Journal of Theoretical and Applied Information Technology
15th September 2017. Vol.95. No.17
© 2005 - Ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION MODEL BASED ON URL AND CONTENT FEATURE APPROACH FOR DETECTION PHISHING WEBSITE IN INDONESIA

[1]FEBRY EKA PURWIANTONO, [2]ARIS TJAHYANTO

Department of Information System, Institut Teknologi Sepuluh Nopember, Indonesia

E-mail: [1]vebryexa@gmail.com, [2]atjahyanto@gmail.com

ABSTRACT

This research proposed a classification model that can be used to detect phishing website accurately. This study takes a case study from Indonesia because data used are sites using Bahasa Indonesia, hosted in Indonesia and frequently accessed by Internet users from Indonesia. Dataset used in this research consisted of approximately 102 authentic websites and 364 phishing websites. The proposed detection technique based on website analysis using the URL and content feature based approach. This classification model combines several heterogeneous features from previous research and proposes new URL and content feature based approach that are expected to improve detection performance when compared with previous research. Moreover, in the proposed classification model created a web crawler to extract feature vectors in this research. This research uses four different algorithms such as Sequential Minimal Optimization (SMO), Naive Bayes, Bagging and Multilayer Perceptron. The result, SMO, Naive Bayes, Bagging and Multilayer Perceptron have accuracy of approximately 89.27%, 93.78%, 95.49% and 92.70%. Algorithm has the best accuracy is Bagging, it will be used in this classification model to compare with classification model in previous research using same dataset. The result, accuracy of classification model in this research outperformed accuracy of classification model in previous research. The classification model in this research outperform 5.79% against classification model in previous research which only yielded 89.70% accuracy.

Keywords: *Classification Model, Detection, Phishing Website, Indonesia, Feature*

1. INTRODUCTION

Phishing website is a site designed by Internet criminals in such a way as to resemble an authentic site (view, content, domain URL or other) to trick a victim (Internet users) by making victim as if accessing a web page from a legitimate source [1]. Website template will be made as closely as possible to authentic site to make sure the victim is on the right site. In addition, there are also phishing website that are designed specifically to provide false information or misleading instructions. If the victim is successfully tricked and submitted the requested information, Internet criminals can easily use the information on legitimate website to perform unwanted activities and of course this will cause significant losses for the victims include financial and data loss.

Online banking and e-commerce website are the most site that targeted of phishing by Internet criminals, because the potential benefits that can be obtained by Internet criminals quite large when compared to other sites. The most popular online banking and e-commerce websites are targeted by Internet criminals including eBay and PayPal [2]. However, not a few sites based on social media are targeted by Internet criminals such as Facebook, Twitter, Instagram and others. In addition to being used for data theft, phishing website is also used to scam Internet users in the name of legitimate sites and spreading computer malware/virus by Internet criminals. data theft, phishing sites also used to perform acts of fraud on behalf of a legitimate site and as the spread of malware/virus komputer by criminals on the Internet.

According to APWG (Anti-Phishing Working Group), public awareness of phishing website is increasing every year, but the number of losses caused and phishing website grow faster. In 4th Quarter 2016 APWG report, Phishing Activity Trends in October 2016 found 89,232 sites detected as phishing website, while in November and December 2016 each found 118,928 and 69,533 sites indicated as phishing website. In the report also found approximately 17 million new malware.

It can cause fear and decrease Internet user confidence in online transctions, whereas online transactions are currently booming in Indonesia. Therefore, Internet users are need a system capable to detect phishing website accurately to prevent and avoid losses caused by phishing website to Internet users. So the data used by researchers are sites using Bahasa Indonesia, hosted in Indonesia and frequently accessed by Internet users from Indonesia as case study. The research questions are how to distinguish phishing sites and authentic sites? and what features are used to detect phishing sites and how to get those features?

Some previous studies used data mining classification (create a classification model) to distinguish phishing website and authentic website. Phishing website detection system emerged as an important mechanism to eradicate phishing website that exist on Internet. Because most of phishing attacks usually steal important information from user by posing as a trustworthy site. Based on previous research, the most detection technique used is website analysis. In the website analysis mentioned that there are several approaches to detect phishing website such as the blacklist, visual similarity, URL and content features, and third-party search engine based approach [2].

Some previous studies were more inclined to use URL dan content features based approach to detect phishing sites. For example Zhang et al [2] create a classification model for detecting phishing websites using 15 feature vectors such as number of dot (.), age of domain, expired of domain and others, then processing data using several classification algorithms such as SMO (Sequential Minimal Optimization), Naive Bayes, Random Forest and Logistic Regression, while Li et al [3] use 12 feature vectors to detect phishing website such as average inbound link, average outbound links, average internal links and others. In the study created a web crawler to extract 12 feature vectors of website into a DOM (Document Object Model) tree before being processed in data mining tools using BVM (Ball-based Support Vector Machine), SVM (Support Vector Machine), Naive Bayes, Simple Logistic and some other classification algorithms.

