A Framework for Vietnamese Email Phishing Detection

Cho Do Xuan, Hoa Dinh Nguyen, Tisenko Victor Nikolaevich

Abstract- Currently, the attacks on network information systems are increasing rapidly in number and level of danger. Phishing email distribution method is widely used by hackers today. This is an attacking technique that exploits human behaviors in the system. This type of attack, though it is not too complicated, becomes very effective for attackers if users are unaware of information security and unable to identify phishing emails. This attack is particularly more commonly effective in developing countries where the information security is still overlooked. As a result, email phishing detection problem has become a hot topic for information security researchers. There have been some published methods to detect phishing emails on given email attacking datasets. However, one of the important issues in email phishing detection relates to the language used in emails. Each particular language used in different emails may lead to a different phising detection approach. In this article, a Vietnamese email phishing detection system is investigated. The research includes a feature selection method and a combination of machine learning algorithms to improve the performance of phishing email detection in Vietnamese language. The proposed method is evaluated using two datasets. The first dataset includes phishing emails from Vietnamese collected from Vietnamese volunteers. The second dataset is the widely used English emails as introduced in [16,17]. The experimental results show that our method is applicable for real Vietnamese email phishing detection systems.

Email phishing detection, Vietnamese language, feature selection, machine learning.

I. INTRODUCTION

The characteristics of email phishing techniques are presented in [1] and [2]. According to this, each email has its own structure, and attackers usually try to alter some parts of those email structures to implement their email phishing attacks. Based on the characteristics of email phishing attacking techniques, it is not easy to detect the phishing attacks. Additionally, the research [2] lists out and classifies some email phishing techniques. The paper [3] has listed the danger and the impact of phishing emails on users. Comparing statistical reports at the same time between 2018 and 2019, it is clear that the danger and the impact of phishing emails on users is constantly increasing. According to the reports from Kaspersky Lab 2019 [5] and Vietnam Computer Emergency Response Team (VNCERT) [6], Vietnam is among the top email phishing targets in Southeast Asia. Some typical attacks by spreading email phishing over the past time in Vietnam are as follows:

- In December, 2014, a businessman in Da Nang was attacked by an e-mail hacker to request a transfer of 18,720 USD in the process of buying asphalt with a partner company in Dubai (UAE).
- In March, 2018, there appeared a series of fake Google phishing emails announcing the winning with great prize values. To receive the prize, the recipient needed to perform the required steps, including the verification of the account via a fake link to steal the user's google account information.
- In July, 2018, there were phishing mails containing information relating to bank accounts from Vietnam Agriculture Bank, VPBank, HSBC Bank, and some other banks. These fake emails required customers to provide online bank access information via a fake website.

From the statistics on the measure of email phishing damage in the world and in Vietnam [5, 6] it can be seen that detecting email phishing techniques are in need of research and development in order to efficiently prevent this type of attacks. However, so far, there are not many researches and commercial products that help detect Vietnamese email phishing. In this paper, the research team propose a model to detect Vietnamese phishing emails based on machine learning algorithms.

II. RELATED WORKS

At present, there are two types of email phishing detection approaches, which are based on either filters or machine learning algorithms [2]. Filter-based methods use predefined signatures obtained from historic data to search for known email phishing. This approach is well known for its fast and accurate detection results. However, email phishing detection method based on filters cannot detect newly unknown attacks, which are not included in training data. Some examples of filter-based email phishing methods include Black list, White list, Pattern Matching, Email verification, Password filter. In order to overcome disadvantages of filter-based method, another approach based on machine learning algorithms can be used. Chandrasekaran et al. [7] propose a method based on the characteristics of the email structures to detect an email phishing. These characteristics are combined with support vector machine (SVM) algorithm to form a complete email phishing detection system. Toolan and Carthy [8] use the C5.0 algorithm to detect email phishing using 5 features extracted from a dataset consisting of 8,000 emails, half of which are phishing emails and the others are normal emails. Jameel et al. [9] propose a method using neural networks to detect email phishing based on 18 features extracted from email subjects and HTML contents.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author

Cho Do Xuan*, Posts and Telecommunications Institute of Technology Hanoi, Vietnam, FPT University Hanoi, Vietnam

