Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages

Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, Jin Song Dong

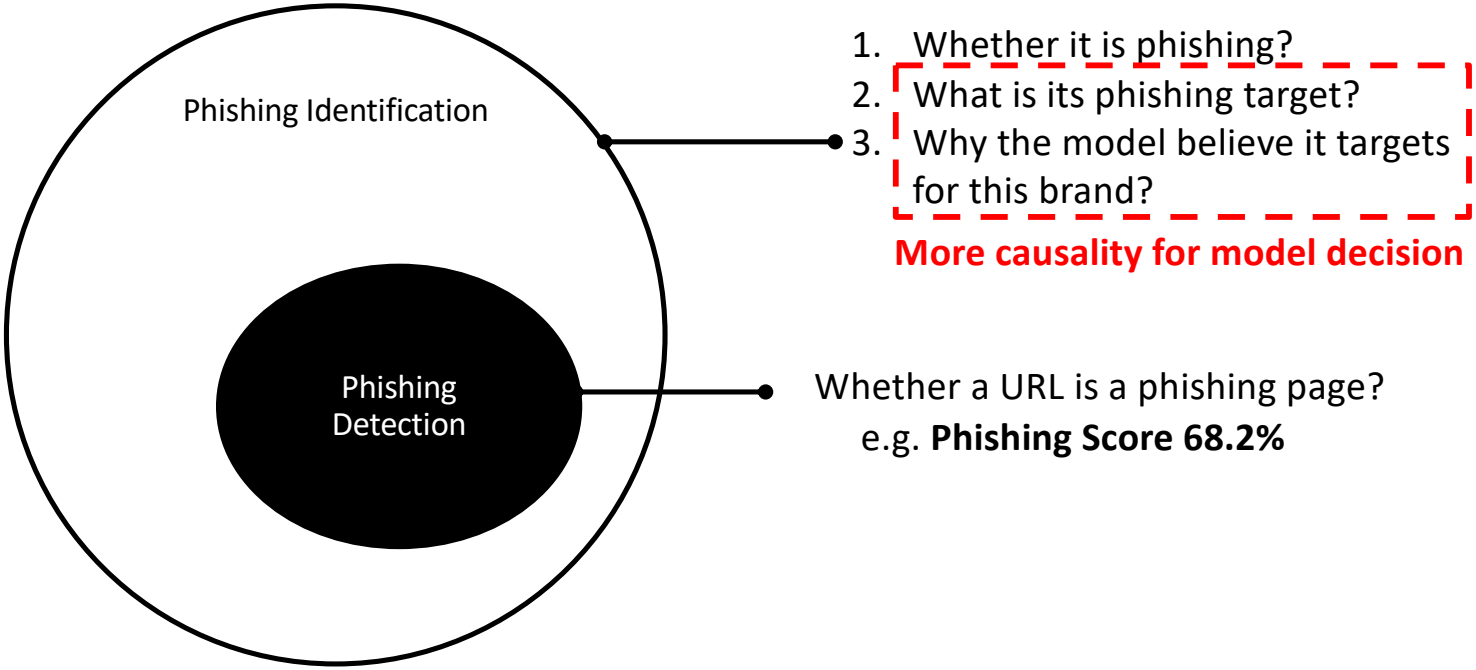
Presenter: Ruofan Liu

Introduction: Phishing Attack

- Definition:
  - A type of cybercrime which steal users credential information by **disguising itself as legitimate entity**

