Applied Information Technology And Computer Science Vol. 5 No. 1 (2024) 034-051

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

Classification of Spear-Phishing Email using Machine Learning Approach

Mohammad Akmal Afif Mohd Zuhdi[1], Isredza Rahmi A Hamid[1]*

[1]Fakulti Sains Komputer dan Teknologi Maklumat,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2024.05.01.003
Received 31 July 2023; Accepted 22 May 2024; Available online 30 August 2024

Abstract: The prevalence of spear phishing email attacks targeting organizations is on the rise, accompanied by an increasing diversity in the techniques employed within spear phishing emails. Although previous research has focused on identifying phishing emails based on their headers, bodies, or attachments, this study aims to tackle spear phishing email classification using a machine learning approach. The research will focus on content-based features rather than headers, bodies, or attachment. This research proposed a new content-based feature called Email_containPosition. The proposed spear phishing email classification model comprises seven phases: raw data acquisition, data pre-processing, feature extraction, information gain, n-fold cross-validation, classification algorithm selection, and model performance evaluation. For this experiment, content-based features extracted from the Enron dataset will be utilized. The model's effectiveness will be assessed using the Random Forest and Naïve Bayes classification algorithms, with evaluation metrics including AUC, precision, F1-score, and recall. Random Forest performed exceptionally well with an Area Under Curve (AUC) score of 0.996, F1-Score of 0.968, precision of 0.969, and recall of 0.967. Naïve Bayes achieved moderate results: AUC 0.742, F1-Score 0.701, precision 0.677, and recall 0.727.

Keywords: Spear Phishing, Email Classification, Machine Learning

1. Introduction

The rapid growth of internet users in recent years has fostered increased connectivity among individuals worldwide. However, along with this connectivity comes the pressing need to address the various risks that can compromise online safety. Phishing schemes have emerged as one such risk, affecting a larger population across different age groups.

Phishing involves the deceptive use of any communication channel to trick targets into revealing important information such as identity card numbers and bank information. Phishing schemes can be categorized into various types such as Short Messaging Service (SMS) phishing (also known as Smishing) and Voice Phishing (Vishing) where attackers exploit SMS or phone calls to deceive their

*Corresponding author: rahmi@uthm.edu.my

targets. Among these types, Spear Phishing stands out as a particularly insidious tactic, where attackers impersonate high-level employees to appear more credible. Spear phishing emails are crafted with careful consideration, often employing urgency and specific keywords, such as "transfer" and "urgent," to manipulate employees [1]. Moreover, these emails adopt convincing format similar to those used by legitimate government organizations.

There are several phishing email classification approaches such as Sender-Based [7] Image-Based[8] and Hyperlink Based[9]. However, despite extensive research in this field, the number of phishing attacks continues to rise. Attackers constantly modify their techniques to bypass anti-phishing measures, including the use of diverse vocabulary in spear-phishing emails. Additionally, most phishing email types exhibit considerable similarities, making it challenging for existing classification techniques to accurately distinguish between different types of phishing attacks. To address this issue, we propose a spear-phishing email classification model. The objectives of this research are:

- to identify spear-phishing email features,
- to design spear-phishing email classification model using machine learning approach
- to evaluate the spear phishing email classification model in terms of F1-Score, recall, AUC, and precision.

The rest of this paper is organized as follows. Section 2 provides an overview of related works in the field of phishing email classification. Section 3 outlines the methodology employed for classification. Subsequently, Section 4 presents the experimental setup and analysis. Finally, Section 5 concludes the research and outlines future directions for further exploration.

2. Related Work

This section discusses phishing, phisher, phishing lifecycle, and type of phishing attack. It also includes information regarding email structure, spear phishing, the motivation behind phishing, approaches for phishing detection, machine learning algorithm, and existing research on phishing classification approaches.

2.1 Phishing

Phishing is a form of social engineering attacks that targets the weakest links in security, which is human. It is a deceptive cyberattack technique where malicious entities attempt to trick users into divulging sensitive information, such as login credentials or financial details. These attacks typically involve impersonating a trustworthy entity, often through emails, websites, or messages, to exploit human vulnerabilities. Phishing is not a new concept because of its predecessor which is phone phreaking. Phone phreaking is a concept that has been around since 1950 [2]. The term "phishing" replaced "phone phreaking" after 1980 [3]. After this period, various incidents by phishing attacks have been documented. One of the incidents was the attack on the Department of Defense (DOD) in December of 1995. During this incident, the attackers achieved a remarkably high success rate of approximately 65 percent, with around 250,000 attempted attacks on DOD computers. Surprisingly, the public was only made aware of these attacks in 1997 when media publications issued warnings about the nature and severity of these phishing attacks.

One of the world's coalitions to combat phishing is Anti-Phishing Working Group (APWG), consist of various enforcement sectors that works together to fight phishing attacks. They produced various publications on the number of unique emails used for phishing, domains that are widely used for phishing, most targeted industries by phishing attacks, and finally reports on method used to conduct these phishing attacks.

2.2 Phisher

The term "phisher" refers to threat actors who employ phishing as a social engineering attacks. These criminals utilize phishing as a method to gather important or sensitive information from their victims, which can be used for impersonations or sold on the black market [7]. Phishers pose a significant threat due to their widespread use of various communication channels to execute their attacks, including fake websites, emails, letters, phone calls, and SMS.

2.3 Phishing Lifecycle

Phishers usually applied several types of approach for various types of attacks. However, phishers try to follow the same methodology to successfully employ their attack. This methodology can be simplified into this phishing lifecycle, which is planning, collecting and fraud[8].

2.3.1 Planning

During the planning phase, the phisher initiates the attack by identifying their target to obtain the desired information. The target must be deemed valuable by the phisher, considering the risk and cost-benefit analysis. Some phishers exploit demographic information such as age, gender, and social interests, which can be collected from social media or public websites. In the case of targeting organizations, phishers may gather information from news sites, reports, or employee social media profiles.

2.3.2 Collecting

In the collecting phase, phisher has already acquired sufficient information to deploy their bait. This bait often takes the form of an email directed at the victims, containing information relevant to them or aligned with the current trends or news. The email may include links that lead the victims to fake website or attachments containing unrecognized malware. Usually, the victims unknowingly enter their credentials information such as bank account, login credentials, and password without hesitation.

2.3.3 Fraud

Once the phisher has successfully obtained information from the victims, it is utilized for impersonate purposes to target additional victims. Moreover, the information may be sold on the black market. If the information obtained includes credentials for organizations or banks, monetary loss may occur.

