



ICLR
International Conference On
Learning Representations



MONASH University



AI2TALE: An Innovative Information Theory-based Approach for Learning to Localize Phishing Attacks

Van Nguyen^{1,2*}, Tingmin Wu², Xingliang Yuan³, Marthie Grobler², Surya Nepal², Carsten Rudolph¹

¹Monash University, ²CSIRO's Data61, ³The University of Melbourne, Australia

* Corresponding author (van.nguyen1@monash.edu)

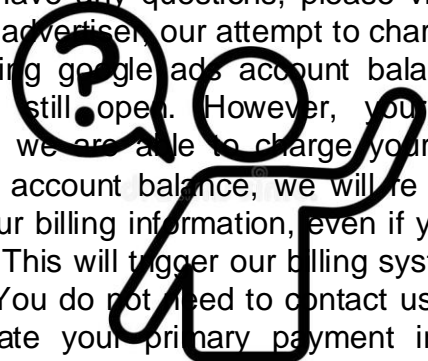


Motivations

- Phishing attacks remain a challenge for detection, explanation, and defense.
 - Conducted in various ways, with email being the most common method.
 - Exploit cognitive principles (e.g., urgency, authority) to trigger psychological reactions in users through persuasion techniques.
- AI-based approaches can detect phishing, reducing the negative effects caused.
 - However, they often lack the intrinsic ability to identify the information causing the data phishing.

An entire email:

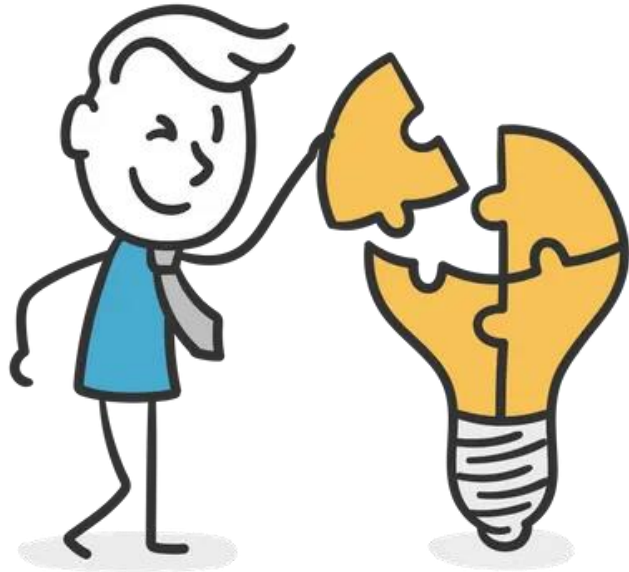
Your payment didn't succeed, so your ads have been suspended. This message was sent from a notification only email address that does not accept incoming email. Please do not reply to this message. If you have any questions, please visit the google ads help center. Hello advertiser, our attempt to charge your credit card for your outstanding google ads account balance was declined. Your account is still open. However, your ads have been suspended. Once we are able to charge your card and receive payment for your account balance, we will re activate your ads. Please update your billing information, even if you plan to use the same credit card. This will trigger our billing system to try charging your card again. You do not need to contact us to reactivate your account. To update your primary payment information, please follow these steps 1. Log in to your account at <http://adwords.google.com> select. 2. Enter your primary payment information. 3. Click "update" when you have finished. Thank you for advertising with google ads. We look forward to providing you with the most effective advertising available. (c) google ads team 2008.



Research Question

In addition to predicting the vulnerability label (phishing or benign) of the data, how can we derive a deep learning-based method that can also automatically identify the most important phishing-relevant information (i.e., sentences) triggering the classification for providing a useful and concise explanation about the vulnerability of the phishing data?
(we refer to this as phishing attack localization)

Contributions



- We address the problem of phishing attack localization to tackle and improve the explainability (transparency) of email phishing detection. Automated AI-based techniques for this problem have not yet been well studied.
- We propose AI2TALE, an innovative deep learning-based framework grounded in information theory and the information bottleneck principle to solve the problem.
- We propose using appropriate measures, including Label-Accuracy and Cognitive-True-Positive for the problem.
- We evaluate our method on seven real-world email datasets, and extensive experiments show its effectiveness and superiority over the baselines.

