



Data Article

Datasets for phishing websites detection

Grega Vrbančič ^{a,*}, Iztok Fister Jr. ^a, Vili Podgorelec ^a

Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška cesta 46, Maribor SI-2000, Slovenia

ARTICLE INFO

Article history:

Received 25 September 2020

Revised 9 October 2020

Accepted 15 October 2020

Available online 23 October 2020

Keywords:

Phishing websites
 Classification
 Computer security
 Optimization

ABSTRACT

Phishing stands for a fraudulent process, where an attacker tries to obtain sensitive information from the victim. Usually, these kinds of attacks are done via emails, text messages, or websites. Phishing websites, which are nowadays in a considerable rise, have the same look as legitimate sites. However, their backend is designed to collect sensitive information that is inputted by the victim. Discovering and detecting phishing websites has recently also gained the machine learning community's attention, which has built the models and performed classifications of phishing websites. This paper presents two dataset variations that consist of 58,645 and 88,647 websites labeled as legitimate or phishing and allow the researchers to train their classification models, build phishing detection systems, and mining association rules.

© 2020 The Author(s). Published by Elsevier Inc.
 This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

* Corresponding author.

E-mail addresses: grega.vrbancic@um.si (G. Vrbančič), iztok.fister1@um.si (I. Fister Jr.), vili.podgorelec@um.si (V. Podgorelec).

Social media: (G. Vrbančič)

Specifications Table

| | |
|--------------------------------|---|
| Subject | Computer Science |
| Specific subject area | Artificial Intelligence |
| Type of data | csv file |
| How data were acquired | Data were acquired through the publicly available lists of phishing and legitimate websites, from which the features presented in the datasets were extracted. |
| Data format | Raw: csv file |
| Parameters for data collection | For the phishing websites, only the ones from the PhishTank registry were included, which are verified from multiple users. For the legitimate websites, we included the websites from publicly available, community labeled and organized lists [1], and from the Alexa top ranking websites. |
| Description of data collection | The data is comprised of the features extracted from the collections of websites addresses. The data in total consists of 111 features, 96 of which are extracted from the website address itself, while the remaining 15 features were extracted using custom Python code. |
| Data source location | Worldwide |
| Data accessibility | Repository name: Mendeley Data Data identification number: 10.17632/72ptz43s9v1 Direct URL to data: https://doi.org/10.17632/72ptz43s9v1 |
| Related research article | Vrbančič, Grega, Iztok Fister Jr, and Vili Podgorelec. "Parameter setting for deep neural networks using swarm intelligence on phishing websites classification." International Journal on Artificial Intelligence Tools 28.06 (2019): 1960008. DOI: 10.1142/S021821301960008X |

Value of the Data

- These data consist of a collection of legitimate, as well as phishing website instances. Each website is represented by the set of features that denote whether the website is legitimate or not. Data can serve as input for the machine learning process.
- Machine learning and data mining researchers can benefit from these datasets, while also computer security researchers and practitioners. Computer security enthusiasts can find these datasets interesting for building firewalls, intelligent ad blockers, and malware detection systems.
- This dataset can help researchers and practitioners easily build classification models in systems preventing phishing attacks since the presented datasets feature the attributes which can be easily extracted.
- Finally, the provided datasets could also be used as a performance benchmark for developing state-of-the-art machine learning methods for the task of phishing websites classification.

1. Data Description

The presented dataset was collected and prepared for the purpose of building and evaluating various classification methods for the task of detecting phishing websites based on the uniform resource locator (URL) properties, URL resolving metrics, and external services. The attributes of the prepared dataset can be divided into six groups:

- attributes based on the whole URL properties presented in [Table 1](#),
- attributes based on the domain properties presented in [Table 2](#),
- attributes based on the URL directory properties presented in [Table 3](#),
- attributes based on the URL file properties presented in [Table 4](#),
- attributes based on the URL parameter properties presented in [Table 5](#), and
- attributes based on the URL resolving data and external metrics presented in [Table 6](#).