Hoa Dinh Nguyen, Posts and Telecommunications Institute of Technology Hanoi, Vietnam

Tisenko Victor Nikolaevich, [3]Peter the Great St. Petersburg Polytechnic University Russia, St.Petersburg, Polytechnicheskaya,

This method is evaluated with five different structures of the neural networks. In [10], Nizamani apply some classification algorithms such as SVM, Naïve Bayes, J48, and CCM using different sets of features to detect email phishing. Kathsirvalavakumar et al. [11] propose a multilayer neural network structure for email phishing detection. A data preprocessing stage is added to this neural network to reduce the number of input features, and hence reduce the computational cost of the system. Other researches presented in [12, 13] focus on email phishing detection methods based on machine learning algorithms, such as SVM, logistic regression, J48, using 47 features. The experimental results are obtained from Weka toolkit show different accuracy rates corresponding to different selected feature sets.

So far, researches on Vietnamese email phishing detection are very limited in terms of number of publications as well as the efficiency. In this research, we proposed a new method for Vietnamese email phishing detection framework. The research is motivated by the results from [13]. The main contribution of the paper is to extract new features from the phishing emails, which are shown to be more suitable to Vietnamese language, and the detection system is evaluated on a new email dataset.

III. VIETNAMESE EMAIL PHISHING DETECTION BASED ON MACHINE LEARNING

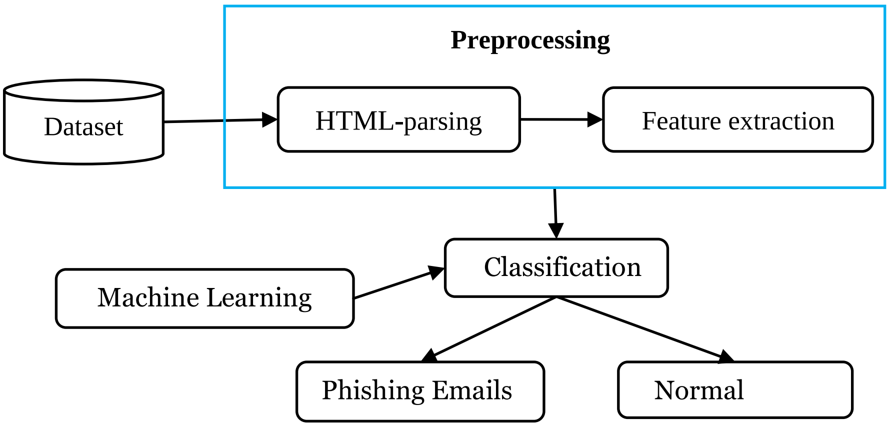


Fig. 1. Vietnamese email phishing detection model based on machine learning

Figure 1 presents a flow chart of the proposed detection model for Vietnamese email phishing. The detailed explanations of all parts of the model are as below

- **Dataset:** in this research, we use both Vietnamese and English emails for phishing detection system. This dataset needs preprocessed before the classification process.
- **Preprocessing:** This stage includes email structure analysis and feature extraction. In this paper, two main parts, which are the header and the body, of Vietnamese emails are analyzed for phishing detection. HTML parsing is used to analyze the email structure into small content pieces. This step peels off each part of the email including: the subject part of the email (including the information of the sender, the recipient), the body of the message (including the JavaScript tags used, URL links, keywords, HTML forms) to support the extraction of corresponding features.
- **Classification:** to classify the emails into either normal or phishing based on the features obtained from previous process. Machine learning algorithms are adopted in this stage to facilitate the classification process.

Feature extraction and selection for Vietnamese emails

In the proposed email phishing detection model presented in Figure 1, there are two main tasks need to focus in the research. They are feature extraction and machine learning algorithm selection. Regarding feature selection step, there are 46 features from 5 components of phishing emails [13] can be used. In this research, due to the characteristics of Vietnamese language, 30 out of 46 features from all 5 email components will be used. Some features relating to Vietnamese content will be separately extracted using Vietnamese language processing tool [6]. The list of features for Vietnamese email phishing detection method is included in Table 1