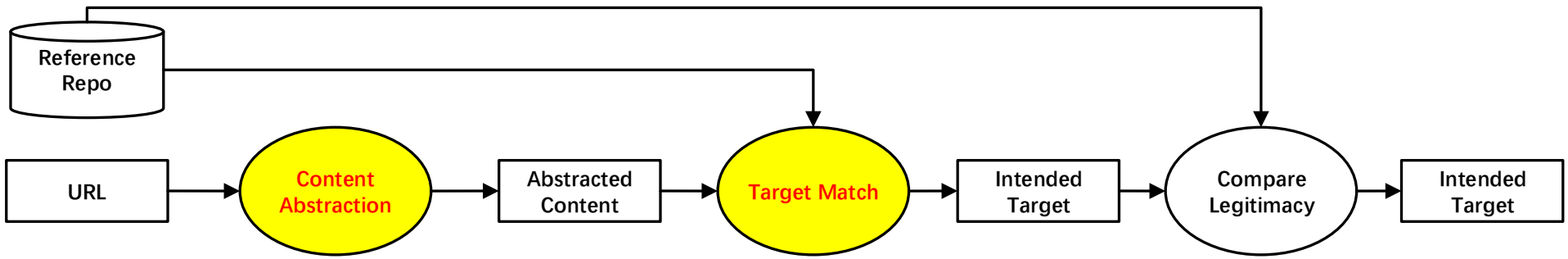
Research Problem

- Accurately **identify** phishing **websites**



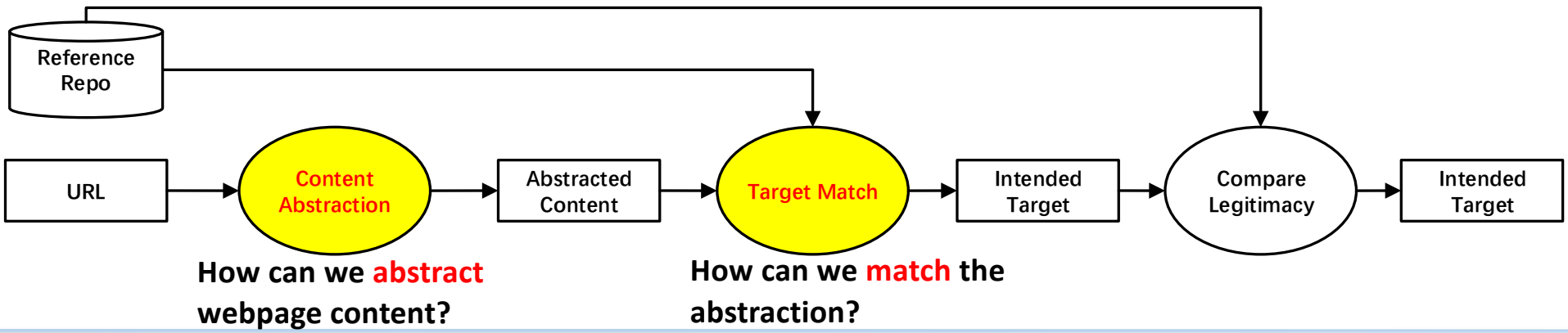
Typical Framework for Phishing Identification

- Key components:
  - A reference repo to store targeted brands webpage signature
  - An approach to abstract webpage content
  - An approach to compare with reference repo

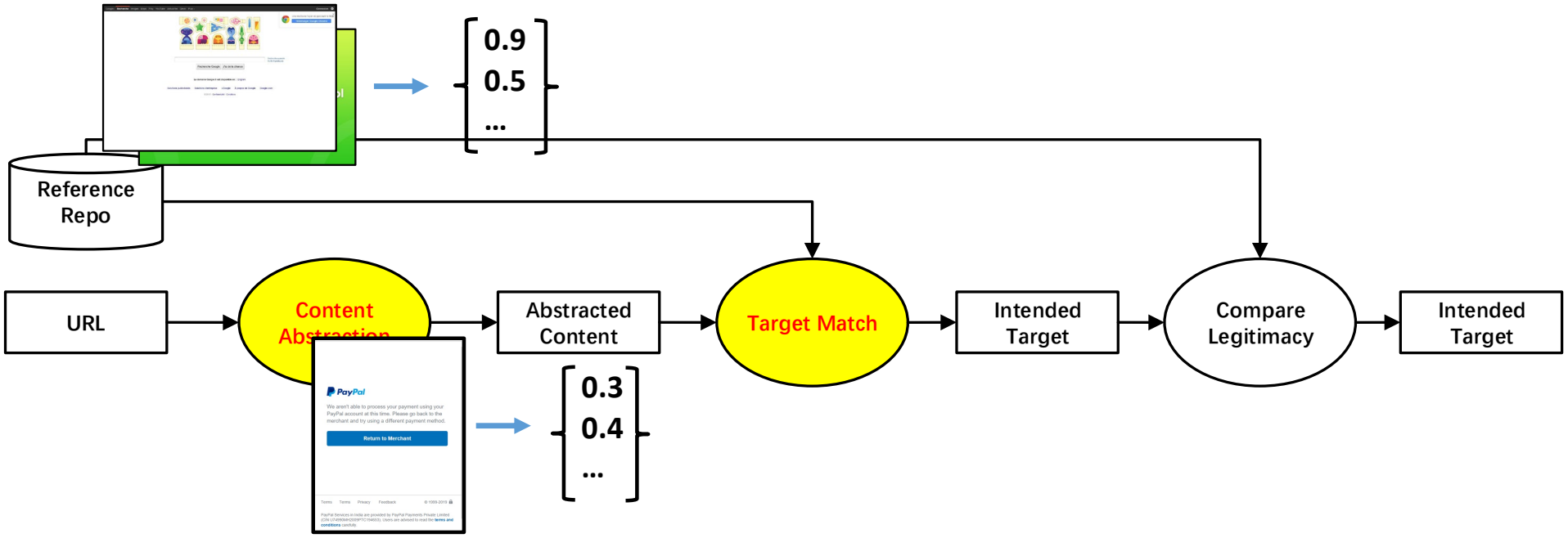


Typical Framework for Phishing Identification

- Key components:
  - A reference repo to store targeted brands webpage signature
  - An approach to abstract webpage content
  - An approach to compare with reference repo



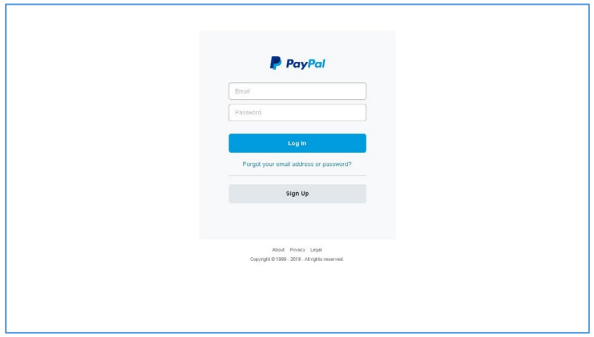
Existing Work 1: EMD [1]



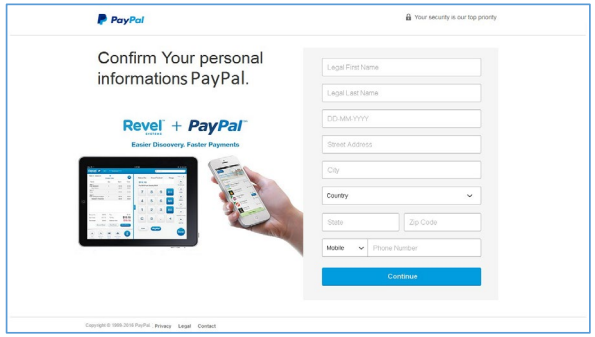
[1] Fu, A. Y., et al. (2006). "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)." IEEE Transactions on Dependable and Secure Computing **3**(4): 301-311.