Table 1

Dataset attributes based on URL.

| Nr. | Attribute | Format | Description | Values |
|-----|----------------------|-----------------------------------|-------------|--------|
| 1 | qty_dot_url | Number of "." signs | Numeric | |
| 2 | qty_hyphen_url | Number of "-" signs | Numeric | |
| 3 | qty_underline_url | Number of "_" signs | Numeric | |
| 4 | qty_slash_url | Number of "/" signs | Numeric | |
| 5 | qty_questionmark_url | Number of "?" signs | Numeric | |
| 6 | qty_equal_url | Number of "=" signs | Numeric | |
| 7 | qty_at_url | Number of "@" signs | Numeric | |
| 8 | qty_and_url | Number of "&" signs | Numeric | |
| 9 | qty_exclamation_url | Number of "!" signs | Numeric | |
| 10 | qty_space_url | Number of " " signs | Numeric | |
| 11 | qty_tilde_url | Number of "˜" signs | Numeric | |
| 12 | qty_comma_url | Number of ";" signs | Numeric | |
| 13 | qty_plus_url | Number of "+" signs | Numeric | |
| 14 | qty_asterisk_url | Number of "*" signs | Numeric | |
| 15 | qty_hashtag_url | Number of "#" signs | Numeric | |
| 16 | qty_dollar_url | Number of "\$" signs | Numeric | |
| 17 | qty_percent_url | Number of "%" signs | Numeric | |
| 18 | qty_tld_url | Top level domain character length | Numeric | |
| 19 | length_url | Number of characters | Numeric | |
| 20 | email_in_url | Is email present | Boolean | [0, 1] |

Table 2

Dataset attributes based on domain URL.

| Nr. | Attribute | Format | Description | Values |
|-----|-------------------------|---------------------------------|-------------|--------|
| 1 | qty_dot_domain | Number of "." signs | Numeric | |
| 2 | qty_hyphen_domain | Number of "-" signs | Numeric | |
| 3 | qty_underline_domain | Number of "_" signs | Numeric | |
| 4 | qty_slash_domain | Number of "/" signs | Numeric | |
| 5 | qty_questionmark_domain | Number of "?" signs | Numeric | |
| 6 | qty_equal_domain | Number of "=" signs | Numeric | |
| 7 | qty_at_domain | Number of "@" signs | Numeric | |
| 8 | qty_and_domain | Number of "&" signs | Numeric | |
| 9 | qty_exclamation_domain | Number of "!" signs | Numeric | |
| 10 | qty_space_domain | Number of " " signs | Numeric | |
| 11 | qty_tilde_domain | Number of "˜" signs | Numeric | |
| 12 | qty_comma_domain | Number of ";" signs | Numeric | |
| 13 | qty_plus_domain | Number of "+" signs | Numeric | |
| 14 | qty_asterisk_domain | Number of "*" signs | Numeric | |
| 15 | qty_hashtag_domain | Number of "#" signs | Numeric | |
| 16 | qty_dollar_domain | Number of "\$" signs | Numeric | |
| 17 | qty_percent_domain | Number of "%" signs | Numeric | |
| 18 | qty_vowels_domain | Number of vowels | Numeric | |
| 19 | domain_length | Number of domain characters | Numeric | |
| 20 | domain_in_ip | URL domain in IP address format | Boolean | [0, 1] |
| 21 | server_client_domain | "server" or "client" in domain | Boolean | [0, 1] |

The first group is based on the values of the attributes on the whole URL string, while the values of the following four groups are based on the particular sub-strings, as presented in [Figure 1](#). The last group attributes are based on the URL resolve metrics as well as on the external services such as Google search index.

The dataset in total features 111 attributes excluding the target *phishing* attribute, which denotes whether the particular instance is legitimate (value 0) or phishing (value 1). We prepared two variations of the dataset, the one where the total number of instances is 58,645 and the balance between the target classes in more or less balanced with 30,647 instances labeled as phishing websites and 27,998 instances labeled as legitimate. The second variant of the dataset

Table 3

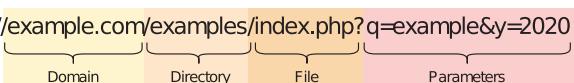
Dataset attributes based on URL directory.

| Nr. | Attribute | Format | Description | Values |
|-----|----------------------------|--------------------------------|-------------|--------|
| 1 | qty_dot_directory | Number of "." signs | Numeric | |
| 2 | qty_hyphen_directory | Number of "-" signs | Numeric | |
| 3 | qty_underscore_directory | Number of "_" signs | Numeric | |
| 4 | qty_slash_directory | Number of "/" signs | Numeric | |
| 5 | qty_questionmark_directory | Number of "?" signs | Numeric | |
| 6 | qty_equal_directory | Number of "=" signs | Numeric | |
| 7 | qty_at_directory | Number of "@" signs | Numeric | |
| 8 | qty_and_directory | Number of "&" signs | Numeric | |
| 9 | qty_exclamation_directory | Number of "!" signs | Numeric | |
| 10 | qty_space_directory | Number of " " signs | Numeric | |
| 11 | qty_tilde_directory | Number of "˜" signs | Numeric | |
| 12 | qty_comma_directory | Number of "," signs | Numeric | |
| 13 | qty_plus_directory | Number of "+" signs | Numeric | |
| 14 | qty_asterisk_directory | Number of "*" signs | Numeric | |
| 15 | qty_hashtag_directory | Number of "#" signs | Numeric | |
| 16 | qty_dollar_directory | Number of "\$" signs | Numeric | |
| 17 | qty_percent_directory | Number of "%" signs | Numeric | |
| 18 | directory_length | Number of directory characters | Numeric | |