The best detection accuracy value was obtained by Zhang et al [2] when using SMO algorithm of 95.38%, while Li et al [3] got the best TP (True Posistive) score of 0.965 when using SVM and Simple Logistic algorithms. TP is defined as proportion of sites that are completely positive among overall sites that show positive test results. However, researchers in that study prefer BVM algorithm which in fact has lower TP value of SVM algorithm and Simple Logistic is approximately approximately 0.964 or -0.001%. It is because calculations performed by BVM algorithm (0.15 seconds) are faster than SVM algorithm (0.35 seconds) and Simple Logistic (30 seconds).

In this study, researchers will also use the URL and content feature based approach and adopt heterogeneous feature selection (data used are sites using Bahasa Indonesia, hosted in Indonesia and frequently accessed by Internet users from Indonesia), because according to [4] heterogeneous feature selection can affect performance of classifier. Heterogeneous feature selection is expected to improve performance of classifier used in this study, resulting in good output. Most of the features to be used in this study are taken from [2] - [3], but some non-heterogeneous features are not selected.

In addition, researchers propose new heterogeneous features that are still based on URL and content approach, create feature vectors extraction based on web crawler and test some algorithms such as SMO, Naive Bayes, Bagging and Multilayer Perceptron to ensure that the classification model created can improve detection performance of phishing website, so accuracy, precision and training time are much faster when compared with previous study [2] that only use basic features.

Software used in data processing and classification modeling in this research is Weka. Weka is open source software that can be used for free to support various standard tasks in data mining such as clustering, association and classification [5]. Weka contains a collection of processes that include a variety of pre-processing techniques and data modeling techniques, which can help researchers test the classification model created in this study. Thus, the results of classification performance can be measured mathematically and validation can be trusted, so the results can be used to support other similar research in the future. Of course if this classification model is implemented, it can avoid and reduce the risk of Internet users exposed to malware attacks or hijacking from phishing website.

2. LITERATURE REVIEW

According to study by Zhang et al [2], the URL and content feature based approach focuses on analyzing the characteristics of the URL and the content of a target website. In that study was created a classification model that could detect phishing websites by involving unique domain features. The proposed model does not depend on prior knowledge or assumptions about authentic sites. The model in that research prioritizes URL and content feature based approach, since it is the most commonly used approach, because it is able to combine and evaluate detection features on a domain. By integrating new features on existing website with some detection feature prediction used in previous research, a feature vector was created for the proposed model that consists of two parts: URL features and web content features.

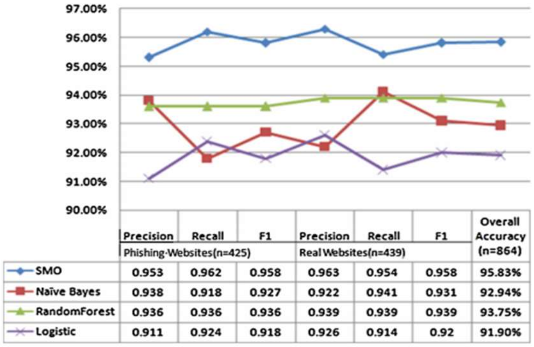
URL features include the following cues extracted from the URL of a target website:

- F1: Whether a URL contains an IP address
  Usually a phishing website contains an IP address. The URL of a target website contains an IP address instead of a domain name, then this feature variable F1 will be assigned a value of 1; otherwise 0.
- F2: Whether a URL contains the symbol '@'
  Phishing websites often insert @ into a URL that takes users to a website different from what Internet criminals expect. If a URL contains the symbol '@', F2 will be assigned a value of 1; otherwise 0.
- F3: Whether the characters in a URL are coded in UNICODE
  In comparison to a truthful website, a phishing website is more likely to use UNICODE in its URL to hide the URL of a truly intended website. F3 will be assigned a value of 1 if the domain name of the URL of a target website contains characters encoded in UNICODE; otherwise 0.
- F4: The number of dots ('.') in a URL
  Previous research [6] suggests that the larger the number of dots in a URL, the higher the possibility the website is a phishing website.
- F5: The number of suffixes in a domain name Users generally catch a glimpse of the first part of a URL but likely miss the remaining part, which actually points to a phishing website.
- F6: Age of a domain name
  Which is represented by the number of days passed since a domain name was registered.
- F7: Expiration time of a domain name
  Which is represented by the number of days remaining before a domain name expires.
- F8: Whether the address of a DNS (Domain Name System) server is consistent with a URL
  DNS server addresses can be obtained through whois domain name queries. If it matches, the value of F8 will be 1; otherwise 0.
- F9: Information about website registration
  F9 to represent whether a domain name is registered (1) or not (0).
- F10: Whether domain registered by organization
  Whether a domain name applicant is an individual (0) or an enterprise (1).
- F11: Whether domain privatized by owner
  F11 is used to represent whether a recorded website name and actual indicated site are consistent (1) or not (0).

