Trustworthiness of Spam Email Addresses using Machine Learning

Francisco Jáñez-Martino
Department of Electrical, Systems and Automation, University of León
Researcher at INCIBE (Spanish National Cybersecurity Institute)
León, Spain
francisco.janez@unileon.es

Rocío Alaiz-Rodríguez
Department of Electrical, Systems and Automation, University of León
Researcher at INCIBE (Spanish National Cybersecurity Institute)
León, Spain
rocio.alaiz@unileon.es

Víctor González-Castro
Department of Electrical, Systems and Automation, University of León
Researcher at INCIBE (Spanish National Cybersecurity Institute)
León, Spain
victor.gonzalez@unileon.es

Eduardo Fidalgo
Department of Electrical, Systems and Automation, University of León
Researcher at INCIBE (Spanish National Cybersecurity Institute)
León, Spain
eduardo.fidalgo@unileon.es

ABSTRACT

Cybercriminals have increasingly used spam email to send scams, phishing, malware and other frauds to organisations and people. They design sophisticated and contextualised emails to make them look trustworthy for users, being the sender addresses an essential part. Although cybersecurity agencies and companies develop products and organise courses for people to detect emails patterns, spam attacks are not totally avoided yet.

This work presents a proof-of-concept methodology to give the user more meaningful information about trustworthiness to detect these harmful emails. For the first time in the literature, we present an email address dataset manually labelled into two classes, low and high quality. Moreover, we extracted 18 handcrafted features based on social engineering techniques and natural language properties. We evaluated four popular machine learning classifiers and obtained the best performance with Naive Bayes, i.e., 88.17% of accuracy and 0.808 of F1-Score. Additionally, we applied the InterpretML framework to find out the most relevant properties to eventually implement an automatic system able to inform about the trustworthiness of email addresses.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Security and privacy** → *Security services*; Human and societal aspects of security and privacy; • **Applied computing** → Document management and text processing.

DocEng '21, August 24–27, 2021, Limerick, Ireland

ACM ISBN 978-1-4503-8596-1/21/08…$15.00
https://doi.org/10.1145/3469096.3475060

KEYWORDS

Spam Email Detection, Machine Learning, Classification Algorithm, Cybersecurity, Feature Extraction

ACM Reference Format:
Francisco Jáñez-Martino, Rocío Alaiz-Rodríguez, Víctor González-Castro, and Eduardo Fidalgo . 2021. Trustworthiness of Spam Email Addresses using Machine Learning. In *ACM Symposium on Document Engineering 2021 (DocEng '21), August 24–27, 2021, Limerick, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469096.3475060

1 INTRODUCTION

The amount of spam emails is steadily increasing and represents more than 50% of emails out of the more than 250 billion daily sent worldwide [8]. Spam email is one of the most used vectors of cyberattacks, which exposes companies and citizens to scams, such as phishing, malware, spoofing or hacking extortion [6]. Europol's European Cybercrime Centre [5] stated that 78% of cyberattacks had their root in spam emails.

Main tools against spam emails are the anti-spam systems, which are generally based on machine learning and natural language processing techniques. However, the spam field has a very dynamic nature that requires a continuous upgrade of the filters [4]. Spammers, i.e., people who send spam emails, design creative strategies for (1) bypassing machine learning filters by contaminating the data [4] and (2) misleading the users by making their harmful emails look more credible and contextualised [10].

Cybersecurity organisations implore users to check the email addresses when receiving suspicious messages and report them. However, employees or citizens have difficulties identifying phishing emails [2]. Spam often contains the company name – or similar – in order to appear legitimate and inspire trust in users. This problem has usually been tackled through black and white listing. But spammers eventually find out a way to bypass such lists, and their emails end up appearing in the mailboxes.

