An integrated Chinese malicious webpages detection method based on pre-trained language models and feature fusion

Yanting Jiang[1,2(✉)] and Di Wu[3,4]

1. Chengdu Aeronautic Polytechnic, Chengdu 610100, China.
2. Sichuan University of Media and Communications, Chengdu 611745, China.
3. Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China.
4. China University of Chinese Academy of Sciences, Shenzhen 518000, China.

jiangyanting@mail.bnu.edu.cn

Abstract. This paper proposed an integrated Chinese malicious webpages detection method. Firstly, we collected and released a Chinese malicious webpages detection dataset called "ChiMalPages" containing URLs and HTML/JavaScript files, and specified the detailed types of malicious pages according to relevant laws. Secondly, we designed a feature template for Chinese webpages and ranked each feature's importance based on information gain of the Random Forest algorithm. Thirdly, we fine-tuned BERT on the external URLs classification task and text on webpages, respectively producing new models "BERT-URL" and "BERT-web-text". The performance of pre-trained models is obviously superior to the baseline models. Finally, we integrated features from manual templates, BERT-URL and BERT-web-text, and the classification F1 score reaches 79.84%, increasing by 7.37% compared with manually designed webpage features. Experiments proved that our method based on BERT is useful and not biased on detailed classes.

Keywords: Chinese malicious webpages detection dataset, URLs, webpages text, pre-trained language models, feature importance ranking

1 Introduction

Internet has been an indispensable part of people's daily life. According to the statistical report from Chinese Internet Network Information Center, by the end of June 2021, the number of Chinese webpages has exceeded 335 billion[1].

However, as Chinese Internet booms, a number of malicious webpages have emerged. "Malicious webpages" is a general concept, which doesn't have precise definition and mainly involves: (1) webpages embedded with virus. (2) phishing webpages disguised as normal ones. (3) webpages including illegal content such as gambling, porn and illegal trades[2]. According to an incomplete survey, there are over 320000 discovered malicious Chinese websites and more pages in 2020[3]. Malicious pages are suspected of cybercrime, being a big threat to people's privacy and property.

Chinese malicious webpages are from websites whose registrars are in China[1]. And automatically detecting them is a challenging task. Firstly, most malicious pages

are short-lived and related datasets are scarce, especially for Chinese webpages[4]. Secondly, webpages' features are complex: URL(Uniform Resource Locator), HTML pages, JavaScript(JS) codes and so on. And many previous works only focused on URLs[5-7]. Thirdly, disguise tricks of malicious pages would evolve and their features are not fixed. Traditional ways to detect malicious pages need improvements.

Recently pre-trained language models have shown their power on sequence modeling and transfer learning, in this paper, we proposed an integrated method to detect Chinese malicious webpages. The main contributions of this paper are as follows:

(1) We opened a malicious Chinese webpages detection dataset "ChiMalPages", which contains URLs, HTML/JS files of each webpage. And we specified detailed types of malicious pages, and cited relevant laws as evidence.

(2) For English webpage features proposed by previous works, we analyzed and ranked each feature's importance based on Information Gain of the Random Forest algorithm, finding some features for English pages may not be suitable for Chinese ones.

(3) We proved the powerfulness of pre-trained models (BERT[8], RoBERTa[9] et al) on the URLs sequence modeling and classification task via transfer learning.

(4) We proposed an integrated method to detect malicious webpages, which combines pre-trained language models with the manual feature template. The experiment showed this integrated method is obviously superior to previous methods.

2 Related work

2.1 Malicious webpages dataset construction

In terms of malicious webpages data, it's not so hard to have access to malicious URLs. For example, "phishtank.org" is a widely-used website to download malicious URLs of English pages. The GitHub repository "CN-Malicious-website-list" also store enough malicious URLs of Chinese webpages[2]. However, it's difficult for researchers to get malicious pages' HTML and JS files[4]. This is mainly because the survival time of most malicious websites is quite short.

The data of related researches are shown in Table 1, from which we can tell that none of datasets are open to the whole research community.

