Pass The Salt 2024

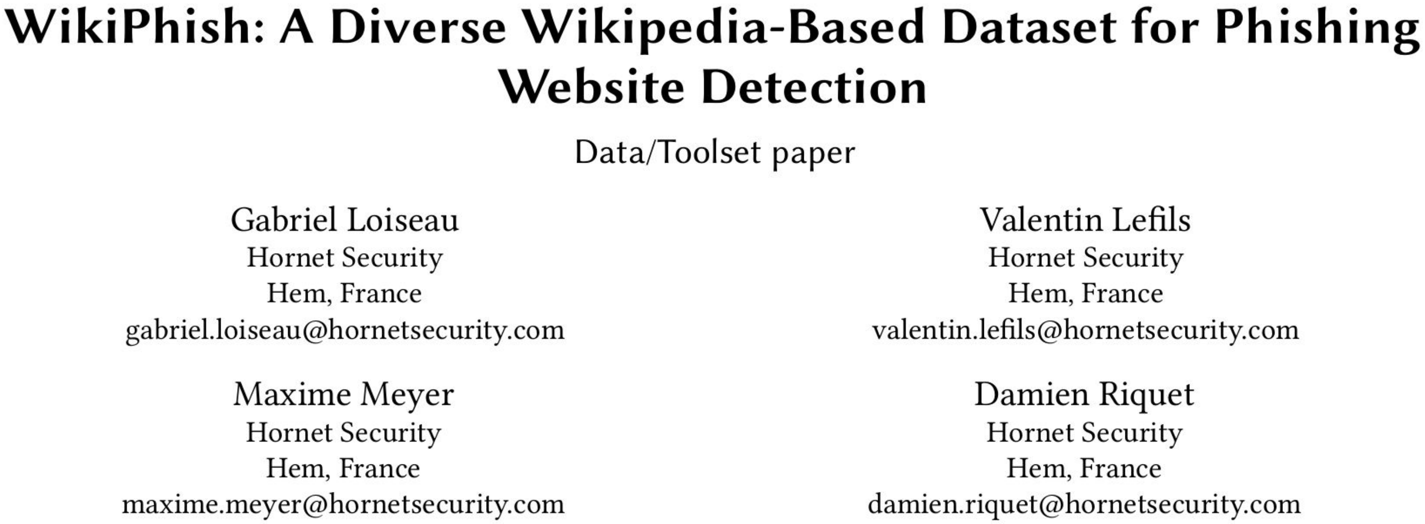
Fighting Phishing by Introducing WikiPhish