The quality of suspicious cyberattacks and criminal activity sites tend to be a major feature for cybercriminals by allowing them to build a confident and fake secure atmosphere for users. Detecting spam email campaigns as soon as possible is essential to reduce the number of people that become victims of them. Law Enforcement Agencies (LEAs) and cybersecurity experts need accurate and time-efficient tools to analyse many documents [1]. Most of the spam filters mainly use the body content from emails. However, features derived from the header data may speed up the entire process, resulting in extremely fast classification models [1]. Hence, our goal is to propose an approach to measure the quality of a spam email to help LEAs and cybersecurity experts to warn of spam campaigns. Experts consider the email address as a header with relevant information to extract [4]. We hypothesise that the addresses that seem to be more trustworthy for users are related to the most harmful spam emails.

We propose a pipeline based on a question vector that feeds up a machine learning algorithm capable of measuring the quality of email addresses. Such question vector recollects the information of email addresses extracted from a well-known spam email dataset. Our classifier outcomes a binary response by labelling an address as high or low quality. This allows people to be aware of suspicious emails when they receive an unexpected message.

2 LITERATURE REVIEW

To the best of our knowledge, no works have tackled the automatic classification of spam email addresses regarding their quality. However, related work, although different, comes from fields like URLs phishing detection. The goal of phishing is similar to the spam, as it seeks to mislead the user using social engineering.

The detection of phishing websites using machine learning algorithms has proven to be more robust than other methods based on black/white lists and several works using URLs to detect phishing websites [13]. Rao et al. [11] extracted features from URLs to feed a phishing classifier, which includes hyperlink properties, obfuscation and third-party based features. Sahingoz et al. [12] designed a detection system which uses 38 Natural Language Processing features and evaluated seven machine learning algorithms, able to increase the detection rate by 10.86%. Their features include word length, word and character counts, regular expressions and adjacent characters. This last feature was also used in [15] to extract local correlations between a character and their adjacent ones.

Later, Li et al. [7] unified features from malicious URLs and increased them up to 62 in order to improve machine learning-based applications. A method based on Convolutional Neural Networks was presented in [14] to create a malicious URLs phishing detector. They used a one-hot vector based on 70 characters including 26 English letters, 10 digits, 33 other characters and the "new line" character. In their survey, Silva et al. [3] evaluated the static features and observed their occurrence in current phishing attacks and identify the similarities, relationship and relevance among them.

3 PROOF-OF-CONCEPT METHODOLOGY

In this work, we proposed a proof of concept that consists of a pipeline with the following steps: 1) defining the classes and features to tackle the problem, 2) building a dataset and 3) using a feature vector to feed up four machine learning classifiers.

For our proof of concept, we followed other works [11–13] and applied traditional algorithms. We selected Naïve Bayes (NB), Support Vectors Machine (SVM), Logistic Regression (LR) and Random Forest (RF) machine learning algorithms as classifiers.

3.1 Quality Definition

The quality of a spam email address is an essential factor to determine the level of trustworthiness that it offers to users. The more trustful an address is, the more harmful and risky the spam email becomes. This meaningful information improves the feedback to people that receive an unwanted and unsolicited email from an unknown sender and allows them to have more awareness to decide whether they should trust the sender. Hence, if an unknown address is classified as high quality category, the email service notifies that the address should be checked thoroughly, since the email may be suspicious of containing high-risk elements for users. Indeed, it can be used as a spam feature detection in combination with other features, such as detection of poisoning text or sender spoofing. In addition, the quality of the addresses can be a useful indicator for cybersecurity experts to identify the most harmful spam campaigns and activate more targeted warnings to people and companies. To evaluate our proof-of-concept pipeline, we separated the addresses into two categories, i.e. low and high quality. **Low quality** contains addresses that are composed of unrelated words, numbers or characters, even being randomly written. **High quality** addresses involves social engineering techniques and imitates conventional, corporate, news, information and companies addresses. They also include popular email services, branch and top-level domains (TLDs). Table 1 presents five examples per quality.

Table 1: Examples of low and high quality spam email addresses.

