

Email phishing attack detection based on BERT transformer model

Haiwei Li^{*a}, Jiahai Yang^a, Yuqi Li^b, Kun Li^b

^aThe Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China;

^bFirst Research Institute of the Ministry of Public Security of PRC, Beijing 100048, China

ABSTRACT

Due to the complexity and diversity of phishing attacks on social work emails, traditional detection methods are no longer able to meet the real-time and effective requirements for detecting phishing attacks on social work emails. Traditional methods of detection mainly rely on rules. With the continuous development of social work email phishing attacks, current social work email phishing attacks can easily avoid known rules, resulting in rules being unable to effectively respond to complex social work email phishing attacks. Traditional detection methods are difficult to effectively identify deceptive content in phishing emails from email workers, resulting in slow response times and potential security risks that can be exploited by attackers before they are known. In response to the poor detection performance of social work email phishing attacks in complex dimensions, this paper proposes a method of using Bidirectional Encoder Representations from Transformer (BERT) training model for detection. BERT model, as a deep learning model, can adapt to new social work email phishing attack methods and variants through continuous training and updates, maintaining the effectiveness and real-time detection; Unlike traditional rule-based matching methods, BERT can combine context for comprehensive analysis, improving the global perspective and accuracy of detection; And due to BERT's ability to handle contextual information, it can make more accurate judgments on emails containing deceptive language or information, thereby reducing false positive rates.

Keywords: Email phishing attack, rule detection, deep learning, natural language processing, BERT

1. INTRODUCTION

Email phishing attack¹ refers to the attack method in which attackers disguise themselves as trusted entities, send fraudulent emails to targets, and lure them to click on malicious links, download malicious attachments, or provide sensitive personal information. With the gradual modernization of domestic and international networks, email phishing attacks are becoming increasingly common and diverse. Attackers often disguise themselves as well-known companies or trusted individuals, sending seemingly legitimate emails to entice recipients to click on malicious links, download malicious attachments, or leak sensitive personal information. With the continuous advancement of technology, the disguise and deception methods of email attacks have become more difficult to identify, posing a serious threat to the information security of organizations and individuals. Therefore, successfully detecting email phishing attacks can prevent the leakage of sensitive information of organizations or individuals, protect organizations and individuals from information security threats, and promote the establishment and improvement of overall security culture.

Traditional methods for detecting email phishing attacks² are typically based on rules, feature engineering, and statistical analysis. They rely on prior knowledge and manually extracted features such as email subject, sender address, attachment type, etc. This method has high real-time performance for traditional email traffic detection, and its features and rule design are relatively transparent, making it easy to understand and debug. However, this type of method requires manual feature extraction and has poor adaptability to new or variant attacks. In the face of complex scenarios or new types of attacks, the false positive rate is relatively high, and the risk of false negatives is also relatively increased. The detection of email phishing attacks based on large model algorithms³ usually uses pre trained language models to understand the content and context of emails, in order to detect email phishing attacks. This type of method can learn the complex semantic and contextual information of email phishing attacks, improve the recognition ability of variant attacks, and does not require manual feature extraction, which can adapt to constantly changing attack patterns. In summary, although traditional methods have certain advantages in processing speed and transparency, their ability to handle complex semantic and variant attacks is relatively weak. However, methods based on large model detection can improve the accuracy and real-time performance of detecting complex semantic and variant attacks. Therefore, in

*lihw@gov110.cn

practical applications, large model detection research has gradually become the mainstream detection method for the increasingly complex and diverse social worker email phishing attack detection.

2. RELATED WORK

Traditional methods are difficult to detect social worker email phishing attacks with complex semantics in multiple dimensions. In response to this phenomenon, domestic and foreign research has gradually begun to apply large-scale model-based methods to detect email phishing attacks.

