NUS
National University
of Singapore

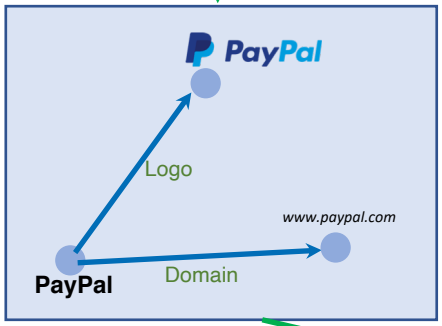
KnowPhish: Knowledge-Aware Deep Learning-Based Phishing Detection

28 June 2023

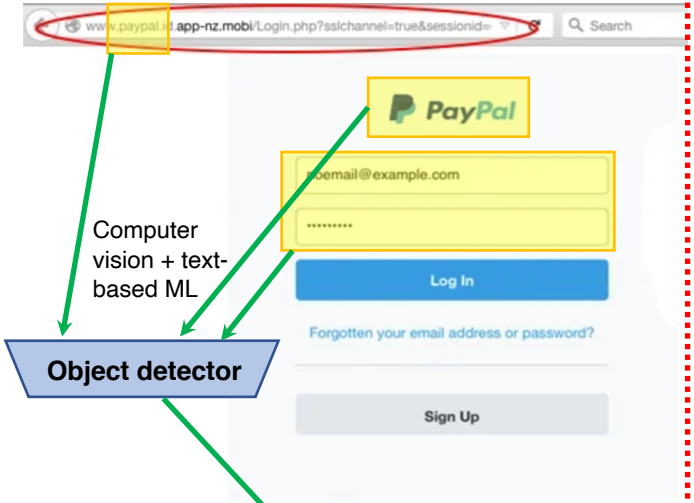
KnowPhish: Model

1. Knowledge Graph

Automated Knowledge Base Construction



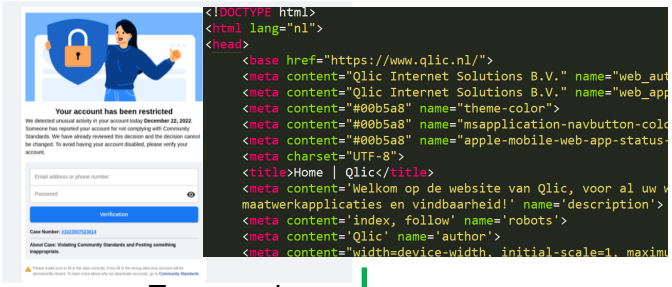
2. Visual Content



Computer vision + text-based ML

Object detector

3. Webpage HTML



Extracted text

```
16 [title] Restrictions Information
27 [div] Your account has been restricted
28 [div] We detected unusual activity in your account today
46 [button] Verification
47 [div] Case Number:
48 [a] #1023007523814
50 [div] About Case: Violating Community Standards and Posti
54 [div] Please make sure to fill in the data correctly, if
56 [font] Community Standards
```

Language model

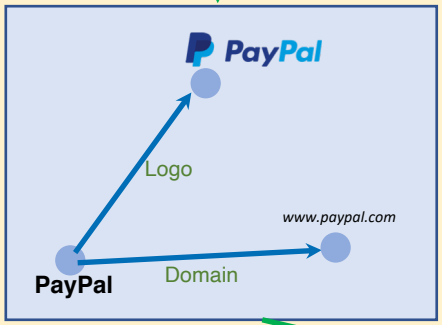
Knowledge-aware phishing detector

Phishing / Benign

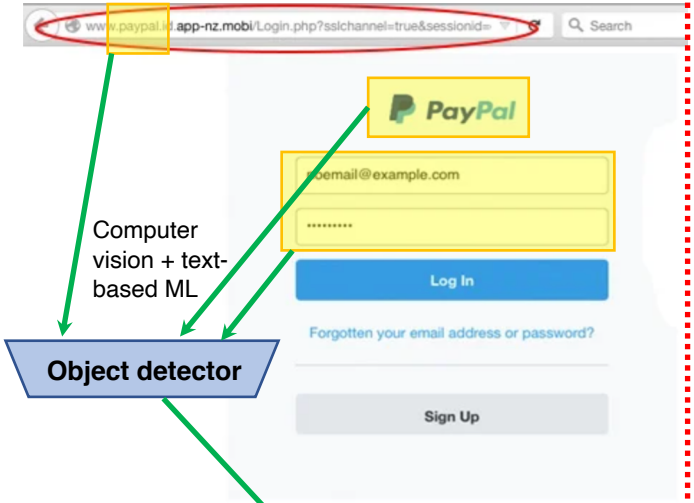
KnowPhish: Model

1. Knowledge Graph

Automated Knowledge Base Construction



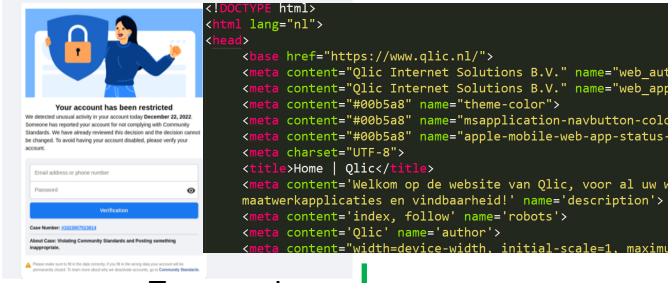
2. Visual Content



Computer vision + text-based ML

Object detector

3. Webpage HTML



Extracted text

```
16 [title] Restrictions Information
27 [div] Your account has been restricted
28 [div] We detected unusual activity in your account today
46 [button] Verification
47 [div] Case Number:
48 [a] #1023007523814
50 [div] About Case: Violating Community Standards and Posti
54 [div] Please make sure to fill in the data correctly, if
56 [font] Community Standards
```