Low Quality
fathers.day.curious.finds@icy7752.com
kn95-masks@updates.cn.com
notification+KHm9BEbYo6kgcABMd0sQaDyGSC9cBaJRB@parliamena.cf
us-concealed-online-partner@via858.com
pop@achieverecruit.xyz
High Quality
AntiVirusProtection@squirreltailoven.org
johnhag224@gmail.com
support@wholesalehilltribesilver.com
frances_martinez98@lesclass.com
mail@amazonsupport.info

3.2 Feature Extraction

Our proposed approach for feature extraction is inspired by Sahingoz et al. [12]. We adapted their URL-based features to the email address structure (Fig. 1) and, with the help of cybersecurity experts from Spanish National Cybersecurity Institute (INCIBE), defined 18

binary questions, whose answers can be used as features as shown in Table 2. An affirmative answer was given the weight of 1 and a negative one of 0. According to their ID, features F1-5, F9-12 and F17 refer to extension or composition properties from some part of the address, whereas features F6-8, F13-16 and F18 are focused on capturing subjective and social engineering information. We used these questions to build our dataset.

Figure 1: Email address components.

Table 2: Features captured from each email address. ID column shows a short name to refer to each feature through Feature-Number. All responses are binary: 1 is an affirmative answer and 0 a negative one, depending on the feature the response means High Quality (HQ) or Low Quality (LQ). SME corresponds to "Small and Medium Enterprise".

ID	Features	Explanation	Affirmative Response
F1	ifusername_tooshort	contains less than 4 chars	LQ
F2	ifusername_random	contains random chars	LQ
F3	ifusername_toonums	contains more than 4 numbers	LQ
F4	ifusername_toochars	contains more than 4 special chars	LQ
F5	ifusername_singleword	is a single word	HQ
F6	ifusername_corporative	imitate a corporative address	HQ
F7	ifusername_onlyname	is only a name or surname	HQ
F8	ifusername_domainrelation	there is a relation among them	HQ
F9	ifdomain_tooshort	contains less than 4 chars	LQ
F10	ifdomain_random	contains random chars	LQ
F11	ifdomain_toonums	contains more than 2 numbers	LQ
F12	ifdomain_chars	contains more than 2 special chars	LQ
F13	ifdomain_popularemail	is a well-know email service provider	HQ
F14	ifdomain_incluidemail	contains the word "mail"	HQ
F15	ifdomain_popularcompany	includes a popular company	HQ
F16	ifdomain_SMEcompany	could be a SME company name	HQ
F17	iftld_manytlds	has more than one tlds	LQ
F18	iftld_unknown	is unknown or unpopular	LQ

3.3 Dataset Creation

To evaluate our classification pipeline, we have extracted the email addresses from a recent and public spam email dataset, i.e., the Bruce Guenter project[1]. This project has been recollecting and publishing spam emails since 1999 and has become the most up-to-date dataset in this field. Hence, we selected the emails from 2019 and 2020 to extract the sender addresses (i.e., the *from* header) from each spam email.

Apart from the authors, we involved five additional people in the labelling tasks in order to reduce uncertainty. Firstly, we prepared a questionnaire to obtain the values for our 18 features, each one corresponding to a question. Although some features, like *ifusername_tooshort* or *iftld_manytlds*, are objective and can be easily extracted automatically, the questionnaire was useful to train the team and give them a clearer perspective to the problem. Other features, such as *ifdomain_SMEcompany* or *iftld_unknown*, are subjective and hard to extract automatically without a baseline. We include the meaning of an affirmative or negative answer (Table 2) per question in order to improve their training to be able to indicate a quality properly. Lastly, the five people tagged subjectively all addresses into two classes, low or high, according to their quality.

We obtained the questionnaire answers and quality class from a total of 6569 email addresses. Since each instance was labelled by five people, the final label was chosen to be the mode of the answers given by them. Finally, our dataset, Email Addresses Quality 6569 (EAQ-6K), which has been made publicly available for research purposes[2], contains 5181 and 1388 low and high quality email addresses, respectively.

