5F-03

Collecting useful features for zero-day malicious emails detection

Sanouphab PHOMKEONA*　　Koji OKAMURA**

* Department of Informatic Science and Electronic Engineering, Kyushu University, Japan

** Research Institute for Information Technology Kyushu University, Japan.

Abstract— *Email is very useful and it is one of a basic element for internet users. In the other hand, about half of email traffics are spam and phishing email in the first quarter of 2018. By using a simple machine learning algorithm such as Support Vector Machine (SVM), Naïve Bayes (NB) or multilayer perception (MPL) neural network, scientist can already separate legitimate normal email from spam email. However, for spear phishing email or zero-day malicious email cases, the classification result from those techniques are still not unreliable. In this paper, we collect as well as select useful features from email's header and body for machine learning. We also introduce a method of using new features by using email subject database in different languages to detect machine-translated phases that is one technique of modern malicious spam does. We aim to use these features to increase an accuracy of zero-day malicious email detection.*

Keywords— *zero-day malicious spam, spam detection, deep-learning, features.*

I. Introduction

Email is not only useful tool to exchange information but it is also used to be a common entry point of attack by cyber villains. While a billion of unsolicited email are sent over the internet every day, most of them are spams or commercial spam which usually contain an advertisement contents. But in some cases, there are malicious emails which contain phishing links or virus-infection attachments that aim to steal the privacy/financial information, or make the devices infected dangerous malwares like ransomwares, spywares, bot-net, cryptocurrency coin-mining, etc. that not only harm a single device but a whole organization network. According to this serious issue, there are many ways to classify those malicious spam and email, and the most common method is by using machine learning algorithms. Above the popular well known algorithms NB, SVN, or K-Nearest Neighbor (KNN), running MPL neural network algorithm on test data seems to be the best way of spam detection in term of efficiency. While most of spam are filtered, there are some particular spam known as spear phishing and zero-day malicious spam that particular designed and AI cannot easily detect. Those spam currently still required security experts to investigate and analyze by doing static analysis, dynamic analysis, or reverse engineering looking for suspicious compositions and risky factors. While AI tasks can be solved by good designing the right set of features, in this paper we design and extract set of features to be closed to the security expert investigated information for machine learning. We aim to use these extracted features to improve the accuracy of zero-day malicious spam by using deep-learning approach in next step.

II. Features Extraction

Features in machine learning is a piece of information included in the representation of the data they are given. It is well-known that a performance of machine learning algorithms depends on the representation of the data they are given. Most of the popular research on spam classification use a set of features that come from email header and body themselves directly, such as domain name, IP address, email title and body, or others features that get from text words analysis. Similar to the security experts do, we collect and investigate more information from header and body part that could be clues to indicate suspicious email and judge them to be malspam or even unknown (zero-day) malspam or not. From analyzed more than 500,000 of emails, we have found that a relationship between those header and body features information are very important. For example, more than 99 percent of work mail are sent only in working time period (8AM-8PM) which is correspond with sender's domain time-zone and language used (in case of other languages than English). It means, Japanese people commonly use Japanese languages to communicate with each other during working time by using local email domain service, while malspam can be any languages, sent in random time and might use domain service from different geolocation. Figure.1 shows how we extract features from email header. Here, from email dataset we first extract 3 main features: source address, timestamp and subject. Then we can extract second stage and third stage features from those main features to in order.