Problem of EMD

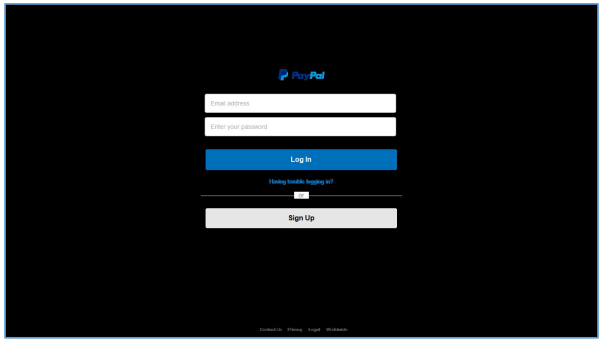
- EMD compares colour distribution of two screenshots
  - Screenshot usually changes from time to time, causing a lot of false negatives



Reference screenshot



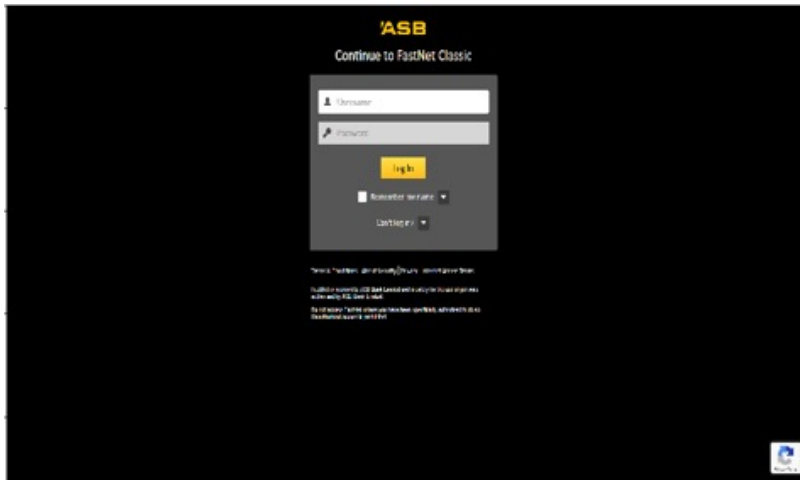
False negatives (change in layout)



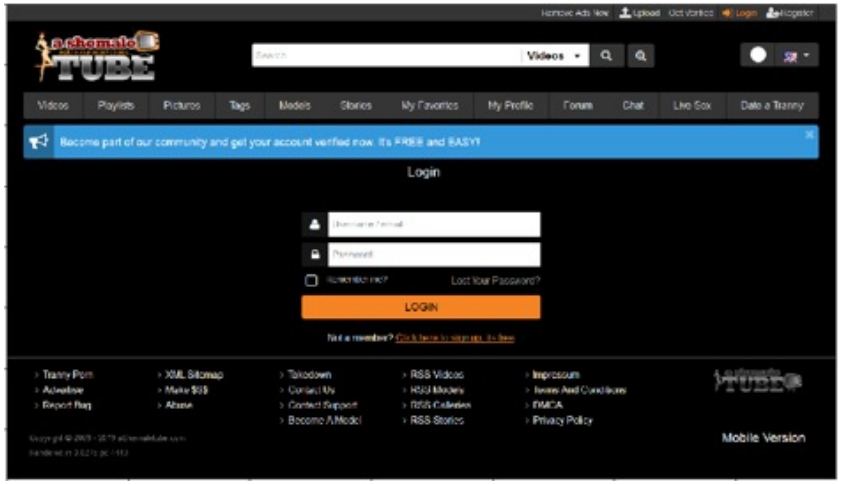
False negatives (change in colour scheme)

Problem of EMD

- EMD compares colour distribution of two screenshots
  - Report false positives as long as colour distributions are similar