A new public dataset based on Wikipedia for legit URLs

Gabriel Loiseau

Hornet Security

WikiPhish Paper



WikiPhish: A Diverse Wikipedia-Based Dataset for Phishing Website Detection

Data/Toolset paper

Gabriel Loiseau
Hornet Security
Hem, France
gabriel.loiseau@hornetsecurity.com

Valentin Lefils
Hornet Security
Hem, France
valentin.lefils@hornetsecurity.com

Maxime Meyer
Hornet Security
Hem, France
maxime.meyer@hornetsecurity.com

Damien Riquet
Hornet Security
Hem, France
damien.riquet@hornetsecurity.com

https://doi.org/10.1145/3626232.3653283

WikiPhish Team

HORNETSECURITY

vade

Gabriel Loiseau
Research Scientist

Valentin Lefils
Data Scientist

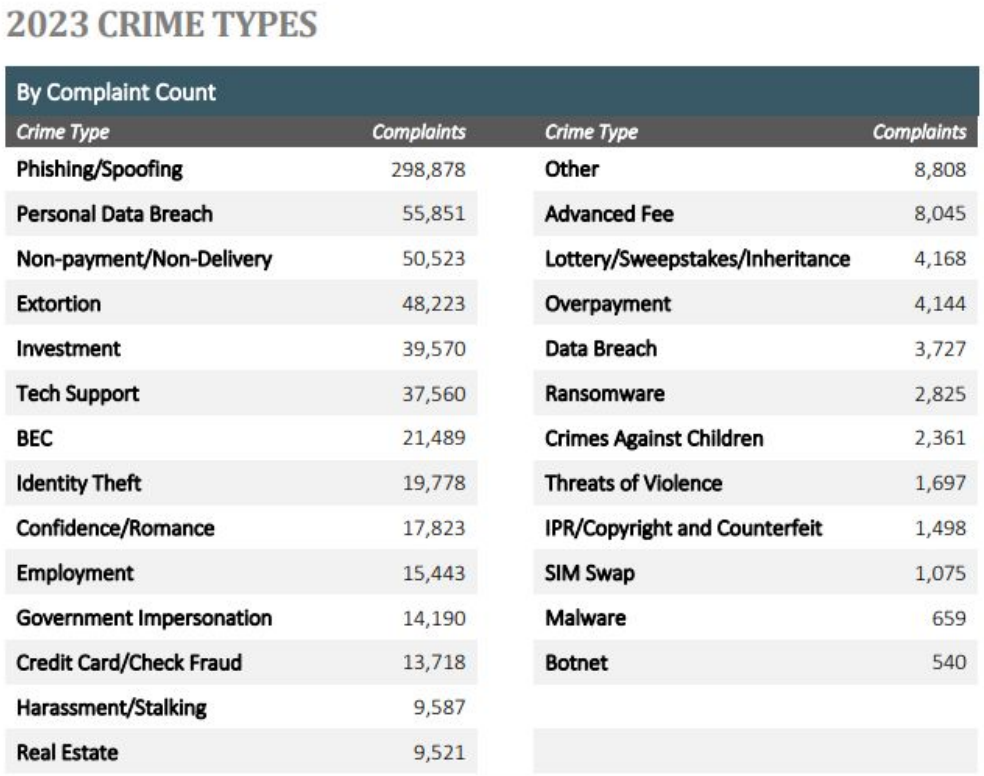
Maxime Meyer
Head of Research

Damien Riquet
Lead Research Engineer

Outline

- Background
- The WikiPhish Dataset
- Importance of Legitimate URL Diversity
- Training and Evaluating Classifiers
- Conclusion

Background: phishing detection using machine learning



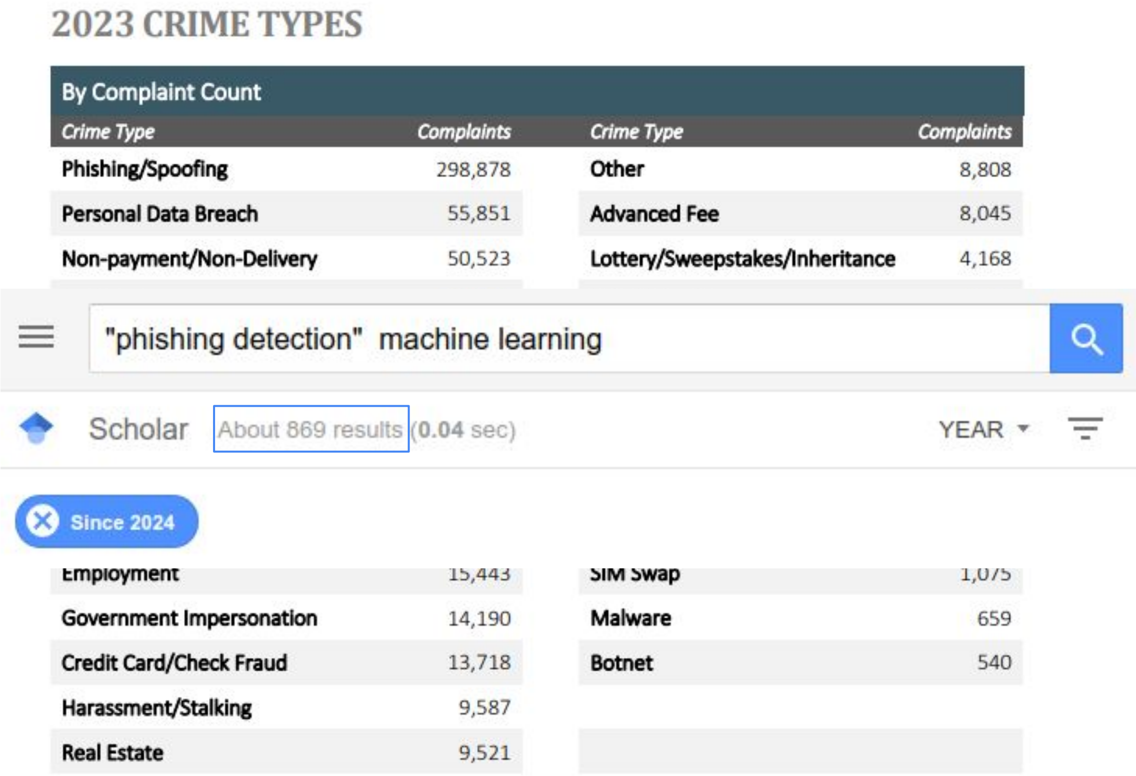
2023 CRIME TYPES

By Complaint Count

Crime Type	Complaints	Crime Type	Complaints
Phishing/Spoofing	298,878	Other	8,808
Personal Data Breach	55,851	Advanced Fee	8,045
Non-payment/Non-Delivery	50,523	Lottery/Sweepstakes/Inheritance	4,168
Extortion	48,223	Overpayment	4,144
Investment	39,570	Data Breach	3,727
Tech Support	37,560	Ransomware	2,825
BEC	21,489	Crimes Against Children	2,361
Identity Theft	19,778	Threats of Violence	1,697
Confidence/Romance	17,823	IPR/Copyright and Counterfeit	1,498
Employment	15,443	SIM Swap	1,075
Government Impersonation	14,190	Malware	659
Credit Card/Check Fraud	13,718	Botnet	540
Harassment/Stalking	9,587		
Real Estate	9,521		

FBI Internet Crime Report 2023

Background: phishing detection using machine learning



2023 CRIME TYPES

By Complaint Count

Crime Type	Complaints	Crime Type	Complaints
Phishing/Spoofing	298,878	Other	8,808
Personal Data Breach	55,851	Advanced Fee	8,045
Non-payment/Non-Delivery	50,523	Lottery/Sweepstakes/Inheritance	4,168
Employment	15,443	SIM Swap	1,075
Government Impersonation	14,190	Malware	659
Credit Card/Check Fraud	13,718	Botnet	540
Harassment/Stalking	9,587		
Real Estate	9,521		