Web content features are automatically extracted from the source code of a website and include the following:

- F12: Whether website contains ICP (Internet Content Provider) license number
  If the website contains ICP license number, then F12 will be assigned a value of 1; otherwise, it will be 0.
- F13: The number of void (null) links on a website
  According to previous study [7], a phishing website tends to have more void links than an authentic website.
- F14: The number of out links on a website
  It is normal for a website to have some out links, but when there are too many, it may increase the probability of a website being a phishing website.
- F15: Whether an e-business website provides e-commerce certificate information
  If a website does not provide any certificate link, the value of F15 will be set to 0; otherwise 1.

Researchers in the study compare 4 algorithms (Figure 1) such as SMO (Sequential Minimal Optimization*)*, Naive Bayes, Random Logistic Regression and Forest. The results of the study noted that SMO algorithm has higher accuracy compared to three other algorithms. SMO has 95.83% accuracy followed by Random Naive Bayes and Forest, Logistic Regression of each 93.75%, 92.94% and 91.90%. Moreover, it has Precision value 0.953, Recall value 0.962 and F-Measure value 0.958. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved, and F-Measure is a measure of a test's accuracy[8].



	Precision	Recall	F1	Precision	Recall	F1	Overall Accuracy (n=864)
	Phishing-Websites(n=425)			Real Websites(n=439)			
SMO	0.953	0.962	0.958	0.963	0.954	0.958	95.83%
Naïve Bayes	0.938	0.918	0.927	0.922	0.941	0.931	92.94%
RandomForest	0.936	0.936	0.936	0.939	0.939	0.939	93.75%
Logistic	0.911	0.924	0.918	0.926	0.914	0.92	91.90%

Figure 1: Performance Comparison of Classification Algorithms for Detection Phishing Website (Zhang et al, 2014)

Which the best performance algorithm (SMO) will be used in the classification model made to compare with the classification model made by [9] and [7] based on hypothesis below:

- H1. The proposed domain-feature enhanced model for the detection of phishing websites will outperform traditional URL and content feature based models with generic website features only in terms of precision.
- H2. The proposed domain-feature enhanced model for the detection of phishing websites will outperform traditional URL and content feature based models with generic website features only in terms of recall.
- H3. The proposed domain-feature enhanced model for the detection of phishing websites will outperform traditional URL and content feature based models with generic website features only in terms of the F1-measure.

In empiric study created by Li et al [3] used 12 types of indicators known as topology feature of website. Topology feature of website is still included in the URL and content feature based approach. The researchers use a web crawler to extract 12 topology features of website into a DOM (Document Object Model) tree. Detailed topology features of website in the study include the following:

- F1: The number of webpages
- F2: The average number of inbound links
- F3: The average number of outbound links
- F4: The average number of internal links
- F5: The average number of images
- F6: The average number of CSS files.
- F7: The average number of JS files.
- F8: The average number of forms
- F9: The average number of input boxes
- F10: The average number of password boxes
- F11: The proportion of form links
- F12: Dynamic webpage proportion

In that study, the researchers compared several algorithms such as Bayes Nets, Naive Bayes, Logistic Regression, RBFN (Radial Basis Function Network), Simple Logistic, Decision Table, Decision Stump, SVM (Support Vector Machine) and BVM (Ball -based Support Vector Machine). However, the unique of the study is researchers choose algorithm with the fastest training time that is BVM regardless of TP (True Positive), FP (False Positive), Presicion, Recall or even F-Measure value. TP was defined as the proportion of sites that were completely positive among all sites showing positive test results, whereas FP was the percentage of all sites that were completely negative among all sites that showed negative test results.

In this study, researchers will also compare some classification algorithms in which algorithm with the best classification performance will be selected and used in this study. The gap in research [2] is not to create a web crawler like research [3] to extract features automatically. Though web crawler is useful to minimize validation errors made by human. In this study researchers will also create a web crawler such as research [3] to extract feature vectors, but use different techniques. Web crawler was created using PHP and API that can be used to extract feature vectors and normalize the feature values that have numeric data type.

Another gap in research [2] - [3] is still using some traditional feature vectors to detect phishing website. These features need to be modified in order to produce a good classification performance. The addition of new features based on URL and content approach is also expected to improve classification performance.

3. METHODOLOGY

3.1 Data Collection

For data collection methods, researchers conducted observation and search data (sites using Bahasa Indonesia, server in Indonesia or frequently accessed by Internet users from Indonesia). Researchers conducted observations and search data via Internet, email, and references from study Zhang et al [2]. The site source of Internet grabs from *https://moz.com/top500* (Moz) and *http://www.alexa.com* (Alexa) for authentic sites, while for phishing site list obtained from PhishTank addressed at *http://phishtank.com* and some website information provider about phishing sites (especially in Indonesia). Results, researchers obtained approximately 466 websites, each consisting of 102 authentic websites and 364 phishing websites. Below is example of authentic website URL has been obtained:

- *https://ibank.bankmandiri.co.id*
- *https://kaskus.co.id*
- *https://paypal.com*

In data collection, each authentic website must have least 1 phishing wesite and below is example of phishing website from authentic website above:

- *http://ablytube.com/clip/Personal*
- *https://kaskusbluemoviess.allalla.com*
- *http://88.198.24.90/~consumired/pos/006b3/*

3.2 Feature Vectors

Feature selection is first step that should be performed in this research before classifying. In this reasearch, researchers will perform heterogeneous feature selection, because according to [4], heterogeneous feature selection can affect performance of classifier, besides heterogeneous feature selection is performed to get appropriate and relevant feature vectors based on URL anda content approach order for detect phishing website accurately.