Table 1. Extracted features for Vietnamese email phishing detection

Feature	No	*Function*	*Type*	*Description*
Body feature	1	body_html	Boolean	Returns True if the email content contains HTML, and False if otherwise.
	2	body_ listword	Integer	Returns the number of following words in the email body: "dear", "suspension", "verify your account", "xác minh tài khoản", "tạm dừng", "gửi", "thân gửi", "bạn thân mến", "chào bạn",…
	3	body_ noCharacters	Integer	Returns total number of characters in the email content.
	4	body_ noDistinctWords	Integer	Returns the number of unique words in the email content.
	5	body_noFunctionWords	Integer	Returns the total number of following functional words in the email body: "account", "access", "bank", "credit", "click", "identity", "inconvenience", "information", "limited", "log", "minutes", "password", "recently", "risk", "social", "security", "service", "suspended", "login", "username", "verify", "notify", "restrict", "hold", "user", "customer", "client", "update", "confirm", "alert", "ssn", "tài khoản", "truy cập", "ngân hàng", "danh tính", "thông tin", "mật khẩu", "rủi ro", "xã hội", "dịch vụ", "bảo mật", "hạn chế", "tạm dừng"…
	6	body_noWords	Integer	Returns the total number of words in email body
	7	body_richness	Float	Returns the ratio of body_noWords over body_noCharacters
Sender feature	8	send_noCharacters	Integer	Returns the number of characters in sender information
	9	send_noWords	Integer	Returns the number of words in sender information
	10	send_diffSenderReplyTo	Boolean	Returns True if the sender is different from the reply-to domain, and False if otherwise.
	11	send_nonModalSenderDomain	Boolean	Returns True if there is a modal domain name in the sender email, and False if otherwise. A modal domain name is defined as a domain referencing to the email body like the domain in the URL
Subject feature	12	subj_listword	Boolean	Returns True if there are following words in the email subject: "verify", "debit", "bank", "xác minh", "vay nợ", "ngân hàng",… and False if otherwise.
	13	subj_reply	Boolean	Returns True if there is the word “Re:” in the email content, and False if otherwise.
	14	subj_noWords	Integer	Returns the number of words in the email subject.
	15	subj_noCharacters	Integer	Returns the number of characters in the email subject.
	16	subj_richness	Float	Returns the ratio of the total number of words over the total number of characters in the email subject
URL feature	17	url_atSymbol	Boolean	Returns True if there is an URL containing “@” in the email, and False if otherwise.
	18	url_ipAddress	Boolean	Returns True if there is an URL containing the IP in the email, and False if otherwise
	19	url_linkText	Boolean	Returns True if there is an URL containing words “Click”, “here”, “login”, “update” in the email, and False if otherwise.
	20	url_maxNoPeriods	Integer	Returns the number of the periods in the link having the highest number of periods.
	21	url_noDomains	Integer	Returns the number of domains found in the URL of the email.
	22	url_noExtLinks	Integer	Returns the number of outside links found in the email. Outside links are the links to the resources outside of the email.
	23	url_noImgLinks	Integer	Returns the number of image links found in the email.
	24	url_noIntLinks	Integer	Returns the number of inside links found in the email.
	25	url_noIpAddresses	Integer	Returns the number of URL containing the IP of a domain.
	26	url_noLinks	Integer	Returns the number of links found in the email body.
	27	url_noPorts	Integer	Returns the number of ports found in the attached URL of the email.
	28	url_nonModalLinks	Integer	Returns the number of links containing words “here”, “click”, “link” and not inside of modal domains
	29	url_ports	Boolean	Returns True if the URL contains a port number
Content type	30	content-type	Boolean	Returns True if the email Content-type = ”multipart/”

