International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

# Email Fraud Detection

Amarnath Pathak[1],Tejeshwani Bharti[2]
Email- 2020aspire09@gmail.com , tejeshwani7@gmail.com
School of Computer Application
Lovely Professional University
9phagwara, punjab00

**Abstract**— Email deception detection is the procedure of recognizing and halting dishonest emails that are intended to pilfer individual or economic details or to distribute malicious software. The procedure usually comprises multiple tiers of separation to pinpoint and sift out dubious emails before they get to the receiver's inbox. One of the most prevalent abstractions exploited in email deception detection is the utilization of artificial intelligence algorithms to scrutinize the substance of emails and distinguish configurations that are linked with deceptive emails. These algorithms can be tutored on extensive data collection of recognized deceptive emails to aid them in detecting fresh occurrences of fraudulent emails with a significant level of precision

**KEYWORDS - Phishing detection, Machine learning, Data Models, Naive Bayes, Logistic Regression, Support Vector Method, KNN, Random Forest, Data Analysis, Classification, Accuracy, Precision, Recall.**

## INTRODUCTION

Spear-phishing is a lucrative form of deception when the offender dupes the recipients and obtains confidential information from them. Recipients of spear-phishing emails may be directed to open an attachment or tap on a hyperlink to a website where they must input sensitive details such as passwords and credit card numbers [1]. The deceiver sends the messages to a large number of users, and although only a small percentage of the recipients may fall for the scheme, the sender can still make a considerable amount of money from it.

With the persistent advancement of email usage and technology, the possibility of losing sensitive information to swindlers has risen. In this investigation, spear-phishing emails are recognized using machine learning techniques[2]. The suggested approach can be perceived as a categorization dilemma with dual categories:

genuine and phishing, with the aim of recognizing phishing emails. In the domain of synthetic intelligence recognized as automatic learning, the system is furnished with the competence to learn without being specifically programmed. Approaches for supervised machine learning are utilized for classification in our design. Using the recognized samples, supervised learning systems forecast the quality of unfamiliar information[23]. These approaches are a subgroup of those applied in automatic learning, which attains knowledge from information step by step.

## I. RELATED WORK

Andronicus et al. utilized a random forest machine learning classifier to classify phishing emails. Their aim was to reduce the number of required criteria while improving classification accuracy. We present a highly effective

method for identifying content-based phishing using a feed-forward neural network[7]. The results show a 98.72% accuracy in categorization, based on a dataset of over 7000 emails and other attributes [3]. The overall accuracy is 99.5%.

Gulches Park et al.'s objective was to identify reliable criteria for distinguishing between genuine and phishing emails. Despite using the same linguistic patterns as legitimate emails, phishing emails fail to differentiate between the subjects and objects of their target verbs[24].

"Email Phishing: An Open Threat to Everyone" discusses various phishing techniques and provides tips for avoiding falling prey to scams.

C. EmilinShyni et al. propose a method that integrates image processing, machine learning, and natural language processing. They utilize 61 distinct attributes in total and achieved a classification accuracy of over 96% using a multi- multi-classifier approach.

Entering the phrase "Detection Phishing Emails Using Features Decisive Values" produces a list of 18 characteristics. The suggested method categorizes each email based on the weighted attributes and the presence of flags. Their research demonstrates that the 18 recoverable features can provide good categorization accuracy when the appropriate attributes are selected.

## II.PROPOSED WORK

classifiers provide 99.99% accuracy, recall, and f-measure rates. The true positive and true negative rates are compared in Table 3. It demonstrates that SVM and Random Forest generate the highest true positive rates. As a result, it is evident that SVM.

## III.CLASSIFIERS Support

**Vector Machines** :

In classification and regression analysis, Support Vector Machines (SVMs), a type of machine learning algorithm, are used. By selecting the ideal border or hyperplane, SVMs may forecast continuous output variables and divide data into discrete groups. The core idea behind SVMs is that after input data is converted into a high-dimensional feature space, classes are split along a linear boundary[11]. By increasing the margin—the separation between the hyperplane and the nearest data points for each class—SVMs choose the best hyperplane. The position and direction of the hyperplane must be determined using the support vectors, or the data points nearest to the hyperplane.

### Naive Bayes

A probabilistic machine learning approach called Naive Bayes is utilized for categorization problems. It is based on the Bayes theorem, which asserts that the likelihood of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis, determines the probability of a hypothesis given the observed evidence. In the classification context, Naive Bayes determines the likelihood that a data instance will fall into a specific class given its feature values[18]. The method makes the naive assumption—hence the name Naive Bayes—that the features are independent of one another.

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

Based on the feature values of the data instance, Naive Bayes determines the probability that the data instance will belong to a specific class in the classification context. Utilizing the Naive Bayes approach.

The non-parametric machine learning technique K-Nearest Neighbors (KNN) is used to solve classification and regression issues. It is predicated on the idea that outcomes produced by comparable data points are typically equivalent[21].

Finding the k nearest neighbors of a given data point in the feature space and using their results to predict the results of the new data point are two of the basic concepts of KNN. The hyperparameter k can be altered using the cross-validation method.