2.4 Types of Phishing Attack

This section discusses various types of phishing attacks such as Short Messaging Service phishing, Voice phishing, and Spear phishing.

2.4.1 Short Messaging Service Phishing

Short Messaging Service (SMS) phishing, also known as Smishing, is a prevalent form of phishing that exploits SMS, a widely used communication method on mobile phones. This phishing attacks involves sending malicious messages containing harmful codes, malicious Uniform Resource Locator (URL) or even malicious Android Package Kit (APK) files [15].

2.4.2 Voice Phishing

Voice phishing or known as Vishing is a type of phishing attack on mobile devices that use Voice over IP (VOIP) technology to gain sensitive information, such as identifications or financial information. The attacker impersonates bank officials or other trusted entities to deceive victims into divulging sensitive information, thereby gaining unauthorized access to their online banking or other sensitive account [16].

2.4.3 Spear Phishing

Spear Phishing is a form of phishing that is highly feared, mainly because it targets specific individuals rather than entire organization or systems [11]. Attackers gather information about the specific person or information, either through data purchased on the black market or research. They then employ social engineering, psychology, or reverse social engineering techniques to tailor their attacks to the individual's characteristics and vulnerabilities.

2.5 Email Structure

Email is commonly utilized as a means of both formal and informal communication. Many organizations continue to rely on email due to its ease of setup and management, inherent to its nature. Email can be categorized into multiple components, as outlined in Table 1.

Table 1: Email Structure

Email Parts	Description
Subject	Title of the email. Title of the email may be forged.
Sender	Email address of the sender. The email address of the sender may appear to be associated with the company, but it is actually from a public email service.
Message / Body	Message for the receiver. Usually contain link that redirect victim to phishing website
Attachment	File attached to the email. The attachment is filename is scrambled and may contain malicious software.

2.6 hishing Detection Approach

This section explains some of the phishing detection methods that exist. Due to the increasing attacks of phishing attacks, there has been few phishing detections approach proposed by various studies based on sender- based [7], image-based [8], and hyperlink-based [9] [12].

2.6.1 Sender-Based Approach

Sender based approach is through detecting whether the sender is a real sender, and if the email is sent from the location stated in the email. Work by [7] identified phishing messages that target banks using Support Vector Machine (SVM) algorithm. They separated phishing and non-phishing email using rulesets. These rulesets are divided into three categories.

- Emails Accounts from Public Email Service – If the email accounts that is used to send the email are from a free public email service such as Google or Yahoo. This coincides with attackers' tendency to include bank names in the email used for phishing. To deceive the victims, attackers usually include the email address of the victims, not the full name used by the banks for communications.
- Sender Geographical Locations – Due to the wide use of the internet, the email sent by the attackers may come from a huge number of different places. While the email may claim to come from the bank in Asia may be from another part of India or vice versa. This approach tries to detect of the origin or domain in the email is
- Authorized Sender – This approach tries to detect if the message is truly from the bank they are said to originate from. This is done by using Sender Policy Framework to verify messages.

This detection approach can be used also on other kinds of phishing messages that target various online retailers and online payment companies. However, this approach is unable to classify phishing on companies that support communication between clients.

2.6.2 Image Based Approach

Image based approach is done by checking either the website is legitimate or fake through image on the website [8]. This is done by checking if the image or logo for the website is consistent with the content of the website. Three processes involved that are:

- Logo Extraction – This is done based on the QR code principle where the image or logo can be extracted from the website from the QR code on the website by using two-dimensional code.
- Logo Recognition – The algorithm will compare the identity of the image or logo with the website. The comparison is done by using CNN-based detection method.
- Logo Identification – Lastly, the domain name is identified by going through dataset in the database created for the research. If the domain is the correct one, then it is a verified image or logo.

Convolutional Neural Network (CNN) automatically learns features from input data which is particularly useful for image-based detection approach which it then uses to create pyramid hierarchy for constructing low-to-high level semantics feature pyramids.

2.6.3 Hyperlink Based Approach

The research proposed by [9] using hyperlinks information to distinguish phishing email to non-phishing email. These hyperlinks are used to generate structure and classes to the research. They used email from a major Australian Bank to create three different datasets and BoosTexter algorithm to create a model with high degree of classification accuracy. Moreover, this model can be enhanced by incorporating more prominent features through future research or utilizing the features identified in this study to facilitate the classification of other research findings. Work by [12] used hyperlink-based approach such as Generic_TLD, URL_Length, Having_Sub_Domain, and Having_Slash.

2.7 Machine Learning Algorithm

This section discusses the machine learning algorithm specifically on Random Forest and Naïve Bayesian.

2.7.1 Random Forest

Random Forest is an ensemble learning algorithm that try to combine multiple prediction from multiple models to gain better accuracy as shown in Figure 1 [10] [14]. This enhances the algorithm's ability to handle new or unseen datasets, as it improves generalization. Additionally, the algorithm can reduce variance and prevent overfitting by employing various techniques. One such technique is Bootstrap Aggregation, where the dataset is used to create multiple subsets for training the model. Each decision tree is trained on a randomly selected subset to ensure more robust results.

Equation 1 shows Random Forest formula where the majority vote is used from the class prediction for the final decision. Let Cˆb(x) present the class prediction of the bth random-forest tree. The equation for Cˆ(B)/rf (x) is obtained through a majority vote among the predictions {Cˆb(x)} for all B trees where,

$$C\hat{}\frac{B}{rf}\,(x) = \; majority\ vote\ \{C\hat{}b(x)\}\frac{B}{1} \qquad Eq.1$$

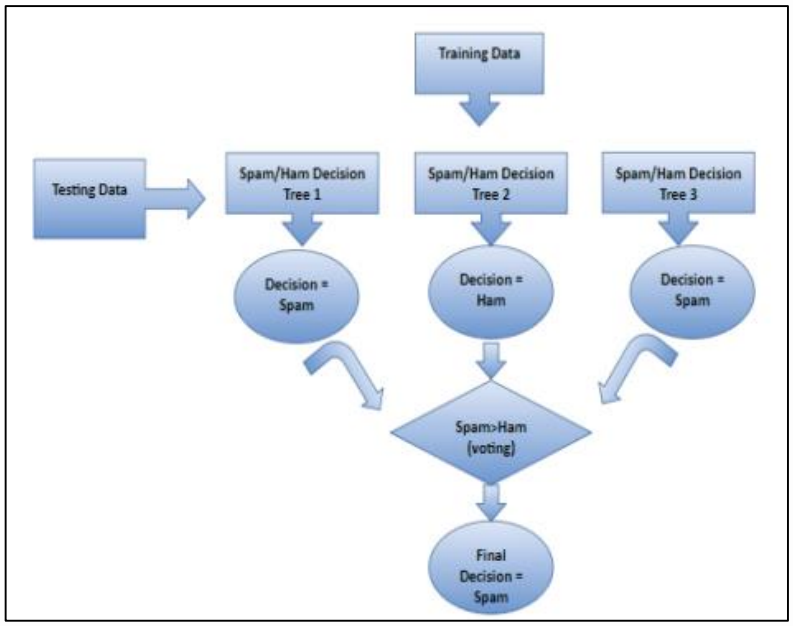

Figure 1: Random Forest [10]