Language model

Knowledge-aware phishing detector

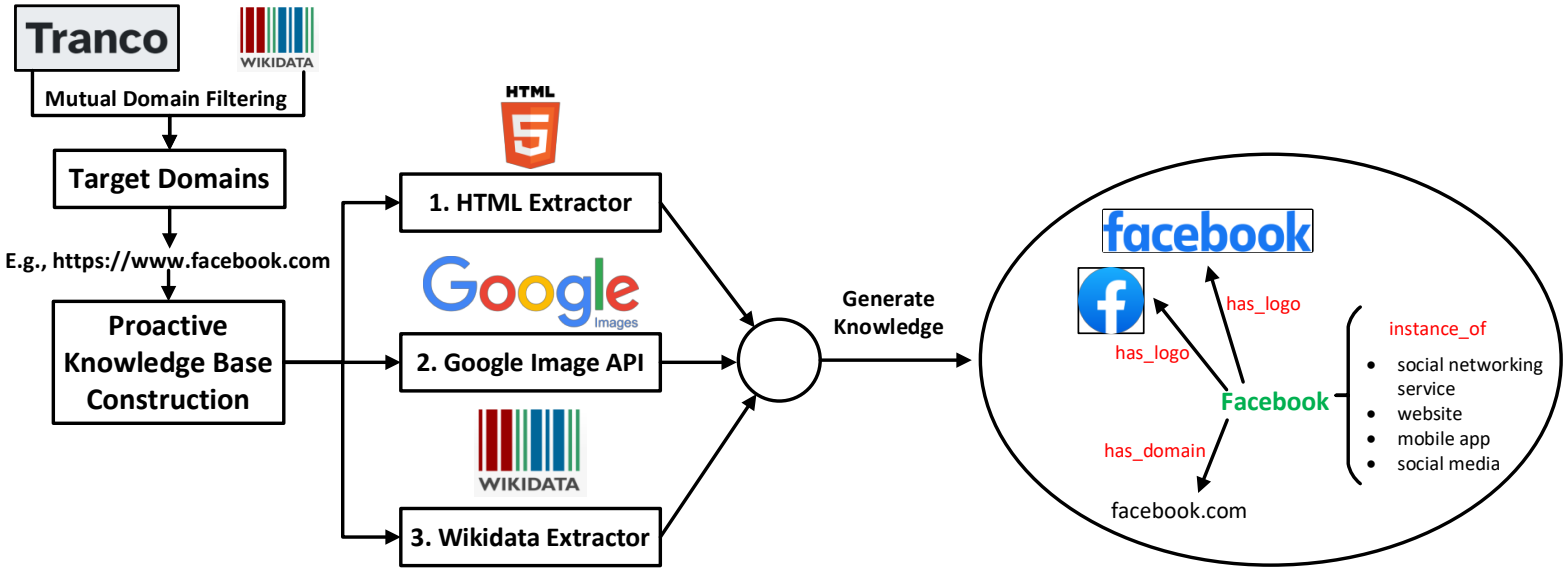
Phishing / Benign

1. Automated Knowledge Base Construction

- **Motivation**: error analysis (on original and GovTech datasets) has generally found that a majority of the mistakes are due to the **brands being absent from our knowledge base**

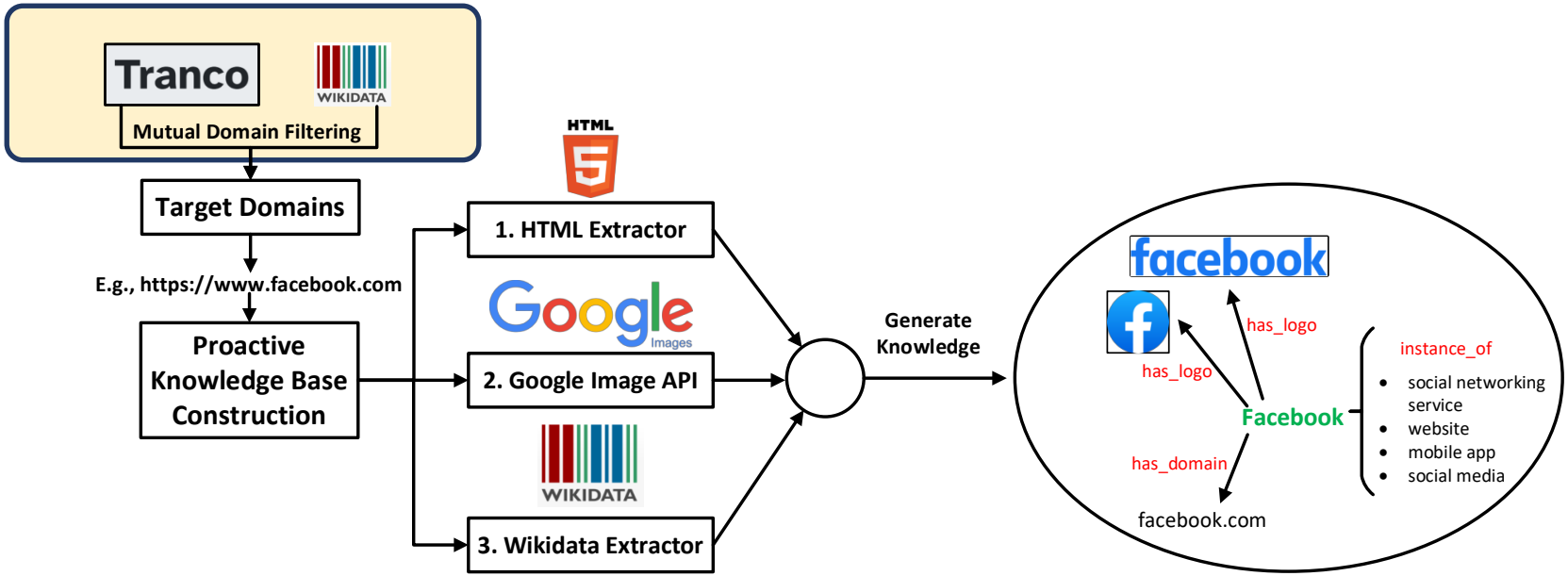
1. Automated Knowledge Base Construction