Heidi et al.³ constructed an AI program using four mainstream large-scale language models (GPT, Claude, PaLM, and LLaMA) to identify relevant user data points for automatic detection of phishing emails, and compared the results with manual detection. Experiments have shown that these language models have strong ability to detect malicious attacks even in subtle phishing emails. Kamble et al.⁴ proposed a model that combines linear constrained attention and deep learning for detecting phishing Uniform Resource Locator (URL). This model can combine character embedding with natural language processing (NLP) functionality, allowing it to utilize deep character relationships while displaying high NLP relevance. The experimental results show that the model can enhance online security and protect sensitive user information, and is better in accuracy than other existing phishing detection systems. It has good application prospects in effectively identifying and preventing phishing attempts. Lee et al.⁵ proposed a multi module-based email phishing detection system. The different modules of the system (structural module, text module, and URL module) can detect phishing attacks in different components of email, with the characteristics of larger coverage, greater flexibility, and less computational complexity. The experimental results indicate that the system achieved a high recall rate of 0.99 with a low false alarm rate of 1/10 K. Sreedhar et al.⁶ proposed a deep learning-based strategy method for identifying malicious phishing attacks. This method trains the model through convolutional neural networks and then uses random forest sampling to extract important features at different levels, thereby improving the detection and accuracy of the model. M. Rabbi et al.⁷ used natural language processing (NLP) and machine learning (ML) based methods to detect phishing emails. Six different ML algorithm functionalities were compared on two public datasets, and the experimental results showed that the combination of natural language processing (NLP) and machine learning (ML) had higher accuracy, precision, and recall in detecting phishing emails. Jain et al.⁸ analyzed phishing website URLs using machine learning algorithms such as decision trees, random forests, and support vector machines. The research results indicate that the above methods have good performance in detecting suspicious URL. Pitre et al.⁹ proposed a detection system based on the integration of blockchain and machine learning models to prevent phishing attacks from being transmitted on the platform. The system uses classification based gradient boosting algorithm and support vector machine algorithm to improve the detection efficiency of phishing attacks. Doda et al.¹⁰ utilized polynomial naive Bayes, recurrent neural networks, and support vector machine algorithms from deep learning on large datasets to improve the detection performance of fraudulent emails and select the most effective spam detection algorithm. Gogoi and Ahmed¹¹ proposed a system for detecting phishing and fraudulent emails based on deep learning. The system achieves high detection accuracy and recall by using advanced pre trained transformer models. Chataut et al.¹² studied the effectiveness of Large Language Models (LLMs) in the key task of detecting phishing emails, and conducted experiments based on three major models: GPT-3.5, GPT-4, and custom ChatGPT. The experimental results showed that LLMs have certain potential in effectively identifying phishing emails and have certain applicability in practical applications of email security.

3. PROPOSED ALGORITHM MODEL

Email phishing attacks, as the name suggests, are cyber-crimes committed through email, where criminals use phishing techniques to forge information and gain the trust and response of victims, thereby stealing personal information, money, and other resources. With the rapid development of the network environment, the current network environment is in a stage of explosive growth of information. Faced with a variety of network data and information, it is difficult for users to accurately judge and distinguish various types of information. Therefore, more and more hackers use the Internet to “phishing” to expand the coverage of network attack, so as to seek more benefits¹³. The schematic diagram of email phishing attack is shown in Figure 1.

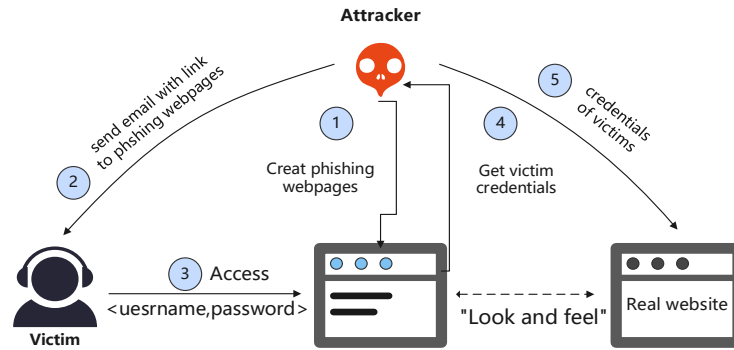


Figure 1. Phishing website attack process.

With the continuous development of technology, phishing email attacks are also constantly upgrading. Currently, methods such as hiding malicious files and phishing links in QR codes, third-party hosting websites, or encrypting and compressing attachments are gradually becoming popular. The protection system that relies on vulnerabilities, malicious files, and phishing techniques in email security operations is gradually becoming ineffective in new types of phishing emails. Traditional machine learning methods that rely on phishing content, such as logistic regression and naive Bayes, are also gradually losing their effectiveness due to the lack of analysis of data source dimensions.

In order to solve the above problems, this paper uses a parameterized BERT¹⁴ model to detect email phishing attacks is shown in Figure 2. The advantages of this method lie in its powerful representation ability and large-scale pre training of deep learning models, which can capture complex semantic information and quickly adapt to new attack methods, thereby providing higher accuracy and generalization ability. At the same time, it automates the processing of unstructured text data, significantly enhancing the effectiveness of network security protection.