2.7.2 Naïve Bayes

Naïve Bayes is a classification algorithm used in machine learning approach. The algorithm works by using probabilistic classifiers built on the principle of Bayes' theorem. The theorem is applied through conditional probability, if the features incorporated in the model hold equal importance and are independent of each other. This approach proves beneficial in text classification tasks, where the model features may follow a multinomial distribution. However, if the features in the model are continuous, a Gaussian distribution can be employed [10].

2.8 Existing Work on Phishing Classification Approach

Previous research has proposed various approaches for spear phishing classification, including hashing algorithm [6], deep learning [8][15] and machine learning techniques [12].

Work by [6] used Kaggle dataset which consist of 3000 emails. The research classifies spear phishing email using email attachment and content-based features. They used hashing algorithms such as MD5, TIGER, and SHA1 to find out if the file has not been tampered with by comparing the original file with the file during transfer. The algorithm used for this is Random Forest (RF) and Support Vector Machine (SVM). The model uses unsupervised machine learning topic mode in Latent Dirichlet Allocation from Gensim Package to generate dominant topic for its content based. This is different from the proposed research that will be using the content of the email only, not the attachment itself. Dadvandipour & Ganie employed Latent Dirichlet Allocation features and achieved a precision rate of 1.00 using SVM and RF algorithm [6].

Work by [18] applied content-based features to classify phishing and legitimate emails through supervised learning using Bayesian classifier. The dataset used is from Gmail that is divided into three datasets of 1000 emails, 1500 emails, and 2100 emails which are evaluated separately. Moreover, the work utilized spam words for frequency as features for classifications of text. However, the research does not specify the spam word features chosen in the report. This research uses Bayesian Classifier algorithm for data mining with accuracy, precision, and recall as the performance evaluation criteria. Rathod & Pattewar employed Bayesian Classifier and dataset 3 achieved the highest amount of accuracy with 0.964, precision of 0.95, and finally recall score of 0.87 [18].

Work by [11] applied deep learning approach to classify the phishing email. The research used dataset from VirusTotal consists of 32,676 emails. They consider using features extracted from email content and attachment. The first level is detection through email header because most of the metadata that is available in the email is in the header such as content-type, sender, and mime version of the email. Malicious email is usually direct where this can be deduced directly using the email header. The second level is the email body. This work features for content-based approach are Bidirectional Encoder Representation from Transformers (BERT) vectorization from a list of blacklisted keywords, URL, and phrases to find the context of the sentences in emails. Furthermore, this work uses XGBoost Meta-Learner as classifier algorithm and by using features extracted from header, body, and attachments. The research will be using Random Forest and Naïve Bayes algorithm for its machine learning. They obtained an AUC rate of 0.9921 incorporating body, header, and attachment-based features.

Work by [10] utilized two methods of classification were employed: machine learning and natural language processing. The research model consisted of two steps. The first step is classification using Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms on the Kaggle and Enron datasets. The results obtained from this step were then subjected to Uniform Resource Locator (URL) filtering. Step two is the URL filtering using three features: URL Blacklisting, Spam Words Triggering, Special Characters Identifications, and URL Filtering. They tested the email classification using the Naïve Bayes and Random Forest algorithms. Junnakar used Gensim Library features an achieved an accuracy rate of 0.97 through the implementation of Naïve Bayes, Decision Tree, RF, SVM.

Work by [13] utilized content and social-based features with dataset that is compiled by the researcher from 15 organizations such as government, NGO, IT, and Enron. These features are divided into subjects (Subject_IsReply, Subjec_hasBank), attachment (Length of attachment name, Atachment size (bytes)), body (Body_numUniqueWords, Body_numNewlines), and finally social (Location, numConnection). Moreover, several algorithms are used for classification methods which are Random Forest, J48 Decision Tree, Naïve Bayesian, Decision Table. Dewan [14] incorporated subject, body, attachment, and social features resulting in an accuracy rate of 0.97.

Table 2 shows the comparison of previous research. Our work differs from previous research in such a way that we apply features from various research and tested it with Random Forest and Naïve Bayes algorithm. The proposed research will also be using Enron dataset that has been used by this research [6] but with features from another research[6][13][17]. The research will focus on content-based features with the addition of a new feature which is position. The evaluation metrics will also be extended to include AUC score.

Table 2: Comparison of Existing Research

Work by	Method	Dataset	Features	Algorithm	Result
Dadvandipour & Ganie [6]	Machine Learning (Attachment and Content Based)	Kaggle	Used topic mode in Latent Dirichlet Allocation from Gensim Package	SVM and RF	SVM: Precision = 1.00 Recall = 0.76 F1-Score = 0.86 RFt: Precision = 1.00 Recall = 0.86 F1 Score = 0.98
Rathod and Pattewar [18]	Content Based Spam Detection in Email using Bayesian Classifier	Gmail	Content-Based Features	Bayesian Classifier	Accuracy = 0.964 Precision = 0.95 Recall = 0.87

Table 2:(cont)

Work by	Method	Dataset	Features	Algorithm	Result
P. Dewan et al [13]	Analyzing social and stylometric features to identify spear phishing emails (Content-Based and Social-Based)	15 Organizations (Government, NGO, IT, and Enron, et cetera)	Subject, Body, Attachment, Social	Random Forest, J48 Decision Tree, Naïve Bayesian, Decision Table	Random Forest: Accuracy = 97.39 FP Rate = 0.029 J48 Decision Tree: Accuracy = 95.84 FP Rate = 0.044 Naïve Bayesian: Accuracy = 54.14 FP Rate = 0.334 Decision Tree: Accuracy = 89.80 FP Rate = 0.090
Muralidharan & Nissim [11]	Deep-learning architecture (Header-based)	VirusTotal	Body, Header, and Attachment	BERT and CNN	BERT & CNN: AUC = 0.9921 F1-Score = 0.9413
Junnarkar et al [10]	Machine Learning and Natural Language Processing (Hyperlink-Based)	Kaggle and Enron	Gensim Library	NB, Decision Tree (DT), RF, and SVM	NB: Precision = 0.93 Recall = 0.85 F1-Score = 0.89 Accuracy = 0.9365 SVM: Precision = 0.73 Recall = 0.86 F1-Score = 0.79 Accuracy = 0.8955 DT: Precision = 0.91 Recall = 0.89 F1-Score = 0.90 Accuracy = 0.9470 RF: Precision = 0.97 Recall = 0.95 F1-Score = 0.96 Accuracy = 0.976