Table 1. Previous Malicious webpages data information

Works	Num of malicious pages	Data open or not	Language	If specify pages' types
Wang [10]	500	Not	Mixed	Not
Gowtham [11]	1764	Not	English	Not
Xu [12]	850	Not	Mixed	Not
Ye [13]	2345	Not	Mixed	Not
Wei [14]	500	Not	Mixed	Not

Hu [15]	400	Not	English	Not
Wu [16]	1456	Not	Mixed	Not
Zhou [17]	249	Not	Chinese	Not
Chen [18]	856	Not	Mixed	Not
Ours	521	Open[1]	Chinese	Specified

2.2 Feature extraction

Feature extraction is a process extracting useful information for malicious pages recognition. The webpage features can be divided into 2 classes: static and dynamic features. Dynamic features mainly involve browser actions, pages skipping relations, HTTP requests and so on[14]. Extracting dynamic features is very difficult and usually needs auxiliary techniques such as virtualization and Honeypot[19]. As a result, we mainly focus on static features. Static features are usually from pages' static information, including hosts information[20], URL features[21], web contents (such as HTML, JS tags)[22]. To extract these features is not so difficult.

Although URLs are not natural languages, they are similar with them in a way because they are both unstructured sequences[23]. URL Word embedding[16], character-level Convolutional Neural Networks (CNN)[24], Bidirectional Long-Short Term Memory (Bi-LSTM)[5] and attention mechanism[6] were applied to perform malicious URL feature selection and detection. However, these works have not combined features from webpage files and URLs.

2.3 Identification methods

The main identification methods of malicious web pages include (1) blacklists, (2) heuristic rules, (3) machine learning algorithms.

The traditional way to identify malicious websites is blacklists[25]. Web browsers collect the human-reported URLs of malicious pages and store them into blacklists, then query whether the URLs of targeted pages are in blacklists. The main drawback of this way is delays in discovering new sites and incomplete coverage. According to Sheng et al[26], for 47%~83% of all the phishing websites, after being found by people, it was over 12 hours before they were stored into blacklists. This means harmful webpages would not be detected until they have caused damage.

The second way is heuristic rules. These heuristic rules often assume that the statistical or binary features of malicious websites are fixed. For large-scale web pages, this method would not only lead to high error rates, but also be hard to update because it relies on the domain knowledges[19].

[1] https://github.com/JiangYanting/Chinese_Malicious_Web_Pages_Dataset_And_Detection

The third way is machine learning algorithms, which regard malicious websites recognition as a supervised classification task. After extracting features from each web pages, Naïve Bayes, SVM, logistic regression[19] and Fully-Connected Networks (FCN)[24, 27] were constructed for classification.

In conclusion, there exist some aspects which need improvements in the previous works. Firstly, it's necessary to construct an open malicious websites recognition dataset for Chinese webpages. There exist huge differences between Chinese and English webpages. For example, Chinese PinYin is extremely rare in English pages' URLs. What's more, specifying the detailed types of malicious pages would make the dataset more fine-grained and practical.Secondly, datasets containing both URLs, HTML and JS files are scarce. It's obvious that judging a webpage benign or malicious only according to its URL is biased. Thirdly, it's noticeable that the natural language in webpages has been ignored in most of previous researches. Fourthly, pre-trained language models were not used to extract the webpages' features, which may be useful to capture potential features of the malicious pages.

3 Method framework

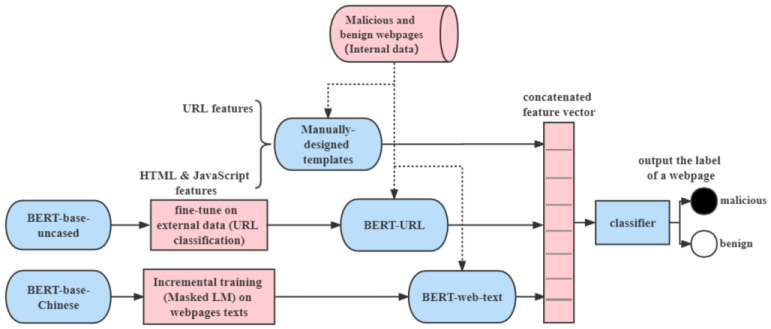

Fig. 1 General framework of malicious webpages detection

The framework of malicious webpages detection method is shown as Fig. 1. We planned to implement transfer learning mechanism of pre-trained language model BERT to detect malicious webpages. The datasets include two parts: external dataset and internal dataset. The purpose of the external dataset is to fine-tune BERT, and the internal dataset is aimed at testing the performance of our method. The feature vector of a webpage is from 3 channels: manually designed feature templates, BERT-URL and BERT-web-text. Each of them would be specified then.