4 EXPERIMENTATION

4.1 Experimental Setup

We carried out our experiments on an Intel® Core™ $i7-7thGen$ with 16G of RAM, under Ubuntu 18.04 OS and Python 3.

The classifier parameters were tuned according to F1-score and accuracy. For NB, we used a Multinomial distribution. We chose a linear kernel for the SVM model, and the cost C value was set 0.1. In the case of LR, we set a $C = 0.1$ and Stochastic Average Gradient (SAG) as solver. For RF, we used 50 estimators and 5 as maximum depth. The rest of the model parameters are left with their default values. Lastly, since our dataset is imbalanced, we have configured a balance weight parameter for all classifiers.

The classifier performance was estimated using 10-fold cross-validation, and it has been reported in terms of accuracy and F1-score.

4.2 Results

The results are shown in Table 3. Naïve Bayes achieved the best performance with an accuracy of 88.17% and F1-score of 0.808 . A reason for this may be that most features are independent among them and Naïve Bayes, rather than the rest, assumes the features are statistically independent. The rest of classifiers reached accuracy values above 80% but lower than Naïve Bayes.

Table 3: Evaluation of four traditional machine learning algorithms. Runtime refers to time per address.

Classifier	Accuracy(%)	F1-Score	Runtime (µs)
NB	**88.17**	**0.808**	0.04
SVM	80.70	0.762	6.9
LR	83.49	0.787	0.16
RF	85.40	0.817	11.1

Moreover, the execution times for Naïve Bayes are also the lowest making this approach fast enough to be incorporated in a cybersecurity application.

4.3 InterpretML report and discussion

InterpretML [9] is a package that includes the state-of-the-art machine learning interpretability techniques. This tool makes possible

[1]http://untroubled.org/spam/ Retrieved July 2021

[2]http://gvis.unileon.es/dataset/email-addresses-quality-eaq-6k/ Retrieved July 2021

to understand a classifier's behaviour in global and individual prediction. The report provided by InterpretML stated the features and classes relationship, including the weights and if there is interference between classes, i.e. if features are balanced. Since Logistic Regression is the only glass-box model available in InterpretML among four classifiers, we show the report obtained with it.

Features F11 (more than two numbers in the domain) and F18 (unknown TLDs) will contribute towards the low quality class. In high-quality cases, the classifier assigns more relevance to F13 and F6, i.e., corporate address that uses popular email service would be the most harmful email addresses.

In addition, this information verifies the relationship between class quality and the binary response to each feature that we assign in Section 3.2. Features F5, F7 and F12 appear to be the least relevant ones. F5 and F7 refer to the username being a single word and a single name, respectively. However, a word or name may be different according to its meaning; thereby, we would need to include semantic information in both features. The Q12 is a special case because there are no affirmative answers from any address (i.e. no address contained more than 2 special characters in its domain).

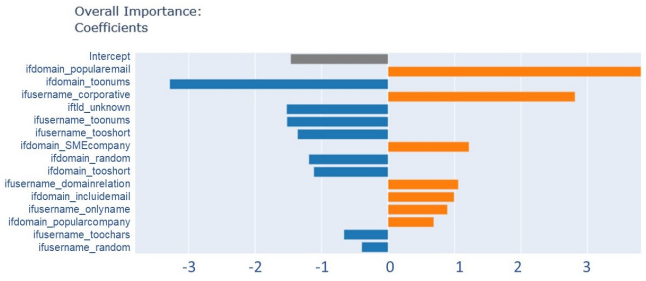

Figure 2: InterpretML report: weight and class relationship of each feature with respect to quality. Axis X contains the weight per features and Axis Y the features.

5 CONCLUSION AND FUTURE WORK

In this work, we proposed a proof of concept methodology to measure the quality of spam email address into two levels, low and high. Our pipeline aims to help both potential victims to detect scam emails more easily and also improve the cybersecurity experts' spam campaigns warnings.