"phishing detection" machine learning

Scholar About 869 results (0.04 sec)

FBI Internet Crime Report 2023

The Dataset

- WikiPhish is a new open-access dataset for phishing website classification

- 110,606 URLs, HTML web pages, and screenshots

- Benign samples: **Wikipedia references**

- Phishing samples: OpenPhish, and PhishTank

WikiPhish Advantages

Diversity

Wide variety of benign and phishing samples

Up-to-date samples

Recent phishing campaigns and benign documents

Transparency

Reproducible and updatable

Data Collection Challenges

Data Collection Challenges

- Evolving ecosystem due to concept drift
  - Phishing detection is truly adversarial

Data Collection Challenges

- Evolving ecosystem due to concept drift
  - Phishing detection is truly adversarial

- Phishing websites are often short-lived
  - Phishers quickly remove their websites
  - URLs are quickly blocked by providers

Data Collection Challenges

- Evolving ecosystem due to concept drift
  - Phishing detection is truly adversarial

- Phishing websites are often short-lived
  - Phishers quickly remove their websites
  - URLs are quickly blocked by providers

- Collecting relevant legitimate documents is difficult
  - New frameworks, development practices

Limitation of Existing Datasets

Limitation of Existing Datasets

- Many datasets are limited to URLs, omitting HTML and visual content
  - Limited features for machine learning

Limitation of Existing Datasets

- Many datasets are limited to URLs, omitting HTML and visual content
    - Limited features for machine learning

- Few datasets serve as benchmarks for phishing detection
    - Studies are conducted using their own dataset

Limitation of Existing Datasets

- Many datasets are limited to URLs, omitting HTML and visual content
  - Limited features for machine learning

- Few datasets serve as benchmarks for phishing detection
  - Studies are conducted using their own dataset

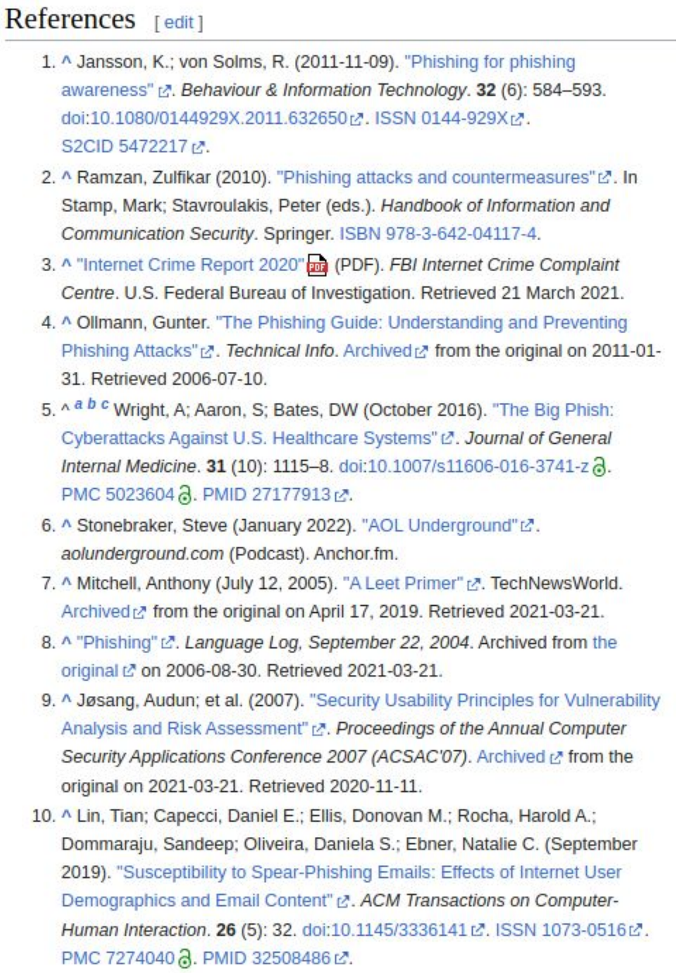
- Standardization is limited, making comparison difficult
  - Dataset sizes, collection protocol, and time periods are different
  - Resulting machine learning models are hardly comparable

WikiPhish Collection (benign)

- Extract URLs from random Wikipedia page references

- Add login page by appending “*/login*” or “*/signin*” subpath

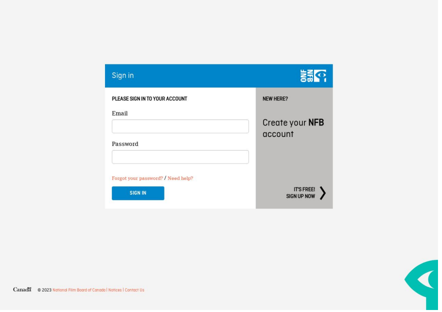
- Includes lesser-known and "out-of-distribution" websites.