3.1 Construction of Chinese malicious webpages detection dataset

The dataset "ChiMalPages" includes two parts: external dataset and internal dataset. The external dataset contains only URLs, whose goal is to fine-tune the pre-trained language models, and enable them to learn URLs representations. The internal dataset contains not only URLs but also HTML and JS files, whose goal is to evaluate the performance of our detection method finally.

Table 2. External and internal data

Data type	Contents	Scale
External data	Pure URLs	27000 benign URLs 27000 malicious URLs
Internal data	URLs, HTML and JavaScript files	954 benign pages, 521 malicious pages

According to similar works in English pages, benign samples are from Chinese websites with high traffic amount, such as "sohu.com", "mail.qq.com", "weibo.com", "pku.edu.cn" and "Tmall.com", which can be collected from websites "www.chinaz.com" and "www.alexa.cn". It's worth mentioning that websites traffic amount itself would not be included in samples, to enable our model to focus on pages' contents rather than predict webpages' traffic amount. Malicious URLs in external data are from "CN-Malicious-website-list"[2]. Malicious samples in internal data were collected from the exposure platform on the website "Security Union"[2] . Many Chinese netizens reported the malicious websites to this platform, which would further check them. Before the malicious pages went invalid, we had collected their URLs, HTML and JS files in time. There are no same URLs in internal and external data.

After analyzing data and consulting relevant lawyers, the 521 collected Chinese malicious pages can be divided into 5 types roughly. And their types, numbers, and laws evidence from the People's Republic of China (PRC) are shown in Table 3.

Table 3. Malicious pages types[3]

Malicious pages types	Num	Descriptions
1. gambling	172	**[Suspicion]** Crime of gambling, opening a casino.
2. porn	84	**[Suspicion]** Crime of organizing or introducing prostitution, crime of spreading obscene articles.
3. phishing	221	**[Details]** Pretend to imitate webpages of Chinese government, education, academic and publication institutes. Imitate webpages of banks, E-mail, e-commerce, phone company, travel agencies and so on. **[Suspicion]**Crime of illegally using network.
4. other pages breaking laws	43	**[Details]** Illegal trading of Taobao account, QQ account and phone number, setting up game servers without permission, selling game cheating tools, selling imitations. **[Suspicion]**Crime of illegally using information network, destroying computer information system, infringing copyright and so on.
5. undesirable pages	33	**[Details]**These pages do not break laws obviously, but they concern excessive ads, popups, forcing downloading software and other problems.

It is worth mentioning that the malicious pages classification is rough. In fact, some malicious pages lie on the border of classes and may have more than one label. For example, some gambling websites also have sexual suggestion, and some phishing

[2] https://jubao.anquan.org/exposure

[3] Some pages belong to more than one class, especially porn and gambling pages.

webpages also carry gambling and porn information on inconspicuous places. And many pages of the above 5 classes have excessive ads and popups, or force clients to download unknown software.

3.2 Manually designed feature template

Many previous researches manually designed features for web pages. Chiew et al[22] have summarized these feature items to a list. According to previous work on English webpages, we constructed a webpage feature template as Table 4 shows:

