



피싱 사이트 피쳐 추출

ALL IT ONE

C O N T E N T S

1. 피쳐 추출(*Based by Address Bar*)
2. 피쳐 추출(*Based by Abnormal*)
3. 피쳐 추출(*Based by Domain*)

피쳐 추출 규칙

Phishing Website 데이터

Phishing Websites Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: This dataset collected mainly from: PhishTank archive, MillerSmiles archive, Google's searching operators.

Data Set Characteristics:	N/A	Number of Instances:	2456	Area:	Computer Security
Attribute Characteristics:	Integer	Number of Attributes:	30	Date Donated	2015-03-26
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	98877

Source:

Rami Mustafa A Mohammad (University of Huddersfield, rami.mohammad@hud.ac.uk, rami.mustafa.a@gmail.com)

Lee McCluskey (University of Huddersfield, t.l.mcccluskey@hud.ac.uk)

Fadi Thabtah (Canadian University of Dubai, fadi@cud.ac.ae)

Data Set Information:

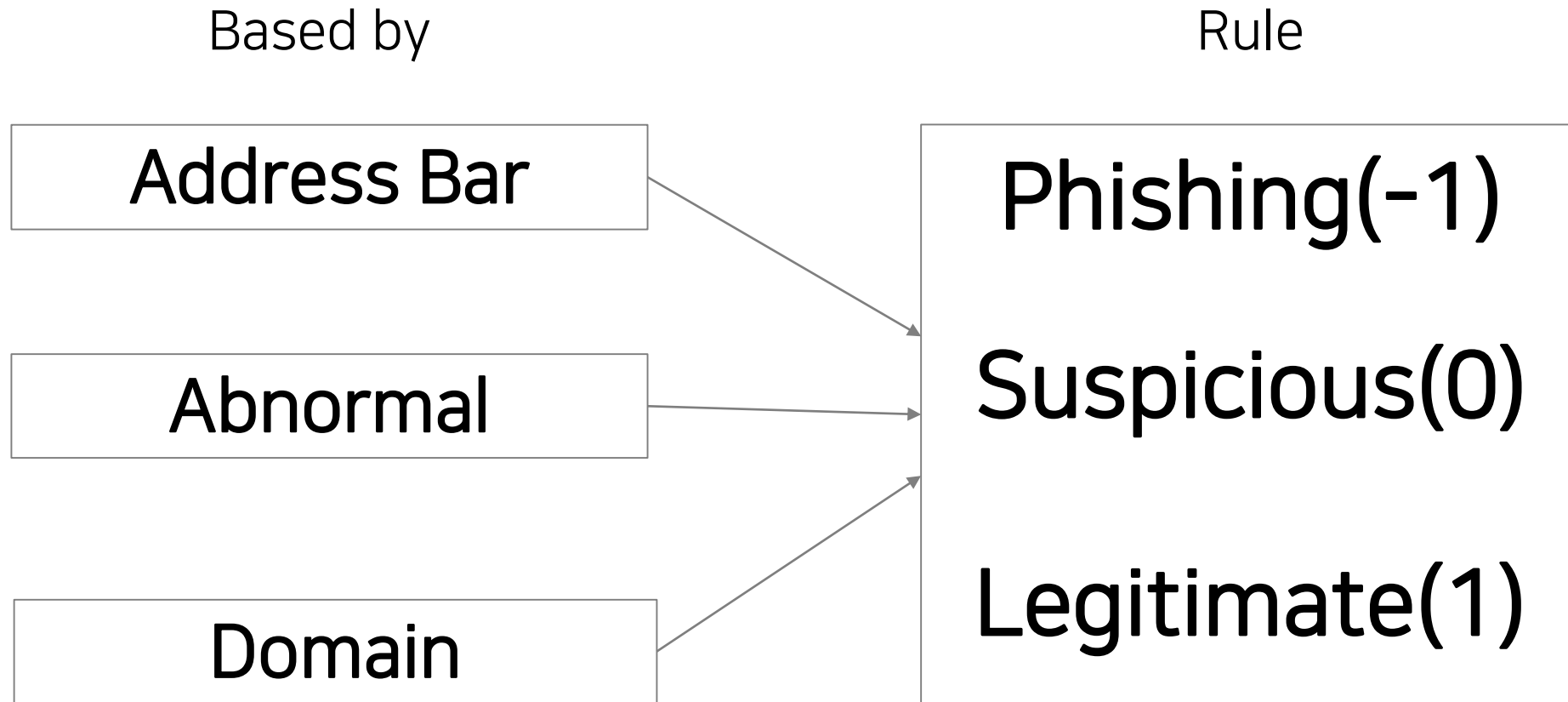
One of the challenges faced by our research was the unavailability of reliable training datasets. In fact this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites have been disseminated these days, no reliable training dataset has been published publically, may be because there is no agreement in literature on the definitive features that characterize phishing webpages, hence it is difficult to shape a dataset that covers all possible features.