The AI2TALE Approach



- **The problem statement**

- We denote an email as $X = [x_1, \dots, x_L]$, a sequence of L sentences, with vulnerability label $Y \in \{0, 1\}$ (i.e., 1: phishing and 0: benign). In phishing attack localization, our goal is to develop an AI-based approach that can detect the vulnerability label Y and automatically identify the important phishing-relevant information (i.e., sentences) \tilde{X} (a subset of X) causing X phishing.
- **Note:** In phishing attack localization, we work in a weakly supervised setting, where training relies solely on Y without requiring the ground truth of the information causing the phishing (i.e., often absent in publicly available phishing-relevant data).

The AI2TALE Approach

- **Phishing-relevant information selection process**

- The selection network ζ learns a set of independent Bernoulli latent variables $\mathbf{z} \in \{0,1\}^L$ representing the importance of the sentences to Y .
- Each element z_i in \mathbf{z} indicates whether x_i is related to Y (i.e., if $z_i = 1$, the sentence x_i is important).
- We model $\mathbf{z} \sim \prod_{i=1}^L \text{Bernoulli}(p_i)$, where $p_i = w_i(X; \alpha)$, and w is a network parameterized by α .
- Using \mathbf{z} , we construct $\tilde{X} = \zeta(X) = X \odot \mathbf{z}$ (i.e., \odot represents the element-wise product).

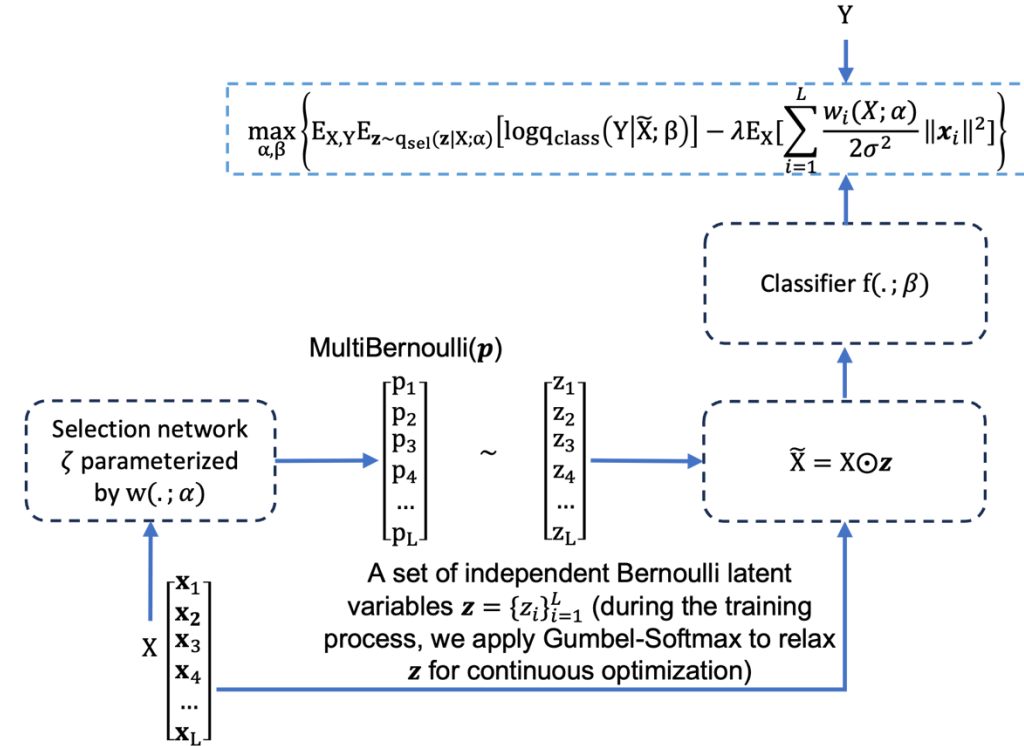


Figure 1. Visualization of our AI2TALE method.