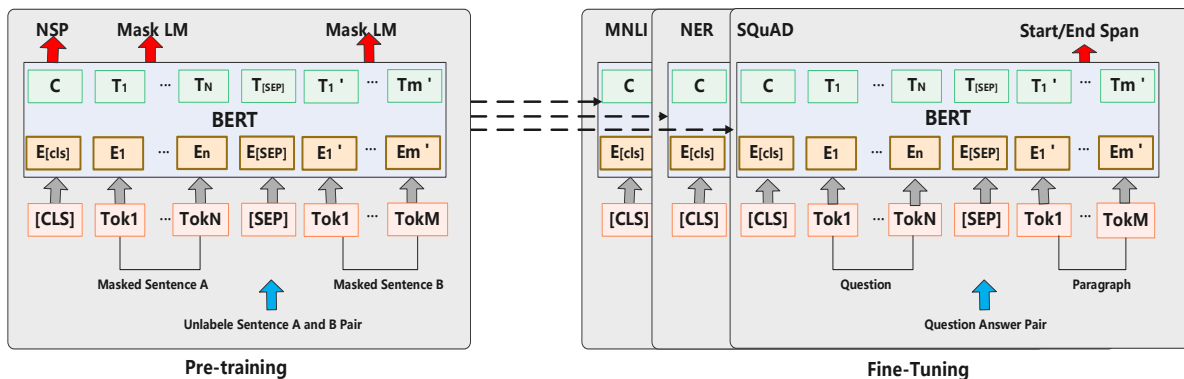


Figure 2. The structure of the proposed mode.

(a) Data preprocessing

Firstly, perform data preprocessing¹⁵ by segmenting the email text into a sequence of words or subwords, and converting each word or subword into its corresponding embedding vector; Then convert the segmented text into an input format that the model can accept, add tags, and fill in to ensure consistent input sequence length. The third step is to convert the processed text data into tensor form for input into the BERT model for processing and learning.

(b) Model training

Bidirectional Encoder Representation from Transformers originates from the Transformer model, which first proposed the use of multi head self-attention mechanism to replace the commonly used recurrent neural network mechanism in the field of natural language processing, enabling parallel computation of the training process and greatly improving efficiency. BERT belongs to a bidirectional Transformer based model that can simultaneously compute word sequences from two directions and establish connections between words. The structure of the BERT model is shown in the Figure 3.

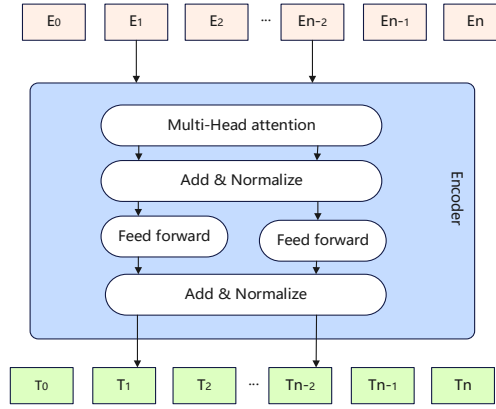


Figure 3. BERT model structure.

BERT contains three layers of encoders internally, with the first layer being the word encoder. BERT also has a pre trained lexicon that can convert input data into digital vectors. The second layer is a position encoder, which performs position encoding on each word in order to distinguish the different meanings of different words at different positions. When the word position is the base digit, it need to use equation (1) for calculation, and when the word position is the even digit, it need to use equation (2) for calculation, where t is the time, j is the word position, and d_E is the dimension of the input sequence.

$$PE_{t,j} = \text{Sin} \left(\frac{t}{10000^{\frac{j}{d_E}}} \right) \quad (1)$$

and

$$PE_{t,j} = \text{Cos} \left(\frac{t}{10000^{\frac{j-1}{d_E}}} \right) \quad (2)$$

The third layer is the sentence encoder, known as segmentation embedding in the BERT model. The sentence encoder can ensure sentence independence and effectively distinguish between the input sentence and the target sentence. BERT combines three layers of encoding as inputs for the entire model, as shown in the figure above where $\{E_0, E_1, E_2, \dots, E_{n-2}, E_{n-1}, E_n\}$ vectors are input vectors.

This model has a multi head attention mechanism, which essentially combines the computational results of multiple independent attention mechanisms (Self Attention). The specific calculation method is shown in equation (3). Among Q , K , and V are three matrices from the same input.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3)$$

In the multi head attention mechanism, each attention mechanism has exactly the same input, and independently outputs a subspace through linear transformation of the input data. Finally, the results of each “head” are connected. This mechanism can effectively prevent overfitting and learning the comprehensive representation information of samples. The calculation formula is shown in equation (4).

$$\text{Attention}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (4)$$

Finally, the Sofmax algorithm is used to normalize the results, which are then multiplied by the matrix V to obtain the weighted sum result. Through the above steps, the BERT model can effectively identify phishing attacks in emails, thereby enhancing network security protection capabilities.

4. EXPERIMENT

The dataset for this experiment originates from four entities engaged in finance, securities, insurance, and banking. We extracted actual email data from three working days for statistical analysis. The specific experimental results are presented in Table 1.