References [edit]

1. ^ Jansson, K.; von Solms, R. (2011-11-09). "Phishing for phishing awareness". *Behaviour & Information Technology*. **32** (6): 584–593. doi:10.1080/0144929X.2011.632650. ISSN 0144-929X. S2CID 5472217.
2. ^ Ramzan, Zulfikar (2010). "Phishing attacks and countermeasures". In Stamp, Mark; Stavroulakis, Peter (eds.). *Handbook of Information and Communication Security*. Springer. ISBN 978-3-642-04117-4.
3. ^ "Internet Crime Report 2020" (PDF). *FBI Internet Crime Complaint Centre*. U.S. Federal Bureau of Investigation. Retrieved 21 March 2021.
4. ^ Ollmann, Gunter. "The Phishing Guide: Understanding and Preventing Phishing Attacks". *Technical Info*. Archived from the original on 2011-01-31. Retrieved 2006-07-10.
5. ^ *a b c* Wright, A; Aaron, S; Bates, DW (October 2016). "The Big Phish: Cyberattacks Against U.S. Healthcare Systems". *Journal of General Internal Medicine*. **31** (10): 1115–8. doi:10.1007/s11606-016-3741-z. PMC 5023604. PMID 27177913.
6. ^ Stonebraker, Steve (January 2022). "AOL Underground". *aolunderground.com* (Podcast). Anchor.fm.
7. ^ Mitchell, Anthony (July 12, 2005). "A Leet Primer". TechNewsWorld. Archived from the original on April 17, 2019. Retrieved 2021-03-21.
8. ^ "Phishing". *Language Log, September 22, 2004*. Archived from the original on 2006-08-30. Retrieved 2021-03-21.
9. ^ Jøsang, Audun; et al. (2007). "Security Usability Principles for Vulnerability Analysis and Risk Assessment". *Proceedings of the Annual Computer Security Applications Conference 2007 (ACSAC'07)*. Archived from the original on 2021-03-21. Retrieved 2020-11-11.
10. ^ Lin, Tian; Capecci, Daniel E.; Ellis, Donovan M.; Rocha, Harold A.; Dommaraju, Sandeep; Oliveira, Daniela S.; Ebner, Natalie C. (September 2019). "Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content". *ACM Transactions on Computer-Human Interaction*. **26** (5): 32. doi:10.1145/3336141. ISSN 1073-0516. PMC 7274040. PMID 32508486.

WikiPhish Collection (benign screenshots)



Sign in

PLEASE SIGN IN TO YOUR ACCOUNT

Email

Password

Forgot your password? / Need help?

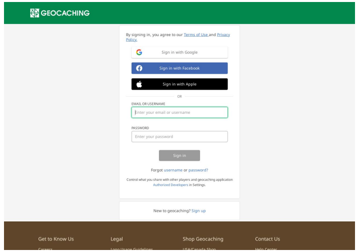
SIGN IN

NEW HERE?

Create your **NFB** account

IT'S FREE! SIGN UP NOW

Canada © 2023 National Film Board of Canada | Notices | Contact Us



GEOCACHING

By signing in, you agree to our Terms of Use and Privacy Policy.

Sign in with Google

Sign in with Facebook

Sign in with Apple

OR

EMAIL OR USERNAME: Enter your email or username

PASSWORD: Enter your password

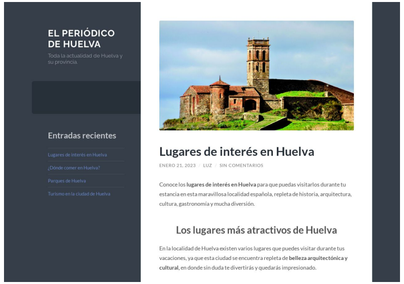
Sign in

Forgot username or password?

Control what you share with other players and geocaching application Authorized Developers in Settings.

New to geocaching? Sign up

Get to Know Us | Legal | Shop Geocaching | Contact Us



EL PERIÓDICO DE HUELVA

Toda la actualidad de Huelva y su provincia.

Entradas recientes

- Lugares de interés en Huelva
- ¿Dónde comer en Huelva?
- Parques de Huelva
- Turismo en la ciudad de Huelva

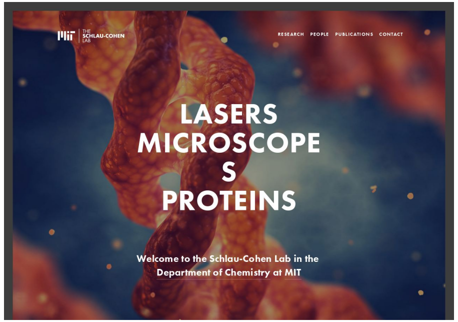
Lugares de interés en Huelva

ENERO 21, 2023 / LUZ / SIN COMENTARIOS

Conoce los **lugares de interés en Huelva** para que puedas visitarlos durante tu estancia en esta maravillosa localidad española, repleta de historia, arquitectura, cultura, gastronomía y mucha diversión.

Los lugares más atractivos de Huelva

En la localidad de Huelva existen varios lugares que puedes visitar durante tus vacaciones, ya que esta ciudad se encuentra repleta de **belleza arquitectónica y cultural**, en donde sin duda te divertirás y quedarás impresionado.

