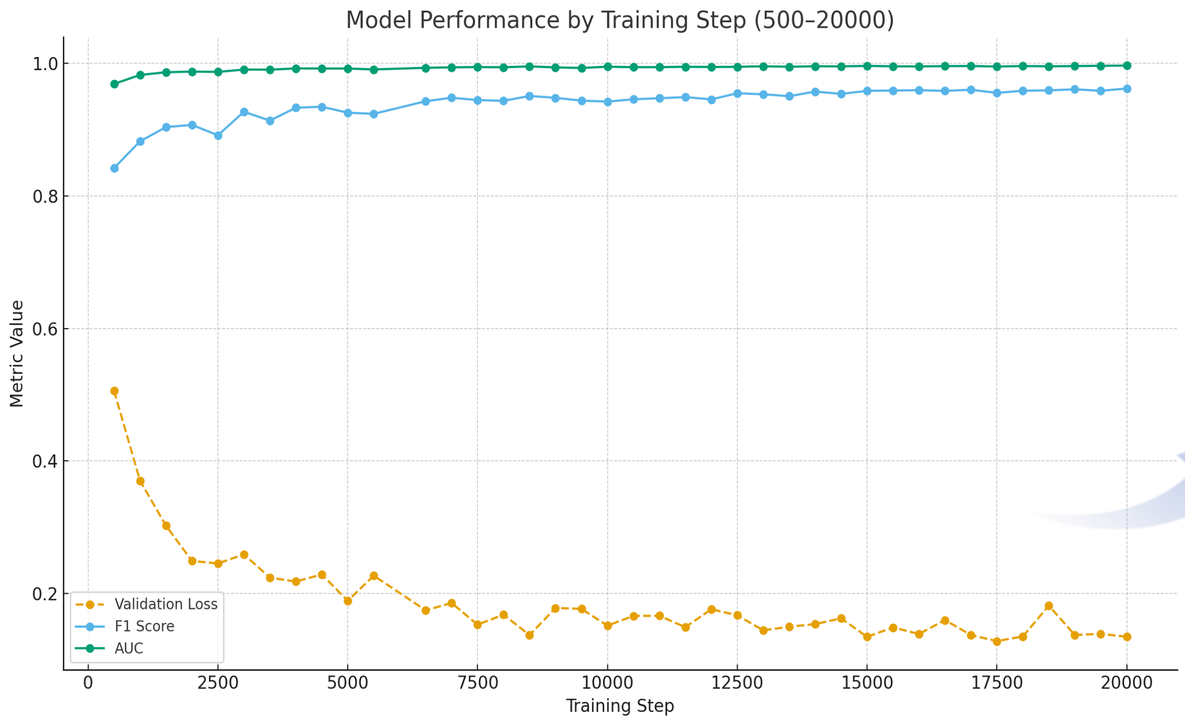
F1 Score: 0.9586

효율성 확보

Model performance

최종 평가 결과

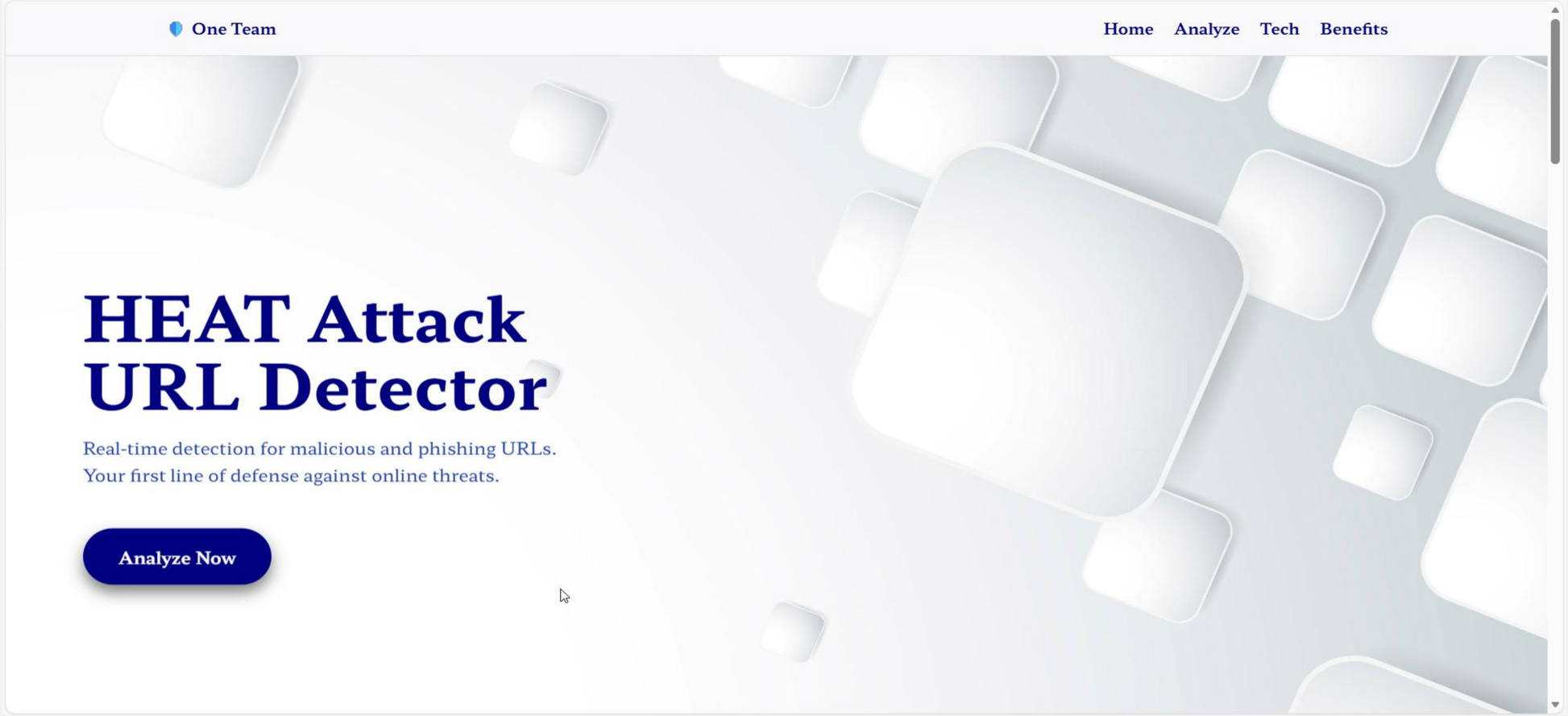
최종 모델 성능 지표



Metric	Score	해석
Accuracy	0.9718	전반적 정확도 우수
Precision	0.9671	악성 탐지 오탐율 낮음
Recall	0.9575	악성 탐지 민감도 높음
F1 Score	0.9621	안정적 조화 성능
AUC	0.9967	판별 경계 매우 우수

Avatar

웹 서비스 구현 및 시연 영상



URL 입력 -> 모델 추론 -> 결과 반환 (정상/악성 표시)

Conclusion

기술적 차별점 및 결론

기존 방식		HEAT 시스템
블랙리스트/규칙기반	탐지 원리	문맥 학습 기반
낮음 (패턴 변형에 취약)	적응성	높음(Transformer 학습)
수동	업데이트	자동화 가능
보통	실시간성	고속 추론 (FP16)

" Transformer 기반으로 변형 URL 탐지율이 크게 향상 "

- ✓ 데이터 품질 -> 모델 안정성 -> 서비스 연동까지 전주기 완성
- ✓ 최종 성능: F1 = 0.9587 / AUC = 0.9960 , 데이터 180만 건 구축 및 실시간 탐지 시스템 구현

감사합니다.

김예지 고가은 곽지현

원팀