The results presented in Table 2 provides existing research findings for email spear phishing classsification. Dadvanpour and Genie [6] employed Latent Dirichlet Allocation features and achieved a precision rate of 1.00 using Support Vector Machine (SVM) and Random Forest (RF) algorithm. Muralidhran and Nissim [11] obtained an AUC rate of 0.9921 incorporating body, header, and attachment-based features. Junnakar et al[10] used Gensim Library features an achieved an accuracy rate of 0.97 through the implementation of Naïve Bayes, Decision Tree, RF, SVM. Dewan et al[13] incorporated subject, body, attachment, and social features resulting in an accuracy rate of 0.97.

3. Methodology

This section provides a detailed explanation of the Spear Phishing Classification Model using a machine learning approach.

3.1 Spear Phishing Email Classification Model

The Spear Phishing Classification Model consists of six phases: raw data, data pre-processing, feature extraction, information gain, 10-fold cross-validation, classification algorithm, and performance evaluation as illustrated in Figure 2.

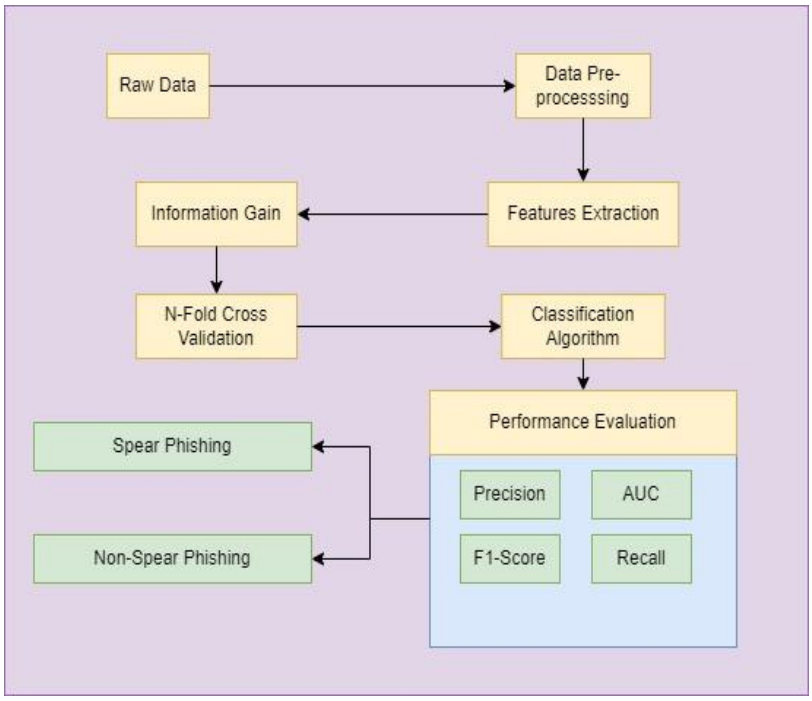

Figure 2: Spear Phishing Email Classification Model

3.1.1 Raw Data

Raw data refers to unfiltered or unprocessed data. In this study, we utilize raw data obtained from the Enron Dataset [6], which has also been utilized by other researchers[6][13][17]as shown in Figure 3 [10]. The dataset contains approximately 33599 emails from the company. While the dataset is publicly accessible, certain names and documents have been withheld to address sensitivity concerns. The dataset consists of message number, subject, and message. The subject only about the email and includes “re:” for reply while message column includes messages, forwarded by whom, time, “to:”, “cc:” and subject.

Message	Subject	Message
0	christmas tree farm pictures	
1	vastar resources , inc .	gary , production from the high island
2	calpine daily gas nomination	- calpine daily gas nomination 1 . doc
3	re : issue	fyi - see note below - already done .
4	meter 7268 nov allocation	fyi .
5	mcmullen gas for 11 / 99	jackie ,
6	meter 1517 - jan 1999	george ,
7	duns number changes	fyi
8	king ranch	there are two fields of gas that i am
9	re : entex transistion	thanks so much for the memo . i
10	entex transistion	the purpose of the email is to recap
11	lst rev dec . 1999 josey ranch nom	fyi
12	2 nd rev dec . 1999 josey ranch nom	- -
13	unify close schedule	the following is the close schedule for
14	meter 1431 - nov 1999	aimee ,

Figure 3: Example of raw data from Enron

3.1.2 Data Pre-processing

Data pre-processing involves filtering and transforming raw data before it can be used in the experiment. The dataset must be cleansed of symbol that is not accepted in the WEKA software, in addition to adjusting the spaces and delimiter for the dataset. Then, we will extract features from the content of the email.

3.1.3 Feature Extraction

The features extracted are based on features used by other researchers [6] [13] [17]. Features with * are features that is used by previous features. We extracted content-based features from Enron emails. Table 3 shows ten features used in our work. Spear phishing emails typically involve replies and contain words commonly used within a company, such as "verify your account", "suspension" and "transfer". These emails often include attachments. While phishing and spear phishing share most features, spear phishing emails possess distinctive characteristics, as outlined in Table 4. The features exclusively present in spear phishing emails are Email_containVerifyYourAccountr, Email_hasAttachment, Subject_isReply, higher number of words to characters ratio in Subject_richness and Body_richness, and a higher number of characters in Subject_num_Characters, Subject_num_words, Body_num_Characters, and Body_num_words.

