International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Identification and Handling of Phishing Email using Natural Language Processing

Nandam Sri Mouli

Student, Rajam, Vizianagaram, 532127, India.

ABSTRACT

Email is preferred by both individuals and businesses as one of the most important forms of communication. Despite the popularity of alternatives like electronic communications, mobile applications, and social networks, email usage and importance have been raising continuously. Phishing is the fraudulent practice of stealing any person's confidential information, such as financial details, login information etc. In which they look completely authentic and genuine and extremely hard for victim to detect whether the mail is spam or not. Moreover, the number of phishing emails is growing daily. So, handling of phishing emails is a crucial research issue. NLP techniques are useful for training data to detect spam content. In this work we studied leading research areas in phishing email detection and features used in spam emails. Feature extraction and selection plays a key role in phishing email detection. This paper also explains how actually the phishing attack happens and obtained results including information about concepts of natural language processing and working procedure.

Keywords: *Phishing Emails, Natural Language Processing(NLP), Machine learning, Email classification,Spam Detection.*

1. Introduction

Phishing is a technique where an attacker develops a false page or fake website that appears to be entirely real but is actually fraudulent. The same is used by the attacker to force users to input their credentials. These days, emails are the primary method used for this. Many misleading websites are available and are utilized by phishers to trick users into opening fraudulent emails in order to steal their personal information such as preparing type of harmful links and also some pop-up's so that the recipient would receive might becomes a phishing email victim. It is a type of trickery in which the attacker appears as someone else and using communication channels to launch an assault while posing as a real person. In present situation we can observe many cybercrimes are happening around the world. In which doing cybercrimes with the help of email phishing is severe issue. NLP techniques are very helpful in identifying the spam content in which, these techniques are crucial in finding the risky content such as phishing emails. This paper provides a deep review of various studies and approaches which are related to machine learning and natural language processing for phishing email detection.

Below steps demonstrates how email phishing works in seven essential steps:

1. Compromising the web server : The attacker's initial move is to access the web server. This is possible with a variety of methods and technologies.
2. Sending the fraud email : The attacker then transfers the email with a harmful link, information, or even fraudulent emails that ask the victim or recipient for personal information.
3. Mail at Recipient : The recipient clicks on a link without realising the email is fake.
4. Access website: The recipient is taken to the fraudulent website after clicking the link.
5. Harmful website appears : The criminal then requests information from the victim by sending them to a false and harmful website.
6. Providing information : The recipient now inputs the requested information while not realising the website is fake and falls victim to mail phishing.
7. Utilize information: After obtaining required information from the victim, the criminal may use that knowledge improperly or even threaten the victim.

2. Literature Survey

In paper[1],G. Mujtaba, L. Shuib, RP. Verma, A. Goyal, and Y. Gigras proposed a system using NLP and machine learning classifiers. This article begins by providing a concise summary of the phishing lifecycle. Where it outlines each phase in the phishing process. It also illustrates how a phishing email operates. This paper also discusses the classification of phishing scams.This research employs machine learning techniques to primarily classify phishing emails. Machine learning classifiers and natural language processing were both employed for classification. Using NLP principles, scikit-learn, and NLTK, text categorization and analysis of phishing datasets has been carried out. There are several different classifiers utilised, including SVC, Decision Tree, and Random Forest K Neighbors Classifiers.

In paper[2], S. Salloum, T. Gaber, S. Vadera and K. Shaalan this study focuses on the most popular NLP methods, including word embeddings and SVMs. We can categorise the emails into genuine and phishing ones using feature selection and feature extraction. Among the supervised machine learning (ML) methods for phishing email detection are Support Vector Machines, Naive Bayes, Decision trees, and Random Forest. K-means clustering is what we utilize for unsupervised. This method divides the dataset into k groups, and K datapoints are chosen at random from each category. Chi-square, Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), etc. These methods are beneficial for feature extraction.

In paper[3], G. Raj, N. Majeed and M. A. Al-Garadi favoured the five e-mail categorization application areas. The indicated application domains contain the most common data sets, feature sets, classification techniques, and performance metrics. The focus is made on various characteristics are the most often used features in email categorization. This paper has a very great work in which many questions regarding research in phishing email are answered.

In paper[4], Borg, A., Boldt, M., Rosander, O. This study intends to explore how classifying incoming customer support emails using an automated machine learning-based classifier can improve classification performance. Sequential and non-sequential models will be the focus of two branches of investigations. they used 10-times 10-fold cross-validation for non-sequential models in order to test them. The F1- score and the Jaccard index will serve as indicators for these models . Using the selected evaluation measure, the trials on the sequential models will determine which combination of corpus, text representation, and LSTM hyperparameters provides the best classification performance.