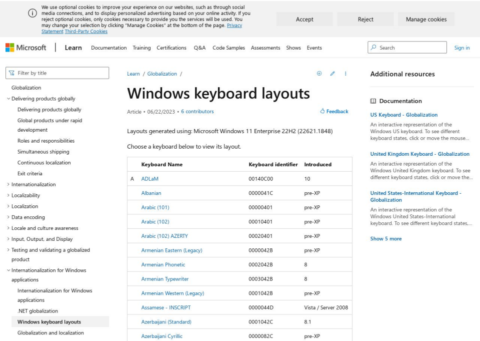


THE SCHLAU-COHEN LAB

RESEARCH PEOPLE PUBLICATIONS CONTACT

LASERS
MICROSCOPE
S
PROTEINS

Welcome to the Schlau-Cohen Lab in the Department of Chemistry at MIT



We use optional cookies to improve your experience on our websites, such as through social media connections, and to display personalized advertising based on your online activity. If you reject optional cookies, only cookies necessary to provide you the services will be used. You may change your selection by clicking "Manage Cookies" at the bottom of the page. Privacy Statement Third-Party Cookies

Accept | Reject | Manage cookies

Microsoft | Learn Documentation Training Certifications Q&A Code Samples Assessments Shows Events Search Sign in

Filter by title

- Globalization
- Delivering products globally
  - Delivering products globally
  - Global products under rapid development
  - Roles and responsibilities
  - Simultaneous shipping
  - Continuous localization
  - Exit criteria
- Internationalization
- Localizability
- Localization
- Data encoding
- Locale and culture awareness
- Input, Output, and Display
- Testing and validating a globalized product
- Internationalization for Windows applications
  - Internationalization for Windows applications
  - .NET globalization
  - Windows keyboard layouts
  - Globalization and localization

Learn / Globalization /

Windows keyboard layouts

Article • 06/22/2023 • 6 contributors Feedback

Layouts generated using: Microsoft Windows 11 Enterprise 22H2 (22621.1848)

Choose a keyboard below to view its layout.

	Keyboard Name	Keyboard identifier	Introduced
A	ADLaM	00140C00	10
	Albanian	0000041C	pre-XP
	Arabic (101)	00000401	pre-XP
	Arabic (102)	00010401	pre-XP
	Arabic (102) AZERTY	00020401	pre-XP
	Armenian Eastern (Legacy)	0000042B	pre-XP
	Armenian Phonetic	0002042B	8
	Armenian Typewriter	0003042B	8
	Armenian Western (Legacy)	0001042B	pre-XP
	Assamese - INSCRIPT	0000044D	Vista / Server 2008
	Azerbaijani (Standard)	0001042C	8.1
	Azerbaijani Cyrillic	0000082C	pre-XP

Additional resources

Documentation

US Keyboard - Globalization

An interactive representation of the Windows US keyboard. To see different keyboard states, click or move the mouse...

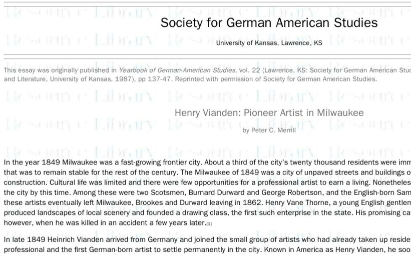
United Kingdom Keyboard - Globalization

An interactive representation of the Windows United Kingdom keyboard. To see different keyboard states, click or move the...

United States-International Keyboard - Globalization

An interactive representation of the Windows United States-International keyboard. To see different keyboard states,...

Show 5 more



Society for German American Studies

University of Kansas, Lawrence, KS

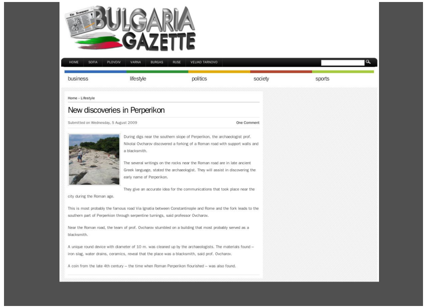
This essay was originally published in *Yearbook of German-American Studies*, vol. 22 (Lawrence, KS: Society for German American Stuc and Literature, University of Kansas, 1987), pp 137-47. Reprinted with permission of Society for German American Studies.

Henry Vianden: Pioneer Artist in Milwaukee

by Peter C. Merrill

In the year 1849 Milwaukee was a fast-growing frontier city. About a third of the city's twenty thousand residents were imn that was to remain stable for the rest of the century. The Milwaukee of 1849 was a city of unpaved streets and buildings o construction. Cultural life was limited and there were few opportunities for a professional artist to earn a living. Nonetheles the city by this time. Among these were two Scotsmen, Burnard Durward and George Robertson, and the English-born Sam these artists eventually left Milwaukee, Brookes and Durward leaving in 1862. Henry Vane Thorne, a young English gentlen produced landscapes of local scenery and founded a drawing class, the first such enterprise in the state. His promising ca however, when he was killed in an accident a few years later.[1]

In late 1849 Heinrich Vianden arrived from Germany and joined the small group of artists who had already taken up reside professional and the first German-born artist to settle permanently in the city. Known in America as Henry Vianden, he soo



BULGARIA GAZETTE

HOME SOFIA PLOVDIV VARNA BURGAS RUSE VELIKO TARNOVO

business lifestyle politics society sports

Home » Lifestyle

New discoveries in Perperikon

Submitted on Wednesday, 5 August 2009 One Comment

During digs near the southern slope of Perperikon, the archaeologist prof. Nikolai Ovcharov discovered a forking of a Roman road with support walls and a blacksmith.

