D-Fence: A Flexible, Efficient, and Comprehensive Phishing Email Detection System

Jehyun Lee†, Farren Tang†, Pingxiao Ye†, Fahim Abbasi†, Phil Hay†, Dinil Mon Divakaran †*

†Trustwave, *NUS

Jehyun.Lee@trustwave.com
Abstract

- Phishing Email: Major Security Concern for Organizations

- Previous works
  - Focusing on specific email component: Evadable by changing attack vector
  - Limited single model performance: Limitation of ML models in nature

- Proposal: Multi-modular phishing email detection system with sophisticated analysis models
  - Structure module: Email headers and HTML structures capturing statistical characteristics.
  - Text module: Text classification with pre-trained text vectorization model (BERT)
  - URL module: Deep-learning-based URL string modelling and classification

- 0.99+ detection sensitivity (Recall) at a low false-positive rate (1 in 10K)
  - Evaluated with 68K of recent phishing email samples and 224K of benign samples
Motivation
Shortcomings in Targeting single email component

- **Email Header Analysis**
  - [+] Useful in detecting (large-scale) spamming of phishing emails
  - [-] Easy to evade in spear phishing

- **Readable text Analysis**
  - [+] Useful in Message-centric phishing
  - [-] Evadable by Image-based emails
  - [-] Bad at short / neutral texts
Motivation (cont’)

Shortcomings in Targeting single email component

- **HTML structure Analysis**
  - [+] Source of phishing techniques
    - e.g., Scripts, Hidden hyperlinks
  - [-] Do not cover Message-centric phishing

- **Embedded URL Analysis**
  - [+] Wide phishing coverage
    - Most of the phishing email has a URL
  - [-] Short living contents

---

Email sample with Various Email Component (HTML Section)

```
Content-Type: text/html; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

&lt;div dir=3D"ltr"><b&gt;&lt;font color=3D"#ff0000">Scheduled delivery pending&lt;/font&gt;&lt;/b&gt;&lt;div dir=3D"ltr"><p>Please visit the website for more information&lt;br&gt;&lt;a href=3D"http://phishing-url.biz"&gt;http://postaloffice.gov&lt;/a&gt;&lt;/p&gt;&lt;/div&gt;&lt;br&gt;&lt;div dir=3D"ltr" class=3D"mail_signature" data-smartmail=3D"mail_signature"><br/&gt;&lt;p style=3D"margin: 0cm 0cm 0pt"&gt;Global Pos Office&lt;/p&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;
```

---
D-Fence: Overview

Three Independent Analysis Modules
- Wide component coverage
- Extensible

No External Information Sources
- Stand-alone solution
- No up-to-date repository required
- No external communications

Flexible model configuration / Update
- e.g., Feature modification, model update, module addition, etc.
D-Fence: Structure Module (1/4)

- Analysis Component
  - Email Header and HTML section

- Feature set
  - 63 Structural features
  - 10 Feature categories

- Classification
  - Probability prediction with a supervised learning model

1. $D_{tr}$
2. HTML
3. Email headers
4. HTML parser
5. Feature extraction
6. Supervised learning
7. Structure model
8. Module prediction

① → ③: Training process
① → ③: Prediction process
D-Fence: Text Module (2/4)

- **Analysis Component**
  - Texts from *text/plain* and *text/html* sections

- **Text Vectorization**
  - Sentences to numeric vectors
  - **BERT**: Bidirectional Encoder Representations from Transformers

- **Classification**
  - Probability prediction with a supervised learning model

---

**Diagram**:

1. **$D_{tr}$**
2. **Language detection**
3. **Pre-trained language model**
4. **Text model**

**Process**:

1. Training process
2. Prediction process

---

1. - 3: Training process
2. 1 - 4: Prediction process
D-Fence: URL Module (3/4)

- **Analysis Component**
  - URL strings in `text/plain` and `text/html` sections

- **Feature set**
  - Encoded characters in a URL string

- **Modelling and Classification**
  - CNN-LSTM
  - Multiple URLs in an email: multiple predictions
  - Classification of an email: Maximum prediction of all embedded URLs

Diagram:
- **Training process (1~3):**
  - 1: Email
  - 2: Embedded URLs with label
  - 3: Module prediction

- **Prediction process (1~3):**
  - 1: Embedded URLs
  - 2: Deep learning module prediction
  - 3: Maximum prediction of all embedded URLs
D-Fence: Meta-classifier (4/4)

- Learning prediction confidence and correlation of the individual module’s prediction
- Training: Prediction values from individual modules for Meta-classifier training set $D_{mt}$
- Prediction: Three module prediction values into one final prediction value
### Evaluation: Enterprise Email Dataset (EES 2020)

- **Email samples from enterprises**
  - Benign emails reviewed by users as Benign
  - Phishing emails detected by multiple solutions
  - Collected in 2018 ~ 2020

- **292K unique samples**
  - Benign: 224K, Phishing: 68K

<table>
<thead>
<tr>
<th>Content</th>
<th>Source</th>
<th>Label</th>
<th>No. of samples</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Any</td>
<td>Benign</td>
<td>212200</td>
<td>94.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>64587</td>
<td>95.59%</td>
</tr>
<tr>
<td></td>
<td>text/plain</td>
<td>Benign</td>
<td>188261</td>
<td>83.99%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>12039</td>
<td>17.82%</td>
</tr>
<tr>
<td></td>
<td>text/html</td>
<td>Benign</td>
<td>136084</td>
<td>60.71%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>59016</td>
<td>87.35%</td>
</tr>
<tr>
<td>HTML</td>
<td>text/html</td>
<td>Benign</td>
<td>173542</td>
<td>77.43%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>62488</td>
<td>92.49%</td>
</tr>
<tr>
<td>URL</td>
<td>All</td>
<td>Benign</td>
<td>197087</td>
<td>87.93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>67559</td>
<td>99.99%</td>
</tr>
<tr>
<td>Total</td>
<td>All</td>
<td>Benign</td>
<td>224137</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phishing</td>
<td>67565</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>291702</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation: Model Selection
AUPRC, and Recall at Fixed False-positive rate 0.001 (10^{-3}). Tested with EES 2020 dataset