Table 4. Feature templates of webpages

No	Feature	Description
0	Num_Dots	Num of dots in URLs.
1	Url_Length	Num of chars in URLs
2-7	Special_chars	Num of "-""~""&""#""_" "@" in URLs
8	Numeric	Num of numeric chars in URLs
9	IpAddress	If IP address is used in URLs
10	top-domain	If top-level domains(.com, .cn, .net) are used in URLs.
11	2th domain	If second-level domains(.gov, .edu) are used in URLs.
12	Sensitive	Num of sensitive words ("secure", "account", "login","signin", "comfirm", "banking") in URLs.
13-24	HTML / JS tags	Num of tags "iframe" "eval" "setTimeout" "setInterval" "window.location" "window.open" "setAttribute""innerHTML""encodeURI""hidden" "display:none""download" in HTML / JS files.
25	External URL	Num of URLs in HTML / JS files whose domains are different from that of this page's URL.
26	JS_proportion	Proportion of JS files size in the webpage folder.
27	Pic-Txt ratio	Ratio of num of images to length of text in webpages.
28	Approval	If website approve information is provided in webpages.
29	If_unicode	If Unicode chars appear in HTML / JS files.

After constructing the template, we computed the importance values of each feature items(No.0 ~ No.29) in the internal data based on the Random Forest(RF) classification algorithm. RF is a kind of ensemble learning, which generates various subsets of training data and trains various sub-classifiers Decision Tree(DT). Beginning with the root node, DT computes the Information Gain(IG) values of each feature item to determine the best partition feature, and partition samples to child nodes. With the partition process, the purity of samples in child nodes gradually improves, until the samples in child nodes belongs to the same class as far as possible. And after the voting of every classification tree, RF model outputs the predicted label[28].

Based on RF, the way to compute importance values is as follows: in a classification tree of RF, a node corresponds to a feature item f_item and the sample set D. And for the sample set D in a node, the num of classes is n, the num of samples is x, and the proportion of Class i (i = 1, 2,…, n) is p_i , then the information entropy of D Entropy_D can be computed as follows:

$$Entropy_D = \sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

The higher Entropy_D is, the harder classification for D is. If the node containing D has m child nodes, and the information entropy of each child node is $Entropy_{sub\_j}$(j=1, 2,…, m)， the num of samples in each child node is x_{sub_j}, then the importance value of the node and corresponding feature item f_item is $\text{IG}_{\text{f\_item}}$:

$$\text{IG}_{\text{f\_item}} = x * Entropy_D - \sum_{j=1}^{m} x_{sub\_j} Entropy_{sub\_j} \quad (2)$$

Similarly, the importance values of every feature in a tree can be gained. Then compute the average values of each feature item on all the trees of a RF model. After normalization, RF gained final importance values of 30 features (see Fig. 2).

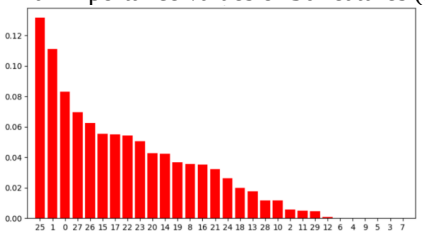

Fig. 2 Feature importance ranking[4]

From Fig. 2 we could tell that the 5 most important features in Table 4 are "External URL""Url_Length""Num_Dots""Pic-Txt ratio" and "JS_proportion". And "Pic-Txt ratio" played an important role in Chinese malicious pages detection in deed, which has not been noticed in the previous works. This indicated that many Chinese malicious webpages may carry their sensitive contents via pictures instead of text, thus escaping being detected. The features in Fig.2 whose importance ranks from 25th-30th are all relevant to URLs (feature No.6, No.4, No.9, No.5, No.3, No.7 in Table 4). These items were all related to URLs and borrowed from works about English malicious webpages detection. However, their importance is so little. This result reflected that the feature template aimed at English webpages cannot be copied directly to Chinese webpages. Besides, the disguise methods of malicious pages would change, so the feature engineering shouldn't never be updated.

3.3 BERT-URL based on external pure URLs classification

In our task, to gain a vector representation good at differentiating benign and malicious URLs, we try to fine-tune "BERT-base-uncased" via external data in Table 2, which contains 27000 benign URLs and 27000 malicious URLs.

[4] Numbers on the horizontal coordinates corresponds to numbers of feature items in Table 4.

For the URL 2-class classification task, BERT extracted the feature vector v of the marker [CLS] on the top layer as the integrated representation of a URL, and then added a 768*n Fully-connected layer W (n is the number of classes). Finally, through a softmax layer, the model output the probabilities that a URL belongs to each class c.