In this dataset, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we propose some new features.

<https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>

피쳐 추출 규칙

■ 규칙



C O N T E N T S

1. 피쳐 추출(Based by Address Bar)

1. 피쳐 추출(Based by Address Bar)

Based by Address Bar

Num	Features
1	Using the IP Address
2	Long URL to Hide the Suspicious Part
3	Using URL Shortening Services "TinyURL"
4	URL's having "@" Symbol
5	Redirection using "///"
6	Adding Prefix or Suffix Separated by (-) to the Domain
7	Sub Domain and Multi Sub Domains
8	HTTPS(Hyper Text Transfer Protocol with Secure Sockets Layer)
9	Domain Registration Length
10	Favicon
11	The Existence of "HTTPS" Token in the Domain Part of the URL

1. 피쳐 추출(Based by Address Bar)

1. Using the IP Address

규 칙		
도메인에 IP 주소가 포함되어 있으면	→	피싱(-1)
그렇지 않으면	→	정상(1)

← → ↻ http://125.98.3.123.asdfas.as.asd.sad.com/fake.html

← → ↻ http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html

1. 피쳐 추출(Based by Address Bar)

2. Long URL to Hide the Suspicious Part

규 칙		
URL 길이 < 54	→	정상(1)
54 ≤ URL 길이 ≤ 75	→	의심(0)
URL 길이 > 75	→	피싱(-1)



http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?
cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5
b74f8dc1e7c2e8dd4105e8@phishing.website.html

1. 피쳐 추출(Based by Address Bar)

3. Using URL Shortening Services "TinyURL"

규 칙		
단축 URL 사용	→	피싱(-1)
그렇지 않으면	→	정상(1)



1. 피쳐 추출(Based by Address Bar)

4. URL's having "@" Symbol

규 칙		
URL에 "@"가 포함되어 있으면	→	피싱(-1)
그렇지 않으면	→	정상(1)

← → ↻ http://somewhere.foo/profile/username@somewhere.foo

← → ↻ http://somewhere.foo/profile/username%40somewhere.foo

1. 피쳐 추출(Based by Address Bar)

5. Redirecting using "//"

규 칙		
URL에 "//"가 7자리 뒤에 있으면	→	피싱(-1)
그렇지 않으면	→	정상(1)

h	t	t	p	:	/	/	n
1	2	3	4	5	6	7	8

h	t	t	p	s	:	/	/
1	2	3	4	5	6	7	8

←
→
↻

1. 피쳐 추출(Based by Address Bar)

6. Adding Prefix or Suffix Separated by (-) to the Domain

규 칙		
도메인 주소에(-)가 포함되어 있으면	→	피싱(-1)
그렇지 않으면	→	정상(1)



1. 피쳐 추출(Based by Address Bar)

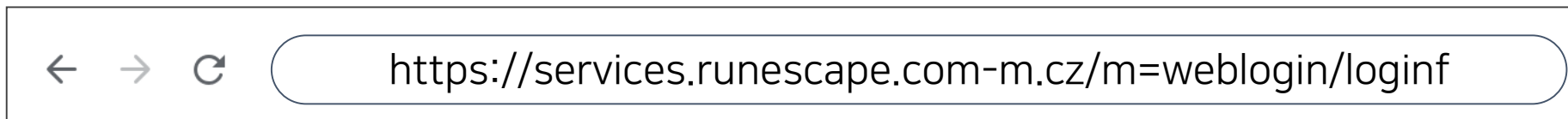
7. Sub Domain and Multi Sub Domains

<http://www.president.go.kr/>

.kr(korea) : country-code top-level domains (ccTLD)

.go(government).kr : second-level domain (SLD)

규 칙(www. 와 ccTLD 를 제거하고)		
Dot(.)의 개수가 1개	→	정상(1)
Dot(.)의 개수가 2개	→	의심(0)
Dot(.)의 개수가 3개 이상	→	피싱(-1)



1. 피쳐 추출(Based by Address Bar)