Researchers conducted a literature study and examined some theories to obtain heterogeneous feature vectors. Existing literature and study are expected to solve problem formulas with regard to difficulty of selecting relevant features as well as improving detection performance. After studying literature and reviewing some theories, the researchers eventually proposed to use some feature vectors that existed in study [2] - [3] and added new heterogeneous feature vectors based on URL and content approach.

In a study conducted by Zhang et al [2], there is some non-heterogeneous feature vectors that are F12 (whether website contains ICP (Internet Content Provider) license number), F15 (whether an e-business website provides e-commerce certificate information) and other. Researchers do not include these feature vectors, because in general these feature vectors are not required in the classification model that will be made in this study. Because the classification model in this study covers overall globally website category in Indonesia (not just for e-commerce website only).

While in study conducted by Li et al [3], researchers took and modified a feature vector that supports practical contribution of this research related to implementation step in the next research which if model/system made implemented, it can prevent Internet users from virus or malware attacks. The feature in question is a feature to detect the average number of JS (JavaScript) files. Because basically more and more JS files on a site, the chance of files are inserted by malware or virus increase.

In addition, the researchers also added new heterogeneous feature vectors based on URL and content approach are length of URL and web page score taken from PageSpeed Insights Google. One of the theoretical contributions of this research is to modify existing features and add new features of detection phishing website based on URL and content feature approach. So, researchers used approximately 11 feature vectors to detect phishing website and below are those feature vectors:

- F1: IP address
  This feature is derived from study [2]. Usually a phishing website contains an IP address. The URL of a target website contains an IP address instead of a domain name, then this feature variable F1 will be assigned a value of 1; otherwise -1.
- F2: Symbol '@'
  This feature is derived from the study [2]. Phishing websites often insert @ into a URL that takes users to a website different from what Internet criminals expect. If a URL contains the symbol '@', F2 will be assigned a value of 1; otherwise -1.
- F3: The number of dots ('.')
  Previous research [6] suggests that the larger the number of dots in a URL, the higher the possibility the website is a phishing website. In study [2] also use this feature.
- F4: The number of affixes
  Internet criminals usually modify the URL of phishing website by adding a few affixes to deceive Internet users as if the website is an authentic website. Affix is categorized into 4 types: prefix, infix, suffix and confix. In study [2] only used suffixes to detect phishing wesite, because the researchers in the study believed that phishing website usually use 2 domain suffixes and users generally catch a glimpse of the first part of a URL but likely miss the remaining part, which actually points to a phishing website. However, researchers will use affixes which include all types of affixes that mentioned above in this study. Examples of affixes in this study are "-" and some domain extensions are considered strange (unnatural).
- F5: Domain age
  Domain age is represented by the number of days passed since a domain name was registered. Research [2] also uses this feature, because basically the younger the age of the domain, its credibility as an authentic website is increasingly questionable.
- F6: DNS (Domain Name Server)
  DNS server addresses can be obtained through whois domain name queries. Research [2] also uses this feature, since most phishing wesite use non-paid hosting (free) which does not need to configure DNS to connect server (it can be said that the website does not have DNS). If it matches, the value of F8 will be -1; otherwise 1.
- F7: Organization
  This feature is derived from research [2]. Usually an authentic website is registered by an institution, company or organization. Whether a domain name applicant is an individual (1) or an organization (-1).
- F8: The number of outbound links

It is normal for a website to have some out links, but when there are too many, it may increase the probability of a website being a phishing website. This feature is used in research [2] - [3].

- F9: The number of JS (JavaScript)
  Most phishing wesite have an unusual number of JS files. This happens because the phishing website uses JS files to spread malware or viruses. In study [3] also used similar features but using different calculations that is the average number of JS, while in this study will be counted the number of JS files and normalized its value.