We present a novel dataset EAQ-6K with sender email addresses from the most recent spam emails of Bruce Guenter dataset. EAQ-6K was built and labelled by five people by answering 18 questions and a class quality question. We used the answers to these questions as features to feed up four machine learning classifiers: NB, SVM, LR and RF.

Experimental results showed that the vector proposed – i.e. the answers to the 18 questions – captures enough information to reach 88.17% of accuracy and 0.808 of F1-Score using the Naïve Bayes classifier. Moreover, our study of feature relevance provides information regarding the most relevant features: *ifdomain_popularemail*, *ifdomain_toonums* and *ifusername_corporative*. It also allows us to give weights for every feature. This way, we could label more email addresses automatically and increase our dataset in order to improve the performance.

We are searching for incorporating information related to semantic and intrinsic information from words using short text classification techniques in future works [1]. Besides, we are exploring different options for the classes, such as adding new labels (e.g. medium quality) or addressing this task as a regression problem.

Finally, due to the report from InterpretML, we will attempt to use the feature weight to be able to label automatically more addresses. However, since Logistic Regression was the only one evaluated, we will also seek to find the features' weights for other classifiers to ensure the consistency in the labels.

ACKNOWLEDGMENTS

This work was supported by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01.

REFERENCES

[1] Mhd Wesam Al-Nabki, Eduardo Fidalgo, Enrique Alegre, and Rocío Alaíz-Rodríguez. 2020. File Name Classification Approach to Identify Child Sexual Abuse. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*. SciTePress, , 228–234.

[2] Moneer Alshaikh, Sean B. Maynard, and Atif Ahmad. 2021. Applying social marketing to evaluate current security education training and awareness programs in organisations. *Computers & Security* 100 (2021), 102090.

[3] Carlo Marcelo Revoredo da Silva, Eduardo Luzeiro Feitosa, and Vinicius Cardoso Garcia. 2020. Heuristic-based strategy for Phishing prediction: A survey of URL-based approach. *Computers & Security* 88 (2020), 101613.

[4] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, 6 (2019), e01802.

[5] Europol. 2019. Spear phishing, a law enforcement and cross-industry perspective. https://www.europol.europa.eu/newsroom/news/europol-publishes-law-enforcement-and-industry-report-spear-phishing. Accessed: 2021-06-01.

[6] Luigi Gallo, Alessandro Maiello, Alessio Botta, and Giorgio Ventre. 2021. 2 Years in the anti-phishing group of a large company. *Computers & Security* , (2021), 102259.

[7] Tie Li, Gang Kou, and Yi Peng. 2020. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Information Systems* 91 (2020), 101494.

[8] Rami Mohammad and A Mohammad. 2020. A lifelong spam emails classification model. *Applied Computing and Informatics* , (01 2020), 10.

[9] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* , (2019), .

[10] Daniela Seabra Oliveira, Tian Lin, Harold Rocha, Donovan Ellis, Sandeep Dommaraju, Huizi Yang, Devon Weir, Sebastian Marin, and Natalie C. Ebner. 2019. Empirical analysis of weapons of influence, life domains, and demographic-targeting in modern spam: an age-comparative perspective. *Crime Science* 8, 1 (2019), 3.

[11] Routhu Rao and Alwyn Pais. 2019. Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications* 31 (08 2019).

[12] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. 2019. Machine learning based phishing detection from URLs. *Expert Systems with Applications* 117 (2019), 345–357.

[13] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre. 2021. Impact of Current Phishing Strategies in Machine Learning Models for Phishing Detection. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor Quintián, and Emilio Corchado (Eds.). Springer International Publishing, Cham, 87–96.

[14] Wei Wei, Qiao Ke, Jakub Nowak, Marcin Korytkowski, Rafał Scherer, and Marcin Woźniak. 2020. Accurate and fast URL phishing detector: A convolutional neural network approach. *Computer Networks* 178 (2020), 107275.

[15] P. Yang, G. Zhao, and P. Zeng. 2019. Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. *IEEE Access* 7 (2019), 15196–15209.