The AI2TALE Approach

- **Mutual information for guiding the selection process**

$$\max_{\zeta} I(\tilde{X}, Y) \quad (1)$$

$$\begin{aligned} I(\tilde{X}, Y) &= \int p(\tilde{X}, Y) \log \frac{p(\tilde{X}, Y)}{p(\tilde{X})p(Y)} d\tilde{X}dY \geq \int p(\tilde{X}, Y) \log \frac{q(Y|\tilde{X})}{p(Y)} dYd\tilde{X} \\ &= \int p(Y, \tilde{X}, X) \log \frac{q(Y|\tilde{X})}{p(Y)} dYd\tilde{X}dX \\ &= E_{X,Y} E_{\tilde{X}|X} [\log q(Y|\tilde{X})] + \text{const} \quad (2) \end{aligned}$$

- To model the conditional variational distribution $q(Y|\tilde{X})$, we introduce a neural network classifier $f(\tilde{X}, \beta)$, that predicts Y given \tilde{X} . Our objective is to optimize both the selection model and classifier by maximizing the following function:

$$\max_{\alpha, \beta} \{E_{X,Y} E_{\mathbf{z} \sim q_{\text{sel}}(\mathbf{z}|X; \alpha)} [\log q_{\text{class}}(Y|X \odot \mathbf{z}; \beta)]\} \quad (3)$$

The AI2TALE Approach

- The joint training process between the classifier $f(., \beta)$ and the selection network $\omega(., \alpha)$ brings benefits for selecting the important and phishing-relevant sentences from phishing emails. However, two potential limitations are observed:
 - It does not theoretically guarantee to eliminate sentences unrelated to the vulnerability label of a given email.
 - The predictions of the classifier $f(., \beta)$ can be based more on the features selected from the selection network $\omega(., \alpha)$ than the underlying information contained in the features.

The AI2TALE Approach

- **Solving the problem of obtaining a superset of phishing-relevant sentences**

- We propose an additional term for training the selection process (model) ζ , derived from the following principle.

$$\max_{\zeta} (I(\tilde{X}, Y) - \lambda I(\tilde{X}, X)) \quad (4)$$

where λ is a hyper-parameter indicating the weight of the second mutual information.

- By minimizing the mutual information between X and \tilde{X} , the selection process prefers to select a smaller subset, excluding sentences irrelevant to the vulnerability label Y of X .

The AI2TALE Approach

- We derive an upper bound of the minimization of the mutual information between X and \tilde{X} :

$$I(\tilde{X}, X) = \int p(\tilde{X}, X) \log \frac{p(\tilde{X}|X)}{p(\tilde{X})} d\tilde{X} dX \leq E_X E_{\tilde{X}|X} \left[\log \frac{p(\tilde{X}|X)}{r(\tilde{X})} \right] \quad (5)$$

for any distribution $r(\tilde{X})$.

$$= E_X E_{\tilde{X}|X} \left[\sum_{i=1}^L \log \frac{p(\tilde{\mathbf{x}}_i|X)}{r(\tilde{\mathbf{x}}_i)} \right] = \sum_{i=1}^L E_X [D_{KL}(p(\tilde{\mathbf{x}}_i|X) || r(\tilde{\mathbf{x}}_i))] \quad (6)$$

where $r(\tilde{\mathbf{x}}_i)$ is the prior distribution, constructed as $r(\tilde{\mathbf{x}}_i) = N(\tilde{\mathbf{x}}_i | 0, \sigma^2)$.