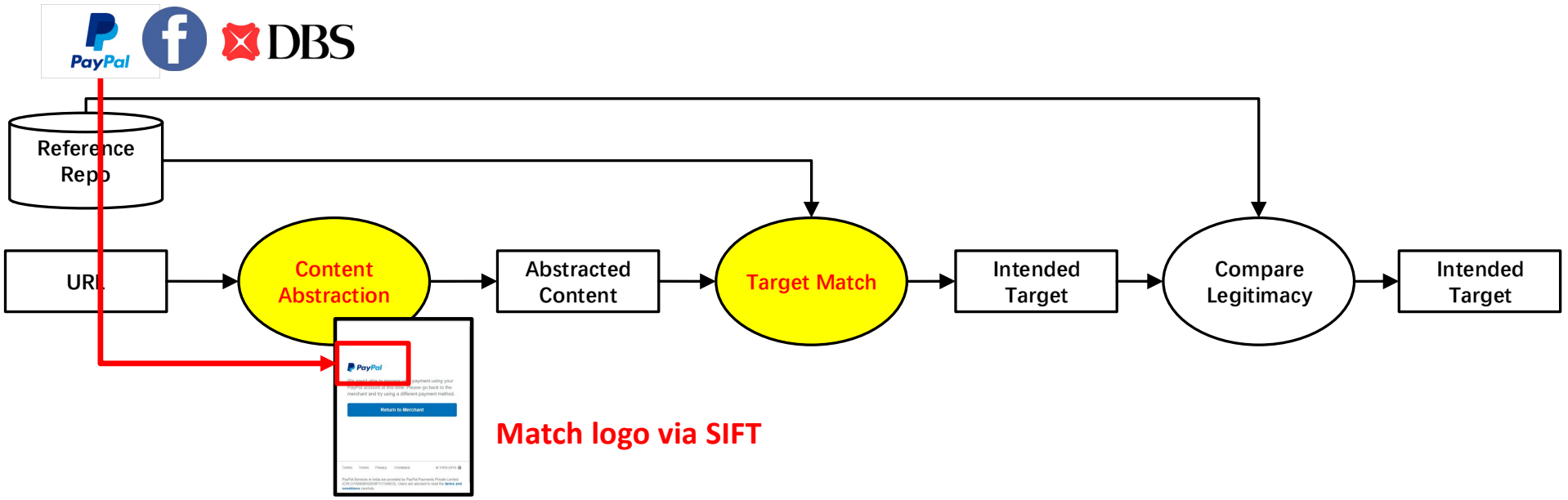


Reference screenshot



False positive

Existing Work 2: PhishZoo[2]

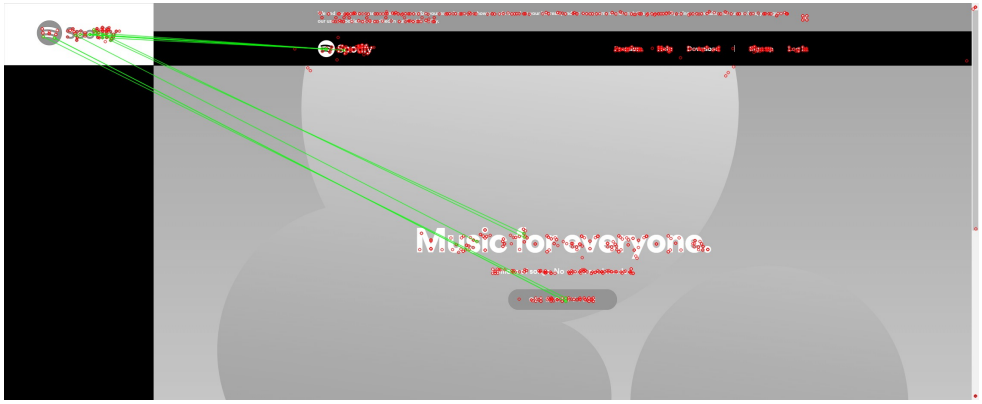


[2] Afroz, S. and R. Greenstadt (2011). PhishZoo: Detecting Phishing Websites by Looking at Them. 2011 IEEE Fifth International Conference on Semantic Computing.

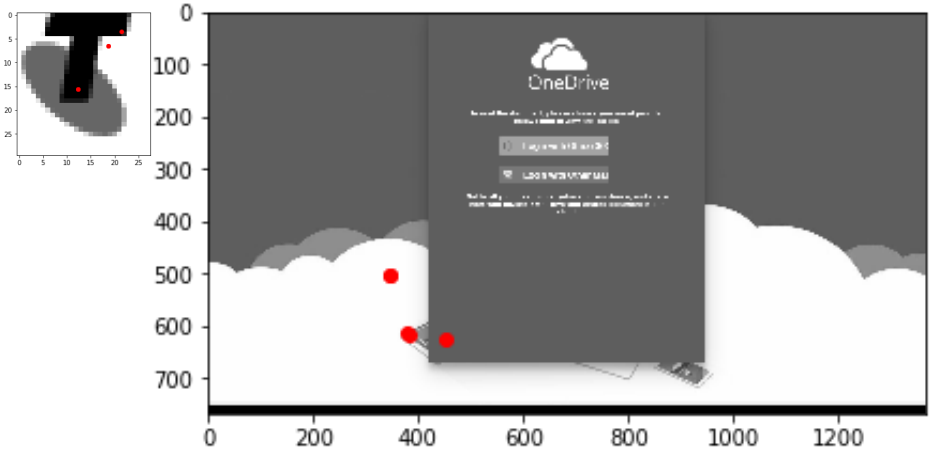
Problem of PhishZoo

- SIFT extract feature points to match a logo on screenshot, but with miss and incorrect matching

