IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

High-Performance Phishing E-mail Detection Using Header Features

[1] Sulaiman A. Maeli, [2] Ajay U. Surwade

[1]Research Scholar, [2]Associate professor

[1,2] School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India

Abstract: E-mail is the most common and used means of communication by individuals or institutions because it enables them to exchange data or information easy and quick. Phishers use e-mail to send phishing messages that enable them to deceive users and steal their personal information or data, including their bank account data, bank card passwords, which causes them a lot of harm and financial losses. Phishing e-mails are increasing menace, so it is necessary to block these e-mails at the Mail Transfer Agent (MTA) or Mail Delivery Agent (MDA) level.

In this study, the architecture of filter has been implemented after extracting important information from e-mail headers. By analyzing the important features of e-mails, they are classified according to rules and conditions framed. This filter is tested on five standard datasets: Enron, Public Phishing Corpus, SPAM Archive, CSDMC2010, and Spam Assassin. The average accuracy for individual dataset is 99.58%, 99.53%, 98.33%, 99.34% and 98.31% respectively and overall average accuracy achieved is 99.01%. Although blocking phishing e-mail based on e-mail header is old technique, but this will certainly reduce the load on different content-based filters. The approach discussed in this paper has significantly improved the accuracy up to 99.01% using header part of e-mail reported so far thus, it is encouraging and this would increase the accuracy of classification of phishing e-mails.

***Index Terms* - E-mail header, OBF, Blacklist, Phishing e-mails, Classification.**

I. INTRODUCTION

E-mail is one of the most common and used means of communication among individuals or institutions because of its ease of use and convenience and has become a popular means of communication. Therefore, Spammers and Phishers exploit their popularity to practice fraud and theft by sending spam and phishing messages to these parties. The receiver of this message interacts with it, whether by clicking on links or sending sensitive and confidential data to the phisher. According to the APWG Phishing Activity Trends Report, the number of phishing attacks has tripled since early 2020, resulting in total losses for both large and small businesses in billions of dollars [1].

The APWG observed almost five million phishing attacks (4,987,809) up to year 2023 in all. This made 2023 the worst year for phishing on record, eclipsing the 4.7 million attacks seen in 2022 [2]. To prevent this, many studies have been conducted to deal with these fraudulent e-mail messages by filtering messages in several different ways before they reach the recipient. The research done so far has focused on the content of message filtering, and header information has been ignored or very little attention has been given to the header of the message. The header of the message contains many elements and attributes that are helpful in classifying e-mails as legitimate or phishing.

Our contribution, an Origin-Based Filter (OBF) has been developed using eight robust rules which are formulated based on observations and investigations carried out on the standard datasets which are publically available. These observations have focuses on the important header elements of e-mail. This OBF was tested on five publicly available datasets such as ‘Enron’, ‘Public_Phishing_Corpus’, ‘SPAM Archive’, ‘CSDMC2010_SPAM’, and ‘SPAM Assassin’.

II. RELATED WORKS

Amir Herzberg, (2009) created two blacklists, a short blacklist collected at local sites and a long blacklist consist of more suspected senders. This has improved lookup at the DNS level [3].

Giovane (2009) designed an e-mail phishing detection system with a self-management architecture. It provides a protection mechanism to internet service providers against phishing attacks [4].

Hajgude and Ragha (2012) proposed a hybrid method to detect phishing mail using a blacklist, whitelist, and heuristic method that performs textual analysis of e-mail content and lexical URL analysis. The DNS in the real link is checked in the blacklist and whitelist. If it is in the blacklist, it has recognized it as phishing and as a legitimate DNS if it is present in the whitelist. However, if it is not present in either a blacklist or whitelist, then it is analyzed using textual and lexical analysis [5].

Kaur and Kalra (2016) developed a hybrid scheme to detect phishing attacks, called the five-tier barrier. They used whitelists along with modules based on heuristic factors to test the URL, which classifies it as phishing or non-phishing [6].

Yasin and Abuhasn (2016) proposed a model to classify e-mails as phishing or non phishing. This model utilizes knowledge discovery and data mining techniques. The construction of the model involved an intelligent preprocessing phase that extracted a set of features from the header, body, and term frequency of the e-mails. Data mining algorithms such as SVM, Bayes Net, Random Forest, J48, and MLP were used to experiment with the model. The classification results achieved enhanced accuracy compared to similar published models [7].