The several writings on the rocks near the Roman road are in late ancient Greek language, stated the archaeologist. They will assist in discovering the early name of Perperikon.

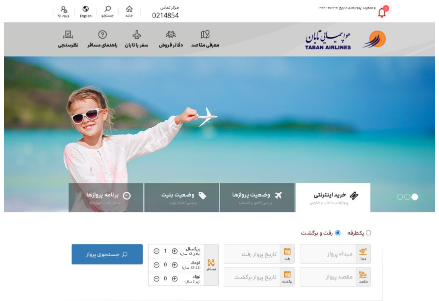
They give an accurate idea for the communications that took place near the city during the Roman age.

This is most probably the famous road Via Ignatia between Constantinople and Rome and the fork leads to the southern part of Perperkion through serpentine turnings, said professor Ovcharov.

Near the Roman road, the team of prof. Ovcharov stumbled on a building that most probably served as a blacksmith.

A unique round device with diameter of 10 m. was cleaned up by the archaeologists. The materials found – iron slag, water drains, ceramics, reveal that the place was a blacksmith, said prof. Ovcharov.

A coin from the late 4th century – the time when Roman Perperikon flourished – was also found.



مرکز تماس 0214854

خانه | جستجو | English | ورود به

هواپیمایی تابان TABAN AIRLINES

معرفی مقاصد | دفاتر فروش | سفر با تابان | راهنمای مسافر | نظرسنجی

خرید اینترنتی | وضعیت پروازها | وضعیت بلیت | برنامه پروازها

○ یکطرفه ● رفت و برگشت

مبداء پرواز | تاریخ پرواز رفت

مقصد پرواز | تاریخ پرواز برگشت

بزرگسال (بالای 12 سال) 1

کودک (2 تا 12 سال) 0

نوزاد (زیر 2 سال) 0

جستجوی پرواز

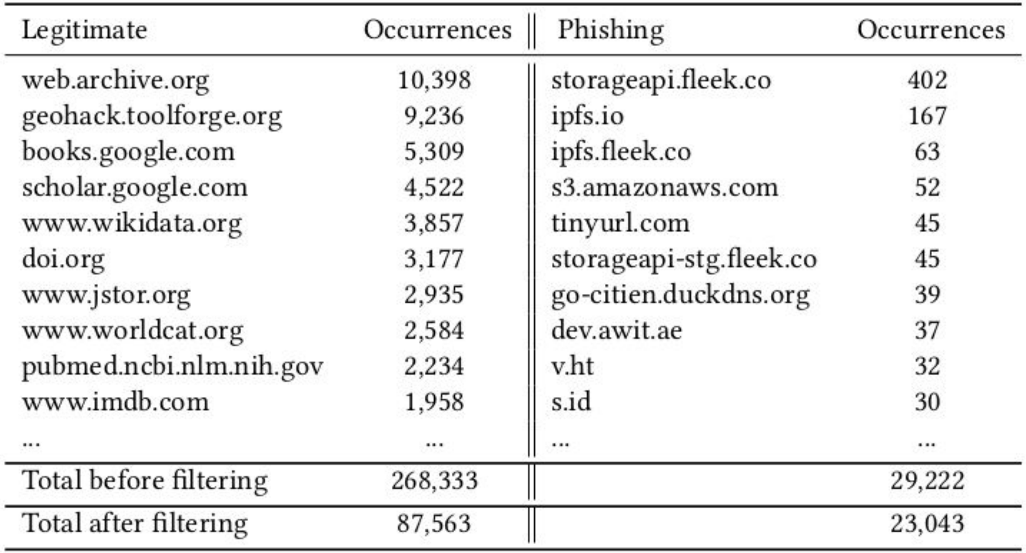
WikiPhish Collection (phishing)

- We monitor the public phishing databases OpenPhish and PhishTank from January 11, 2023, to October 22, 2023

- We collect every page as soon as they are listed, and only add them to the dataset if they are further validated by moderators

WikiPhish Collection (preprocessing)

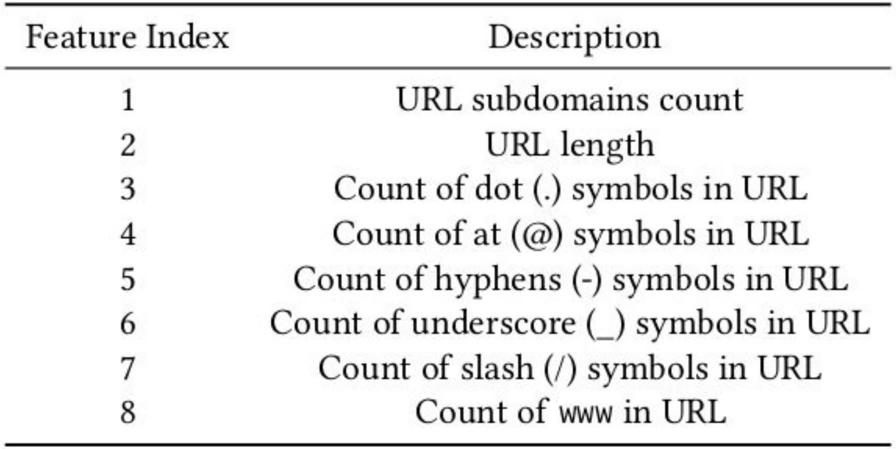
- We limit the number of items per Fully Qualified Domain Name (FQDN) to 10 to limit redundancy and enhance diversity