- **Motivation**: error analysis (on original and GovTech datasets) has generally found that a majority of the mistakes are due to the **brands being absent from our knowledge base**



1. Automated Knowledge Base Construction

- **Mutual domain filtering**: use Tranco to find top domains, and also use Wikidata to find well-known banks / financial institutions / etc (or subclasses of these)



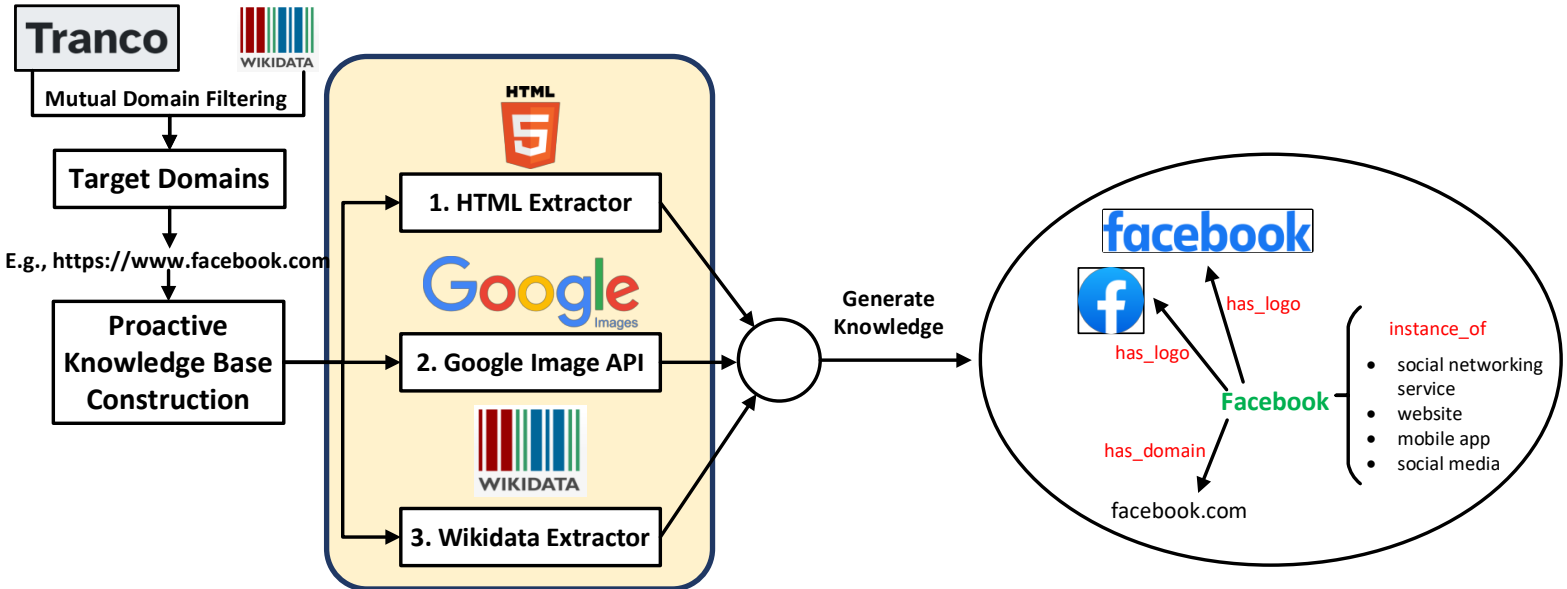
Categories of Phishing Targets

1. **Transaction**: bank/financial institution/credit institution/payment system/digital wallet/…
2. **Social Media**: social networking service/online video platform/social media/…
3. **Postal Service**: package delivery/postal service/
4. **E-Commerce**: online shop/online marketplace/…
5. **Government**: central government/federal government/…
6. **Online Service**: webmail/internet service provider/internet hosting service/online service/website/web service/web search engine
7. **University**: university/private university/research university/…
8. **Software?**: open-source software/mobile app/…