Sheu et al. (2017) proposed a method based on machine learning and data mining techniques by analyzing the common rules between various attributes of the e-mail header spam and developed a spam filtering mechanism for classification based on the header section [8].

Gascon et al. (2018) developed a spoofing e-mail-detection approach based on sender traits in the structure of an e-mail, such as personal preferences, e-mail client, and infrastructure. It can learn the sender's profile and recognize spoofed e-mails without depending on their content or server-side implementations [9].

Ghogare et al. (2018) proposed a framework for feature selection and spam classification by extracting sender e-mails and using it for classification [10].

Krause et al. (2019) proposed a new approach for spam detection based only on metadata features derived from the e-mail header section using a static set of engineered and automatically extracted features [11].

Surwade (2019) addressed the phishing e-mail problem by studying literature surveys of phishing e-mail filtering techniques. This research gap has been identified as an effective and adaptive technique for phishing e-mail filtering [12].

Kulkarni et al. (2020) investigated attributes of the header part of an e-mail using five Feature Selection techniques with Five Machine learning classifiers: Support Vector Machine, Random Forest, Naïve Bayes, NBTree, and Decision Tree [13].

Bijalwan (2020) suggested the use of blacklists for network traffic to filter infected packets. This method was used to detect malware and botnets. After the packet filtering process, all suspicious IP addresses were classified as blacklists [14].

Surwade (2020) proposed an origin- based filter (OBF) that blocks phishing e-mails by extracting information from the e-mail message headers and analyzing it according to a set of rules and conditions using a blacklist approach. A public phishing corpus dataset was used for testing. Improved accuracy has been reported [15].

III. Methodology

The e-mail message is composed of two parts. The first part is the message header, containing message information like ‘To’, ‘From’, ‘Reply-To’, ‘Return-Path’, ‘Subject’, etc. The second part is the e-mail body that contains the actual message in text format [16]. Accordingly, our system has been designed to deal with the header part to classify e-mail messages as phishing or non- phishing messages. The architecture of the system is illustrated in Fig. 1.

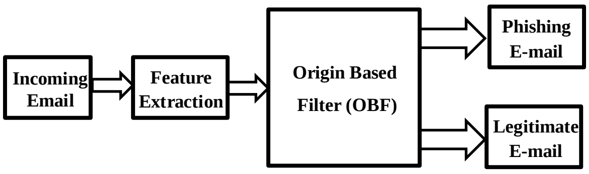

Figure1. System Architecture for Anti Phishing System

The header elements of the phishing and non-phishing e-mails were analyzed. An origin-based filter (OBF) was designed, which includes the eight important rules described below. These rules were then used for further classification of the e-mails.

Rule-1: This rule is checking an e-mail for its format, whether it is having valid format or invalid format of an e-mail address.
¬ regex1(FROM)
where, ‘regex1’ is a regular Expression for verifying the ‘from’ field of e-mail address.

Rule-2: This rule is checking, if ‘From’ field and ‘Reply-To’ field has different e-mail addresses, then it is phishing e-mail
otherwise, if it is same, it is non–phishing e-mail.
‘From’ field and ‘Reply-To’ field has different domain name. (FROM <Domain_name> ≠ REPLY-TO< Domain_name>).

Rule-3: This rule is checking, if ‘From’ field and ‘Return-Path’ field has different e-mail addresses, then it is phishing e-mail
otherwise, if it same, it is non–phishing.
‘From’ field and ‘Return-Path’ fields has different domain name.
(FROM <Domain name> ≠ RETURN-PATH <Domain_name>)

Rule-4: This rule is checking, if ‘Reply-To’ field or ‘Return-Path’ field or both fields are empty, then it is phishing e-mail
otherwise, it is non–phishing.
‘Reply-To’ or ‘Return-Path’ or both are empty.
len (REPLY-TO) == 0, OR
len (RETURN-PATH)== 0, OR
(len(REPLY-TO) == 0 AND RETURN-PATH== 0)

Rule-5: This rule is checking, if ‘From’ field and ‘message id’ field has different e-mail addresses, then it is phishing e-mail
otherwise, if it same, it is non–phishing. ‘From’ field and ‘message id’ has different domain name.
(FROM <Domain_ name> ≠ Message_ID <Domain_ name>)

Rule-6: This rule is checking, if ‘Subject’ field is blank or empty then it is phishing e-mail otherwise it is non–phishing.
‘Subject’ field is blank or empty.
(len(SUBJCT) = 0)

Rule-7: This rule is checking, if all words in ‘Subject’ field is in upper case then it is phishing e-mail otherwise it is non –

phishing. All subject words are in uppercase.
is_uppercase (SUBJCT)

Rule-8: This rule is checking, if ‘Subject’ field containing special characters then it is phishing e-mail otherwise it is non–

phishing. The ‘Subject’ field containing special characters.
(regex.search (SUBJCT) ≠ ∅)
where, ‘regex’ is a regular expression, checking special characters in subject field.