Matching ratio: 0.4, miss-match



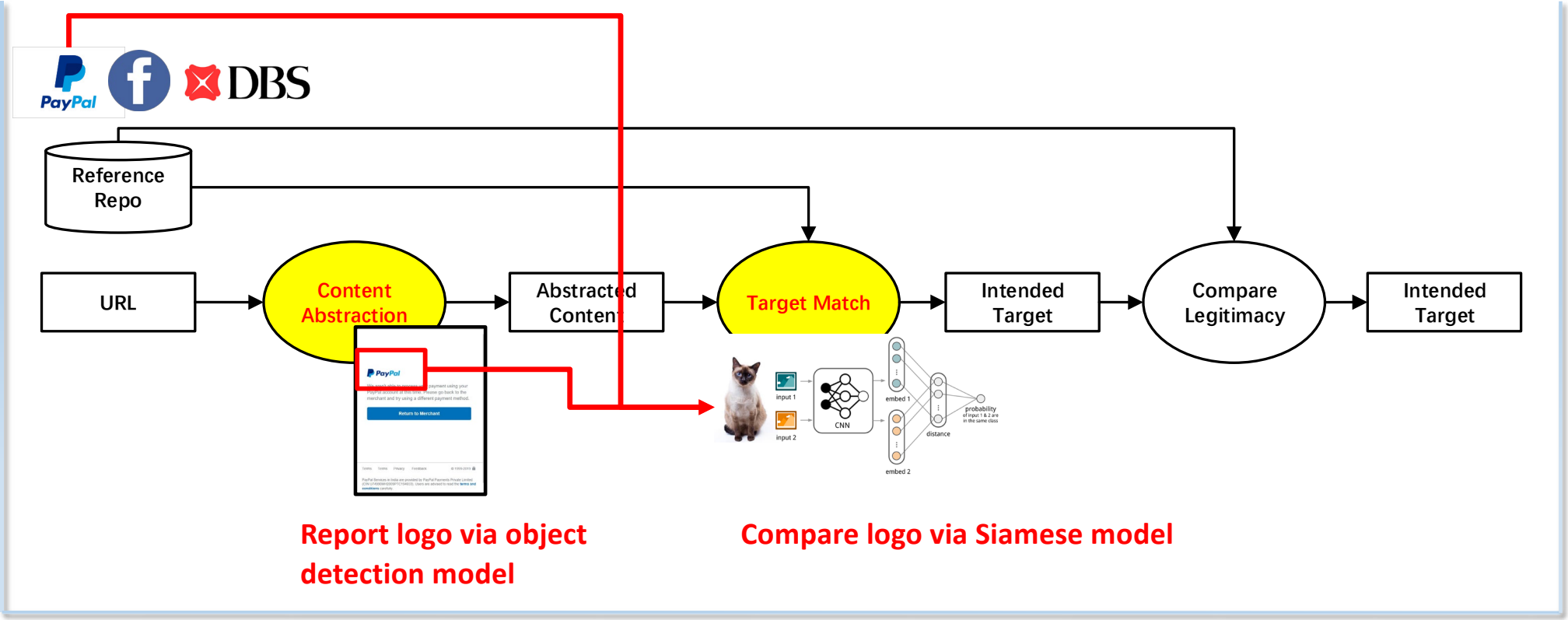
Matching ratio: 0.83, incorrect-match



Problem of PhishZoo

- SIFT extract feature points to match a logo on screenshot, but with long running time (On average 18 seconds per URL webpage)

Phishpedia Approach Overview



Phishpedia: Explainable Annotation

Phishpedia

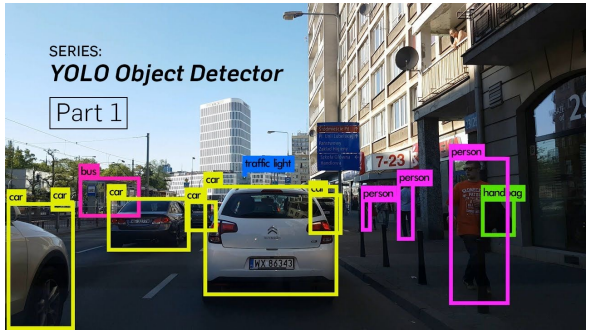
- **Perfect logo identification** with customized object detection approach.
- **Low runtime overhead** even with visual analysis.
- **Low bias in phishing dataset**, no need to train on any pre-collected phishing datasets.
- **Visualization annotation** on phishing screenshot, improving user confidence.

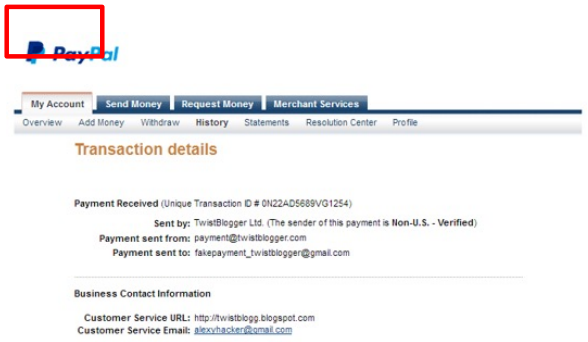
Challenges of Phishpedia

- Identifying perfect logo
- Training model to learn similar logos: representation learning

C1: Identifying perfect logo

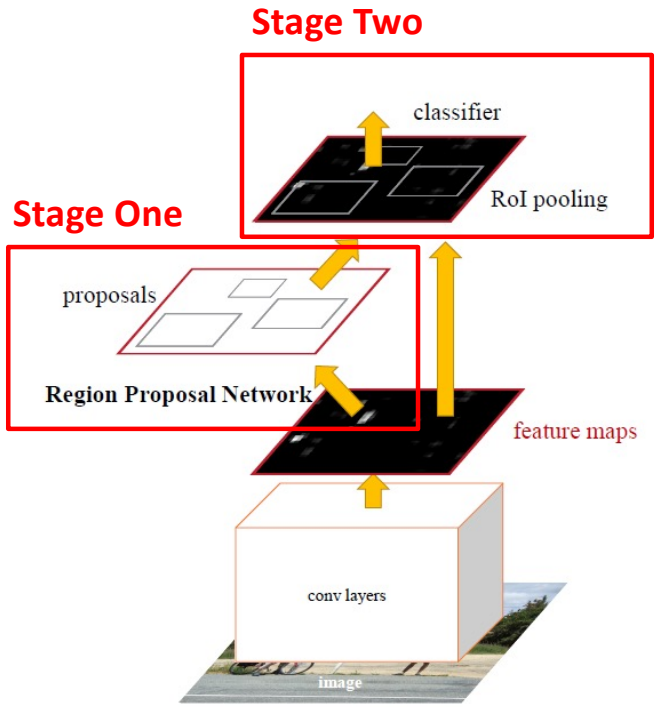
- Logo detection as an object detection problem.
- Off-the-shelf one-stage object detection face challenges of sometimes **generating only partial logo**.