<table>
<thead>
<tr>
<th>Structural Module</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>AUPRC</td>
<td>Recall</td>
<td>Train (s)</td>
<td>Test (ms)</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.9993</td>
<td>0.9933</td>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.9994</td>
<td>0.9884</td>
<td>10</td>
<td>0.01</td>
</tr>
<tr>
<td>SVM (SVC)</td>
<td>0.9969</td>
<td>0.9618</td>
<td>919</td>
<td>0.55</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.8940</td>
<td>0.0</td>
<td>2</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Module</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (BERT+)</td>
<td>AUPRC</td>
<td>Recall</td>
<td>Train (s)</td>
<td>Test (ms)</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.9757</td>
<td>0.7796</td>
<td>61</td>
<td>0.01</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.9746</td>
<td>0.6995</td>
<td>560</td>
<td>0.02</td>
</tr>
<tr>
<td>SVM (SVC)</td>
<td>0.8310</td>
<td>0.0776</td>
<td>48392</td>
<td>8.44</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.7353</td>
<td>0.0</td>
<td>3</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>URL Module</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>AUPRC</td>
<td>Recall</td>
<td>Train (s)</td>
<td>Test (ms)</td>
</tr>
<tr>
<td>CNN</td>
<td>0.9406</td>
<td>0.5775</td>
<td>302</td>
<td>0.76</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.9149</td>
<td>0.5787</td>
<td>7728</td>
<td>14.41</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>0.9851</td>
<td>0.7648</td>
<td>4247</td>
<td>7.85</td>
</tr>
</tbody>
</table>

Models for Cost-efficient Configuration Analysis
Models for Best-Accuracy Evaluation
## Evaluation: Comparison with Baselines

10-Cross-fold validation (90:10 splits). Recall at $10^{-3}$ FPR

<table>
<thead>
<tr>
<th>System</th>
<th>AUPRC ($\sigma$)</th>
<th>Recall ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legacy structure features+RF</td>
<td>0.9985 (0.0002)</td>
<td>0.9663 (0.0051)</td>
</tr>
<tr>
<td>Text Word2Vec+LSTM</td>
<td>0.8313 (0.0074)</td>
<td>0.1365 (0.0023)</td>
</tr>
<tr>
<td>URL CNN-LSTM</td>
<td>0.9851 (0.0031)</td>
<td>0.7648 (0.0353)</td>
</tr>
<tr>
<td><strong>Our proposals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined structure features+RF</td>
<td>0.9993 (0.0003)</td>
<td>0.9933 (0.0017)</td>
</tr>
<tr>
<td>Text BERT+RF</td>
<td>0.9757 (0.0039)</td>
<td>0.7796 (0.0038)</td>
</tr>
<tr>
<td>D-Fence</td>
<td>0.9997 (0.0001)</td>
<td>0.9935 (0.0013)</td>
</tr>
</tbody>
</table>
### Evaluation: Recall at $10^{-4}$ FPR

EES 2020 Dataset. Best Accuracy Configuration.

<table>
<thead>
<tr>
<th>Module</th>
<th>AUPRC</th>
<th>Recall ($10^{-3}$ FPR)</th>
<th>Recall ($10^{-4}$ FPR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure module</td>
<td>0.9994</td>
<td>0.9878</td>
<td>0.9428</td>
</tr>
<tr>
<td>Text module</td>
<td>0.9192</td>
<td>0.6182</td>
<td>0.2710</td>
</tr>
<tr>
<td>URL module</td>
<td>0.9492</td>
<td>0.8806</td>
<td>0.7721</td>
</tr>
<tr>
<td>D-Fence</td>
<td>0.9995</td>
<td>0.9932</td>
<td>0.9844</td>
</tr>
</tbody>
</table>

4% more detection e.g., 1K more phishing emails in our test set.
Cost Reduction: Structural Module

Feature set Reduction

- Feature selection by Feature Category
  - 10 Feature categories
  - e.g., Msg-ID features, Link features, ... etc.

- Test on $2^{10}$ Feature set combinations
  - A point + on plot indicates one combination

- Cost-Efficient Features
  - Less feature extraction time but high accuracy

~50% Prediction time reduction from Reduced feature set with keeping 95%+ Recall at FPR $10^{-3}$
Cost Reduction: URL Module

Hyper-parameter tuning: Simpler/Faster Neural network

- Shorter training Epoch
  - Advantage: Shorter training time
  - Cost: Loss in accuracy

- Higher Max Pooling
  - Advantage: Shorter training/prediction time

- CNN (without LSTM layer)
  - Advantage: Faster training/prediction
  - Cost: Large loss in accuracy
Cost-Efficient Configuration
Combinations of the module configurations

- Text module fixed as the fastest configuration. (100 words analysis)
- A pair of points (purple and green): one config combination

~20% of Training time reduction from mainly Deep-learning for URL
~10% of Prediction time reduction from URL and Structure module with 0.95+ Recall at 10^{-5} FPR
Conclusions

D-FENCE: Flexible Multi-modular phishing email detection system

- Wide component coverage with comprehensive detection: little evasion surface
- Low False-detection powered by independent analysis modules supplementing each other
- Evaluated with near 300K of real-world Enterprise email dataset

Cost-efficient Configuration

- Synergetic configuration: Better than combination of the best individual configurations
- Training time reduction without harming accuracy
Thank You

Q & A