Table 4

Dataset attributes based on URL file name.

| Nr. | Attribute | Format | Description | Values |
|-----|-----------------------|--------------------------------|-------------|--------|
| 1 | qty_dot_file | Number of "." signs | Numeric | |
| 2 | qty_hyphen_file | Number of "-" signs | Numeric | |
| 3 | qty_underscore_file | Number of "_" signs | Numeric | |
| 4 | qty_slash_file | Number of "/" signs | Numeric | |
| 5 | qty_questionmark_file | Number of "?" signs | Numeric | |
| 6 | qty_equal_file | Number of "=" signs | Numeric | |
| 7 | qty_at_file | Number of "@" signs | Numeric | |
| 8 | qty_and_file | Number of "&" signs | Numeric | |
| 9 | qty_exclamation_file | Number of "!" signs | Numeric | |
| 10 | qty_space_file | Number of " " signs | Numeric | |
| 11 | qty_tilde_file | Number of "˜" signs | Numeric | |
| 12 | qty_comma_file | Number of "," signs | Numeric | |
| 13 | qty_plus_file | Number of "+" signs | Numeric | |
| 14 | qty_asterisk_file | Number of "*" signs | Numeric | |
| 15 | qty_hashtag_file | Number of "#" signs | Numeric | |
| 16 | qty_dollar_file | Number of "\$" signs | Numeric | |
| 17 | qty_percent_file | Number of "%" signs | Numeric | |
| 18 | file_length | Number of file name characters | Numeric | |



https://example.com/examples/index.php?q=example&y=2020

Fig. 1. Separation of the whole URL string into sub-strings.

is comprised of 88,647 instances with 30,647 instances labeled as phishing and 58,000 instances labeled as legitimate, the purpose of which is to mimic the real-world situation where there are more legitimate websites present. The distribution between the classes of both dataset variants is presented in [Figure 2](#).

Table 5

Dataset attributes based on URL parameters.

| Nr. | Attribute | Format | Description | Values |
|-----|-------------------------|--|-------------|--------|
| 1 | qty_dot_params | Number of "." signs | Numeric | |
| 2 | qty_hyphen_params | Number of "-" signs | Numeric | |
| 3 | qty_underline_params | Number of "_" signs | Numeric | |
| 4 | qty_slash_params | Number of "/" signs | Numeric | |
| 5 | qty_questionmark_params | Number of "?" signs | Numeric | |
| 6 | qty_equal_params | Number of "=" signs | Numeric | |
| 7 | qty_at_params | Number of "@" signs | Numeric | |
| 8 | qty_and_params | Number of "&" signs | Numeric | |
| 9 | qty_exclamation_params | Number of "!" signs | Numeric | |
| 10 | qty_space_params | Number of " " signs | Numeric | |
| 11 | qty_tilde_params | Number of "˜" signs | Numeric | |
| 12 | qty_comma_params | Number of "," signs | Numeric | |
| 13 | qty_plus_params | Number of "+" signs | Numeric | |
| 14 | qty_asterisk_params | Number of "*" signs | Numeric | |
| 15 | qty_hashtag_params | Number of "#" signs | Numeric | |
| 16 | qty_dollar_params | Number of "\$" signs | Numeric | |
| 17 | qty_percent_params | Number of "%" signs | Numeric | |
| 18 | params_length | Number of parameters characters | Numeric | |
| 19 | tld_present_params | TLD ¹ present in parameters | Boolean | [0, 1] |
| 20 | qty_params | Number of parameters | Numeric | |

Table 6

Dataset attributes based on resolving URL and external services.

| Nr. | Attribute | Format | Description | Values |
|-----|------------------------|--|-------------|--------|
| 1 | time_response | Domain lookup time response | Numeric | |
| 2 | domain_spf | Domain has SPF ² | Boolean | [0, 1] |
| 3 | asn_ip | ASN ³ | Numeric | |
| 4 | time_domain_activation | Domain activation time (in days) | Numeric | |
| 5 | time_domain_expiration | Domain expiration time (in days) | Numeric | |
| 6 | qty_ip_resolved | Number of resolved IPs | Numeric | |
| 8 | qty_nameservers | Number of resolved NS ⁴ | Numeric | |
| 9 | qty_mx_servers | Number of MX ⁵ servers | Numeric | |
| 10 | ttl_hostname | Time-To-Live (TTL) | Numeric | |
| 11 | tls_ssl_certificate | Has valid TLS ⁶ /SSL ⁷ certificate | Boolean | [0, 1] |
| 12 | qty_redirects | Number of redirects | Numeric | |
| 13 | url_google_index | Is URL indexed on Google | Boolean | [0, 1] |
| 14 | domain_google_index | Is domain indexed on Google | Boolean | [0, 1] |
| 15 | url_shortened | Is URL shortened | Boolean | |
| 16 | phishing | Is phishing website | Boolean | [0, 1] |