Table 1: Example of Prefixation Result

http://88.198.24.90/~consumired/pos/006b3/	*1 http://88.198.24.90/~consumired/pos/006b3/*
http://ablytube.com/clip/Personal	*1 http://ablytube.com/clip/Personal*
https://ibank.bankmandiri.co.id	*-1 https://ibank.bankmandiri.co.id*
https://kaskusbluemoviess.allalla.com	*1 https://kaskusbluemoviess.allalla.com*
https://kaskus.co.id	*-1 https://kaskus.co.id*
https://paypal.com	*-1 https://paypal.com*

- F10: The length of URL
  Researchers propose this feature because typically phishing website has unusual URL lengths and encoded into UNICODE [2].
- F11: Page score of the website
  Phishing website usually take more loading time when accessed, because basically phishing website contains many scripts, JS, pop-ups or malware. In research [10], it is said that loading speed of the website (response time and latency) can affect the ranking of website page. So it can be concluded that most web page that has good score must have fast loading time. PageSpeed Insights Google is one of the features of Google Inc. which can be used to calculate page score on a website by taking into account the criteria mentioned above.

For all features that have numeric data type (not boolean) will be normalized in feature extraction to determine relationship between each feature value and produce a simple feature value, so it is possible to optimize classification performance.

3.3 Data Preprocessing

Data preprocessing is a process/step to make raw data into quality data (good input for data mining tools) [11]. This step is performed to prepare and support processing data in the next step. Data preprocessing in this study there is 2 stages that are prefixation and feature extraction.

3.3.1 Prefixation

First stage of data preprocessing in this study is prefixation. Prefixation is a process to built a word by adding an affix to basic form and attach it in front of basic form [12]. This study does not use letter prefixes, but number prefixes (1 and -1) which symbolize phishing website and authentic website. In data collection, each word or line represents a URL of site. Each URL of site will be prefixed 1 and -1 based on its type as described previously (1 for phishing website and -1 for authentic website). This prefix is added manually by researchers and Table 1 is example of prefixation result.

3.3.2 Feature extraction

Second stage is feature extraction. Feature extraction is feature retrieval of a form in which value obtained will be analyzed for next process [13]. In this study feature extraction is performed to extract 11 heterogeneous feature vectors from data that has been collected in this study into ARFF file format (Attribute-Relation File Format) containing headers (relation and attribute) and data (feature value) to be directly processed using data mining tool (Weka), Feature extraction in this study also has another function that is to normalize all heterogeneous feature vector values of attributes of numeric data type.

Normalization is performed to produce a simpler feature vector value so as to optimize classification performance. For example, F5 (domain age), which the minimum age is 0 day and its maximum age is 11,618 days. When observed distance of minimum and maximum age is very far about 11,618 days. The difference is very far can affect relationship between value of feature vector, so it will affect classification performance. Therefore, it is necessary to normalize to optimize classification performance and simplify value of feature vector without changing relationship between feature values. In equation (1) is formula used in this study to normalize feature vector values:

$$N = \frac{n - min}{max - min} \qquad (1)$$

Which:
- N : Normalization value
- n : Feature value
- min : Minimum value of feature
- max : Maximum value of feature

For extraction feature in this study, the researchers created a web crawler based on PHP and API (Application Programming Interface). API method is generally used within code snippets along with other methods of the API of interest [14]. In this study, API is used to detect some features such as F5 (domain age), F6 (Domain Name Server), F8 (the number of JS) and other features. For the result of feature extraction can be seen in Figure 2.

```
@relation phishing

@attribute ip { 1,-1 }
@attribute symbol_at { 1,-1 }
@attribute dots numeric
@attribute affixes numeric
@attribute age numeric
@attribute dns { 1,-1 }
@attribute organization { 1,-1 }
@attribute outbound_links numeric
@attribute js numeric
@attribute length_url numeric
@attribute page_score numeric
@attribute status {'phising','authentic'}

@data
-1,-1,0,0.0384615384615,0.383542778447,1,-
1,0.168779714739,0.162162162162,0.0128865
979381,0.72,authentic'
1,-1,0.166666666667,0.0384615384615,0,-
1,1,0.00237717908082,0,0.0425257731959,0,'
phising'
```

Figure 2: Extraction Feature Result

3.4 Classification Model for Detection Phishing Website

In a study learned by Zhang et al [2], the first step to build the classification model is determines feature vectors based on URL and content feature approach, then selects several classifiers and compares the results of its detection performance. Which classifier have the best performance would be selected and applied to classification model for comparison with performance results in studies [9] and [7] that use traditional features for detection of phishing sites.

3.4.1 Classification model

The classification model for detection phishing website in this study refers to research created by Zhang et al [2]. However, in data preprocessing of this study, prefixes and extraction feature based on web crawler were not performed in study [2]. Surely it is novelty of this study. The making of feature extraction refers to research conducted by Li et al [3] to support researchers when processing big data. The design of classification model made in this study (Figure 3) was adopted from study created by Catal et al [15], Lee et al [16] and Thirumala et al [17] which used multiple classifiers, then selected the best performance classifier in model validation stage.