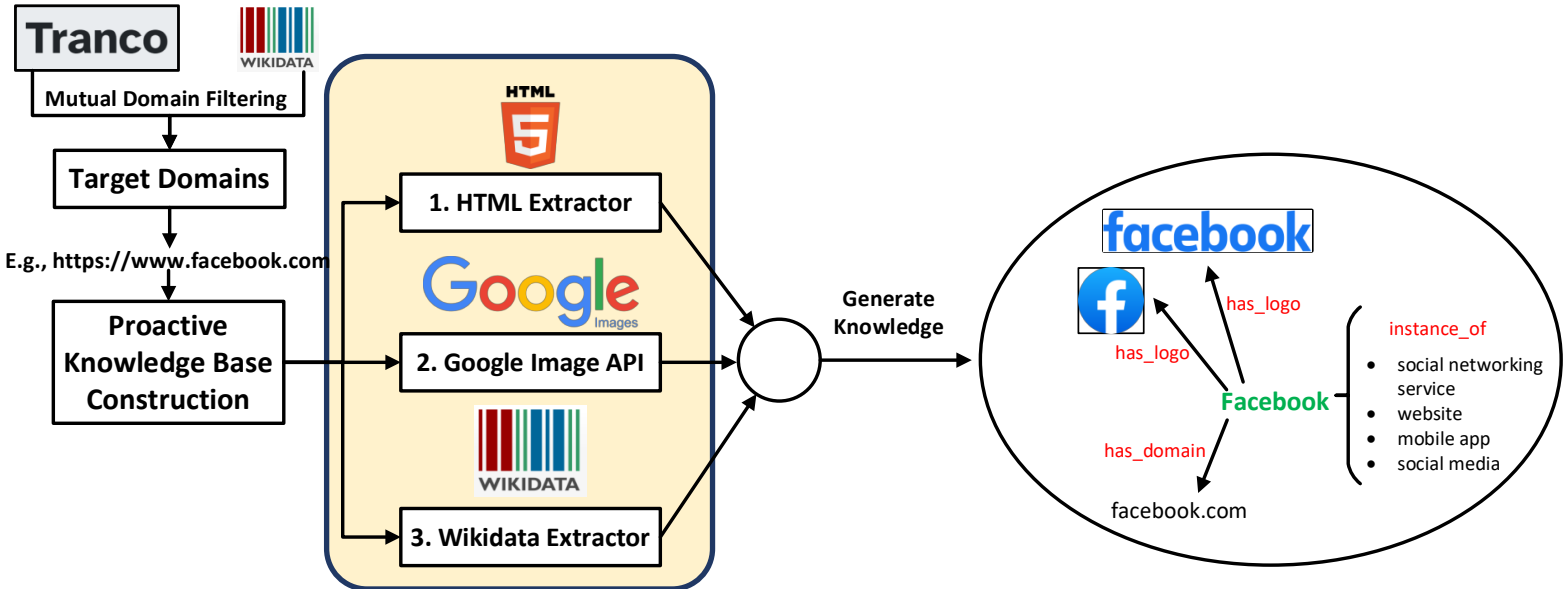
1. Automated Knowledge Base Construction

- **Logo collection**: search for logo images for these brands from various sources (e.g. Google Image API / Wikidata), and verify that they belong to the correct domain (e.g. facebook.com)



1. Automated Knowledge Base Construction

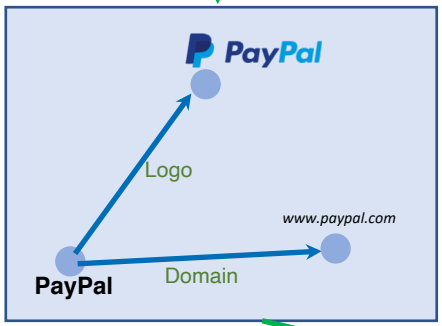
- **Current status**: still updating domain collection pipeline to make it collect local brands / government pages. Currently, many local brands are missed



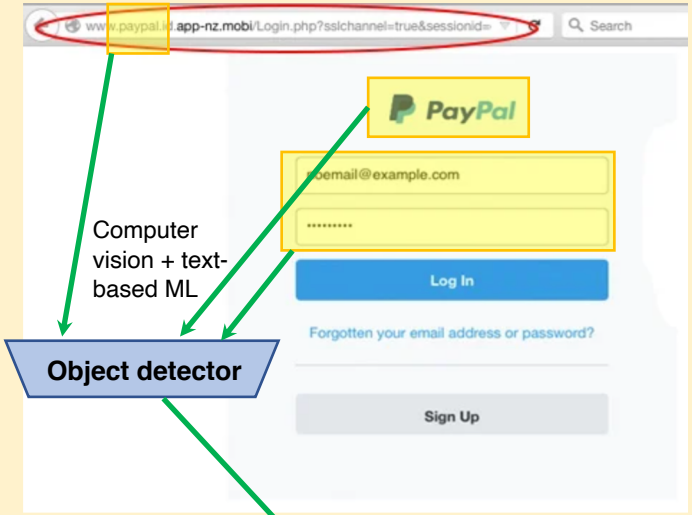
KnowPhish: Model

1. Knowledge Graph

Automated Knowledge Base Construction



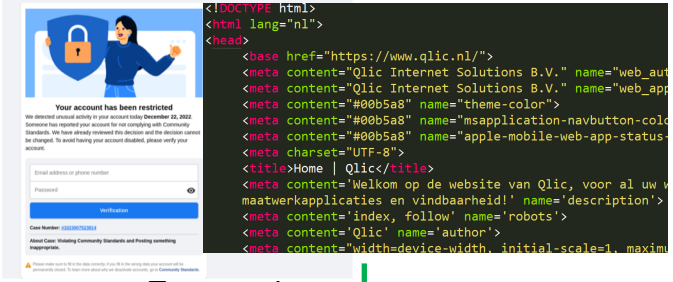
2. Visual Content



Computer vision + text-based ML

Object detector

3. Webpage HTML



Extracted text

```
16 [title] Restrictions Information
27 [div] Your account has been restricted
28 [div] We detected unusual activity in your account today
46 [button] Verification
47 [div] Case Number:
48 [a] #1023007523814
50 [div] About Case: Violating Community Standards and Posti
54 [div] Please make sure to fill in the data correctly, if
56 [font] Community Standards
```