Table 3: Feature list

Features	Description	Value
Email_containVerifyYour Account	The body of the email contain the word verify your account [13] *	Absent or present: R_1 = {0,1}
Email_hasAttachment	The email has attachment [6] *	Absent or present: R_3 = {0,1}
Email_containPosition	The body of the email contain position word with lowercase or uppercase: (Manage, Bos, HR, Administrator, Admi, Presiden, Vice Presiden, Director, General Manager, Senior Manager, Head of Department, Executive Directo, Managing Director, Partner, CEO, CFO, COO, CTO, CIO, CMO, CHRO, CLO, CSO)	Absent or present: R_4 = {0,1}
Subject_isReply	The subject of the email is a reply [13] *	Absent or present: R_5 = {0,1}
Subject_richness	The ration of characters to word in the email subject [13] *	Numbers: R_6 = {n}
Subject_num_Characters	The number of characters in the subject of the email [13] *	Numbers: R_7 = {n}
Subject_num_Words	The number of words in the subject of the email [13] *	Numbers: R_8 = {n}
Body_richness	The ratio of characters to word in the email body [17] *	Numbers: R_9 = {n}
Body_num_of_Words	The number of words in the email body [17] *	Numbers: R_{10} = {n}
Body_num_of_Character	The number of characters in the email body [17] *	Numbers:R_{11} = {n}

Table 4: Difference of feature for phishing and spear phishing

Features	Non-Spearphishing	Spear Phishing
Email_containVerifyYourAccount	0-1	0-1
Email_hasAttachment	0-1	0-1
Email_containPosition	0-1	0-1
Subject_isReply	0-1	0-1
Subject_richness	0.2 – 0.22	0.14 – 0.16
Subject_num_Characters	> 1000	1000 – 2000
Subject_num_Words	< 20	< 20
Body_richness	0.16-0.18	0.18-0.2
Body_num_of_Words	> 200	200 – 400
Body_num_of_Character	< 1000	1000 – 2000

Based on Table 4, spear phishing and non-spear phishing share a lot in common value in the features. However, they differ in values ranges for subject_richness, subject_num_characters, body_richness, body_num_of_words, and body_num_of_character features. For instance, in the case of subject_richness, spear phishing exhibits smaller value ranges from 0.14 to 0.16, whereas non-spear phishing ranges from 0.2 to 0.22.

3.1.4 Information Gain

Information gain is a technique used to rank features by using what is called decrease of entropy [2]. This decreased in value is caused by using different features used for the dataset. While entropy is an evaluation technique used for evaluating features by measuring disorder caused in spear phishing email [2].

3.1.5 N-Fold Cross Validation

Training and testing are part of machine learning process. Training is done by using a subset of the dataset to train the machine while testing is to test the machine performance by using the subset of data not used in the training process. This phase will be using N-Fold Cross Validation for training and testing. The reason why N-fold Cross Validation favoured is due to how it functions, this technique divides the dataset into certain number of "folds" which is a group of data that are divided with no bias. These folds will then be used to train the dataset leaving only one for testing. Then, the technique will rotate between the folds for training and testing in Orange to ensure there are no bias in the training and testing. The number of folds used is 10-fold. This is good to prevent overfitting or underfitting the predictions with reduced bias.

3.1.6 Classification Algorithm

A classification algorithm is utilized to determine the grouping of each data point in the dataset based on predetermined categories. In this experiment, Random Forest and Naïve Bayes classification algorithm were employed to access the performance of classification. Random Forest was chosen due to its utilization as an ensemble machine learning model, which employs multiple independent decision trees for more accurate predictions. This approach reduces data variance and mitigates overfitting of the dataset as demonstrated in Equation 2.

Let Cˆb(x) present the class prediction of the bth random-forest tree. The equation for Cˆ(B)/rf (x) is obtained through a majority vote among the predictions {Cˆb(x)} for all B trees where,

$$C\hat{}(B)/rf\ (x) = majority\ vote\ \{C\hat{}b(x)\}\ B/1 \qquad \text{Eq.2}$$

Naïve Bayes, on the other hand, is selected for its efficient handling of high-dimensional data points and faster training speed. It is considered comparable to advanced models in text classification, particularly when coupled with effective data pre-processing, as demonstrated in Equation 3. The calculation of interest involves the conditional probability of the observation, denoted as *P(x1, x2, …, xn | yi)* while the prior while the prior probability of the event is referred to as the marginal probability *P(yi)* where,

$$P(yi \mid x1, x2, \dots, xn) = P(x1, x2, \dots, xn \mid yi) * P(yi)/P(x1, x2, \dots, xn) \qquad \text{Eq.3}$$

3.2 Performance Evaluation Metrics

The performance evaluation metrics that will be used to evaluate the proposed model are:

- Precision: To identify how many spears phishing email can be classified correctly from all emails by using true positive (TP) and false positive (FP) as shown in Equation 4. Higher precision value indicated the machine low rate of false positive in classifying spear phishing email. Precision is calculated by True Positive (TP) divided by total of True Positive (TP) and False Positive (FP).

$$Precision = TP/(TP + FP) \quad \text{Eq.4}$$

- Recall: To identify how many spears phishing email can be classified from the whole email by using true positive and false negative (FN) as shown in Equation 5. Higher recall value indicates bigger proportion spear phishing email is classified correctly from all actual spear phishing email. Recall or True Positive Rate (TPR) is calculated by True Positive (TP) divided by total of True Positive (TP) and False Negative (FN).

$$TPR = TP/(TP + FN) \quad \text{Eq.5}$$

- F1 Score: To calculate the accuracy of the model by using both precision and recall. This is done to reduce inaccuracies if the model has class imbalance. The calculation involves precision, recall, true positive, false positive, and false negative as shown in Equation 6. Higher value indicates higher accuracy and more proportion of classifying spear phishing email. F1-Score is calculated by multiplication of Precision and Recall multiplied by 2, divided by total of Precision and Recall. Another formula for calculating F1-Score is by dividing True Positive (TP) with total of True Positive with 0.5 multiplied by total of False Positive (FP) and False Negative (FN).

$$F1 - SCORE = 2 * \frac{Precision*Recall}{Precision+Recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad \text{Eq.6}$$

- AUC: To calculate AUC by measuring Receiver Operating Characteristic (ROC) curve. The calculation involves True Positive Rate (TPR), False Positive Rate (FPR) as depicts in Equation 7. Higher value indicates model is better at classifying spear phishing to non-spear phishing email. AUC is calculated whereby $[TPR(i+1) + TPR(i)]/2$ represents the average of true positive between two consecutive threshold which is $(i\ and\ i+1)$ multiplied by $[FPR(i+1) - FPR(i)]$ that represents the difference in False Positive Rate between two consecutive threshold which is $(i\ and\ i+1)$. $\sum$ denotes the summation operation.

$$AUC = \sum\left(\frac{[TPR(i+1)+TPR(i)]}{2}\right) * (FPR[i+1+FPR[i]) \quad \text{Eq.7}$$

3.3 Hardware and Software Requirement

Table 5 presents the hardware requirement to conduct this work. The following are the specifications of the personal laptop Asus TUF A-15. The software used in this experiment is Orange, an open-source program utilized for mining data using diverse machine learning algorithms. Orange is constructed using the Python language.