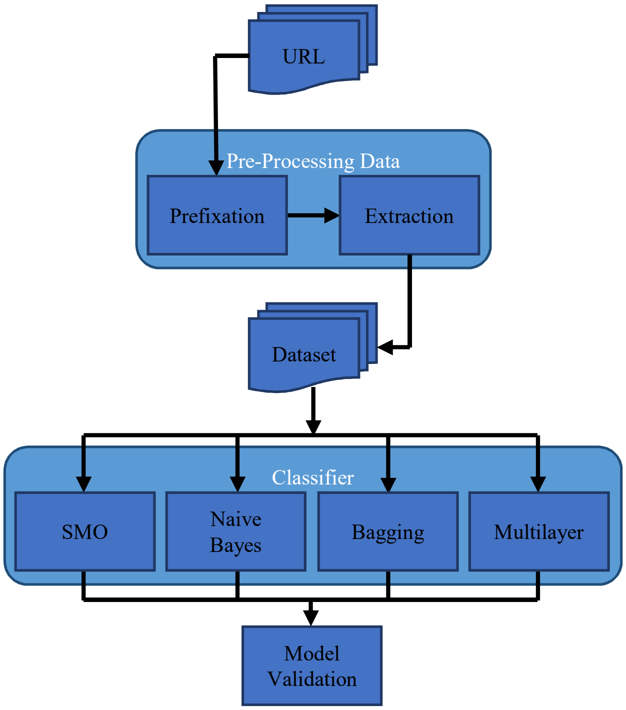


Figure 3: Classification Model for Detection Phishing Website

3.4.2 Classification perfomance

Similar to research [2], this study will also use Precision (P), Recall (R) and F-Measure (F) to evaluate performance of classification model made. But the novelty from this study is add another aspect to evaluate performance of classification model made is accuracy and time. If refers to what has been explained before, then in this research there are 4 possible results obtained from the classification of True Positive (TP), False Negative (FN) and False Positive (FP) and True Negative (TN).

According to [2], Precision is referred to as a positive predictive value which is a percentage of the true prediction depicted by *TP / (TP + FP)*, the Recall value is the actual positive proportion in the tested population data which is written as *TP / (TP + FN)*, whereas F-Measure is the mean value of a combination of Precision and Recall which can be calculated using the formula *2 (P x R) / (P + R)*. The value of Precision, Recall and F-Measure ranges from 0 to 1. If the accuracy value is close to

Precision, Recall and F-Measure (more or less), it is certain that the accuracy value is valid and reliable. In Table 2 is its confusion matrix.

Table 2: Confusion Matrix

Prediction	Result	
	1	0
1	*TP*	*FP*
0	*FN*	*TN*

Zhang et al [2] have hypothesized the proposed domain-feature enhanced model for the detection of phishing websites will outperform traditional URL and content feature based models with generic website features in terms of Precision., Recall and F-Measure. While in this study, researchers have hypothesized that classification model created will produce good detection performance in terms of Precision, Recall, F-Measure, accuracy and time.

3.4.3 Classification algorithms

In this research will be used 4 different classification algorithms in trial phase include:

- SMO (Sequential Minimal Optimization)
  SMO algorithm is used because it can solve the QP (Quadratic Programming) problems that arise during SVM (Support Vector Machine) training which in this research will be used big data which allows occurrence of errors when manipulating the matrix. In addition to research conducted by Zhang et al [2], the algorithm yields the best accuracy value.
- Naive Bayes
  In similar study about detection of phishing website [2] and [3], Naive Bayes classifier is the most commonly used algorithm. It can not be separated from Naive Bayes function as a classifier that can be used to predict something based on existing data using probability and statistical methods including to predict whether websit includes phishing or authentic website.
- Bagging
  Catal et al [15] create sentiment classification model of customer review on blogs, forums and social media in Turkey using Bagging algorithm. Algorithm Bagging is used in the model because it is able to provide a decision using multiple value combined into a single prediction.
- Multilayer Perceptron
  In this study, the researchers proposed using other algorithm called Multilayer Perceptron-based on ANN (Artificial Neural Networks) like a study learned by Lee et al [16] that use similar algorithm based on NN (Neural Network) to build a classification model that capable to predict activators on CAR (Constitutive Androstane Receptor) and offering structural information about ligand/protein interactions in liver. Therefore, researchers want to try to use similar algorithms but with different case studies.

4. RESULT

The data used in this study consisted of 102 authentic websites and 364 phishing websites. There are two tests in this research. The first test is trial of classification algorithms. This test aims to get classification algorithm with the best detection performance. In this first test, there are 4 phases, such as trial of SMO (Sequential Minimal Optimization), Naive Bayes, Bagging and Multilayer Perceptron algorithm. Which algorithm with the best detection performance results will be used in this study.

The second test is trial of classification model. This test was conducted to measure performance of classification model in the previous study [2] when using basic features (non-heterogeneous features) and same dataset like this study, so the results of trial can be used as comparison to assess the detection performance of classification model that has been made in this study. To support this second test, researchers have created a special web crawler to extract the feature vectors in the study [2].

4.1 Trial of Classification Algorithms

4.1.1 SMO (Sequential Minimal Optimization)

Table 3 shows the result of SMO algorithm based on confusion matrix.