Two-stage Object Detection Model

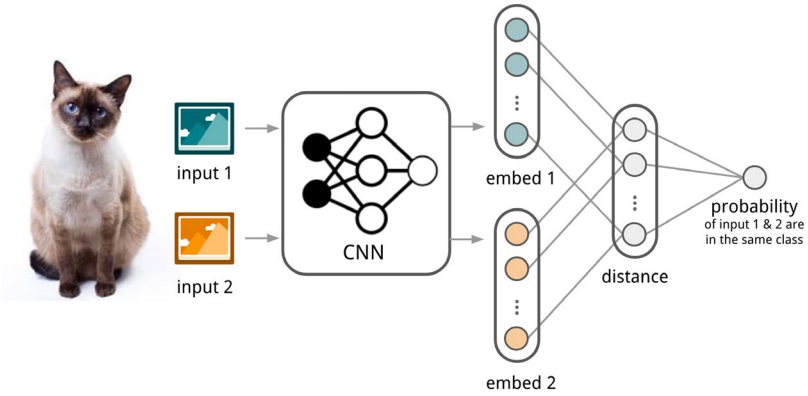
- Stage one:
  - Detect **region** containing objects (i.e., logo and input) on the screenshot first.
- Stage two:
  - Refine the reported region
  - Classify the region content
    - Logo
    - Input

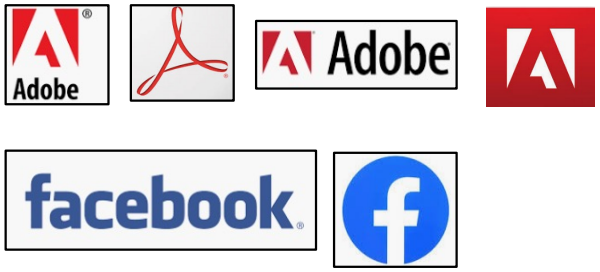


[3] Ren, S., et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems.

C2: Training model to learn similar logo

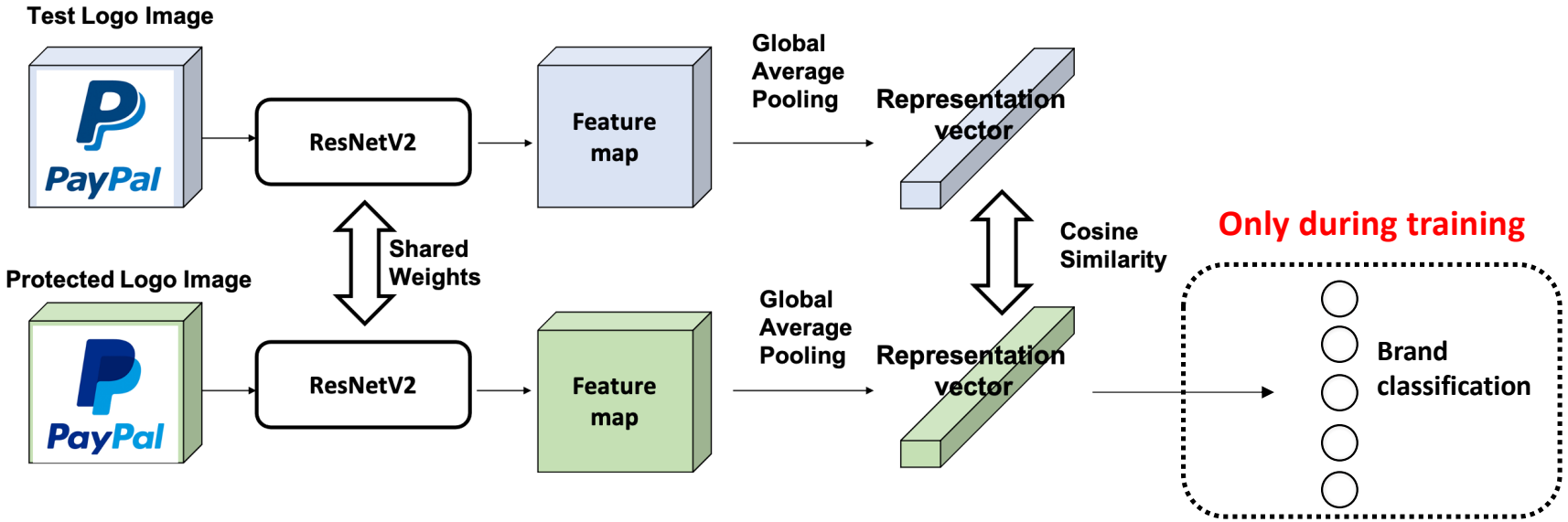
- Traditional way of training Siamese model does not work.
  - Logo variants are very different to generalize.
  - Forcing the model to train different logo variants end up overfitting