- The Kullback-Leibler divergence $D_{KL}(p(\tilde{\mathbf{x}}_i|X) || r(\tilde{\mathbf{x}}_i))$ can be approximated as:

$$\frac{w_i(X; \alpha)}{2\sigma^2} \|\mathbf{x}_i\|^2 + \left(\log \sigma + \frac{1}{2} \sigma^2 \right) + \text{const} \quad (7)$$

- Combining $I(\tilde{X}, Y)$ and $I(\tilde{X}, X)$ by $\max(I(\tilde{X}, Y) - \lambda I(\tilde{X}, X))$, we get a unified training objective:

$$\max_{\alpha, \beta} \left\{ E_{X,Y} E_{\mathbf{z} \sim q_{\text{sel}}(\mathbf{z}|X; \alpha)} [\log q_{\text{class}}(Y|X \odot \mathbf{z}; \beta)] - \lambda E_X \left[\sum_{i=1}^L \frac{w_i(X; \alpha)}{2\sigma^2} \|\mathbf{x}_i\|^2 \right] \right\} \quad (8)$$

The AI2TALE Approach

- **Solving the problem of encoding the label via selections instead of meaningful information**
 - We propose learning the classifier model $f(., \beta)$ disjointly to approximate the conditional distribution of Y given X_R (i.e., $X_R = X \odot \mathbf{r}$ with $\mathbf{r} \sim \text{MultiBernoulli}(0.5)$ (i.e., $\mathbf{r} \sim B(0.5)$)).
 - This approach allows the classifier to incorporate both selection network information and data-driven updates to its parameters.
 - The training procedure maximizes:

$$\mathbb{E}_{X,Y} \mathbb{E}_{\mathbf{r} \sim B(0.5)} [\log q_{\text{class}}(Y|X \odot \mathbf{r}; \beta)] \text{ (9)}$$

- To ensure continuous and differentiable sampling from a Multi-Bernoulli distribution, we apply the temperature-dependent Gumbel-Softmax trick during training.

Algorithm

• Training

- AI2TALE is trained to learn and identify the most important and phishing-relevant sentences in emails without requiring any information about the ground truth of phishing-relevant sentences.
- We sequentially update the classifier and selector using Eqs. (8) and (9).

• Testing

- We can pick out the most important and phishing-relevant x_i using the magnitude $w_i(X; \alpha)$ from the selection network ζ .

Algorithm 1: The algorithm of our proposed AI2TALE method for the phishing attack localization problem.

Input: An email dataset $S = \{(X_1, Y_1), \dots, (X_{N_S}, Y_{N_S})\}$ where each email $X_i = [\mathbf{x}_1, \dots, \mathbf{x}_L]$ consisting of L sentences while its label $Y_i \in \{0, 1\}$ (i.e., 1: phishing and 0: benign). We denote the number of training iterations nt ; the mini-batch size m ; the trade-off hyper-parameter λ . We randomly partition S into three different sets including the training set S_{train} (for training the model), the validation set D_{val} (for model selection during training), and the testing set D_{test} (for evaluating the model).

- 1 We initialize the parameters α and β of the selection model ζ (i.e., parameterized by $\omega(\cdot, \alpha)$) and the classifier model $f(\cdot, \beta)$, respectively.
- 2 **for** $t = 1$ **to** nt **do**
- 3 Choose a mini-batch of embedded emails denoted by $\{(X_i, Y_i)\}_{i=1}^m$.
- 4 Update the classifier's parameter β via minimizing the following cross-entropy loss $\mathbb{E}_{X,Y} \mathbb{E}_{\mathbf{r} \sim B(0.5)} [\mathcal{L}_{ce}(Y, f_\beta(X \odot \mathbf{r}))]$ using Adam optimizer (Kingma & Ba, 2014).
- 5 Update the classifier's parameter β and the selection model parameter's α via minimizing the following objective function $\mathbb{E}_{X,Y} \mathbb{E}_{\mathbf{z} \sim q_{sel}(\mathbf{z}|X;\alpha)} [\mathcal{L}_{ce}(Y, f_\beta(X \odot \mathbf{z}))] + \lambda \mathbb{E}_X [\sum_{i=1}^L \frac{\omega_i(X;\alpha)}{2\sigma^2} \|\mathbf{x}_i\|^2]$ using Adam optimizer.
- 6 **end**

Output: The trained model for phishing attack localization.