During training, the model would adjust W and parameters of 12 layers of BERT to maximize the probability corresponding to the true label.

In addition, we compared the external URLs classification performance of BERT with baseline models: the feature template in Table 4, URL char-level word2vec average vector, Fasttext and RoBERTa-base. The experimental results are shown in chapter 4.1.

3.4 BERT-web-text based on Masked Language Model

To make BERT-base-Chinese encode the style of webpage language, we further trained it via Masked Language Model (Masked LM)[8], which learns semantic information through a cloze task. After Masked LM, BERT-base-Chinese produces new model "BERT-web-text".

4 Experimental results and analysis

4.1 External URLs classification

We divided external data (27000 benign URLs and 27000 malicious URLs in Table 2) into the training, validation, and test data according to the proportion 8:1:1. All the training process followed the principle of Early Stopping to prevent overfitting. As for BERT, RoBERTa and DistilBERT, the initial learning rate is 2e-5, batch size is 16, dropout probability is 0.1. The URLs binary classification result is shown in Table 5.

Table 5. External URLs classification performance

Models	Accuracy on test data
URL Features in Table 4 except No.3~7 and No.9[5]	76.17%
word2vec(300-dimension)	78.39%
Fasttext(300-dimension)	84.97%
BERT-base-uncased	**87.01%**
RoBERTa-base	86.65%
DistilBERT-base-uncased	86.63%

As Table 5 shows, BERT-base-uncased performed best among 6 models and it produced the fine-tuned model "BERT-URL". To verify the influence to URL representation after fine-tuning on the URL classification task, we extracted 768-dimension URL vectors from different BERT models, and visualized them by the PCA (Principal Component Analysis) dimension reduction algorithm. We selected 100 malicious URLs and 100 benign URLs from test data for an example. The visualization result was shown as Fig. 3 and Fig. 4[6].

[5] These feature items have little importance according to the analysis of chapter 3.2.

[6] The blue dots represent benign URLs. The red dots represent malicious URLs.

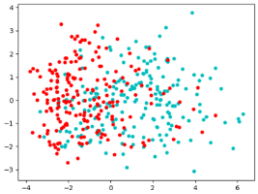

Fig.3 URL vectors BERT-base-uncased

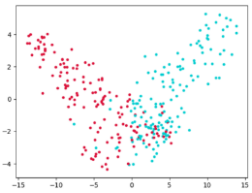

Fig. 4 URL vectors from BERT-URL

We can find that from the initial model "BERT-base-uncased" to "BERT-URL", URLs belonging to the same class tend to cluster, and distance between different clusters are lengthened. This phenomenon indicated that when doing the external URL classification, the parameters of BERT itself were also optimized.

4.2 Chinese malicious webpages detection on the internal data

We conducted classification experiments on internal data containing 954 benign webpages and 521 malicious webpages as Table 2 shows.

Random Forest(RF) and Fully-Connect Network(FCN) served as classifiers and the better one would be chosen. After 10-fold cross-validation, the detection results are the average on 10 test sets. And classifiers reaching the best performance were recorded. The results are in Table 6:

Table 6. Chinese malicious webpage detection results on internal data

No	Feature extraction method	Acc(%)	F1(%)
1	Table 4 except feature No.3~7 and No.9 which have little importance (hereafter abbreviated as **Template**)	80.61	72.47
2	URL features from **Template**	75.80	65.23
3	URL features from BERT-base-uncased	76.67	63.91
4	BERT-URL	81.09	72.25
5	**Template +** BERT-URL	82.58	75.04
6	BERT-URL + BERT-base-Chinese	84.67	78.41
7	BERT-URL + BERT-web-text	85.22	78.87
8	**Template** + BERT-URL + BERT-web-text	**85.76**	**79.84**

Some feature extraction methods are interpreted as follows：

"**Template**" means the webpages feature template based on Table 4, excluding feature items No.3~7 and No.9 which have little importance according to the analysis of chapter 3.2. "BERT-URL" was from BERT-base-uncased fine-tuned on external URL classification task, as chapter 4.1 described. "BERT-web-text" was from BERT-base-Chinese further pre-trained described in chapter 3.4. Inspired by Sun et al[29], if text length of a page is over 510, BERT extracted the representation of first 250 characters and the last 250 characters.