In paper[5], Srinivasarao, U., & Sharaff, A. This method for extracting the sentiment from email patterns presents a classification-based clustering methodology. Python platform is used to implement email-based categorization and clustering. The available massive dataset is sorted into interested, disinterested, and individual email patterns throughout the categorization procedure. Typically, a variety of strategies are used to classify emails based on spam, however in this method, LDA and SVR classifiers are used to achieve a classification based on patterns. There are three distinct classes that have been identified: interested, disinterested, and individual email patterns.

3. Data Collection

The below data is collected from the references.

4. Methodology

Researchers have put a lot of effort into studying various methods for email categorization, detection, and prevention. We concentrate on categorising phishing emails using machine learning methods. For categorization, 5,574 labelled as spam or ham , actual, and unencoded English messages from the "The Short message service Spam Collection " dataset were utilised . Machine learning classifiers and natural language processing were both employed for classification. Using NLP principles, scikit-learn, and NLTK, text categorization and analysis of phishing datasets has been carried out. Different classifiers are employed, including SVC, Decision Trees, and Random Forest KNeighbors Classifiers.

Natural language processing(NLP) is described as an area of artificial intelligence that enables computers to connect with people. Stop words, punctuation, special characters, tokenization, stemming, tagging, language detection, and semantic connection identification are all examples of fundamental NLP activities.It is sometimes described as a way for software to automatically handle natural language. Stop word elimination, tokenization, part-of-speech tagging, stemming, punctuation, special characters, language detection, and semantic connection identification are examples of basic NLP activities. Scikit-learn also defines many clustering, regression, and classification techniques. There are several supervised and unsupervised learning methods available in this Python library . It is based on a number of technologies that you may already be acquainted with, including NumPy, pandas, and Matplotlib. An important tool NLTK is also hailed as a fantastic toolkit for interacting with natural language. By creating Python applications on this platform, it is applicable to work with data that is presented in human language. NLTK has very adaptive and easy to understand interfaces to more than 50 corpora and lexical resources, as well as a number of text processing packages.

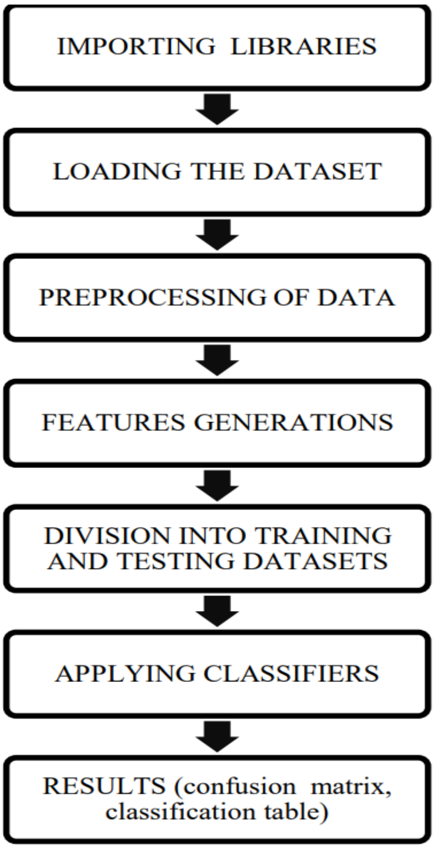

Figure 1: working procedure

I. **Import Required Libraries:** The first action is to import the required libraries. All the libraries which are required need to be imported.

II. **Load the Dataset**: The dataset has been read using Python pandas

III. **Preprocessing of Datasets**: Preprocessing datasets is the initial stage in the classification process. This involves changing the entire text to lower case, eliminating punctuation, digits, and site addresses, as well as stop words, tokenization, and stemming.

IV. **Features generation:** A crucial phase of categorization. Using domain expertise, it is useful to obtain features presented in the dataset, and machine learning algorithms will exploit those features. The features are presented as tokens created in the previous stage. From these features, a feature set is produced that includes the most prevalent features. The feature set may include characteristics that are meaningless or have very short durations; these features should be deleted for better outcomes.

V. **Creation of Datasets for the Model's Training and Testing:** To create training and testing datasets, the feature set is distributed according to whatever ratio we deem appropriate. The classifiers or models are trained using the training dataset. While the generated model is validated and the results are computed using a testing dataset.

VI. **Applying Classifiers:** Importing the classifiers and methods that will be utilized for classification is necessary. There have been used a number of sklearn classifiers.

VII. **Results:** Compiling a classification report and confusion matrix, as well as calculating the accuracy rate of each classifier.

The classifiers used in this methodology are:

1. K Nearest Neighbors :

The KNN is a widely used supervised learning algorithm that often aids in categorization. Here, it is assumed that related elements remain near to one another. To determine the degree of similarity, similarity measurements are used, most frequently the Euclidean distance. Since tuning parameters and model parameters are not built, KNN implementation is simple. Since the KNN is a non-parametric method, no fundamental assumptions about the distribution of the data are necessary. The larger and more dimensional the dataset, the slower the algorithm will operate.