KNN can be used to address classification and regression issues. In this instance, the great majority of the k nearest neighbors are used to classify the new data point.

### Random Forest

Random Forest is a supervised machine-learning technique for classification and regression issues. The construction of several decision trees, each trained using a random subset of the input variables and training data, is the fundamental concept behind Random Forest. The final forecast was made using an ensemble technique. To arrive at the ultimate prediction, the projections from each tree are combined[13].

With Random Forest, classification and regression issues can be solved. The majority of the predictions from the trees are utilized to categorize the data. A final forecast is obtained using regression and then averaged.

## IV. FEATURE OF PROJECT

*1) Locally Linear Embedding (LLE)*

The locally linear embedding (LLE) method of nonlinear dimensionality reduction is used in both manifold learning and data visualization [19]. Due to the fundamental idea, even when data points are restricted to a smaller geographic area, the local structure of the data points of a nonlinear manifold can be preserved[10].

The main objective of LLE is to find a set of weights that minimizes the reconstruction error while linearly rebuilding each data point from its neighbors. The weights are constrained by the requirement that the reconstructed data points exist in a lower-dimensional space.

LLE develops in two stages. It begins by identifying the k closest neighbors of each data point in the high-dimensional space [25]. In order to reduce rebuilding, a weight matrix is computed in the second stage. *2). Linear Discriminant Analysis (LDA)*

The supervised machine learning method of linear discriminant analysis (LDA) is used to solve classification problems [15]. It is predicated on the idea that data can be split into numerous classes using a linear boundary. The aim of LDA is to find a linear combination of input variables that maximizes class separation while minimizing variation within each class. By maximizing the difference between between-class variance and within-class variance, the linear combination is discovered.

By re-projecting the data onto the linear combination that divides the classes, LDA can also be used to reduce the number of classes. The resulting projection may be used to visualize data or further study.

LDA is nevertheless subject to a variety of restrictions. It is assumed that the data has a normal distribution and that the covariance matrices for the classes are identical [22]. If the classes aren't isolated from one another if the data structure is sophisticated, it might not function correctly. If there are too many input variables in relation to the number of observations, overfitting may also result.

One of the applications of LDA that is frequently utilized is credit scoring [20]. Additional uses include the use of speech, images, and medical diagnostics. It is a helpful tactic for dealing with classification issues when the classes can be distinguished clearly and there aren't many input variables.

## V.DATA SET

The collection consists of 1605 emails in total, 414 of which are spam and 1191 of which are fraudulent. The phished emails are a collection of emails from various sources, whereas the ham emails are collected from a publicly accessible dataset.

## VI.RESULT AND INTERFACE

Five classifiers are given portions of the extracted feature-based dataset, and the resulting classifications are recorded [16].

Using the 10fold cross-validation process, the initial data sample was split into a training set and a test set.

*Precision:*

The precision performance metric in machine learning is used to assess a classification model's accuracy [14]. The total number of accurate estimates from the model can be used to determine the proportion of true positive forecasts. The ratio of true positives to all true and false positives is another term for it.

Accuracy is defined in mathematics as the total of true positives and false positives.

$$Precision = \frac{TP}{TP + FP}$$

*Recall:*

Recall is a machine learning performance metric that is used to evaluate the efficacy of categorization models. It establishes
the proportion of accurate predictions to successful examples in the dataset. To put it another way, it is the proportion of real positives to the total of real positives plus real negatives.

Recall is described in formal mathematics as follows: The recall equation is True Positives / (False Negatives + True Positives)[12].

Situations that the model correctly identified as belonging to the positive category are considered positive.

$$Recall = \frac{TP}{TP + FN}$$

*F-measure:*

The F-measure, often known as the F1 score, is a harmonic mean of recall and precision. It is a single performance metric that adds the recall and accuracy of a classification model to provide a single score.

The F1 score is equal to 2 plus that much when the precision and recall values are combined.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

*True Negative Rate*:

A performance indicator called True Negative Rate (TNR), commonly referred to as specificity, measures how well a model predicts the negative class in binary classification tasks.

After accounting for all of the actual negative cases in the dataset, TNR is calculated as the percentage of true negative predictions[6]. Another approach to measure it is the ratio of true negatives to the sum of true negatives and false positives.

$$TN = \frac{N_h}{H}$$

*False Positive Rate*

False Positive Rate (FPR), a performance indicator for binary classification issues, gauges how well a model foretells the positive class.

The false positive ratio (FPR) is defined as the sum of all true negative cases to all false positive predictions (FPP)[5]. To put it another way, the ratio of false positives to the total of true negatives and false positives tells us how accurate a test is.

FPR is described mathematically as follows:

This formula is used to calculate the false positive rate or FPR:

$$FP = \frac{N_f}{H}$$

*True Positive Rate:*

The true positive rate (TPR), also known as sensitivity, is a measurement of how well a model predicts the positive class in binary classification tasks.

