HYBRID PHISHING EMAIL DETECTOR BASED ON EMAIL FEATURES AND CONTENTS

Chih-Chun Yeh, Institute of Information Management, National Taiwan University, 1 Sec. 4 Roosevelt Road, Taipei, Taiwan, ROC, 1123tisbiq@gmail.com
Yong-Xuan Chen, Institute of Information Management, National Yang Ming Chiao Tung University, 1001 Ta-Hsueh Road, Hsin-Chu, Taiwan, ROC, yongxuanchen0106@gmail.com
Angel Lee, Heinz College of Information Systems and Public Policy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA, Lee.Angel.DS@gmail.com
Shao-Ci Wang, Institute of Information Management, National Yang Ming Chiao Tung University, 1001 Ta-Hsueh Road, Hsin-Chu, Taiwan, ROC, amy83121822@gmail.com
Cooper Cheng-Yuan Ku, Institute of Information Management, National Yang Ming Chiao Tung University, 1001 Ta-Hsueh Road, Hsin-Chu, Taiwan, ROC, cooperku@nycu.edu.tw

ABSTRACT

Cybercrime events have been increasing quickly in recent years. Therefore, how to detect them becomes essential in many enterprises. According to related research, the major trend of phishing email detection methods is using only email context or functionality. In this manuscript, we consider using email contents and features with the specifically designed extractor to capture critical parameters to train the selected machine-learning algorithms. Then the individual models for email contents and features are built. The prediction results are combined by an OR operator to create a hybrid detector. The performance includes accuracy 0.9977, precision 0.9907, recall 1.0000, and F-measure 0.9953.

Keywords: Phishing emails, Email contents, Email features, Hybrid detector, Machine learning

INTRODUCTION

Nowadays, phishing emails are getting sophisticated, making them hard to detect. Although protection mechanisms and security awareness continue to improve, phishing emails with reasonable scenarios, changing formats, and cheating methods, especially spear phishing attacks, still cause much trouble for many organizations and individual users. Therefore, this study aims to design an efficient hybrid detection method focusing on English phishing emails to minizine the impact and harm of these malicious contents.

METHODOLOGY

We briefly introduce the proposed hybrid phishing email detection framework, as shown in Fig. 1. It includes two modules, i.e., the email content prediction model and the feature-based prediction model.

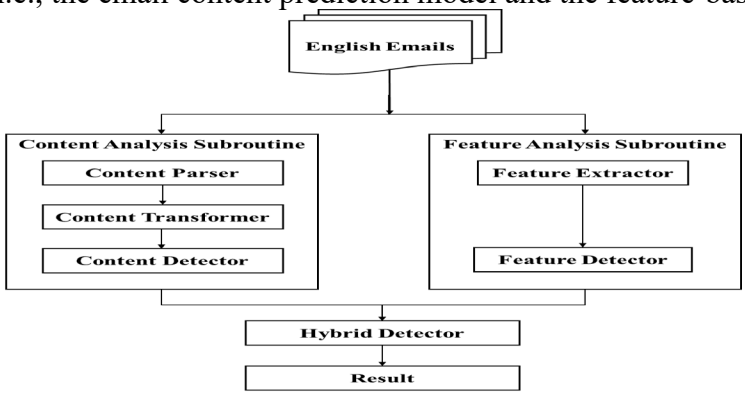

Fig. 1 Phishing email detection framework

First, the content parser and transformer filter and analyze the email contents to retrieve the critical characteristics of phishing. These parser techniques, such as segmentation, lemmatization, and stemming, and the transformer tasks of TF-IDF, Doc2Vec, and Latent Dirichlet Allocation, are all embedded in our model to improve the prediction performance. In addition, the logistic regression method is adopted based on the performance consideration. Then, the feature extractor retrieves the critical parameters that indicate high maliciousness. We select the random forest for this module because it functions best. Finally, a mix-up scheme of OR operation implements the conclusive decision.

EXPERIMENTAL RESULTS

In this section, we analyze the performance of our hybrid detector with the best combination of context and feature selection algorithms. Four raw datasets are chosen for these experiments, i.e., the Nazario phishing corpus (Nazario, 2005), the dataset used in IOP 2019 paper (Yang et al., 2019), the SpamAssassin corpus (Apache, 2020), and the Enron email dataset (Kunegis, 2020). The phishing emails are randomly selected from the Nazario phishing corpus and the dataset of IOP 2019 paper. The legitimate emails are randomly chosen from the dataset of IOP 2019 paper, the SpamAssassin Corpus, and the Enron email dataset. After removing the duplicates, non-English, and invalid emails, our experimental dataset consists of 9183 emails with 7180 legitimate emails and 2003 phishing emails. Table 1 indicates the detail.

Table 1 Experimental dataset

Corpus	Phishing Emails	Legitimate Emails
Nazario phishing corpus	1416	0
Dataset used in IOP 2019 paper	587	4383
SpamAssassin corpus	0	2665
Enron email datasetMax	0	132
Total number (%)	2003 (21.8%)	7180 (78.2%)

CONCLUSION AND FUTURE WORK

This abstract proposes a multi-stage hybrid approach that aims at detecting phishing emails based on techniques related to text processing, feature extraction, machine learning, and improved classification algorithms. The primary aspect is to obtain more characteristic attributes for phishing detection models to achieve the optimized precision, recall, accuracy, and F-measure score. As a prospect of future research objectives, the use of languages different from the datasets in this study will be very interesting to test. Also, we will implement other approaches to detect phishing emails based on natural language processing techniques. It may make the results of the proposed phishing detection model more interpretable.

REFERENCES

Kunegis, J. Enron Email DatasetMax. Accessed in 2020 at: http://konect.cc/networks/enron/.

Nazario, J. The online phishing corpus. Collected in 2005 at: http://monkey.org/~jose/wiki/doku.php.

The Apache Spamassassin Public Corpus. Accessed in 2020 at: http://spamassassin.apache.org/downloads.cgi?update=201504291720

Yang, Z., Qiao, C., Kan, W. & Qiu, J. Phishing Email Detection Based on Hybrid Features. IOP Conference Series: Earth and Environmental Science, 2019, 252, 042051.