Experiments

- **Datasets**

- We conducted experiments on seven diverse real-world email datasets including IWSPA-AP¹, Nazario Phishing Corpus², Miller Smiles Phishing Email³, Phish Bowl Cornell University⁴, Fraud emails⁵, Cambridge⁶, and Enron Emails⁷.

- **Baselines**

- The baselines of our method are recent, popular, and state-of-the-art interpretable machine learning approaches falling into the category of intrinsic interpretable models including L2X (Chen et al., 2018), INVASE (Yoon et al., 2019), ICVH (Nguyen et al., 2021), VIBI (Bang et al., 2021), and AIM (Vo et al., 2023b).

¹<https://github.com/BarathiGanesh-HB/IWSPA-AP/tree/master/data/> ²<https://monkey.org/~jose/phishing/> ³<http://www.millersmiles.co.uk/archives.php>

⁴<https://it.cornell.edu/phish-bowl> ⁵<https://www.kaggle.com/datasets/rtatman/fraudulent-email-corpus> ⁶A private dataset ⁷<https://www.cs.cmu.edu/~enron/>

Experiments

- **Measures**

- We introduce and utilize two main metrics: **Label-Accuracy** and **Cognitive-True-Positive**.
- **Label-Accuracy**
 - Measure whether the selected sentences are important and accurately predict the true vulnerability label (i.e., phishing or benign) of the emails.
- **Cognitive-True-Positive**
 - Investigate if the selected sentences also reflect the human cognitive principles, exploiting psychological triggers to deceive recipients used in phishing emails (i.e., Reciprocity, Consistency, Social Proof, Authority, Liking, and Scarcity).
- **Note:** We evaluate the model performance based on the most important sentence from each email. Higher values of Label-Accuracy and Cognitive-True-Positive indicate better model performance in selecting crucial, label-relevant information and reflecting cognitive principles.

Experiments

• Quantitative results

- AI2TALE achieves a significantly higher performance, with improvements ranging from approximately 1.5% to 3.5% (on average of Label-Accuracy and Cognitive-True-Positive) compared to the baselines.
- The additional measures on F1-score, false positive rate (FPR), and false negative rate (FNR) further highlight its superiority over the baselines.

Table 1. The performance of our AI2TALE method and the baselines for the Label-Accuracy (Label-Acc) and Cognitive-True-Positive (Cognitive-TP) measures, as well as their combined average results (denoted as Average) (the best results in bold).

Methods	Label-Acc	Cognitive-TP	Average
INVASE (Yoon et al., 2019)	98.30%	97.20%	97.75%
ICVH (Nguyen et al., 2021)	96.72%	98.10%	97.41%
L2X (Chen et al., 2018)	98.25%	97.20%	97.73%
VIBI (Bang et al., 2021)	96.65%	94.99%	95.82%
AIM (Vo et al., 2023b)	98.40%	97.10%	97.75%
AI2TALE (Ours)	99.33%	98.95%	99.14% $\uparrow \sim (1.5\% \rightarrow 3.5\%)$

Table 2. The performance of our AI2TALE method and the baselines for additional measures, including F1-score, false positive rate (FPR) and false negative rate (FNR) (the best results in bold).