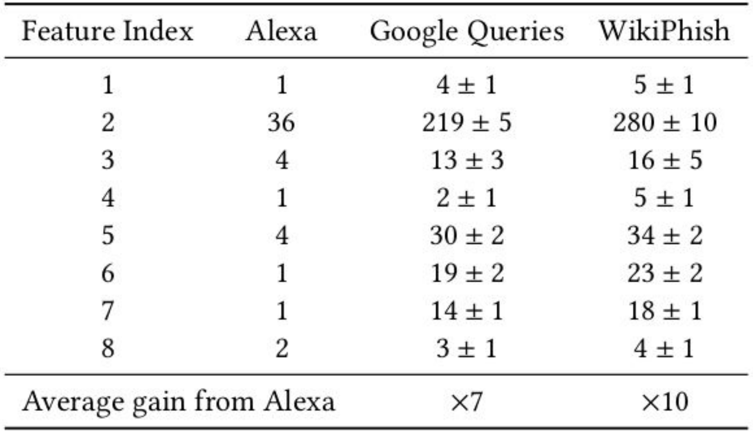
Legitimate	Occurrences	Phishing	Occurrences
web.archive.org	10,398	storageapi.fleek.co	402
geohack.toolforge.org	9,236	ipfs.io	167
books.google.com	5,309	ipfs.fleek.co	63
scholar.google.com	4,522	s3.amazonaws.com	52
www.wikidata.org	3,857	tinyurl.com	45
doi.org	3,177	storageapi-stg.fleek.co	45
www.jstor.org	2,935	go-citien.duckdns.org	39
www.worldcat.org	2,584	dev.awit.ae	37
pubmed.ncbi.nlm.nih.gov	2,234	v.ht	32
www.imdb.com	1,958	s.id	30
...	...	...	...
Total before filtering	268,333		29,222
Total after filtering	87,563		23,043

Importance of Legitimate URL Diversity

- Comparison with collection strategies of two other datasets (Apruzzese et al. 2022, Ariyadasa et al. 2021)
  - Top, middle, and bottom part of **Alexa**
  - Keywords used in Google search engine: “**Google Queries**”



Feature Index	Description
1	URL subdomains count
2	URL length
3	Count of dot (.) symbols in URL
4	Count of at (@) symbols in URL
5	Count of hyphens (-) symbols in URL
6	Count of underscore (_) symbols in URL
7	Count of slash (/) symbols in URL
8	Count of www in URL