Table 3: Confusion Matrix of SMO Algorithm

Prediction	Results	
	Phishing	Authentic
Phishing	351	13
Authentic	37	65

The accuracy and training time of SMO algorithm are approximately 89.27% and 0.23 seconds. Table 4 shows the result of SMO algorithm performance in terms of Precision (P), Recall (R) and F-Measure (F).

Table 4: SMO Algorithm Performance

Class	P	R	F
Phishing	0.905	0.964	0.934
Authentic	0.833	0.637	0.722

4.1.2 Naive Bayes Algorithm

Table 5 shows the result of Naive Bayes algorithm based on confusion matrix.

Table 5: Confusion Matrix of Naive Bayes Algorithm

Prediction	Results	
	Phishing	Authentic
Phishing	344	20
Authentic	9	93

The accuracy and training time of Naive Bayes algorithm are approximately 93.78% and 0.04 seconds. Table 6 shows the result of Naive Bayes algorithm performance in terms of Precision (P), Recall (R) and F-Measure (F).

Table 6: Naive Bayes Algorithm Performance

Class	P	R	F
Phishing	0.975	0.945	0.96
Authentic	0.823	0.912	0.865

4.1.3 Bagging Algorithm

Table 7 shows the result of Bagging algorithm based on confusion matrix.

Table 7: Confusion Matrix of Bagging Algorithm

Prediction	Results	
	Phishing	Authentic
Phishing	357	7
Authentic	14	88

The accuracy and training time of Bagging algorithm are approximately 95.49% and 0.34 seconds. Table 8 shows the result of Bagging algorithm performance in terms of Precision (P), Recall (R) and F-Measure (F).

Table 8: Bagging Algorithm Performance

Class	P	R	F
Phishing	0.962	0.981	0.941
Authentic	0.926	0.863	0.893

4.1.3 Multilayer Perceptron Algorithm

Table 9 shows the result of Multilayer Perceptron algorithm based on confusion matrix.

Table 9: Confusion Matrix of Multilayer Perceptron Algorithm

Prediction	Results	
	Phishing	Authentic
Phishing	350	14
Authentic	20	82

The accuracy and training time of Multilayer Perceptron algorithm are approximately 92.70% and 2.55 seconds. Table 10 shows the result of Multilayer Perceptron algorithm performance in terms of Precision (P), Recall (R) and F-Measure (F).

Table 10: Multilayer Perceptron Algorithm Performance

Class	P	R	F
Phishing	0.946	0.962	0.954
Authentic	0.854	0.804	0.828

4.2 Classification Model

Table 11 shows the result of second test (trial of classification model) for study [2] using SMO algorithm and same dataset.

Table 11: Confusion Matrix of Classification Model in Previous Study using SMO Algorithm

Prediction	Results	
	Phishing	Authentic
Phishing	356	8
Authentic	40	62

SMO algorithm was chosen in the second test, because study [2] says that classification model used on it has the best classification performance when using SMO algorithm. The accuracy and training time resulted by classification model are approximately 89.70% and 0.29 seconds. Table 12 shows the result of classification model performance in terms of Precision (P), Recall (R) and F-Measure (F).

Table 12: Classification Model Performance in Previous Study

Class	P	R	F
Phishing	0.899	0.978	0.937
Authentic	0.886	0.608	0.721

5. ANALYSIS

In trial of classification algorithms appear that classifier Naive Bayes has the fastest training time when compared with other algorithms that is only about 0.04 seconds ahead 0.19, 0.30 and 2.51 from the SMO, Bagging and Multilayer Perceptron algorithms. In addition to classifier Naive Bayes also has quite good accuracy of about 93.78%. However, Bagging has the highest accuracy value which is about 95.49% difference of 1.71%, 2.79% and 6.22% from Naive Bayes, Multilayer Perceptron and SMO.

Therefore, Bagging algorithm will be used in this classification model to be compared with second test results (trial of classification model). Table 13 shows that the classification model made in this study

Table 13: Comparison of Classification Performance Results

Name	Precision	Recall	F-Measure	Accuracy	Time
This Study	0.954	0.955	0.954	95.49%	0.34 s
Previous Study	0.896	0.897	0.890	89.70%	0.29 s

outperform the classification model in [2] using only the basic feature vectors. The classification model in this study outperform in some aspects such as accuracy, precision, recall, f-measure except time each 5.79%, 0.064, 0.058, 0.058 and -0.05.

Contribution of this research there are two that is theoretical contribution and practical contribution. The theoretical contribution of this research is to propose a classification model for detection phishing website in Indonesia based on URL and content feature approach which can distinguish phishing site and authentic site with good performance. In theoretical contribution, researchers modify some of the existing feature vectors are JS and affixes, besides researchers also proposed new features that are the length of URL and page score of the website to improve classification performance. So that classification model can be used in other research to make detection phishing website system more specific (for example: phishing online banking detection system, phishing social media detection system, phishing e-business website detection system in Indonesia or others).