Table 5: Hardware Specification

Hardware	Description			
Asus TUF A-15	Processor	System Type	Installed RAM	Windows Version
	AMD Ryzen 5 4600H Series	64-bit operating system	8.00GB	Windows 11 Home Single Language

4. Performance Analysis

This section describes on spear phishing and non-spear phishing and experimental setup using Random Forest and Naïve Bayes. We discuss on feature ranking, feature matrix construction, and result of the experiment.

4.1 Experimental Setup

The experiment was conducted using the Enron Dataset [13], as shown in Table 6. The dataset comprises 33599 emails, divided into two classes: 16428 spear phishing email and 17171 non-spear phishing email.

Table 6: Dataset

Dataset	Spear Phishing	Non-Spear Phishing	Total
Enron Dataset	16428	17171	33599

4.2 Feature Ranking

The features are selected by using Information Gain as a means of ranking the features. Information Gain is a technique used to calculate entropy to select which feature contribute the most to classification model. By determining the most relevant features, improvement can be made to the performance of the model. Table 7 show the ranking of all 10 features including feature position that is proposed in this research. The most relevant features is Email_hasAttachment while Email_containVerifyYourAccount is the least relevant feature.

Table 7: Ranking of Features

Features	Information Gain Value
Email_hasAttachment	0.071
Body_richness	0.049
Subject_num_Words	0.018
Subject_num_Characters	0.015
Subject_isReply	0.009
Body_num_of_Words	0.008
Body_num_of_Character	0.007
Subject_richness	0.002
Email_containPosition	0.001
Email_containVerifyYourAccount	0.000

4.3 Constructing Feature Matrix

We build the feature matrix of the 10 features $F_{i,}$ i = 1,….,10, i for all the spear phishing and non-spear phishing. Note that some features are binary while others are records. The features are summarized in Table 8.

Table 8: Feature Summarized

Features	Description	Value
F_1	Email_containVerifyYourAccount	$R_1 = \{0,1\}$
F_2	Email_hasAttachment	$R_3 = \{0,1\}$
F_3	Email_containPosition	$R_4 = \{0,1\}$
F_4	Subject_isReply	$R_5 = \{0,1\}$
F_5	Subject_richness	$R_6 = \{n\}$
F_6	Subject_num_Characters	$R_7 = \{n\}$
F_7	Subject_num_Words	$R_8 = \{n\}$
F_8	Body_richness	$R_9 = \{n\}$
F_9	Body_num_of_Words	$R_{10} = \{n\}$
F_{10}	Body_num_of_Character	$R_{11} = \{n\}$

Let $E = \{e_{1,} e_2, ……., e_{||E|}\}$ and $F = \{ f_1, f_2, ……., f_{|F|}\}$ denotes all the email and feature vector space respectively. With this, |E| is a total of emails and |F| refer to the number of feature vector. Let a_{ik} be the value of *k*th feature of *i*th emails. Due to that, the presentation of each email is $A_i = \{a_{i1}, a_{i2,} ……a_{i[e]}\}$, and each email $\{a_{ik}\}$ where i=1,2,……..,|F| and k=1,2,…..,|E|. Every email consists of A={Email_containVerifyYourAccount, Email_hasAttachment, Email_containPosition, Subject_isReply, Subject_richness, Subject_num_Characters, Subject_num_Words, Body_richness, Body_num_of_Words, Body_num_of_Character}. Finally, the dataset is run through Orange and tested using Random Forest and Naïve Bayes.

4.4 Result and Discussion

The experiment is tested on 33599 Enron dataset, which is then constructed two sets of experiment. Data for Experiment 1 consists of nine content-based features that are Email_containVerifyYourAccount, Email_hasAttachment, Subject_isReply, Subject_richness, Subject_num_Characters, Subject_num_Words, Body_richness, Body_num_of_Words and Body_num_of_Character. While data for experiment 2 contains ten features Email_containVerifyYourAccount, Email_hasAttachment, Subject_isReply, Subject_richness, Subject_num_Characters, Subject_num_Words, Body_richness, Body_num_of_Words, Body_num_of_Character and newly proposed features Email_containPosition. Then, we used 10-fold cross validation to train the Random Forest and Naïve Bayes algorithm.

Table 9 demonstrates the improvements achieved by incorporating the "Email_containPosition" feature in addition to the previous research features for both Random Forest and Naïve Bayes algorithms. For the Random Forest algorithm, there is an improvement in the F1-Score, with an increase of 0.003 points. Additionally, the Precision score sees 0.002 points increase, and the Recall score improves by 0.003 points. In the case of Naïve Bayes, there is a sole improvement observed in the Precision score, which sees an increase of 0.003 points. These results indicate that the inclusion of the "Email_containPosition" feature has a positive impact on the performance of the Random Forest algorithm, leading to enhanced classification accuracy. However, for Naïve Bayes, the improvement is limited to the Precision score only.

Table 9: Comparison between previous research with new feature

Experiment	Random Forest				Naïve Bayes			
	AUC	F1-Score	Precision	Recall	AUC	F1-Score	Precision	Recall
1	0.996	0.965	0.967	0.964	0.743	0.703	0.674	0.733
2	0.996	0.968	0.969	0.967	0.742	0.701	0.677	0.727

4.5 Area Under Curve (AUC) Result

Data in Figure 4 shows the AUC result for classification model of Enron Dataset tested on Random Forest (RF) and Naïve Bayes (NB) algorithm. The AUC result is 99.60% for RF and 74.20% NB. This shows that RF has a higher AUC rate than NB where RF is more accurate than NB in classifying spear phishing emails. Higher AUC also shows that the spear phishing email classification model perform better in terms of differentiate between spear phishing and non-spear phishing emails.