Machine learning algorithms

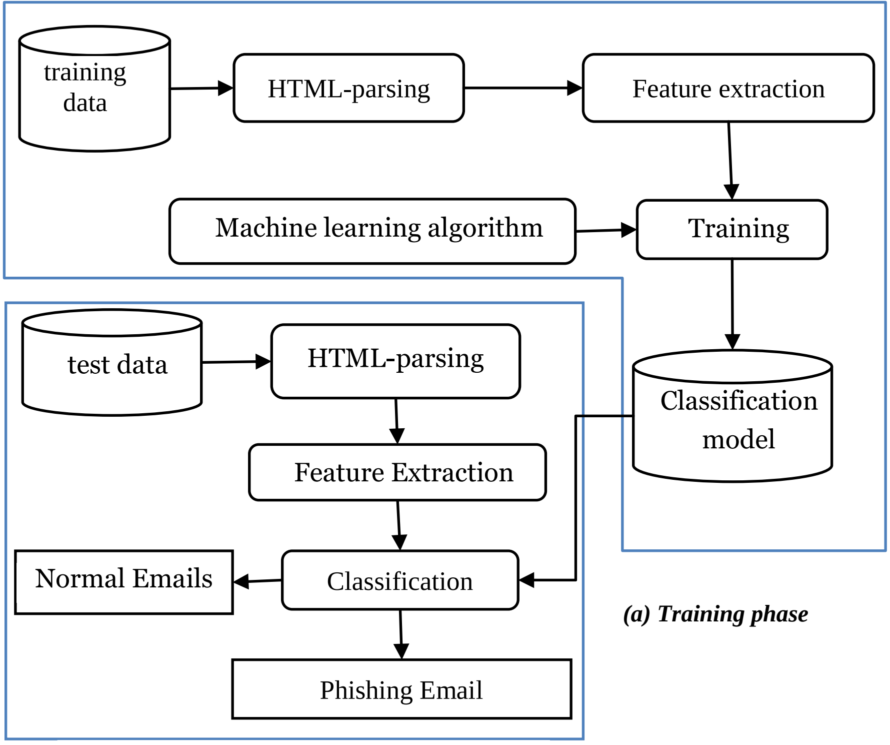
There are two main types of commonly used machine learning algorithms named supervised learning and unsupervised learning. Besides, there is also some additional methods, such as semi-supervised learning and reinforcement learning. Each machine learning algorithm has its own pros and cons. In this research, some well-known supervised machine learning algorithms are investigated for email phishing detection system. Those algorithms include k - Nearest Neighbor (kNN), Random Forest (RF), Decision tree [14, 15]. The parameters of these machine learning methods will be adjusted to observe the change in the phishing detection efficiency.

IV. EXPERIMENTAL RESULTS

Experimental setup

The experiments include two phases: (a) training phase, and (b) testing phase.

- Training phase: the emails from input dataset are first preprocessed to extract feature vectors. During the training step, machine learning models are trained to classify the email accordingly to the output labels based on the extracted features.
- Testing phase: The testing dataset is preprocessed in the same way as in the training phase. The trained model will classify the preprocessed data and assign corresponding label to each testing email.



(a) Training phase

(b) Testing phase

Fig. 2. Experimental model for Vietnamese email phishing detection

Datasets and experimental environments

Datasets

Table 1. Details about training and testing datasets

Dataset	*Number of samples*	*Total*	
	Phishing Emails in both Vietnamese and English	Normal Email in both Vietnamese and English	
Training	6536	19125	25661
Testing	2800	5400	8200

In this research, normal English emails are collected from [16, 17], while normal Vietnamese emails are collected from real known volunteers. About 4336 English phishing emails are collected from [16, 17], and about 5000 Vietnamese phishing emails are manually collected.

In the experiments, the email phishing detection system is evaluated using two scenarios with 46 and 30 features. All three supervised machine learning algorithms kNN, random forest, decision trees are adopted. The parameters in each of machine learning algorithm are set as follows.

- kNN: the parameters can be adjusted in this method are the number of neighbors, k, and the neighbor weighting method. Some values of k are selected as {1, 3, 5, 7, 9, 12, 15, 18}. Two weighting scenarios are used: distance-based weighting and uniform weighting.
- Random Forest: in this method, the number of decision trees plays an important role in the classification rate. Some numbers of decision trees are investigated as {10, 15, 20, 25, 30, 35, 40, 45}.
- Decision trees: in this method, the algorithm CART is used, and two ranking criteria are used as entropy and gini.