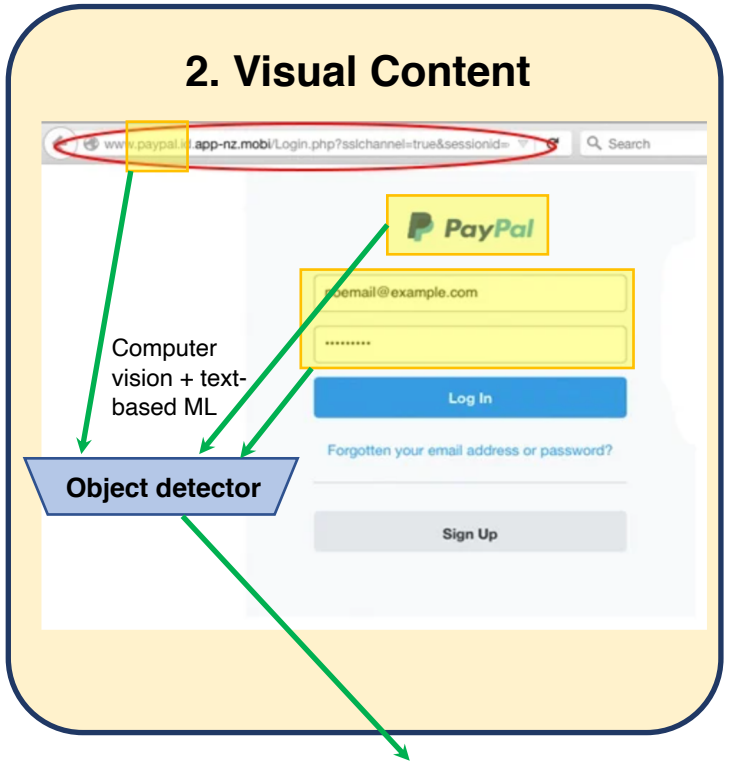
Language model

Knowledge-aware phishing detector

Phishing / Benign

2. Visual Content Component

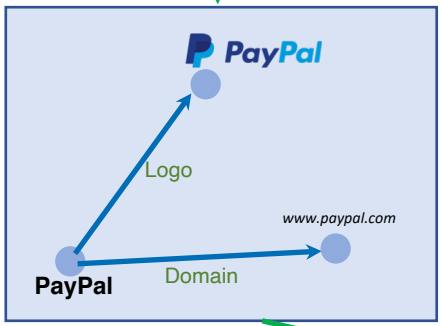
- Uses computer vision-based detectors to identify key objects on the page: logos and password fields
- **Recent improvements:**
  - Use newer object detection model
  - Can make use of multiple logos now
  - Make logo classification model more robust to bounding box variations (in progress)



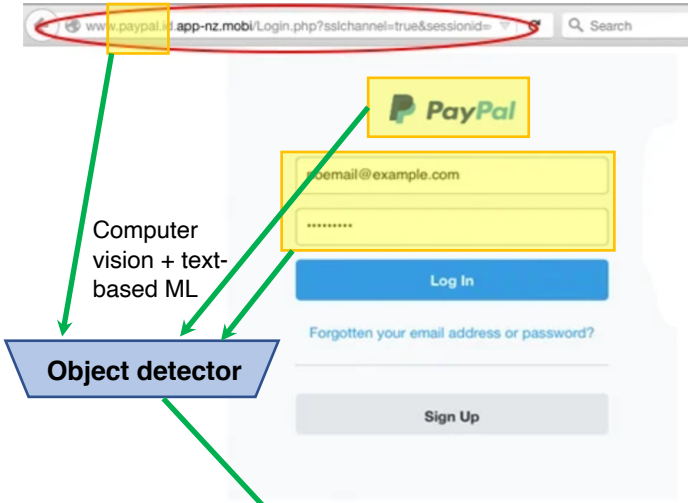
KnowPhish: Model

1. Knowledge Graph

Automated Knowledge Base Construction



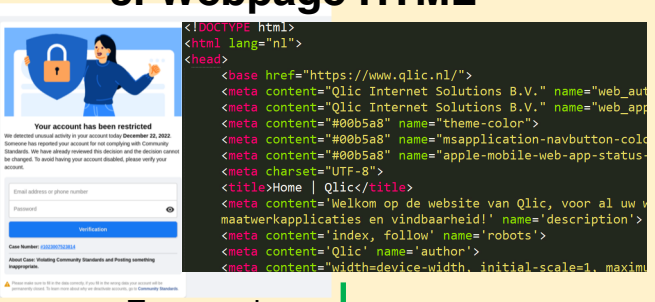
2. Visual Content



Computer vision + text-based ML

Object detector

3. Webpage HTML



Extracted text

```
16 [title] Restrictions Information
27 [div] Your account has been restricted
28 [div] We detected unusual activity in your account today
46 [button] Verification
47 [div] Case Number:
48 [a] #1023007523814
50 [div] About Case: Violating Community Standards and Posti
54 [div] Please make sure to fill in the data correctly, if
56 [font] Community Standards
```

Language model

Knowledge-aware phishing detector

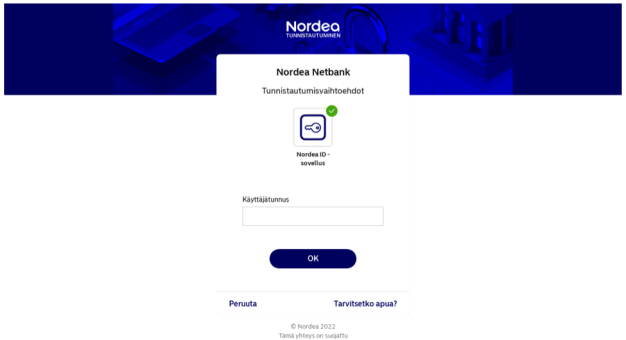
Phishing / Benign

3. Webpage HTML Component

- **Motivation**: important information (esp. password fields / “credential requiring") can be sometimes be only visible in the HTML