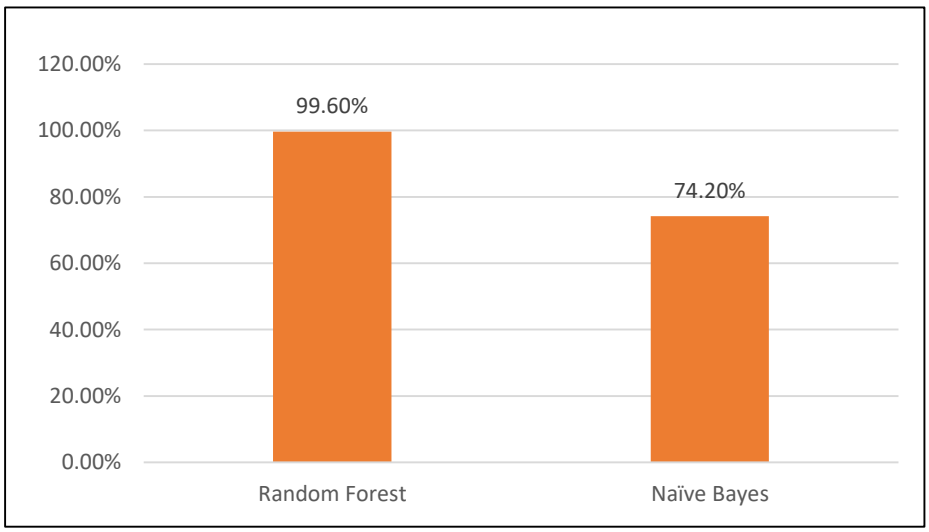

Figure 4: AUC Result

4.6 F1-Score Result

Figure 7 shows F1-Score for classification model for Enron Dataset tested on RF and NB algorithm. The results show that F1-Score for RF is 96.80% while NB is 70.10%. This means that RF is more accurate for classifying spear phishing emails than Naïve Bayes. Furthermore, this indicates that the spear phishing email classification model has higher performance using RF algorithm.

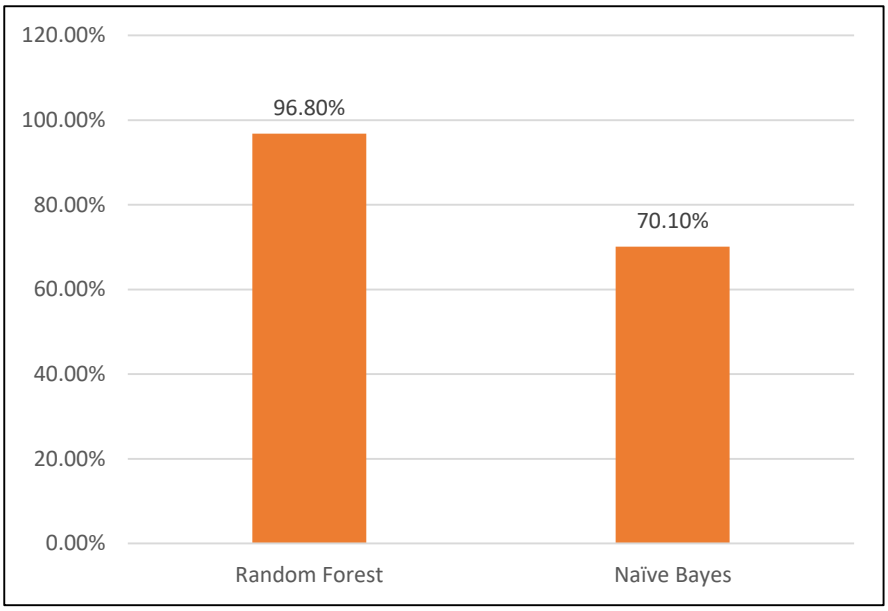

Figure 7: F1-Score Result

4.7 Precision Result

Figure 8 shows precision for classification model for Enron Dataset tested on RF and NB algorithm. The results show that RF has a higher percentage than NB which is 96.90% compared to 67.70%. This means that RF is more precise in classifying spear phishing email than Naïve Bayes. More precise means there are a smaller number of false positives is spear phishing email classification.

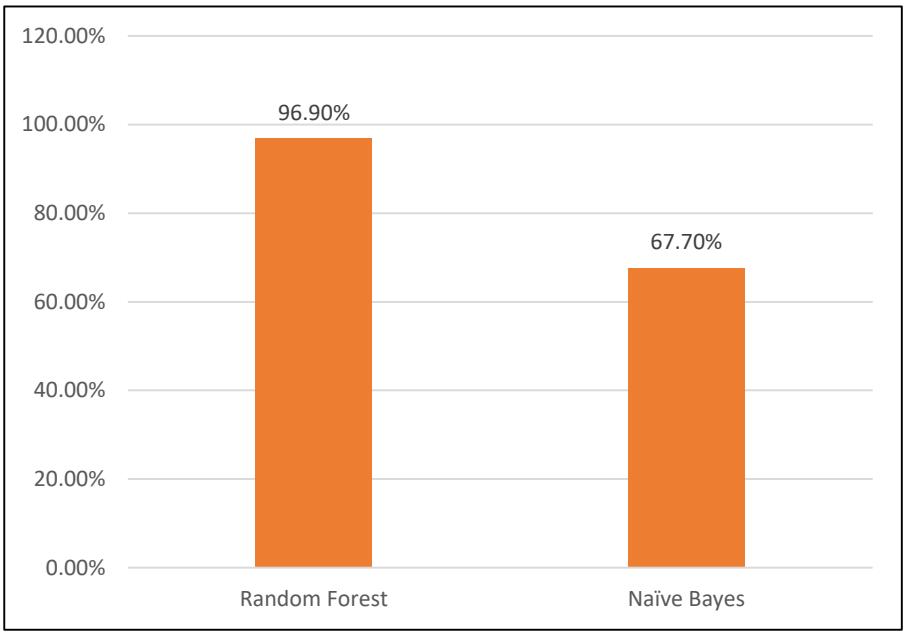

Figure 8: Precision Result

4.8 Recall Result

Data in Figure 9 shows Recall result for classification model of Enron Dataset that is tested by RF and NB algorithm. This shows RF with 96.70% recall percentage and Naïve Bayes with 72.70%. This shows that RF has a higher percentage than NB meaning that RF has higher amounts of positive predictions.