The system is deployed in windows 10 operation system. All the codes are written in Python language with the use of some libraries as: BeautifulSoup, Scikit-Learn, Pandas, Numpy, Scipy

Experimental results and discussions

Performance evaluation metrics

Detection accuracy: this metric is calculated based on the percentage of correct decision among all testing samples

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (1)$$

where TP is the true positive representing the number of correct alarms for phishing emails; FN is the false negative representing the number of phishing emails are misclassified into normal ones; TN is the true negative representing the normal emails are correctly labeled; FP is the false positive representing the number of normal emails have been misclassified as phishing emails.

Confusion matrix: is a table describing the performance of the classification model on the testing dataset for which the true labels are known. It allows the visualization of the performance of an algorithm. The confusion matrix is described in Table 3.

Table 2. Confusion matrix

	Phishing emails detected	*Normal email detected*
Ground truth phishing emails	TP	FN
Ground truth normal emails	FP	TN

Precision: is the percentage of the true positive among all testing samples labeled as phishing emails (TP+FP). Higher value of precision means the decision of phishing email is more reliable.

$$precision = \frac{TP}{TP+FP} \times 100\% \qquad (2)$$

Recall: is the percentage of the true positive among all phishing emails needed to be labeled (TP+FN). Higher value of recall means the rate of missing phishing emails is low.

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (3)$$

F1-score: is the harmonic mean the precision and the recall. Higher value of F1 means the classifier is better.

$$F1 = \frac{2 \times precision \times Recall}{precision + Recall} \qquad (4)$$

FPR: is the false prediction rate, which is calculated as:

$$FRP = \frac{FP}{FP+TN} \times 100\% \qquad (5)$$

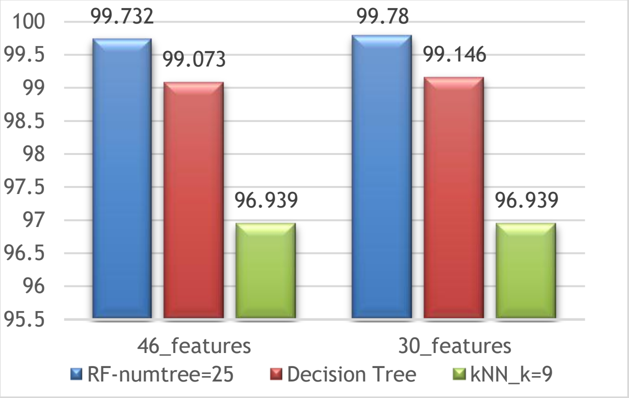


Fig. 3. The performance of Vietnamese email phishing detection system using 46 and 30 features as inputs to the machine learning algorithms

The email phishing classification performance based on both 46 and 30 extracted features of the system is illustrated in Figure 4. In this experiment, the number of decision trees in the Random forest algorithm is set at 25, the ranking metric for the decision tree algorithm is entropy, and the number of neighbors for kNN algorithm is set at 9. The experimental results show that, 30 features used in this research is a little bit more useful for Vietnamese email phishing detection than 46 features introduced in [13].