Table 1. Comparison of experimental data results.

Unit name	Industry	Total number of emails	The alarm of rule detection engine	The accuracy rate of rule detection engine	The alarm of AI detection engine	The accuracy of AI detection engine	The improvement rate of AI engine combined with rule detection
A	Security	198667	60325	100.00%	62366	99.99%	3.38%
B	Finance	6321	265	98.15%	4330	99.84%	1533.96%
C	Insure	222774	10	66.67%	171	97.71%	1610.00%
D	Bank	55256	3361	98.71%	5311	98.52%	58.02%

Through the analysis of the actual email detection environment of the above four important industry units, it can be seen that the upgrade of the AI detection model through sample analysis has well covered different templates of black and gray production phishing emails and the detection of variants that bypass detection. After the AI model upgrade, the detection rate of phishing emails for three customers has significantly improved. The AI detection engine of one of the customers has achieved a detection accuracy rate of 97.71%, which is much higher than the traditional engine detection of 66.67%. The reason is that the AI model has good generalization ability, which can draw inferences from one instance to others based on training samples. It can not only generalize to phishing emails, but also achieve good results with a small number of samples for normal emails and business emails.

5. CONCLUSION

The large model proposed in this article can improve the detection capability of email phishing attacks in important industry units. The experimental results show that the upgrade of the AI detection model effectively covers different templates of phishing emails of the same type in black and gray production, as well as variant detection that bypasses detection. Based on the training samples, it can generalize not only to phishing emails but also to normal and business emails, achieving good results with a small number of samples. However, the detection effect based on the combination of traditional rules and large models is better than that of AI detection model algorithms. How to combine rules and large models is the focus of future research.

REFERENCES

- [1] Althobaiti, K., Wolters, M. K., Alsufyani, N. and Vaniea, K., "Using clustering algorithms to automatically identify phishing campaigns," *IEEE Access* 11, 96502-96513 (2023),
- [2] Do, N. Q., Selamat, A., Lim, K. C., Krejcar, O. and Ghani, N. A. M., "Transformer-based model for malicious URL classification," *2023 IEEE International Conference on Computing (ICOCO)*, 323-327 (2023).
- [3] Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P. S., "Devising and detecting phishing emails using large language models," *IEEE Access* 12, 42131-42146 (2024).
- [4] Kamble, N. and Mishra, N., "Securing cyberspace: unveiling phishing attacks through deep neural networks for enhanced detection," *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 1-5 (2024).
- [5] Lee, J., Tang, F., Ye, P., Abbasi, F., Hay, P. and Divakaran, D. M., "D-Fence: A flexible, efficient, and comprehensive phishing email detection system," *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, Vienna, Austria, 578-597 (2021).

- [6] Velpula, S., Parise, R., Vamsi, N. K. and Chaitanya, S. K., "Phishing attack detection using convolutional neural networks," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 1381-1385 (2023).
- [7] Rabbi, M. F., Champa, A. I. and Zibran, M. F., "Phishy? Detecting phishing emails using ML and NLP," 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA), 77-83 (2023).
- [8] Jain, N., Jaiswal, P., Sharma, S., Sharma, K. and Sharma, V., "A machine learning based approach to detect phishing attack," 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 305-309 (2023).
- [9] Pitre, V., Joshi, A. and Das, S., "Blockchain and machine learning based approach to prevent phishing attacks," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), 1-6 (2023).
- [10] Dodda, R., Maddhi, S., Thuraab, M. S., Reddy, A. N. and Chandra, A. S. M., "NLP-Driven strategies for effective email spam detection: A performance evaluation," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), 275-279 (2023).
- [11] Gogoi, B. and Ahmed, T., "Phishing and fraudulent email detection through transfer learning using pretrained transformer models," 2022 IEEE 19th India Council International Conference (INDICON), 1-6 (2022).
- [12] Chataut, R., Gyawali, P. K. and Usman, Y., "Can AI keep you safe? A study of large language models for phishing detection," 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), 548-554 (2024).
- [13] Bitaab, M., et al., "Scam pandemic: How attackers exploit public fear through phishing," 2020 APWG Symposium on Electronic Crime Research (eCrime), 1-10 (2020).
- [14] Giri, S., Banerjee, S., Bag, K. and Maiti, D., "Comparative study of content-based phishing email detection using global vector (GloVe) and bidirectional encoder representation from transformer (BERT) word embedding models," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 1-6 (2022).
- [15] Yadav, J., Kumar, D. and Chauhan, D., "Cyberbullying detection using pre-trained BERT model," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 1096-1100 (2020).