In the above eight rules, **‘¬’** denotes logical negation, **‘V’** denotes logical OR, **‘≠’** denotes not equal to, **‘len(SUBJCT)’** denotes the length of the string SUBJCT field, **‘is_uppercase (SUBJCT)’** denotes a function that returns True if SUBJCT is in all uppercase letters, **‘Λ’** denotes logical AND, and **‘∅’** denotes the empty set.
By connecting above conditions logically using OR operator, it becomes the following expression for classification of e-mail as phishing or non-phishing.

If ((¬ regex1(FROM)) ∨ (FROM <Domain_name> ≠ RETURN-PATH <Domain_name>) ∨ (FROM <Domain name> ≠ REPLY-TO <Domain_name>) ∨ (len(REPLY-TO) = 0) ∨ (len(RETURN-PATH) = 0) ∨((len(REPLY-TO)=0) ∧ (len (RETURN-PATH)=0)) ∨(len(SUBJCT) = 0) ∨ ((SUBJCT) is_uppercase)) ∨(regex.search (SUBJCT) ≠ ∅) ∨ (FROM <Domain_name> ≠ Message_ID <Domain_name>)).

If all or one of the following rules is proved or tested to be true, then the e-mail message is classified as ‘phishing mail’, The domain name or IP address is extracted from it and then added it to the ‘Black-listed’ domain names or IP addresses. Otherwise, the e-mail needs to be checked by using some other Content based filters. A working diagram of the OBF filter is shown in Fig. 2.

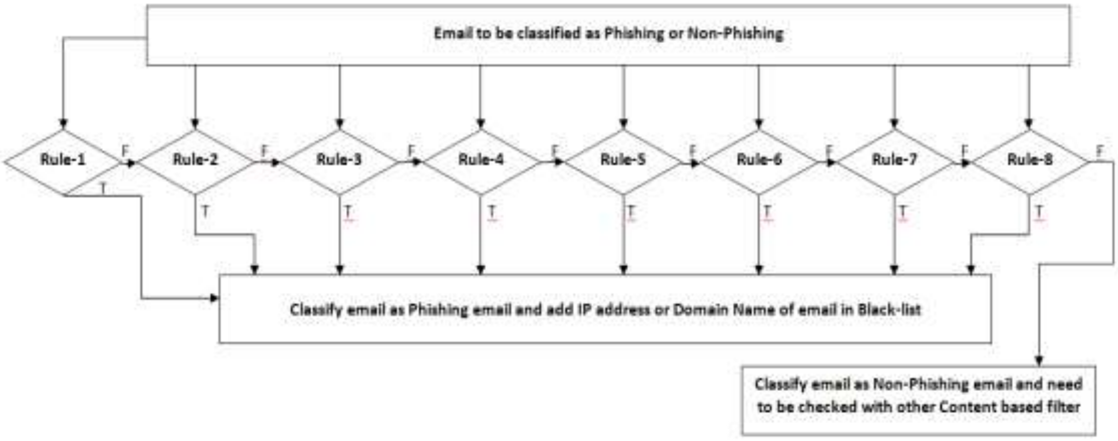
Email to be classified as Phishing or Non-Phishing
Rule-1
F
T
Rule-2
F
T
Rule-3
F
T
Rule-4
F
T
Rule-5
F
T
Rule-6
F
T
Rule-7
F
T
Rule-8
F
T
Classify email as Phishing email and add IP address or Domain Name of email in Black-list
Classify email as Non-Phishing email and need to be checked with other Content based filter