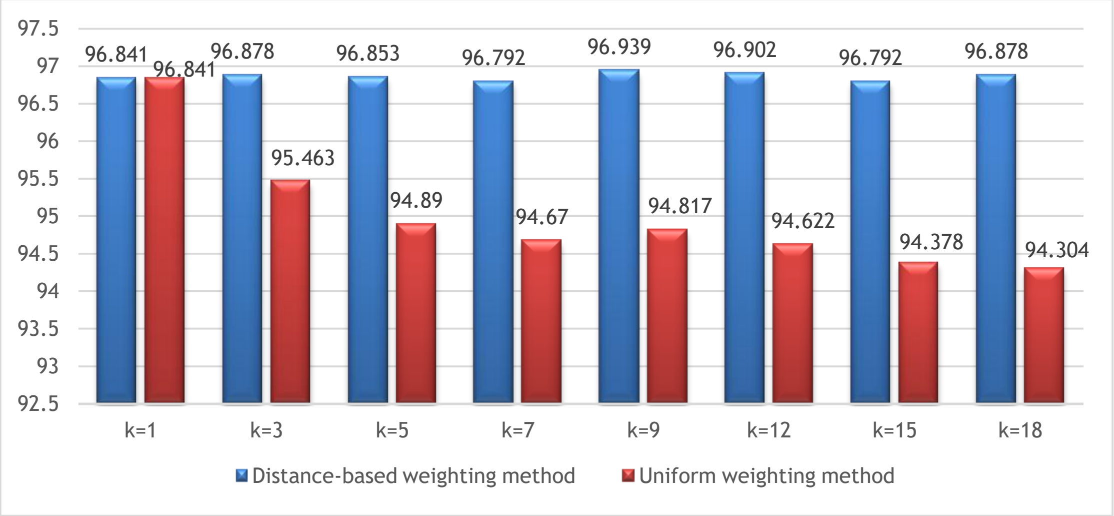


Fig. 4. Experimental results for kNN with different k values, and 30 features. Both distance-based weighting and uniform weighting methods are used

The email phishing detection performance of the kNN algorithm with different number of nearest neighbors applied on 30 extracted features is presented in Figure 4. The results are collected from two weighting scenarios: distance-based weighting and uniform weighting. It is shown that the distance-based weighting approach is better than the uniform weighting method.

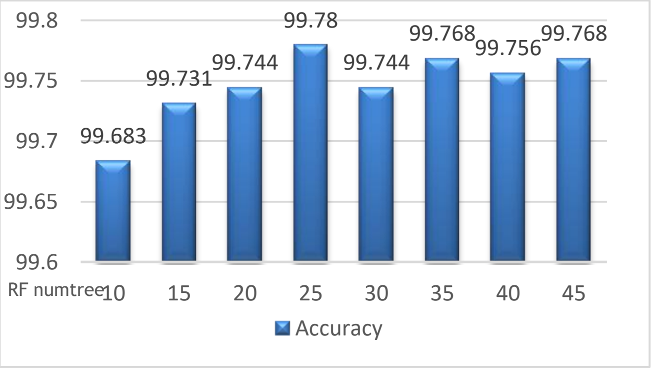


Fig. 5. Experimental results of the Random forest algorithm with different number of trees applied on 30 features

Figure 5 illustrates the email phishing detection performance of the Random forest applied on the proposed 30 features. Different numbers of trees are also investigated. The results show that the Random forest algorithm has the highest accuracy rate at 99.78% when the number of trees is 25, while the lowest accuracy rate is at 99.683% when the number of trees is 10.

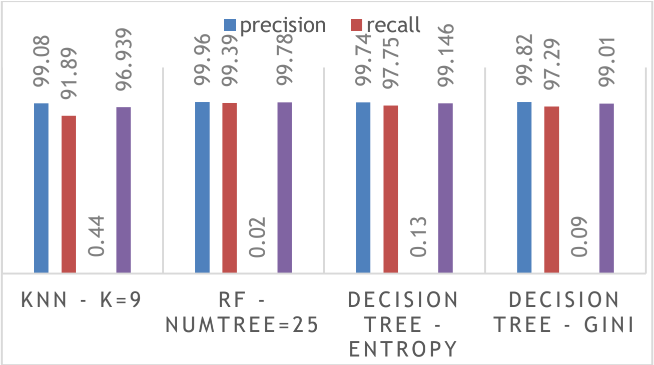


Fig. 6. Comparison between three machine learning algorithms applied on 30 features

The results presented in Figure 6 show that:

- The decision tree algorithm is, in general, better than the kNN algorithm, but is not as good as the Random forest. Regarding the ranking criteria, investigated with both entropy and gini ranking scores, the entropy is better in terms of the accuracy. However, gini criterion is better than entropy in terms of the precision rate.
- The experimental results also show that the use of just 30 out of 46 features introduced in [13] can still produce the same detection performances. In some cases, the performance of 30 features is even better. This is very meaningful in real life implementation since the computational cost as well as the complexity of the system can be significantly reduced. Figure 8 below illustrates the importance of 30 selected features for Vietnamese email phishing detection. It can be seen that the features relating to lingual information have higher importance indices than others. This is one reason why 30 features adopted in this research can help provide a reliable Vietnamese email phishing detection performance.

However, the false positive rates of all three investigated machine learning models are still very high. This is the issue that needs more improvement effort in our future work.