There are some findings according to Table 6. Firstly, when feature vectors have low dimensions, the classification performance of RF is superior to that of FCN. While the number of dimensions exceeds 1500(No.6 and No.7 in Table 4), FCN becomes better.

Secondly, comparing experiment No.2 ~ No.4, we found that the acc and F1 of BERT-URL increase by 5.29% and 7.02% compared with those of manually designed URL features. This indicated that after being fine-tuned on the external URLs, BERT-URL has the ability to distinguish malicious and benign URLs to a certain extent.

Thirdly, comparing experiment No.1 with No.7, although losing the HTML and JS features, the feature extraction method base on BERT-URL and BERT-web-text is still obviously superior to the feature **Template.** This result proved the usefulness of pre-trained language models on the malicious webpages detection task.

Fourthly, comparing experiment No.1 with No.8, after integrating the features from **Template**, BERT-URL and BERT-web-text, the performance reached the best level, indicating the feature fusion method we proposed plays a prominent part.

Finally, comparing experiment No.4 with No.7, No.5 with No.8, adding the webpages features from BERT-web-text can increase F1 by 6.62% and 4.8%. These experiments showing that natural language features in webpages cannot be neglected.

We selected one of the 10-fold cross-validation, and analyzed its detection results on detailed malicious webpages classes as Table 7 shows.

Table 7. detailed malicious pages detection performance based on Feature fusion

Webpages types	Num	Num of True Positive
Benign webpages	96	86
Malicious-porn pages	9	8
Malicious-gambling pages	20	16
Malicious-phishing pages	20	16
Other pages breaking laws	10	8

From Table 7 we could tell that the feature fusion method performed well on various malicious pages, which proving this detection method is comprehensive and not biased.

4.3 Webpages high-frequency words analysis

In addition, we counted the high-frequency words in malicious webpages and drew a wordcloud picture as Fig. 5. It's worth mentioning that regardless of context, some high-frequency words in malicious pages are hardly blamed in deed. For example, "成人" usually means "a mature, fully developed person legally responsible for his actions", while it may also serve as an euphemism for "porn" in some cases. As a result, taking context into consideration, BERT-web-text is more suitable for detecting sensitive contents than the simple character matching method based on sensitive vocabulary.

Fig. 5 Wordcloud of malicious webpages

5 Conclusion

Pre-trained language models have shown their power on sequence modeling[30]. This paper applied the pre-trained language models to malicious webpages detection, proposed the feature fusion method for Chinese malicious pages detection. Firstly, this paper released an open Chinese malicious webpages detection dataset, which contains external and internal data, specifying the detailed types of malicious pages and their legal risks. Secondly, we measured and ranked each feature's importance based on Information Gain as Fig. 2 shows, which can optimize the feature engineering. Thirdly, we fine-tuned BERT on external URLs classification and webpages text, producing new models "BERT-URL" and "BERT-web-text". The classification result showed the URL-encoding power of BERT obviously exceeds manual feature items, word2vec and Fasttext, even slightly exceeding RoBERTa and DistilBERT. Finally, combining features, the detection accuracy and F1 score reached 85.76% and 79.84%, respectively increasing by 5.15% and 7.37% compared with those of the machine learning method based on manual feature templates. These experiments proved that BERT works well on malicious webpages detection. Besides, BERT can tune webpage representation via transfer learning as webpages evolve. It extends the application fields of pre-trained models on cyber-security, and saves the cost of feature engineering, improving the efficiency of system update.