3. Webpage HTML Component

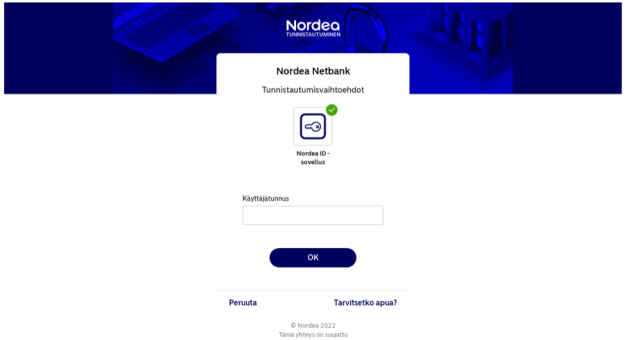
- **Approach**: train a language model (e.g. BERT) to predict whether a webpage is a “credential requiring page” (CRP)
- To train this approach, we manually label pages based on whether they are:
  - Explicit credential requiring: form asking for credentials
  - Implicit credential requiring: buttons leading to page that asks for credentials
  - Non credential requiring




```
7 [title] Nordea - Tunnistautuminen
25 [p] Ladataan...
27 [p] JavaScript must be enabled to use this page.
31 [span] Tärkeää tietoa:
36 [span] Lue lisää
41 [span] Nordea
50 [div] Tunnistautuminen
53 [h1] Nordea Netbank
57 [legend] Tunnistautumisvaihtoehdot
68 [span] Nordea ID -sovellus
78 [label] Käyttäjätunnus
80 [span] Vain numerot on sallittu.
81 [button] OK
84 [button] Vaihda tunnistautumistapaa
85 [button] Peruuta
87 [button] Tarvitsetko apua?
92 [p] Ladataan...
94 [div] © Nordea 2022
```

3. Webpage HTML Component

- **Current progress**: CRP classifier has accuracy of ~99% on our original dataset; 90% on GovTech dataset.
- Currently:
  - Integrate CRP classifier with the rest of the phishing classifier, which should (hopefully) improve its detection accuracy
  - Investigate more into reason for lower accuracy on GovTech dataset




```
7 [title] Nordea - Tunnistautuminen
25 [p] Ladataan...
27 [p] JavaScript must be enabled to use this page.
31 [span] Tärkeää tietoa:
36 [span] Lue lisää
41 [span] Nordea
50 [div] Tunnistautuminen
53 [h1] Nordea Netbank
57 [legend] Tunnistautumisvaihtoehdot
68 [span] Nordea ID -sovellus
78 [label] Käyttäjätunnus
80 [span] Vain numerot on sallittu.
81 [button] OK
84 [button] Vaihda tunnistautumistapaa
85 [button] Peruuta
87 [button] Tarvitsetko apua?
92 [p] Ladataan...
94 [div] © Nordea 2022
```

Accuracy on Phishpedia + APWG datasets

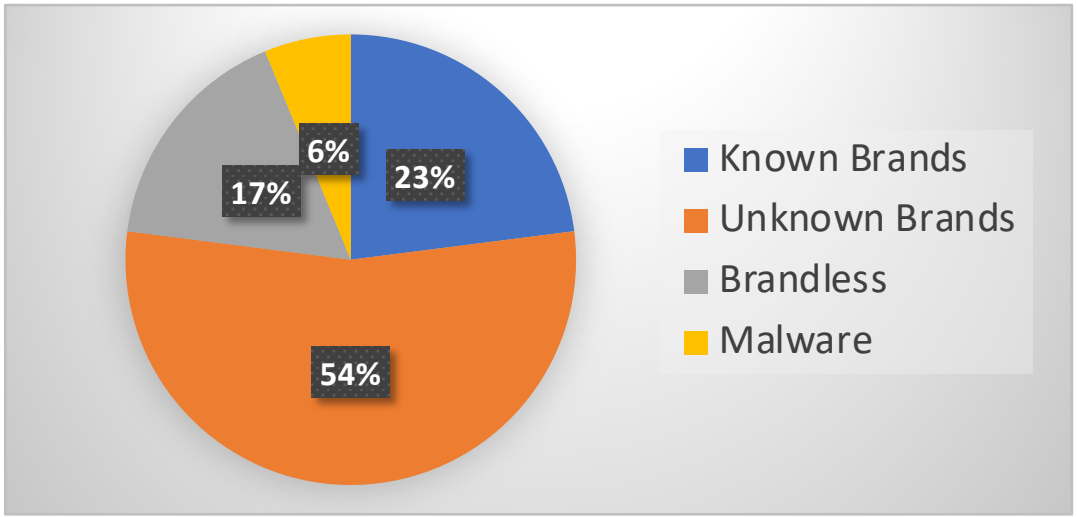
Model	Phishpedia Dataset (30k + 30k)				Our Dataset (24k + 6k)			
	Acc	F1	Prec	Recall	Acc	F1	Prec	Recall
YOLOv8-x + OCR Aided Siamese	0.9304	0.9278	0.9440	0.9121	0.9373	0.8225	0.9350	0.7342

Analysis of GovTech Data

- Total of 577 webpages
  - 186 phishing pages tagged by URLScan (~183 are DBS-related phishing)
  - 161 phishing pages labelled manually or by VirusTotal
  - 230 non-phishing pages (labelled manually)