When compared to all of the actual positive cases in the dataset, TPR is the percentage of predictions that were correct. We define phrasing as the ratio of actual positives to the sum of real positives + real negatives in another way.

The following is the TPR mathematical definition:

TPR is calculated as the sum of true positives and false negatives.

$TP = \frac{N_p}{P}$

## DATA VISUALIZATION

Data visualization is the process of representing data graphically to gain insights and communicate findings to others[8]. It involves creating visual representations of data using charts, graphs, maps, and other visual tools to help identify patterns, relationships, and trends.

Effective data visualization is important because it can help to make complex data more understandable and accessible to a wider audience. It can also help to identify outliers and anomalies in the data, as well as highlight important features and trends.

The F-measure, true negative, true positive, and false positive are examples of assessment metrics[9].

We discovered that the random forest approach outscored all others with an F1 score of 0.94 based on our examination of the email fraud dataset. Furthermore, we demonstrated the importance of using cross-validation techniques, such as K-fold crossvalidation, to make sure that our models generalize well to new data.

Our research suggests that machine learning could be a useful tool for identifying email fraud, but it is important to carefully assess the

Our research suggests that machine learning could be a useful tool for identifying email fraud, but it is important to carefully assess the

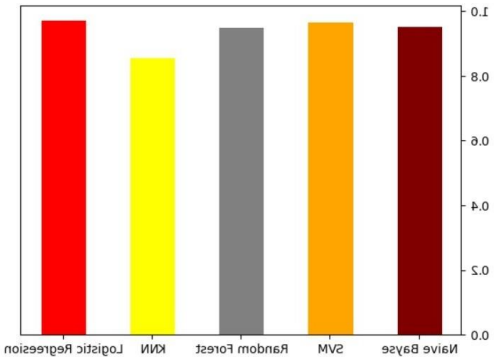
**Figure 2: Classification Accuracy of 5 ML Classifiers**

**Table 2: Comparison of Accuracy**

| Classifier | Accuracy (%) |
|---|---|
| SVM | 0.96 |
| Random Forest | 0.93 |
| Logistic | 0.96 |
| Naive Bayes | 0.94 |

## CONCLUSION

Due to the potential harm that phishing and other email scams may bring to money and reputation, email fraud detection is a




crucial topic of research in the field of cybersecurity.

SVM, logistic regression, Naive Bayes, random forest, KNN, and LDA were some of the machine learning techniques examined in this study for email fraud detection[17]. We talked a lot about choosing a model, creating features, and evaluating it with metrics like true negative, true positive, false positive.

### Result

We discovered that the **Scalable Vector method** that is (SVM) approach outscored all others with an accurate score of 1.0 which is based on our examination of the email fraud dataset. We also demonstrated the value of using cross validation techniques[4].

## References:

[1] Savaliya, Bansi R., and C. George Philip. "Email fraud detection by identifying email sender." *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. IEEE, 2017.

[2] Kumar, Koushal, and Bhagwati Prasad Pande. "Applications of machine learning techniques in the realm of cybersecurity." *Cyber Security and Digital Forensics* (2022): 295-315.

[3] Shane, Scott, and Daniel Cable. "Network ties, reputation, and the financing of new ventures." *Management Science* 48.3 (2002): 364-381.

[4] Browne, Michael W. "Cross-validation methods." *Journal of mathematical psychology* 44.1 (2000): 108-132.

[5] Bose, Prosenjit, et al. "On the false-positive rate of Bloom filters." *Information Processing Letters* 108.4 (2008):210-213.

[6] Ralph M. Richart wrote "Evaluation of the true false negative rate in cytology." American Journal of Obstetrics and gynecology 89.6 (1964): 723–726

[7] Schmidt, Paige M., et al. "Evaluation of euthanasia and trap–neuter–return (TNR) programs in managing free-roaming cat populations." *Wildlife Research* 36.2 (2009): 117-125.

[8] Schmidt, Paige M., et al. "Evaluation of euthanasia and trap–neuter–return (TNR) programs in managing free-roaming cat populations." *Wildlife Research* 36.2 (2009): 117-125.

[9] Post, Frits H., Gregory Nielson, and Georges-Pierre Bonneau, eds. "Data visualization: The state of the art." (2002).

[10] Chen, Tsong Yueh, Fei-Ching Kuo, and Robert Merkel. "On the statistical properties of the f-measure." *Fourth International Conference on Quality Software, 2004. QSIC 2004. Proceedings.* IEEE, 2004.

[11] Chang, Hong, and Dit-Yan Yeung. "Robust locally linear embedding." *Pattern recognition* 39.6 (2006): 1053-1065.

[12] Conte, S. D. "The numerical solution of linear boundary value problems." Siam Review8.3 (1966): 309-321

[13] Browne, Michael W. "Cross-validation methods." Journal of mathematical psychology 44.1 (2000): 108-132.

[14] Huicho, Luis, Miguel Campos-Sanchez, and Carlos Alamo. "Meta-analysis of urine screening tests for determining the risk of urinary tract infection in children: CME REVIEW ARTICLE." The Pediatric infectious disease journal 21.1 (2002): 1-