Methods	F1-score (i) \uparrow	F1-score (ii) \uparrow	FPR \downarrow	FNR \downarrow
INVASE (Yoon et al., 2019)	98.313%	98.299%	2.353%	1.048%
ICVH (Nguyen et al., 2021)	96.732%	96.725%	3.355%	3.195%
L2X (Chen et al., 2018)	98.261%	98.249%	2.253%	1.248%
VIBI (Bang et al., 2021)	96.626%	96.649%	2.504%	4.194%
AIM (Vo et al., 2023b)	98.406%	98.399%	1.853%	1.348%
AI2TALE (Ours)	99.324%	99.325%	0.451%	0.899%

Experiments

● Qualitative results

An entire email:

Bulk attention! Your discover account will close soon! Dear member, we have faced some problems with your account, so please update the account. If you do not update will be closed. To update your account, just confirm your information. (it only takes a minute). It's easy. 1. Click the link below to open a secure browser window. 2. Confirm that you're the owner of the account, and then follow the instructions."

Ground truth: Phishing

Model: Phishing

1st sentence: implies a sense of urgency (concern) via problems with your account while "Dear member" aims to establish a connection with the recipient and imply that the message comes from a trusted source. The phrase "please update the account" creates a sense of familiarity and consistency.

2nd and 3rd sentences: aim to minimize perceived effort, making the recipient more likely to comply without hesitation.

1st selected sentence

2nd selected sentence

3rd selected sentence

An entire email:

Your payment didn't succeed, so your ads have been suspended. This message was sent from a notification only email address that does not accept incoming email. Please do not reply to this message. If you have any questions, please visit the google ads help center. Hello advertiser, our attempt to charge your credit card for your outstanding google ads account balance was declined. Your account is still open. However, your ads have been suspended. Once we are able to charge your card and receive payment for your account balance, we will re activate your ads. Please update your billing information, even if you plan to use the same credit card. This will trigger our billing system to try charging your card again. You do not need to contact us to reactivate your account. To update your primary payment information, please follow these steps 1. Log in to your account at http ad words google com select. 2. Enter your primary payment information. 3. Click "update" when you have finished. Thank you for advertising with google ads. We look forward to providing you with the most effective advertising available. (c) google ads team 2008.

Ground truth: Phishing.

Model: Phishing.

1st sentence: exemplifies key tactics in the phishing email by creating a sense of urgency and alarm, prompting recipients to act quickly.

2nd and 3rd sentences: use authoritative language to appear legitimate and pressure the recipient into acting without careful consideration. 2nd sentence exploits the fear of an account issue, while 3rd sentence suggests a specific action, a common phishing tactic to collect sensitive data

Experiments

- **Human evaluation**

- Evaluate whether the most important selected sentences in phishing emails by AI2TALE are perceived as convincing information for users.
- **Participation:** 25 university students and staff (i.e., lecturers, professors, engineers, research scientists, and research fellows).
- **Note:** To maintain objectivity in the results, no information was provided about the source of the selected sentences.

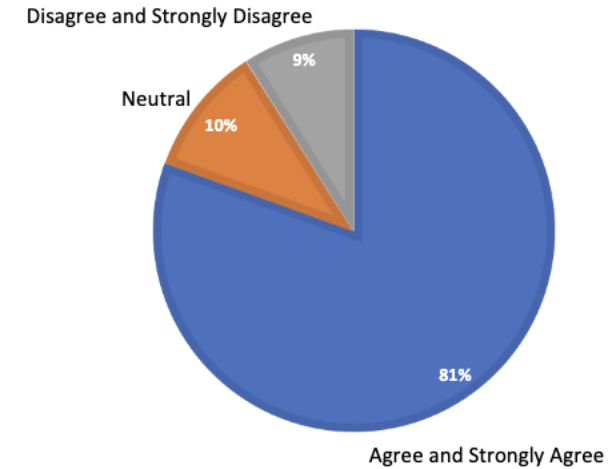


Figure 2. Human evaluation on the importance of the top-1 selected information (i.e., a sentence) from each email (by our AI2TALE method) in affecting and persuading users to follow the instructions from the email. We evaluate the selected sentences of 10 different phishing emails (randomly chosen from the testing set).

THANKS FOR YOUR ATTENTION

Question & Answer