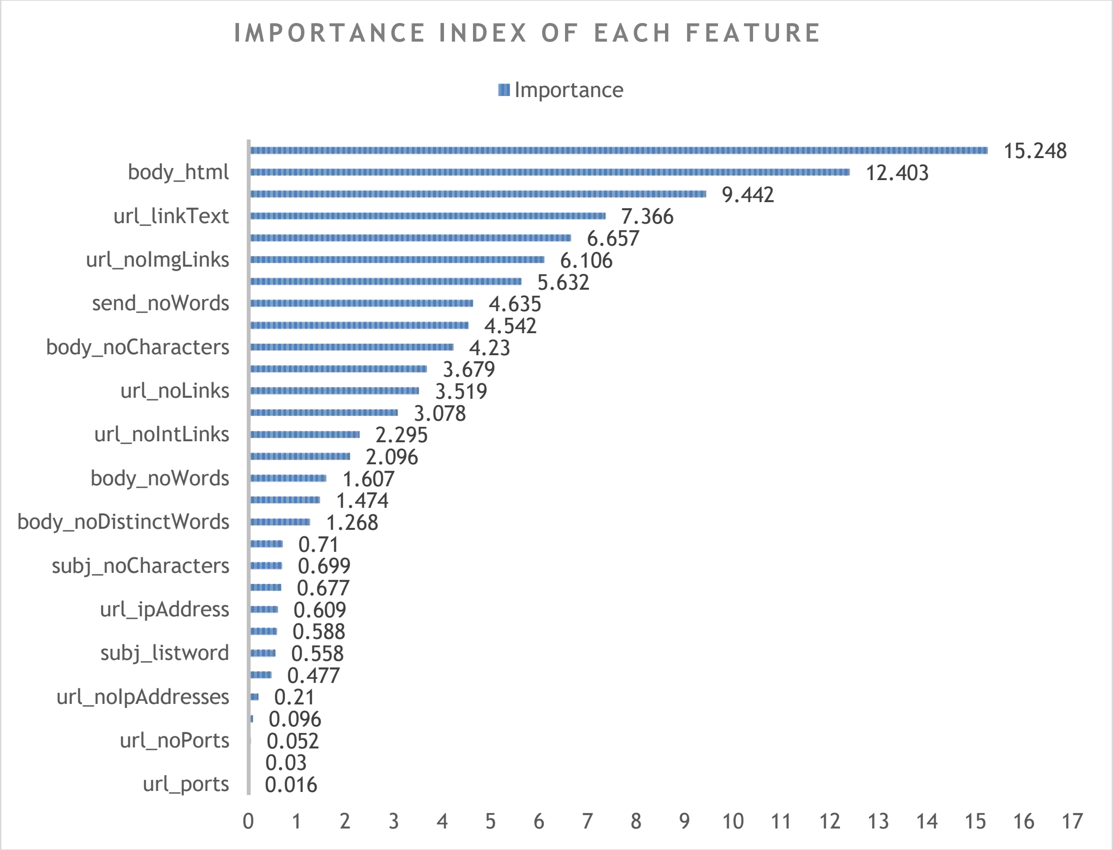


Fig. 7. The importance of 30 selected features for Vietnamese email phishing detection.

V. CONCLUSIONS

Email phishing attacks are becoming a permanent danger to all organizations and countries around the world. Especially in Vietnam, where users are still limited in awareness of information security issues. Therefore, our research proposal in this article can help pave a path for researches and analysis related to phishing email detection in Vietnamese language. A framework for Vietnamese email phishing detection systems has been presented, which includes two main stages. The first stage mentions about a new set of features extracted from the Email. This feature set is more compact than commonly used feature set, i.e. 30 features compared to 46 features. This is meaningful in real implementation since it helps reduce the computational cost and the complexity of the detection systems. The second stage includes the use of some machine learning algorithm for email phishing detection. Some practical machine learning algorithms have been investigated.

The empirical results show that phishing email detection systems may be used with different number of features extracted from the email written in different language. In Vietnamese languare, 30 features introduced in this research is applicable. Additionally, machine learning algorithms combinding with natural language processing techniques can be useful sources to improve the performance of phishing detection frameworks.