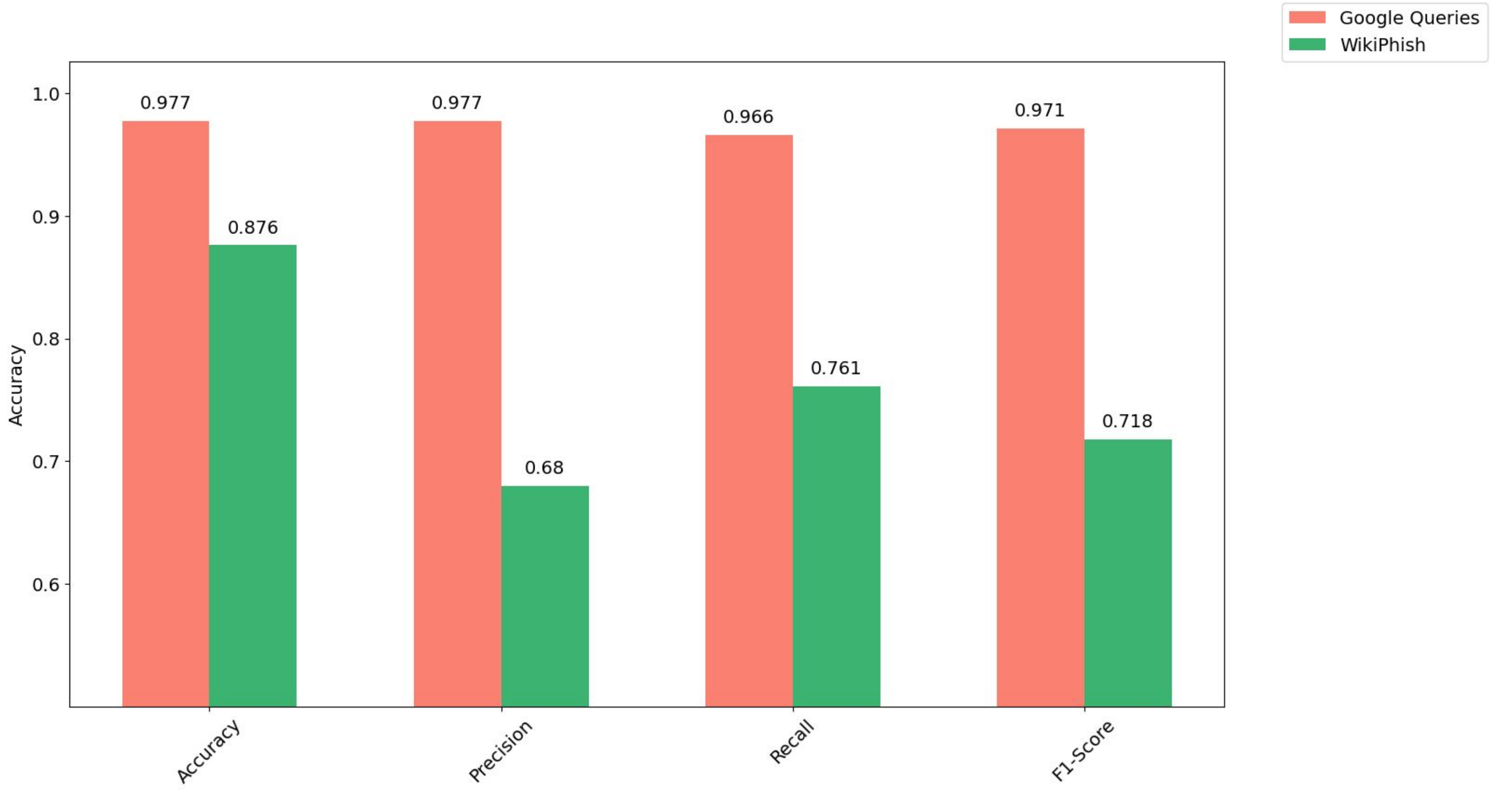


Feature Index	Alexa	Google Queries	WikiPhish
1	1	4 ± 1	5 ± 1
2	36	219 ± 5	280 ± 10
3	4	13 ± 3	16 ± 5
4	1	2 ± 1	5 ± 1
5	4	30 ± 2	34 ± 2
6	1	19 ± 2	23 ± 2
7	1	14 ± 1	18 ± 1
8	2	3 ± 1	4 ± 1
Average gain from Alexa		×7	×10

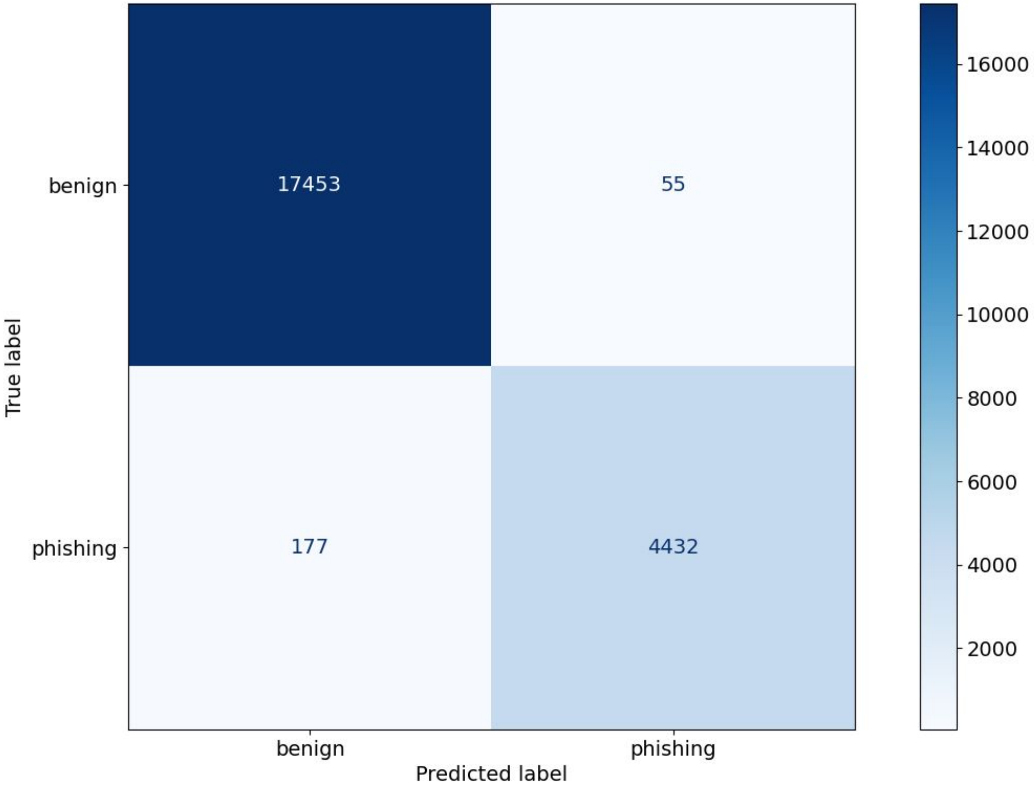
Training and Evaluating Classifiers

- Training and evaluation of a Random Forest model
- 8 URL features, and 11 HTML features
    - 1) Train on older datasets, evaluate on WikiPhish
    - 2) Train and evaluate on WikiPhish

Train on older datasets, evaluate on WikiPhish



Train and evaluate on WikiPhish



Conclusion

- WikiPhish is a new dataset for phishing website detection
- It emphasizes diversity, up-to-date samples, and transparency
- It provides a valuable benchmark for research and development of new phishing detectors

Thank you!