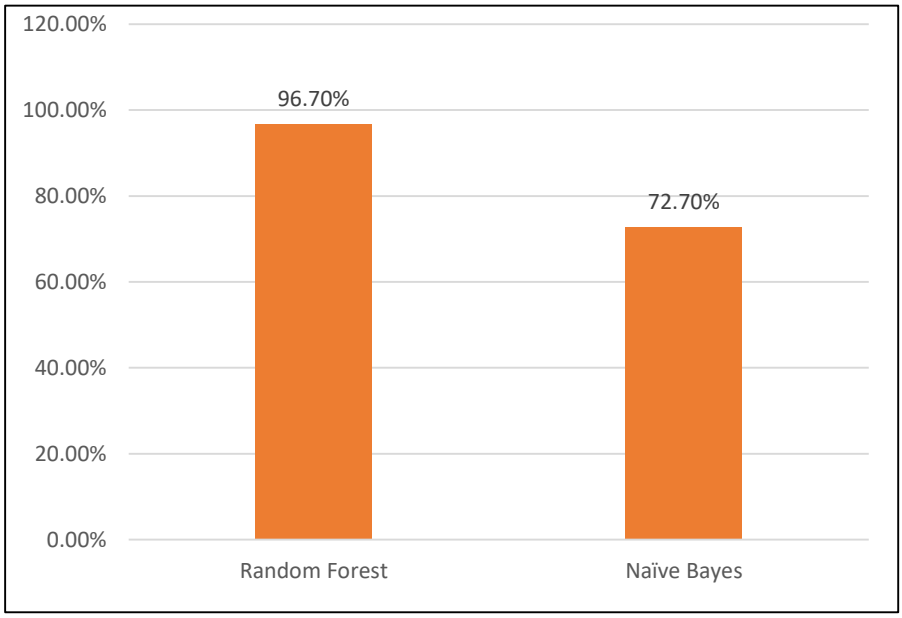

Figure 9: Recall Result

5. Conclusion

Spear phishing is a wide-spread threat in the field of cyber security where information leaks and financial loss is rampant. The attackers devised various ways for information and financial gain, by using emails as this is the main means of communication between the employees, companies, and organizations. For the conclusion of this research, spear phishing email classification using machine learning approach using Random Forest and Naïve Bayes algorithm achieved its objectives to identify spear phishing email features, to design a spear phishing email classification model, and finally to evaluate its performance using AUC, precision, F1-score, and recall. The proposed research used content-based features that consist of 10 features with 33599 emails containing both spear phishing and non-spear phishing. Features are then extracted with previous research and a newly proposed feature which is Email_containPosition. The features are then ranked using Information Gain to find the feature with high degree of relevance to the dataset. This model can contribute to improving the current spear phishing email classification solution. Two experiments are conducted where the first experiment uses nine features from the previous research and the second experiment include the newly proposed Email_containPosition. The model shows an improvement for Random Forest in the second experiment that combines previous research features with newly proposed features which is Email_containPosition where classification is improved with F1-Score seeing 0.003 points increase, Precision score seeing 0.002 points increase, Recall score seeing 0.003 points increase. As for Naïve Bayes in the second experiment, only Precision score seeing 0.003 points increase. With the end of this research, we intend to find and extract new features that can be used by other machine learning algorithms to gain an understanding of the performance of the proposed feature.

Acknowledgements

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support throughout the process of conducting this project.

References

[1] X. Ding, B. Liu, Z. Jiang, Q. Wang, and L. Xin, "Spear Phishing Emails Detection Based on Machine Learning," in 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), May 2021, pp. 354–359. doi: 10.1109/CSCWD49262.2021.9437758.

[2] B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: taxonomy of methods, current issues and future directions," Telecommunication Systems, vol. 67, no. 2, pp. 247–267, 2018.

[3] G. Ollmann, The phishing guide. Next Generation Security Software Limited, 2004.

[4] F. Quinkert, M. Degeling, J. Blythe, and T. Holz, "Be the Phisher – Understanding Users' Perception of Malicious Domains," in Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, Oct. 2020, pp. 263–276. doi: 10.1145/3320269.3384765.

[5] R. M. Mohammad, F. Thabtah, and L. McCluskey, Tutorial and critical analysis of phishing websites methods, vol. 17. Elsevier Ireland Ltd, 2015, pp. 1–24. doi: 10.1016/j.cosrev.2015.04.001.

[6] S. Dadvandipour and A. G. Ganie, "Analyzing and predicting spear-phishing using machine learning methods," Multidiszciplináris tudományok, vol. 10, no. 4, pp. 262–273, 2020, doi: 10.35925/j.multi.2020.4.30.

[7] F. Sanchez and Z. Duan, "A Sender-Centric Approach to Detecting Phishing Emails," in 2012 International Conference on Cyber Security, 2012, pp. 32–39. doi: 10.1109/CyberSecurity.2012.11.

[8] W. Yao, Y. Ding, and X. Li, "Deep Learning for Phishing Detection," in 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), 2018, pp. 645–650. doi: 10.1109/BDCloud.2018.00099.

[9] J. Yearwood, M. Mammadov, and A. Banerjee, "Profiling Phishing Emails Based on Hyperlink Information," in 2010 International Conference on Advances in Social Networks Analysis and Mining, 2010, pp. 120–127. doi: 10.1109/ASONAM.2010.56.

[10] A. Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar, and D. Karia, "E-mail spam classification via machine learning and natural language processing," in Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021, Feb. 2021, pp. 693–699. doi: 10.1109/ICICV50876.2021.9388530.

[11] T. Muralidharan and N. Nissim, "Improving malicious email detection through novel designated deep-learning architectures utilizing entire email," Neural Networks, Sep. 2022, doi: 10.1016/j.neunet.2022.09.002.

[12] N. A. Afandi and Isredza Rahmi A Hamid, "Covid-19 Phishing Detection Based on Hyperlink Using K-Nearest Neighbor (KNN) Algorithm", *aitcs*, vol. 2, no. 2, pp. 287–301, Nov. 2021.

[13] P. Dewan, A. Kashyap, and P. Kumaraguru, "Analyzing social and stylometric features to identify spear phishing emails," in eCrime Researchers Summit, eCrime, 2014, vol. 2014-January, pp. 1–13. doi: 10.1109/ECRIME.2014.6963160.

[14] APWG, "PHISHING ACTIVITY TRENDS REPORT 3rd Quarter 2022," 2022.

[15] S. Mishra and D. Soni, "Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis," Future Generation Computer Systems, vol. 108, pp. 803–815, Jul. 2020, doi: 10.1016/j.future.2020.03.021.

[16] H. Shahriar, T. Klintic, and V. Clincy, "Mobile Phishing Attacks and Mitigation Techniques," Journal of Information Security, vol. 6, no. 3, pp. 206–212, 2015, doi: 10.4236/jis.2015.63021.

[17] F. Toolan and J. Carthy, "Feature selection for Spam and Phishing detection, " *2010 ECrime Researchers Summit*, 2010, pp. 1–12. https://doi.org/10.1109/ecrime.2010.5706696

[18] S. B. Rathod and T. M. Pattewar, "Content based spam detection in email using Bayesian classifier," *2015 International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 1257–1261. https://doi.org/10.1109/ICCSP.2015.7322709