REFERENCES

1. Mahmoud Khonji, Youssef Iraqi.: Phishing Detection: A Literature Survey. IEEE Communications Surveys & Tutorials 15 (4), 2091 - 2121 (Fourth Quarter 2013).
2. Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani.: A Survey of Phishing Email Filtering Techniques. IEEE communications surveys & tutorials 15 (4), 2070 – 2090 (fourth quarter 2013).
3. APWG Phishing Activity Trends Report, 1st 2rd Quarters 2019. https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf. Last accessed 26 Oct 2019
4. Spam and phishing in Q2 2019. https://securelist.com/spam-and-phishing-in-q2-2019/92379/. Last accessed 26 Oct 2019.
5. Việt Nam among top targets for phishing in Southeast Asia, https://vietnamnews.vn/society/535097/viet-nam-among-top-targets-for-phishing-in-southeast-asia.html#3obGRzMAbOJfEPDv.97. Last accessed 26 Oct 2019.
6. A Vietnamese Text Processing Toolkit. https://github.com/phuonglh/vn.vitk. Last accessed 26 Oct 2019.
7. Chandrasekaran, M., Narayanan, K., Upadhyaya.: Phishing email detection based on structural properties. 9th Annual NYS Cyber Security Conference, pp 1-7, Albany, (14 June 2006).
8. Fergus Toolan, Joe Carthy.: Phishing detection using classifier ensembles. eCrime Researchers Summit, pp. 1-9, USA (21 Oct. 2009).
9. Noor Ghazi Mohammed Jameel, Loay Edwar George.: Detection of phishing emails using feed forward neural network. International Journal of Computer Applications 77 (7), 10- 16, (2013).
10. Sarwat Nizamani, Nasrullah Memon, Mathies Glasdam;, Dong Duong Nguyen.: Detection of fraudulent emails by employing advanced feature abundance, Egyptian Informatics Journal 15 (3), 169-174, (2014).
11. T. Kathirvalavakumar, K. R. Kavitha, Rathinasamy Palaniappan.: Efficient Harmful Email Identification Using Neural Network, British Journal of Mathematics & Computer Science 7 (1), 58- 67, (2015).
12. Mahmoud Khonji, Youssef Iraqi, Andrew Jones.: Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach, International Journal for Information Security Research 2 (1), 236- 245, 2012.
13. Detecting Phishing Emails Using Machine Learning Techniques. https://www.meu.edu.jo/libraryTheses/590422b4d5dd8_1.pdf. https://monkey.org/~jose/phishing/. Last accessed 26 Oct 2019.
14. Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).
15. Thomas G. Dietterich. Ensemble Methods in Machine Learning. International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.
16. Index of /~jose/phishing. https://monkey.org/~jose/phishing/. Last accessed 26 Oct 2019.
17. Enterprise Open-Source Spam Filter. http://spamassassin.apache.org/publiccorpus/. Last accessed 26 Oct 2019.

AUTHORS PROFILE

First Author Profile Dr. **Do Xuan Cho** is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam

In 2008, received a bachelor's degree in the Saint Petersburg Electrotechnical University "LETI" on a specialty "Computer science and computer facilities", Russia. In 2010, graduated a masters from the Saint Petersburg Electrotechnical University "LETI" on a specialty "Computer science and computer facilities", Russia. In 2013, received a PhD in the Saint Petersburg Electrotechnical University "LETI", on a specialty CAD. Russia. Area of scientific interests - modeling, control systems, algorithmization

Email:chodx@ptit.edu.vn and ***chodx@fe.edu.vn***

Second Author: Profile Dr. Hoa Dinh Nguyen earned bachelor and master of science degrees from Hanoi University of Technology in 2000 and 2002, respectively. He got his PhD. degree in electrical and computer engineering in 2013 from Oklahoma State University. He is now a lecturer in information technology at PTIT. His research fields of interest include dynamic systems, data mining and machine learning.

Email: **hoand@ptit.edu.vn**

Third Author: My position is the professor of Institute of computer sciences and technologies in Peter the Great Saint-Petersburg Polytechnic Ubiversity. I have received the degree Doctor of Technical Sciences in 1998 in accordance of scientific speciality "Systems of automatic Desing" in SPbPY. The area of scintific interest is use of new type of fuzzy logics in different applications. I think that we could cooperate intensively in future..

Email: ***v_tisenko@mail.ru***