8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

규 칙		
https를 사용 and 신뢰할 수 있는 인증 기관 and 인증서가 1년 이상	→	정상(1)
https를 사용 and 신뢰할 수 없는 인증 기관	→	의심(0)
그 밖에 다른 경우	→	피싱(-1)

이름	종류	만료
AAA Certificate Services	인증서	2029. 1. 1. 오전 8:59:59
Actalis Authentication Root CA	인증서	2030. 9. 22. 오후 8:22:02
AddTrust Class 1 CA Root	인증서	2020. 5. 30. 오후 7:38:31
AddTrust External CA Root	인증서	2020. 5. 30. 오후 7:48:38
Admin-Root-CA	인증서	2021. 11. 10. 오후 4:51:07
AffirmTrust Commercial	인증서	2030. 12. 31. 오후 11:06:06
AffirmTrust Networking	인증서	2030. 12. 31. 오후 11:08:24
AffirmTrust Premium	인증서	2040. 12. 31. 오후 11:10:36
AffirmTrust Premium ECC	인증서	2040. 12. 31. 오후 11:20:24
Amazon Root CA 1	인증서	2038. 1. 17. 오전 9:00:00
Amazon Root CA 2	인증서	2040. 5. 26. 오전 9:00:00
Amazon Root CA 3	인증서	2040. 5. 26. 오전 9:00:00

발급자 이름	
국가	US
조직	Amazon
일반 이름	Amazon Root CA 1
일련 번호	06 6C 9F CF 99 BF 8C 0A 39 E2 F0 78 8A 43 E6 96 36 5B CA
버전	3
서명 알고리즘	SHA-256(RSA 암호화)(1.2.840.113549.1.1.11)
매개변수	없음
다음 전에 유효하지 않음	2015년 5월 26일 화요일 오전 9시 0분 0초 대한민국 표준시
다음 후에 유효하지 않음	2038년 1월 17일 일요일 오전 9시 0분 0초 대한민국 표준시

1. 피쳐 추출(Based by Address Bar)

9. Domain Registration Length

규 칙		
도메인 만료 기간이 1년 이하이면	→	피싱(-1)
그렇지 않으면	→	정상(1)

<https://후이즈검색.한국/kor/main.jsp>

```
# KOREAN(UTF8)

도메인이름      : president.go.kr
등록인         : 대통령비서실
책임자         : Domain Administrator
책임자 전자우편 : postmaster@president.go.kr
등록일         : 2000. 06. 29.
최근 정보 변경일 : 2016. 11. 24.
사용 종료일    : 2020. 06. 29.
정보공개여부   : N
등록대행자     : (주)후이즈
DNSSEC         : 미서명
```

```
QUERY:acesseactualizacao.com
Domain Name: ACESSEATUALIZACAO.COM
Registry Domain ID: 2350888200_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.godaddy.com
Registrar URL: http://www.godaddy.com
Updated Date: 2019-01-09T13:18:52Z
Creation Date: 2019-01-09T13:18:51Z
Registry Expiry Date: 2020-01-09T13:18:51Z
Registrar: GoDaddy.com, LLC
Registrar IANA ID: 146
Registrar Abuse Contact Email: abuse@godaddy.com
Registrar Abuse Contact Phone: 480-624-2505
Domain Status: clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited
Domain Status: clientRenewProhibited https://icann.org/epp#clientRenewProhibited
Domain Status: clientTransferProhibited https://icann.org/epp#clientTransferProhibited
Domain Status: clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited
Name Server: NS05.DOMAINCONTROL.COM
Name Server: NS06.DOMAINCONTROL.COM
DNSSEC: unsigned
URL of the ICANN Whois Inaccuracy Complaint Form: https://www.icann.org/wicf/
>>> Last update of whois database: 2019-01-24T11:12:27Z <<<
```

1. 피쳐 추출(Based by Address Bar)

10. Favicon

규 칙		
Favicon이 외부 도메인에서 로드 되면	→	피싱(-1)
그렇지 않으면	→	정상(1)



```
<link rel="shortcut icon" type="image/x-icon" href="/favicon.ico">
```


1. 피쳐 추출(Based by Address Bar)

11. The Existence of "HTTPS" Token in the Domain Part of the URL

규 칙		
도메인에 https가 포함되어 있으면	→	피싱(-1)
그렇지 않으면	→	정상(1)



C O N T E N T S

2. 피쳐 추출(Based by Abnormal)

2. 피쳐 추출(Based by Abnormal)