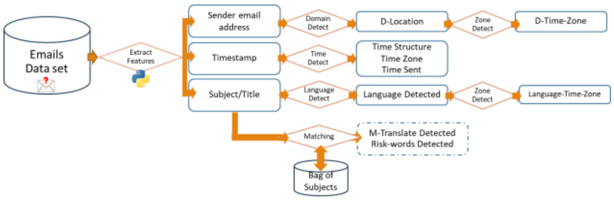

Figure 1: Flow chart of features extraction from email header

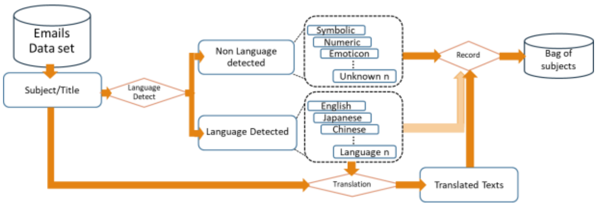

Figure2. Flow chart of building email title database

Table 1. Email datasets

Name	Amount	Type	Description	Reference
SPAM archive	4,567,714	spam	Spam emails from 2012-June 2018	http://untroubled.org/spam/
Zmal	281	malicious spam	Malicious spam and phishing emails (2017-2018)	Cyber Security Center, Kyushu university
Enron dataset	517,401	legitimate	May 7, 2015 Version of dataset	https://www.cs.cmu.edu/~./enron/
TM	4,251	legitimate	Emails collect between 2012-2018	-

Finally, email subject database is created for matching propose to receive machine translate detected features and risk-words detected features. Several services such as whoisip, language detect and Yandex-translation services are used in this step. From email body part, we collect features using free online services such as virustotal.com, url-query.net and aguse.com by upload and check whatever it is links, pictures or attachment files that came with email.

III. Email Datasets

A performance of formulating spam filter depends on an input data we provide. In table 1. shows email dataset that we currently use and extracted for features. Most of the data set are download from untroubled.org, enron email dataset and got from the Kyushu university zero-day malicious email investigation and analysis lab which contain spam, legitimate, and also malicious/phishing spam emails from 2012 to a recent time in English, Japanese, Chinese, Lao and other languages. Thus, it is still small number of data and we are looking for more on both malicious spam and legitimate dataset especially in many kind of languages used other than English for increase a capacity and diversity of email title database.

IV. Eperiment and results

In our experiment, to verify an accuracy rate of our method of zero-day malspam detection, we use the extracted 27 features and use the testing datasets different from the training datasets by on TensorFlow. From Table 2. shows the results of both spam and normal email type detection are very similar and close to 78% accuracy even using the dataset that have never be trained to the system before. Thus, the number of dataset in this experiment still small due to the limitation of hardware performance at first. Currently the bigger set of data is under processing by using a high performance hardware ITO system of Kyushu university.

Table 2. Experiment and results

Email type	Training set	Testing set	Recall	Preci-sion	TP	TF
Spam	4700	1200	0.7782	0.7866	0.7866	0.2133
Normal	4700	1200	0.7843	0.7758	0.7758	0.2241

To extract some features such as machine translated detected and risk words detected features, the email title database is required to have as much as translated text words in many languages as possible. One more thing, the effectively of translation tool is also important. Now, the database we provide only supports in 4 languages included English, Japanese, Chinese and Lao using Yandex-Translate API. By randomly check the translation results, we found that some title still not translated correctly, at least compare to the popular transition service such as google translate, so it is possible that this issue could be affect the overall result.

V. CONCLUSION

In this paper, we proposed a method by using new features extracted from email headers and deep-learning approach to detect malspam. We have successful extracted 27 features included machine translation detected and risk words detected features from email's header and body part by using several tools. Our quick experiment results show the accuracy rate of malspam detection is about 78%. Thus, the experiment need to improving on dataset capacity and a comparison between several machine learning algorithms such as SVN, NB, KNN, etc..

Acknowledgment

This research was supported by Strategic International Research Cooperative Program, Japan Science and Technology Agency (JST) and JSPS KAKENHI Grant Number JP16K00480..

References

[1] S.Phomkeona and K.Okamura. "A Design Method for Zero-day Malicious Email Detection Using Email Header Information Analysis (EHIA) and Deep-Learning Approach" CSS2018, Nagano, Japan.

[2] S. Phomkeona, K. Edwards, Y. Ban, K. Okamura. "Zero-day Malicious Email Behavior Investigation and Analysis". Asia Pasific Advance Network Workshop (APAN44 Dalian). Aug, 2017.

[3] Omar Al-Jarrah, Ismail Khaterz and Basheer Al-Duwairi, "Identifying Potentially Useful Email Header Features for Email Spam Filtering", ICDS 2012: The Sixth International Conference on Digital Society.

[4] Ali Shafigh Aski and Navid Khalilzadeh Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques" Pacific Science Review A: Natural Science and Engineering, Volume 18, issue 2, July 2016, Pages 145-149