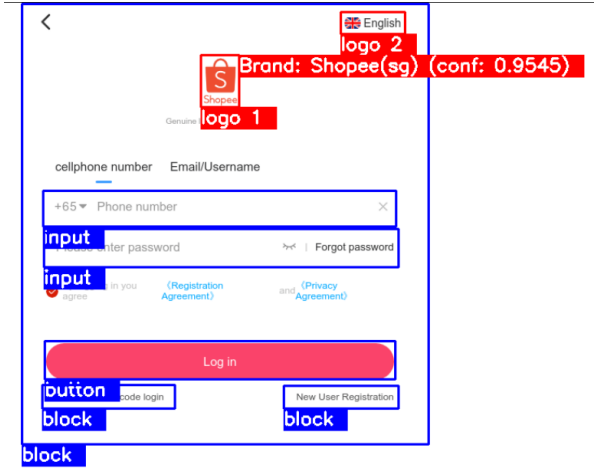
Brand coverage in GovTech dataset

- The local brands in the GovTech dataset are currently not very well covered by our knowledge base
  - (We are still updating the knowledge collection process to properly cover local brands and government-related entities)



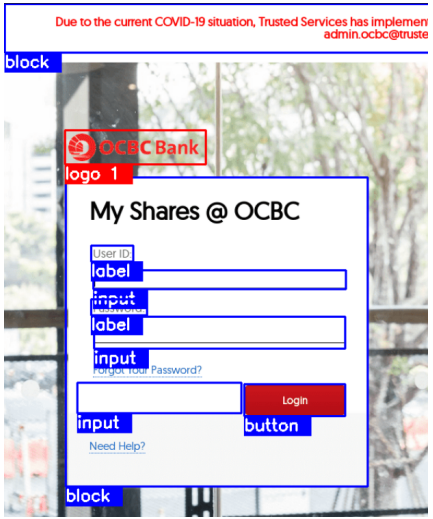
Known Brands (size=37)

- Known brands are the brands that can be found in Wikidata
- Our current database enables us detect 20 phishing samples, where 15 are true positives, 1 is a QR code page, 4 are false positives



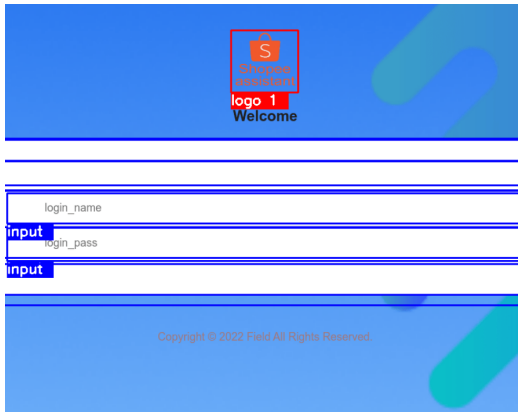
Known Brands (size=37)

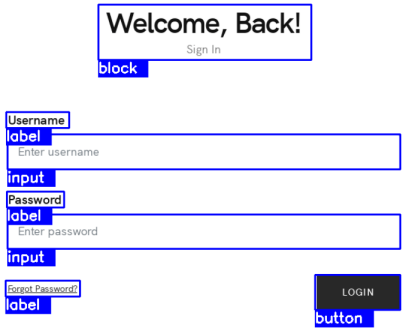
- Known brands are the brands that can be found in Wikidata
- 17 false negatives because of 1) noisy logo, 2) incorrect logo bounding box, 3) same SLD, 4) large domain rank

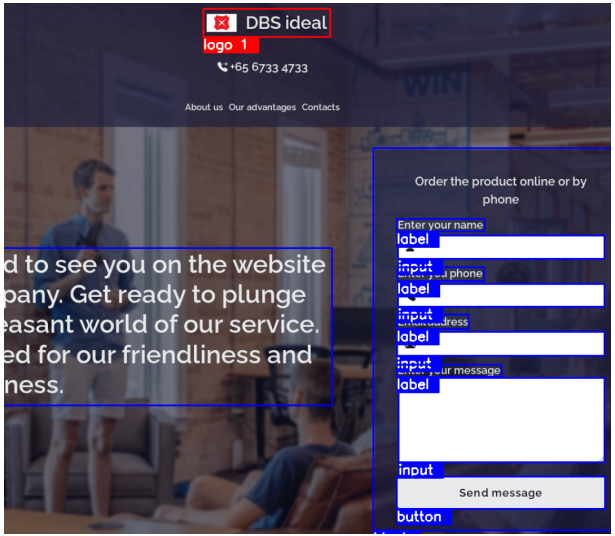


Known Brands (size=37)

- Known brands are the brands that can be found in Wikidata
- 17 false negatives because of 1) noisy logo, 2) incorrect logo bounding box, 3) same SLD, 4) large domain rank