Figure2. Working of Origin based Filter

The Python program has been written using the eight rules which are discussed in methodology section. This program is checking the conditions as mentioned in each rule and individually and decision is made as shown in Fig.-2. If condition in ‘Rule-1’ is ‘False’ the control is passed to ‘Rule-2’ and so on till ‘Rule-8’. When the condition is found ‘False’ in the eight rule the e-mail should be checked using Content based Filter (Which would be implemented using ML techniques).
Otherwise, if condition is found ‘True’ in any one of the eight rules, the program is classifying that e-mail as ‘Black-list’. Once e-mail address is classified as ‘Black-list’ it is added to Black-listed addresses. Thus, day-by-day or filter-by-filter the list of Black-list addresses will become richer and richer. When large number of Black-list e-mail are added in this list, it will become richest in nature and it would certainly reduce the load on Content based Filter since, correct classification will be done at Origin based Filter only. The overall

accuracy reported by this filter is 99.01%. Very fewer cases will be then needed to filter using Content based Filter. This program is tested on the standard datasets such as Enron, Public Phishing Corpus, SPAM Archive, CSDMC2010_SPAM, Spam-Assassin.

Enron is a public collection of real e-mails gathered and organized by the CALO Project (A Cognitive Assistant that Learns and Organizes). The corpus contains approximately 500,000 messages from approximately 150 users, most of which are from Enron Senior Management[17]. Folders with spam e-mails with a total of 32988 messages were selected for classification [18].

Public Phishing Corpus is a public-phishing dataset containing 4554 phishing messages. It consists of total 4554 e-mail stored with E-mail Message format (.eml) in four folders such as 'phishing0' contains 414 e-mails, 'phishing2' contains 1423 e-mails, 'phishing3' contains 2279 e-mails and '20051114' folder contains 438 e-mails. Each '.eml' file contains one e-mail message [19].

SPAM Archive comprises public spam messages collected by Bruce Guenter since early 1998, and the dataset files are updated every month by adding new spam e-mails [20]. The collection of messages for two months was selected [21].

CSDMC2010_SPAM is dataset included a two-part training part containing 1378 spam and 2949 ham, out of which only 1378 spam e-mails were used for this experiment. Another testing part contained 4292 e-mails without known class labels; therefore, these 4292 e-mails were not used for this experiment [22].

Spam Assassin is public e-mail corpus with a total of 6047 messages with about a 31% spam ratio. Spam folders 20030228_spam and 20030228_spam_2, with a total of 1897 spam messages were selected for experiments [23].

IV. Results and Discussion

The OBF developed using the methodology described in section-3 tested using known spam messages of the standard datasets mentioned above. The above experiments were conducted using the following standard datasets: Enron, Public Phishing Corpus, SPAM Archive, CSDMC2010_SPAM, and SPAM Assassin. In the following section, each standard dataset is discussed along with the results. The results of the classification of e-mails using the proposed OBF are shown in tables 1-5.

Table1. Classification of e-mails in Enron dataset

Enron SPAM Dataset					
Folder Name	**Correct Classification as Spam**	**Misclassification as Legitimate**	**Total**	**Accuracy**	**Misclassification Rate**
BG	9866	134	10000	98.66%	1.34 %
GP	13719	0	13719	100%	0 %
SH	9266	3	9269	99.96%	0.032 %
Total	**32851**	**137**	**32988**	**99.58%**	**0.415 %**

Table 2. Classification of e-mails in Public Phishing Corpus dataset

Public Phishing Corpus					
Folder Name	**Correct Classification as Phishing**	**Misclassification as Legitimate**	**Total**	**Accuracy**	**Misclassification Rate**
Phishing 0	412	2	414	99.51%	0.483 %
20051114	436	2	438	99.54%	0.456 %
Phishing 2	1417	6	1423	99.57%	0.421 %
Phishing 3	2268	11	2279	99.51%	0.482 %

Total	**4533**	**21**	**4554**	**99.53 %**	**0.461 %**

Table 3. Classification of e-mails in SPAM Archive dataset

SPAM Archive Dataset					
Folder Name	**Correct Classification as Spam**	**Misclassification as Legitimate**	**Total**	**Accuracy**	**Misclassification Rate**
01/2020	2603	40	2643	98.48%	1.513 %
02/2020	7366	129	7495	98.27%	1.721 %
Total	**9969**	**169**	**10138**	**98.33 %**	**1.666 %**

Table 4. Classification of e-mails in CSDMC2010 dataset

CSDMC2010 Spam Dataset					
Folder Name	**Correct Classification as Spam**	**Misclassification as Legitimate**	**Total**	**Accuracy**	**Misclassification Rate**
Spam	1369	9	1378	99.34%	0.653 %