Training Siamese Model

- Train model in classification way
- Use the model in Siamese style

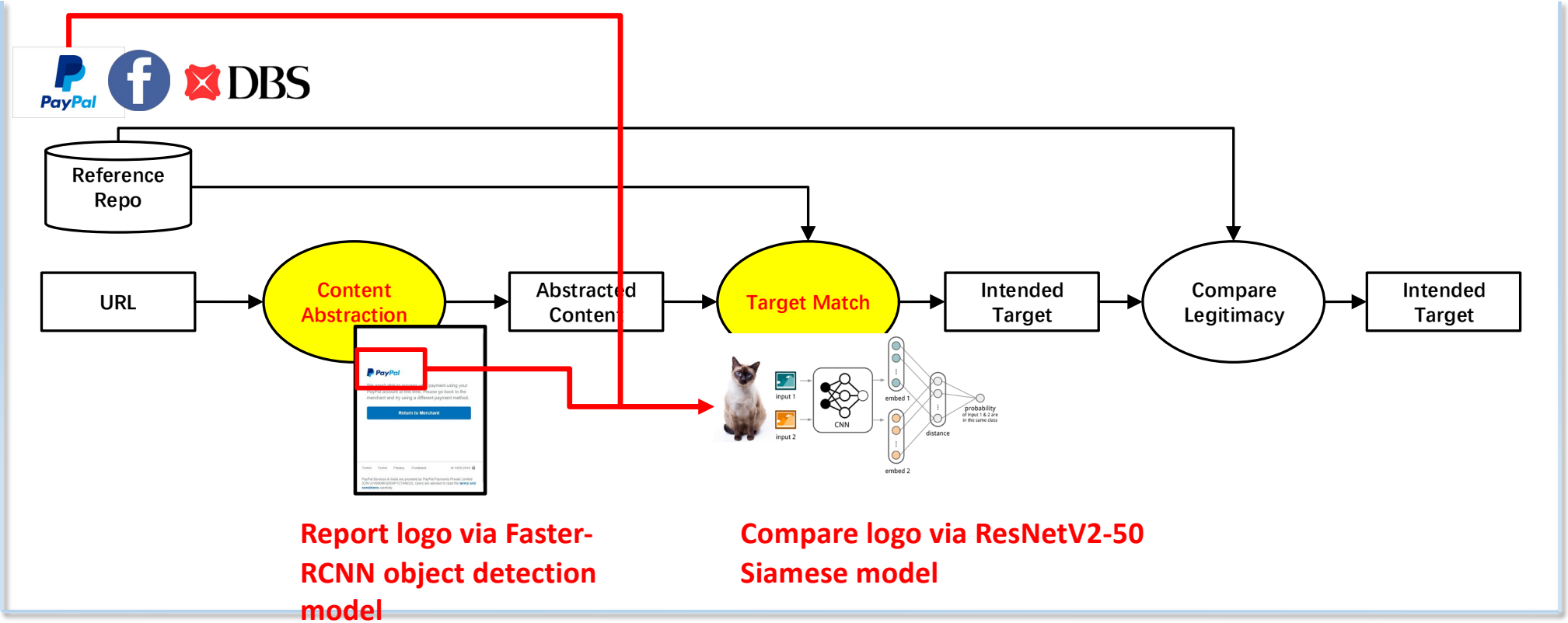


Training Siamese Model

- Use ResnetV2-50 [4] model architecture.
- Apply two-stage training:
  - Stage 1: pretrain a ResnetV2-50 classification model to classify 2341 brand of logos on Logo2K+ dataset [5].
  - Stage 2: finetune the pretrained model to classify 181 target brands.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630-645). Springer, Cham.
[5] Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., Wang, H., & Jiang, S. (2020). Logo-2K+: A large-scale logo dataset for scalable logo classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 6194-6201).

Phishpedia Approach Overview



Experiment

- RQ1: Whether Phishpedia can identify the phishing URL effectively?
- RQ2: How good is the performance for individual component (i.e. Logo detector and Siamese)?
- RQ3: Whether Phishpedia can catch more phishing in the wild?

Experiment Setup

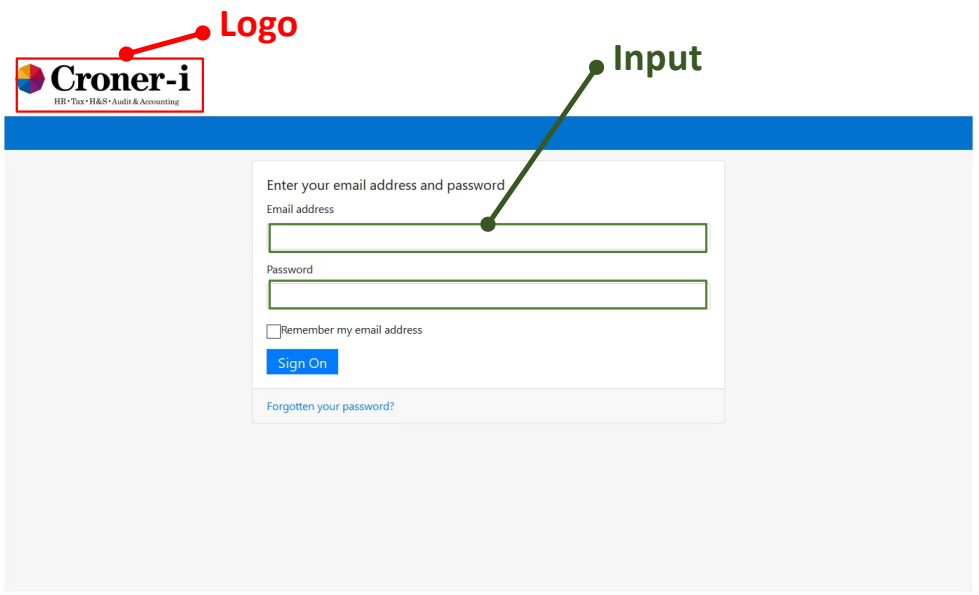
- 29,496 phishing webpages
  - From Openphish
- 29,951 benign webpages
  - Top Alexa
- For each phishing/benign page, we collect
  - Meta-data (i.e. phishing target)
  - Screenshot
  - HTML code