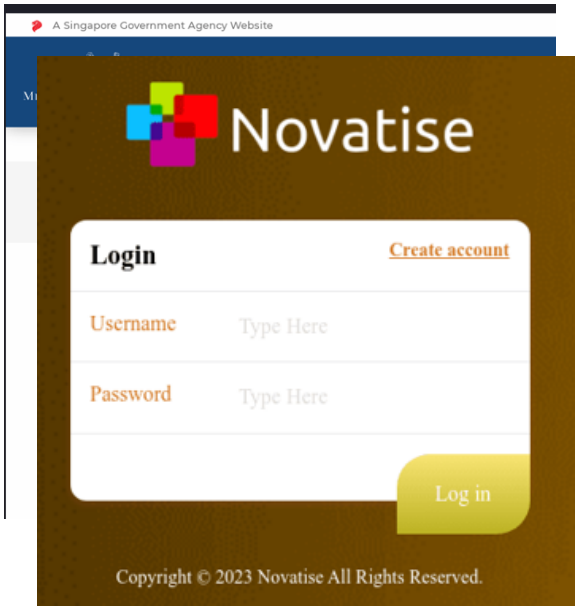


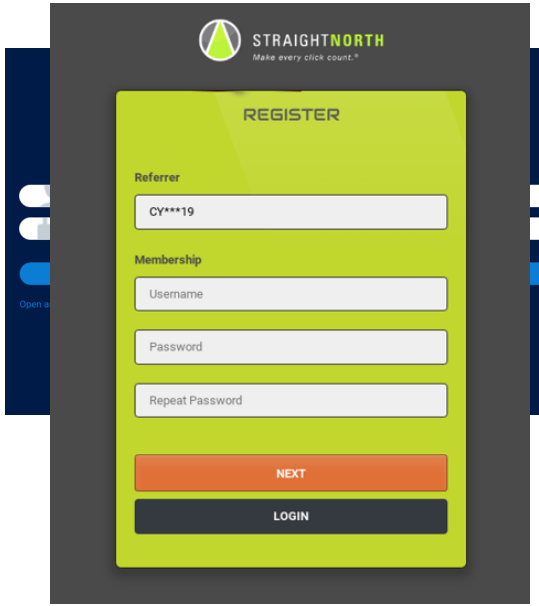


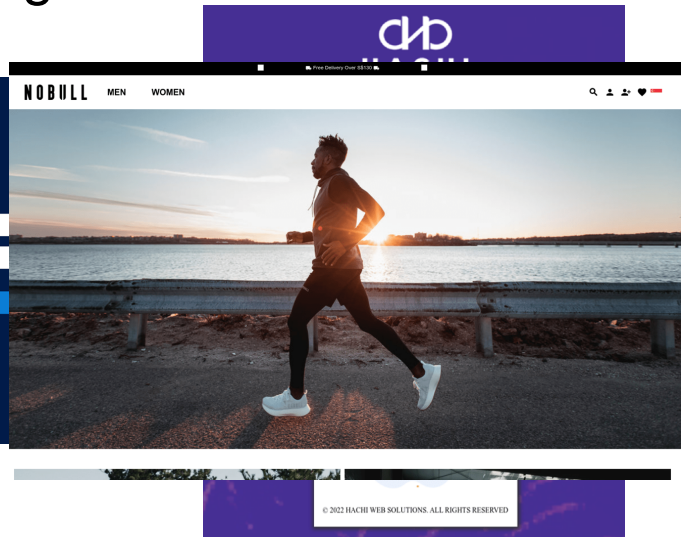


Unknown Brands (size=87)

- Unknown brands are the brands that cannot be found in Wikidata, and cannot be identified by Google Vision service.

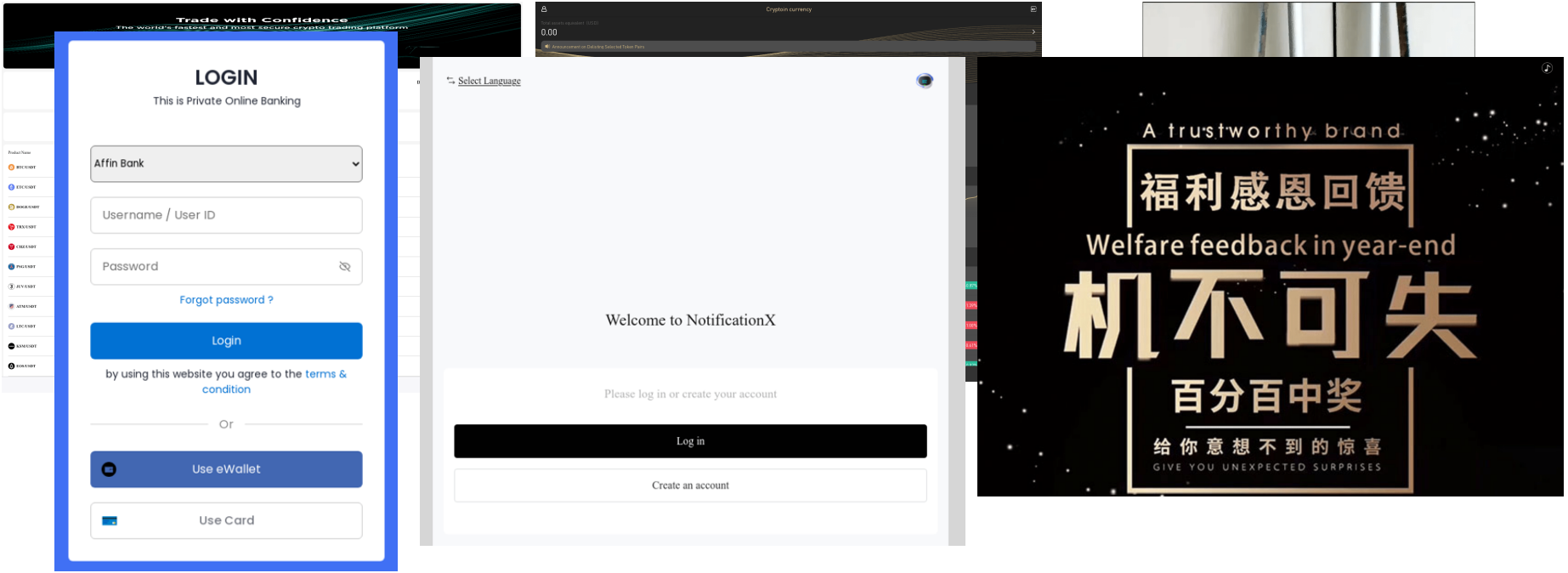






Brandless (size=27)

- Brandless means there is no brand showing in the screenshot.



Thank you!