Table 5. Classification of e-mails in Spam Assassin dataset

Spam Assassin Spam Dataset					
Folder Name	**Correct Classification as Spam**	**Misclassification as Legitimate**	**Total**	**Accuracy**	**Misclassification Rate**
20030228_spam	493	7	500	98.60 %	1.4 %
20030228_spam_2	1372	25	1397	98.21 %	1.789 %
Total	**1865**	**32**	**1897**	**98.31%**	**1.686 %**

The Enron consists of 32988 spam messages, out of which 32851 messages are correctly classified as ‘Phishing’, while 137 messages have been misclassified as ‘Legitimate’ messages. Thus, OBF has achieved 99.58% average accuracy on Enron Spam messages and misclassification is 0.4153%.

The OBF has correctly classified 4533 phishing e-mails as ‘Phishing’ while, it has misclassified 21 e-mail messages as ‘Legitimate’. Thus, it has achieved 99.53% average accuracy and misclassification is 0.4611%.

On SPAM Archive dataset which contains 10138 messages out of which OBF has correctly classified 9969 messages as ‘Phishing’ while, 169 messages are misclassified. It has achieved 98.33% accuracy of classification and misclassification is 1.66 %.

The CSDMC2010 Spam dataset consists of 1378 messages out of which 1369 messages have correctly classified by OBF as ‘Phishing’ while 9 messages are misclassified by it as ‘Legitimate’. Thus, it has achieved 99.34% accuracy and misclassification is 0.653%.

The Spam Assassin dataset consists of 1897 messages out of which 1865 messages are correctly classified as ‘Phishing’ while, 32 messages have been misclassified as ‘Legitimate’ by the OBF. Thus, it has achieved 98.31% accuracy and misclassification is 1.686%.

The OBF described in this paper works on header part of e-mail has achieved the average accuracy of 99.01% on all five datasets and average misclassification rate is 0.976 % which considerable very less.

The effect of accuracy of these eight rules which are discussed in methodology sections works positively and it has helped OBF to detect majority of e-mail as Phishing while, there is very low misclassification rate. This OBF when deployed at MTA or MDA level will certainly reduce the load on other filter based on Content of e-mail (i.e. Content based Filter). The result reported in this paper is encouraging which would increase the accuracy of classification of phishing e-mails.

V. Conclusion

Because phishing e-mail is an increasing menace, it is important to block it. In this paper, architecture is proposed for detecting and blocking phishing e-mails. This architecture consists of an origin-based filter and content-based filter. The origin-based filter was implemented using a Python script and tested on five standard datasets.

The accuracy calculated for these standard datasets is as for Enron dataset accuracy is 99.58%, for Public Phishing corpus it is 99.53%, for SPAM Archive accuracy achieved is 98.31%. The accuracy for CSDMC2010 achieved is 99.34%, and for Spam Assassin datasets accuracy is 98.31%. The rules explained in the methodology section were constructed based on observations from all standard datasets. All these rules are coupled with logical ‘OR’ which reduces the chances of misclassification of phishing e-mails to non-phishing e-mails. Thus, overall average accuracy achieved is 99.01% and very low rate of average misclassification 0.976%.

The header part of the e-mails was used to classify e-mails as phishing or non-phishing sounds good. This would certainly reduce the load on the other filters with which it could be coupled further. The phishing e-mail classification described in this paper is designed with intent of achieving classification accuracy and we achieved average accuracy 99.01% for five standard datasets. The information extracted from header part outperform for accurate classification of e-mails.

References

[1] ‘Phishing Activity Trends Report, 4th Quarter 2021’’ (2022). Phishing Working Group (APWG). Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2021.pdf.

[2] ‘Phishing Activity Trends Report,4th Quarter 2023’’ (2024). Phishing Working Group (APWG). Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf.

[3] A. Herzberg, (2009). "Combining Authentication, Reputation and Classification to Make Phishing Unprofitable," in Proc. IFIP Int. Inf. Security Conf. (SEC) (pp.13-24). Available at: https://doi.org/10.1007/978-3-642-01244-0_2.

[4] Giovane C. M. Moura and A. Pras (2009). "Scalable Detection and Isolation of Phishing," in Proc. 3rd Int. Conf. Autonomous Infrastructure, Management and Security (AIMS) (pp.195–198). Available at: https://doi.org/10.1007/978-3-642-02627-0_20.