■ Based by Abnormal

Num	Features
1	Request URL
2	URL of Anchor
3	Links in <Meta>, <script> and <Link> tags
4	Server Form Handler (SFH)
5	Abnormal URL

2. 피쳐 추출(Based by Abnormal)

1. Request URL

규 칙		
외부 요청 < 22%	→	정상(1)
22% <= 외부 요청 < 61%	→	의심(0)
외부 요청 >= 61%	→	피싱(-1)

2. 피쳐 추출(Based by Abnormal)

2. URL of Anchor

규 칙		
Anchor의 개수 < 31%	→	정상(1)
31% ≤ Anchor의 개수 ≤ 67%	→	의심(0)
Anchor의 개수 > 68%	→	피싱(-1)

Anchor 1

Phishing.co.kr

Anchor 2

- ①
- ②
- ③
- ④

2. 피쳐 추출(Based by Abnormal)

3. Links in <Meta>, <Script> and <Link> tags

규 칙(도메인과 태그의 도메인이 다를 경우)		
태그의 수 < 17%	→	정상(1)
17% ≤ 태그의 수 ≤ 81%	→	의심(0)
태그의 수 > 81%	→	피싱(-1)

<Meta> : HTML 문서의 내용, 작성자 등의 내용을 담는 태그

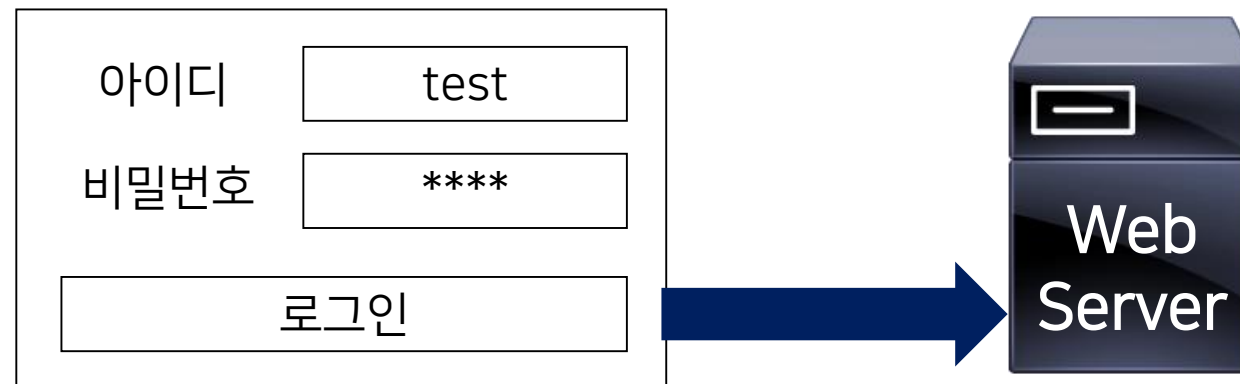
<Script> : 스크립트를 작성하기 위한 태그

<Link> : 외부 CSS 등을 가져오기 위해 사용하는 태그

2. 피쳐 추출(Based by Abnormal)

4. Server Form Handler(SFH)

| 규 칙 | | |
|-------------------------|---|--------|
| "about: blank" or Empty | → | 피싱(-1) |
| 다른 도메인에 전송 | → | 의심(0) |
| 그 외의 경우 | → | 정상(1) |



2. 피쳐 추출(Based by Abnormal)

5. Abnormal URL

| 규 칙 | | |
|----------------------------|---|--------|
| 호스트 이름이 url에 포함되어 있지 않을 경우 | → | 피싱(-1) |
| 그렇지 않으면 | → | 정상(1) |

| | |
|-----------|-----------------------------|
| 도메인이름 | : samsung.co.kr |
| 등록인 | : 삼성전자주식회사 |
| 등록인 주소 | : 경기도 수원시 영통구 삼성로 129 (매탄동) |
| 등록인 우편번호 | : 16677 |
| 책임자 | : 삼성전자주식회사 |
| 책임자 전자우편 | : ssdomain@samsung.com |
| 책임자 전화번호 | : 02-727-7203 |
| 등록일 | : 1995. 10. 26. |
| 최근 정보 변경일 | : 2018. 04. 17. |
| 사용 종료일 | : 2020. 10. 15. |
| 정보공개여부 | : Y |
| 등록대행자 | : <u>(주)후이즈</u> |
| DNSSEC | : 미서명 |