References

1. CNNIC. The 49th China Statistical Report on Internet Development. http://www.cnnic.cn/hlwfzyj/hlwxzbg/, last accessed 2022/03/28.
2. Zzhihao. CN-Malicious-website-list. https://github.com/zzhihao2017/CN-Malicious-website-list, last accessed 2022/03/28.
3. National Internet Emergency Center. 2020 China Internet Network Security Report. https://www.cert.org.cn/publish/main/17/index.html, last accessed 2022/03/28.
4. WAN M, YAO H. GAN model for malicious web training data generation. *Computer Engineering and Applications*, (6), 1-10 (2020).
5.Wang H , Yu L, Tian S W, et al. Bidirectional LSTM Malicious webpages detection algo rithm based on convolutional neural network and independent recurrent neural network. *Applied Intelligence*, (8), 3016-3026(2019).
6. Peng Y, Tian S, Yu L. A Joint Approach to Detect Malicious URL Based on Attention Mechanism. *International Journal of Computational Intelligence and Applications*, 18(3), (2019).
7. Sahoo D, Liu C, Hoi S C H. Malicious URL Detection using Machine Learning: A Survey. arXiv e-prints,1701-7179(2017).
8. Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints, 1810-4805(2018).
9. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints, 1907-11692(2019).
10. Tao W, Yu S, Xie B. A Novel Framework for Learning to Detect Malicious Web Pages. *2010 International Forum on Information Technology and Applications*,(2), 353-357(2010).
11. Gowtham, R., Krishnamurthi, et al. A comprehensive and efficacious architecture for de

tecting phishing webpages. *COMPUTERS AND SECURITY*, 40, 23-37(2014).
12. Xu L. A research of phishing detection technology based on deep learning. ChengDu: Uni versity of Electronic Science and Technology of China (2017).
13. Ye Z. Designing and application of a large-scale and fast malicious web page recognition method based on combinaton of Kafka and spark-streaming. Nanjing: Nanjing Univerisity of Posts and telecommunications (2019).
14. Wei X, Cheng W. Malicious web page recognition based on feature fusion and machine learning. *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, (5):95-104 (2019).
15. Hu Z, Wang C, Wu J, et al. Malicious Websites Identification based on Hyperlink Analysis and Classification rule. *Journal of Information Resources Management*, (1):105-113(2019).
16. Wu H. Research and implementation of activaye defense technology for malicious crawlers. Beijing: Beijing University of posts and telecommunications (2019).
17. Zhou W. Machine learning based malicious webpage analysis. Shanghai: Shanghai Jiaotong University (2019).
18.Chen B, Song L. Malicious Webpage Detection Method for Webpage Content Link Hierarchy Semantic Tree. *Computer Engineering and Applications*, (11):90-97 (2020).
19. Sha H, Liu Q, Liu T, et al. Survey on Malicious webpage detection research. *Chinese Journal of Computers*, (3):529-542(2016).
20. Seifert C, Komisarczuk P, Welch I, et al. Identification of malicious web pages through analysis of underlying DNS and web server relationships. *IEEE Conference on Local Computer Networks*, 935-941(2008).
21. Spirin N, Han J. Survey on web spam detection. *ACM SIGKDD Explorations Newsletter*, 13(2), 50(2012).
22. Chiew K L, Tan C L, Wong K S, et al. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, (484), 153-166 (2019).
23. Sahoo D, Liu C, Hoi S C H. Malicious URL Detection using Machine Learning: A Survey}. arXiv e-prints, 1701-7179(2017).
24. Le H, Pham Q, Sahoo D, et al. URLNet: Learning a URL Representation with Deep Learn ing for Malicious URL Detection. arXiv e-prints, 1802-3162(2018).
25. Peng P, Yang L, Song L. Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines. *The Internet Measurement Conference*, 478-485(2019).
26. Sheng S, Wardman B, Warner G, et al. An empirical analysis of phishing blacklists. *6th Conference on Email and Anti-Spam, CEAS 2009*, Mountain View, CA, United states (2009).
27. Saxe J, Berlin K. eXpose: A Character-Level Convolutional Neural Network with Embed dings For Detecting Malicious URLs, File Paths and Registry Keys. arXiv e-prints, 1702-8568 (2017).
28. Zhou Z. Machine Learning. Beijing: Tsinghua University Press, 178-181 (2016).
29. Sun C, Qiu X, Xu Y, et al. How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*, 194-206 (2019).
30. Yu, P. and X. Wang, BERT-Based Named Entity Recognition in Chinese Twenty-Four Histories. 17th International Conference of Web Information Systems and Applications, WISA 2020, 289-301(2020).