[5] J. Hajgude and L. Ragha, (2012) "Phish mail guard: Phishing mail detection technique by using textual and URL analysis," World Congress on Information and Communication Technologies (pp. 297-302) Available at: https://doi.org/10.1109/WICT.2012.6409092.

[6] Kaur D. and Kalra S. (2016). Five-tier barrier anti-phishing scheme using hybrid approach, Information Security Journal: A Global Perspective, 25(4-6), 247-259, Available at: https://doi.org/10.1080/19393555.2016.1215573.

[7] Adwan Yasin and Abdelmunem Abuhasan (2016). An Intelligent Classification Model for Phishing E-mail Detection, International Journal of Network Security and Its Applications, 8(4), 55-72, Available at: https://doi.org/10.5121/ijnsa.2016.8405.

[8] Sheu J-J, Chu K-T, Li N-F, Lee C-C (2017) An efficient incremental learning mechanism for tracking concept drift in spam filtering, PLOS ONE, 12(2), e0171518, Available at: https://doi.org/10.1371/journal.pone.0171518.

[9] Hugo Gascon, Steffen Ullrich, Benjamin Stritter et al. (2018). Reading Between the Lines: Content-Agnostic Detection of Spear-Phishing E-mails, RAID 2018, 11050, 69-91.

[10] Ghogare, Pramod, Surwade, Ajay, and Patil, Manoj (2018). Effective E-mail Spam Filtering Using Origin Based Information, International Journal of Computer Sciences and Engineering, 6(11), 359-362, Available at: 10.26438/ijcse/v6i11.359362.

[11] Krause, T., Uetz, R., & Kretschmann, T. (2019). Recognizing Email Spam from Meta Data Only. 2019 IEEE Conference on Communications and Network Security (CNS), (pp. 178-186). IEEE, Available at: https://doi.org/10.1109/CNS.2019.8802827.

[12] Surwade, A.U. (2020). Phishing e-mail is an increasing menace, International Journal of Information Technology, 12(3), 611–617. Available at: https://doi.org/10.1007/s41870-019-00407-6 .

[13] Kulkarni Priti, Saini Jatinderkumar, and Acharya Haridas (2020). Effect of Header-based Features on Accuracy of Classifiers for Spam Email Classification, International Journal of Advanced Computer Science and Applications, 11(3), Available at: 10.14569/IJACSA.2020.0110350.

[14] Anchit Bijalwan (2020). Botnet Forensic Analysis Using Machine Learning, Hindawi's Journal of Security and Communication Networks, Volume 2020, 1-9, Available at: https://doi.org/10.1155/2020/9302318.

[15] Surwade, A. U. (2020). Blocking Phishing e-mail by extracting header information of e-mails. 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), (pp.151-155). IEEE, Available at: https://doi.org/10.1109/ICSIDEMPC49020.2020.9299596.

[16] Sallab, Ahmad, Rashwan, Mohsen (2012). E-mail classification using deep networks, Journal of Theoretical and Applied Information Technology, 37, Available at: http://www.jatit.org/volumes/Vol37No2/12Vol37No2.pdf

[17] W. W. Cohen (2015), "Enron Email Dataset", [Online] Available from https://www.cs.cmu.edu/~enron/ [Accessed on June 2024].

[18] Enron Data Set (2004), "Enron Email Dataset," [Online] Available from https://www.cs.cmu.edu/~enron/ [Accessed on July 2024].

[19] Vit Listik (2015), "Phishing corpus", [Online] Available from https://academictorrents.com/details/a77cda9a9d89a60dbdfbe581adf6e2df9197995a [Accessed on June 2024].

[20] Wang, D., Irani, D., & Pu, C. (2013). A study on evolution of email spam over fifteen years. 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, (pp. 1-10). IEEE.

[21 Untroubled (1998), "SPAM Archive", [Online] Available from http://untroubled.org/spam/ [Accessed on June 2024].

[22] J. D. Wilson (2018), "SPAMData", [Online] Available from https://github.com/jdwilson4/Intro-to-Machine-Learning/tree/master/Data/SPAMData [Accessed on June 2024].

[23]Apache SpamAssassin Project (2004), "SpamAssassin Public Corpus," [Online] Available from https://spamassassin.apache.org/old/publiccorpus/ [Accessed on June 2024].