2. Experimental Design, Materials and Methods

In the process of preparing the phishing websites datasets variants presented in [2], we followed common steps which were also used in the dataset preparation process of similar datasets presented by Mohammad et al. [3] and Abdelhamid et al. [4].¹²³⁴⁵⁶⁷

¹ Top-Level Domain² Sender Policy Framework³ Autonomous System Number⁴ Name Server⁵ Mail eXchanger⁶ Transport Layer Security⁷ Secure Socket Layers



Fig. 2. The distribution between classes for both dataset variations. The *dataset_full* denotes the larger dataset, while the *dataset_small* denotes the smaller dataset variation. The target class 0 denotes legitimate websites while the target class 1 denotes the phishing websites.

In the manner of such preparation process, we firstly collected a list of a total of 30,647 confirmed phishing URLs from the Phishtank [5] website. On the other hand, the list of legitimate URLs was obtained from Alexa ranking website⁸ from which we gathered 58,000 legitimate website URLs. Additionally, we have also obtained the list of 27,998 community labeled and organized URLs [1], which are the URLs pointing to the objectively reported news and are in that manner also legitimate.

From the URL lists of phishing and legitimate websites, we prepared, as already presented, two variants of the dataset. The smaller, more balanced dataset *dataset_small* comprises instances of extracted features from Phishtank URLs and instances of extracted features from community labeled and organized URLs representing legitimate ones. On the other hand, the larger, more unbalanced dataset consists of all of the instances from the *dataset_small* and the additional instances of extracted features from Alexa top sites URL list.

The complete process of extracting the features from the list of collected website addresses was conducted automatically, using a Python script. The extracting process is outlined in [Algorithm 1](#). Such procedure was conducted in total two times, each time given different set of website addresses as already described. The final outcome reflects in two csv files containing extracted features. The csv files are handy and easy to work with various tools and programming libraries.

⁸ <https://www.alexa.com>

Algorithm 1 Feature extraction process

Input: URLs
Input: signs
Output: dataset.csv

1: $i \leftarrow 0$
2: $totalURLs \leftarrow \text{length}(URLs)$ ▷ Get the number of URLs in array.

3: **while** $i < totalURLs$ **do**
4: $url \leftarrow URLs(i)$
5: $countsUrl \leftarrow \text{getCounts}(url, signs)$ ▷ Get signs and character counts.

6: **for** substring in $\text{splitURL}(url)$ **do** ▷ Iterate through the sub-strings of URL.
7: $countsSubstring \leftarrow \text{getCounts}(\text{substrings}, signs)$ ▷ Get signs and character counts.

8: **end for**

9: $measuredFeatures \leftarrow \text{fetchFeatures}(url)$ ▷ Get features from external services.
10: $\text{toCsv}(countsUrl, countsSubstring, measuredFeatures)$ ▷ Append row to csv.

11: $i \leftarrow i + 1$
12: **end while**

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

Authors acknowledge the financial support from the [Slovenian Research Agency](#) (Research Core Funding No. [P2-0057](#)).

References

- [1] C. Lab, Others, Url testing lists intended for discovering website, [Censorship](#) (2014). <https://github.com/citizenlab/test-lists>
- [2] G. Vrbancic, I.J. Fister, V. Podgorelec, Parameter setting for deep neural networks using swarm intelligence on phishing websites classification, *Int. J. Artif. Intell. Tools* 28 (6) (2019) 28, doi:[10.1142/S021821301960008X](https://doi.org/10.1142/S021821301960008X).
- [3] R.M. Mohammad, F. Thabtah, L. McCluskey, An assessment of features related to phishing websites using an automated technique, in: *Internet Technology And Secured Transactions, 2012 International Conference for, IEEE, 2012*, pp. 492–497.
- [4] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based associative classification data mining, *Expert Syst. Appl.* 41 (13) (2014) 5948–5959, doi:[10.1016/j.eswa.2014.03.019](https://doi.org/10.1016/j.eswa.2014.03.019).
- [5] OpenDNS, PhishTank data archives, 2018, Available at <https://www.phishtank.com/>, Accessed: 2018-01-17