Experiment Setup

- Reference logo database
  - Cover top 181 brands that are mostly attacked from Openphish data
  - Used in Siamese training

Experiment Setup

- 30649 labelled legitimate website (logo & input box annotation)
  - Used in object detector training



RQ1: Overall Phishpedia Performance

Approach	Detection Accuracy		Identification Accuracy	Runtime Overhead (s)
	Precision	Recall		
EMD (*TDSC'06*)	52.0%	76.2%	27.7%	0.19
PhishZoo (*ICSC'11*)	68.9%	81.8%	28.5%	18.2
LogoSense (*Computers & Security*)	20.5%	26.9%	37.8%	27.2
Phishpedia	**98.2%**	**87.1%**	**99.2%**	**0.19**

RQ2: Individual component performance

- Logo detector

Object Class	Logo	Input Boxes	Overall (mAP)
Training AP	52.7	73.5	63.1
Testing AP	49.3	70.0	59.7

- Siamese matching accuracy
  - 93.5%

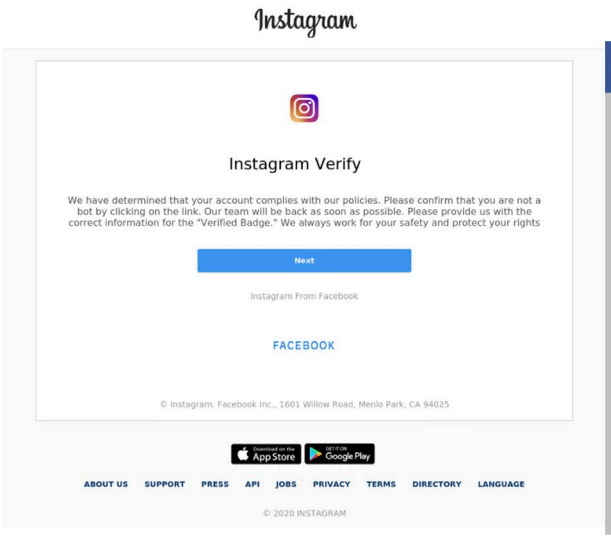
RQ3: Whether Phishpedia can catch more phishing in the wild?

- We deploy Phishpedia on emerging domains fed from CertStream for one month
- Phishpedia report **1820** as phishing, out of which **1704** are real phishing, **1133** are zero-day
- Other tools report too many false positives

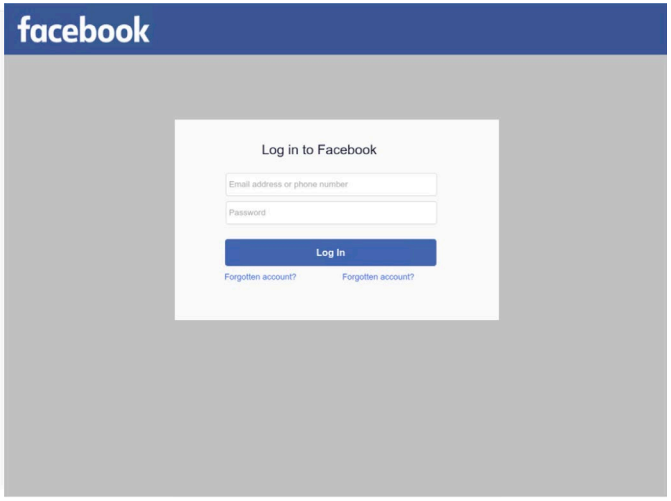
Tool	Category	#Reported Phishing	#Top Ranked Samples	#Real Phishing	#Zero-day Phishing
EMD	Identification	299,082	1000	3	2
Phishzoo	Identification	9,127	1000	8	5
Phishpeida	Identification	1,820	1000	939	623

RQ3: Whether Phishpedia can catch more phishing in the wild?

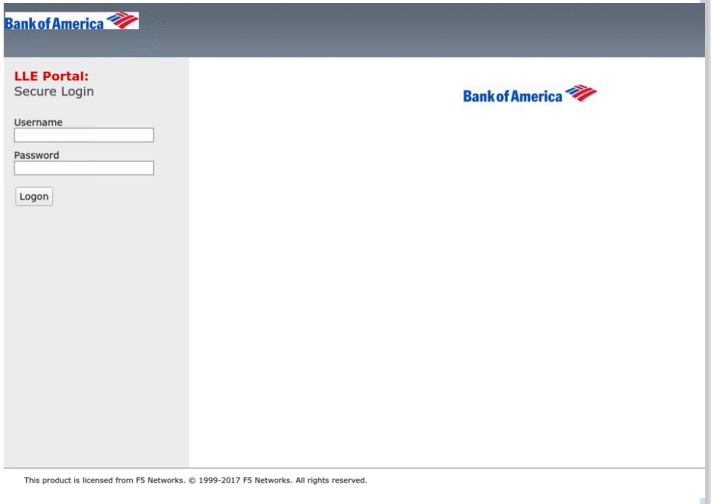
- Found phishing examples



verifiedbadgeforInstagram.ml



www.grab-pay.xyz



www.accountmgmtpl2.bamlqa.com

Takeaways

- Phishpedia is a phishing identification technique helping explain phishing causality.
  - **Visualization annotation**
- Technically, Phishpedia makes the following contributions:
  - **Perfect identity logo identification**
  - **Accurate logo matching model**
  - **Low runtime overhead**
  - **Low bias in phishing dataset**


THANK YOU!