While the practical contribution of this research is to support researchers in further research in the development of phishing site detection system, because this classification model can be implemented into a service, so that information provided can prevent Internet users from malware attacks or hijacking from phishing sites and reduce the risk of financial and data loss caused by phishing website.

6. CONCLUSION

In this study, Bagging is algorithm that has the best classification performance when compared with other algorithms. The average of accuracy, precision, recall, f-measure and time of Bagging algorithm are 95.49%, 0.954, 0.955, 0.954 and 0.34 seconds, so it outperformed some aspects except time. Heterogeneous feature selection and new features proposed based on URL and content approach proved to improve classification performance. The features are the length of URL and web page score. This classification model outperform in some aspect such as accuracy, precision, recall, f-measure except training time to classification model in previous study that only uses basic features.

For further research that might be possible is how to reduce training time. For reducing training time it is possible to use another Naive Bayes classifier family which Naive Bayes in this study has the fastest training time is 0.04 seconds, but only has 93.78% accuracy. Moreover, the classification model that has been made can be implemented in other research to make the phishing website detection system more specific (e.g. phishing e-commerce website detection system, phishing online banking site detection system or others).

Limitations in this research are sites using Indonesian language, hosted in Indonesia, and frequently accessed by Internet users from Indonesia and not implementing the classification model. The classification model can be implemented into a service to detect phishing sites in standalone site or using API, so it can avoid and reduce the risk of Internet users exposed to malware attacks or hijacking from phishing website.

ACKNOWLEDGMENTS

This research is supported by Research Grant (No 566/PKS/ITS/2017). Hopefully this good cooperation can be continued so as to contribute to the research development and strengthening of national innovation system.

REFRENCES:

[1] R. S. Rao and S. T. Ali, “PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” *Elev. Int. Conf. Commun. Netw. ICCN 2015 August 21-23 2015 Bangalore India Elev. Int. Conf. Data Min. Warehous. ICDMW 2015 August 21-23 2015 Bangalore India Elev. Int. Conf. Image Signal Process. ICISP 2015 August 21-23 2015 Bangalore India*, vol. 54, pp. 147–156, Jan. 2015.

[2] D. Zhang, Z. Yan, H. Jiang, and T. Kim, “A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites,” *Inf. Manage.*, vol. 51, no. 7, pp. 845–853, Nov. 2014.

[3] Y. Li, L. Yang, and J. Ding, “A minimum enclosing ball-based support vector machine approach for detection of phishing websites,”

Opt. - Int. J. Light Electron Opt., vol. 127, no. 1, pp. 345–351, Jan. 2016.

[4] H. Sulistiani and A. Tjahyanto, "Heterogeneous Feature Selection for Classification of Customer Loyalty Fast Moving Consumer Goods (Case Study: Instant Noodle)," *J. Theor. Appl. Inf. Technol.*, vol. 94, Dec. 2016.

[5] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Int. Conf. Comput. Model. Secur. CMS 2016*, vol. 85, pp. 662–668, Jan. 2016.

[6] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach," *IEEE Trans. Neural Netw.*, vol. 22, no. 10. pp. 1532–1546, Oct. 2011.

[7] J. F.-T. HE Gao-Hui, "Phishing Detection System Based on SVM Active Learning Algorithm," *Comput. Eng.*, vol. 37, no. 19, pp. 126–128, 2011.

[8] X. Zhang, X. Feng, P. Xiao, G. He, and L. Zhu, "Segmentation quality evaluation using region-based precision and recall measures for remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 102, pp. 73–84, Apr. 2015.

[9] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and Nunamaker, *Detecting fake websites: The contribution of statistical learning theory*. MIS Quarterly: Management Information Systems, 2010.

[10] M. Almulla, H. Yahyaoui, and K. Al-Matori, "A new fuzzy hybrid technique for ranking real world Web services," *Knowl.-Based Syst.*, vol. 77, pp. 1–15, Mar. 2015.

[11] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, May 2017.

[12] S. Ben Hedia and I. Plag, "Gemination and degemination in English prefixation: Phonetic evidence for morphological organization," *J. Phon.*, vol. 62, pp. 34–49, May 2017.

[13] X. Gao, Q. Sun, and H. Xu, "Multiple-rank supervised canonical correlation analysis for feature extraction, fusion and recognition," *Expert Syst. Appl.*, vol. 84, pp. 171–185, Oct. 2017.

[14] H. Eyal Salman, "Identification multi-level frequent usage patterns from APIs," *J. Syst. Softw.*, vol. 130. pp. 42–56, Aug. 2017.

[15] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Appl. Soft Comput.*, vol. 50. pp. 135–141, Jan. 2017.

[16] K. Lee, H. You, J. Choi, and K. T. No, "Development of pharmacophore-based classification model for activators of constitutive androstane receptor," *Drug Metab. Pharmacokinet.*

[17] K. Thirumala, A. C. Umarikar, and T. Jain, "A new classification model based on SVM for single and combined power quality disturbances," *IEEE*, Feb. 2017.