| | |
|-----------|---------------------------------|
| 도메인이름 | : kisa.or.kr |
| 등록인 | : 한국인터넷진흥원 |
| 등록인 주소 | : 전라남도 나주시 진흥길 9(빛가람동) 한국인터넷진흥원 |
| 등록인 우편번호 | : 58324 |
| 책임자 | : 도메인관리자 |
| 책임자 전자우편 | : kisairm@kisa.or.kr |
| 책임자 전화번호 | : 061-820-1163 |
| 등록일 | : 1996. 07. 20. |
| 최근 정보 변경일 | : 2018. 12. 28. |
| 사용 종료일 | : 9999. 12. 31. |
| 정보공개여부 | : Y |
| 등록대행자 | : <u>한국인터넷진흥원</u> |
| DNSSEC | : 서명 |

C O N T E N T S

3. 피쳐 추출(Based by Domain)

3. 피쳐 추출(Based by Domain)

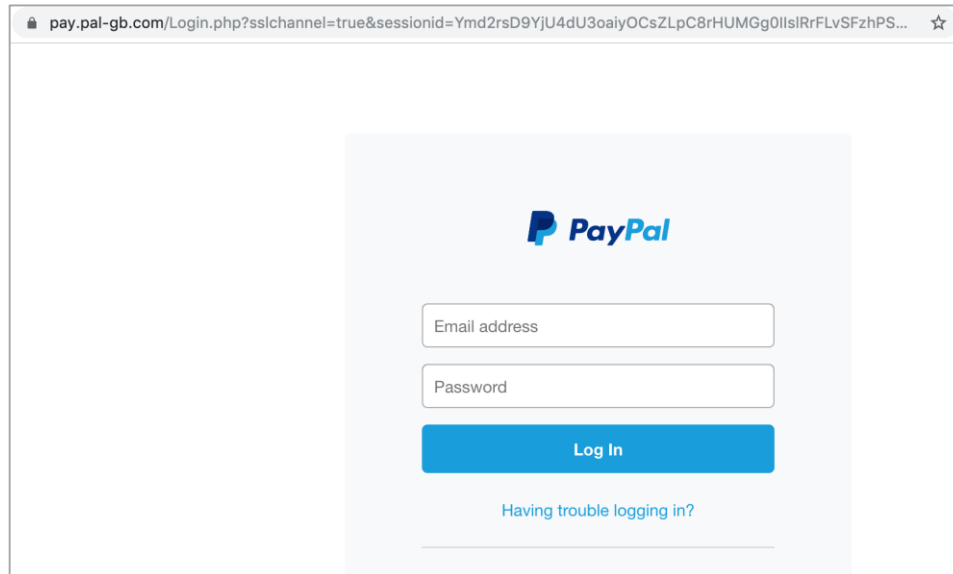
■ Based by Domain

| Num | Features |
|-----|----------------------------------|
| 1 | Age of Domain |
| 2 | DNS Record |
| 3 | PageRank |
| 4 | Number of Links Pointing to Page |

3. 피쳐 추출(Based by Domain)

1. Age of Domain

| 규 칙 | | |
|-------------------------|---|--------|
| 도메인 등록 기간 \geq 6개월 이상 | → | 정상(1) |
| 그렇지 않으면 | → | 피싱(-1) |



3. 피쳐 추출(Based by Domain)

■ ■ 2. DNS Record

| 규 칙 | | |
|------------------|---|--------|
| whois 검색 결과가 없으면 | → | 피싱(-1) |
| 그렇지 않으면 | → | 정상(1) |

3. 피쳐 추출(Based by Domain)

3. PageRank


| 규 칙 | |
|----------------|----------|
| PageRank < 0.2 | → 피싱(-1) |
| 그렇지 않으면 | → 정상(1) |

Domain Analysis For:
 **naver.com**

[Download PDF](#)

Date: August 14 2019

Google PageRank: 8/10
cPR Score: 9.0/10

Domain Analysis For:
 **tokokainbandung.com**

[Download PDF](#)

Date: August 14 2019

Google PageRank: 0/10
cPR Score: 0.1/10

<https://checkpagerank.net/>

3. 피쳐 추출(Based by Domain)

■ ■ 4. Number of Links Pointing to Page

| 규 칙 | | |
|---------------|---|--------|
| 외부 링크 = 0 | → | 피싱(-1) |
| 0 < 외부 링크 ≤ 2 | → | 의심(0) |
| 외부 링크 > 2 | → | 정상(1) |



감사합니다.
Thank you