2. Decision Tree:

A supervised learning method called a decision tree may be used to solve classification and regression problems, however it is typically used to address classification issues. It is known as a decision tree because, like a tree, it begins with a root node and grows on subsequent branches to form a tree-like structure.By using a function to determine how much each split in accuracy will cost us. This procedure is recursive in nature since the groups produced will be subdivided using the same technique. The split that costs the least is picked. Due to this method, this algorithm is also known as the Greedy Algorithm since we have a greater motivation to decrease the price. The foundation node becomes the best predictor/classifier as a result.

3. Logistic regression :

The efficient supervised machine learning technique logistic regression is used for binary classification problems. The probabilities used to describe the likely results of a particular test are modelled using a logistic function. In essence, logistic regression uses a logistic function to classify the data and remove outliers. Regression using a multinomial logistic model may represent situations with more than two discrete possible outcomes. If the training data are learnt too carefully, a model cannot match fresh data or forecast observations that have not yet happened. The phrase for this circumstance is overfitting. The logistic function will be used to reduce overfitting.

4. Random Forest :

A popular machine learning method that uses supervised learning is called Random Forest. It may be used to solve classification and regression issues in ML. It has been shown to support the idea of ensemble learning, which is the act of combining several classifiers in order to solve complex problems and improve model performance. A classifier called Random Forest may be used to increase the predicted accuracy of a dataset by combining many decision trees on different subsets of the provided information. The more trees there are in the forest, the better the accuracy and the less overfitting.

5. Naive Bayes:

The Bayes Theorem is that the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilised for a spread of classification problems. there's less training data needed. It manages data that's continuous and discrete. With reference to the quantity of predictors and data points, it's quite scalable. it's quick and may be utilised to make predictions in the present.

6. SVM:

One of the finest supervised learning methods for classification and regression issues is the support vector machine (SVM). However, it is mostly employed in ML to solve classification issues. The SVM method seeks to construct the optimum line that can classify an n-dimensional space. A hyper plane is the name of this finest line. The new data point will be placed in the appropriate category using this hyper plane. SVM uses the extreme points and vectors to help build the hyperplane. Support vectors are the term for these extreme examples, and the SVM technique is named after them. Due to the positive and negative hyperplanes, the method is more precise.

5. Results and Discussion

Based on the studies done all classifiers that are used perform well in identifying phishing emails where KNN will give low accuracy and Naïve Bayes and SVM will provide good accuracy.

The confusion matrix is used to summarise how well the classification algorithms work. The confusion matrix demonstrates how our model becomes perplexed when making predictions. It provides a clearer picture of the kinds of errors that our model is committing.

We will obtain Classification report shows the model's accuracy, recall, F1 and support scores The proportion of TP to that of TP and FP is referred to as precision. It indicates what percentage of all favourably labelled situations are in fact accurate. Recall is the classifier's capacity to identify all of the favourable cases. It indicates the proportion of correctly identified events that were genuinely positive. It is the proportion of TP to TP plus FN. The mean of recall and accuracy is used to determine the F1-score. All these evaluation metrics are useful in making classification report.

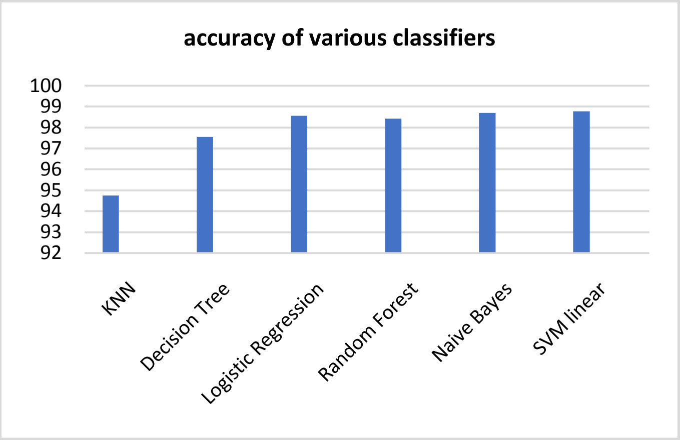

Figure 3. Showing calculated accuracy rates

	Precision	**Recall**	**F1-Score**	**Support**
0(ham)	0.99	1.00	0.99	1208
1(spam)	0.99	0.93	0.96	185
Micro average	0.99	0.99	0.99	1393
Macro average	0.99	0.96	0.98	1393
Weighted average	0.99	0.99	0.99	1393

Table 1. Classification report

Actual/predicted	***ham***	***spam***
ham	*1206*	*2*
spam	*13*	*172*

Table 2. Confusion matrix

6. Conclusion

Email phishing is the practice of deceiving a company's or another entity's mail recipient in order to get sensitive personal information or any confidential information by sending bogus emails and leading the recipient to believe that they were sent from a reliable source. In order to combat such a serious problem, user education and awareness are essential. This article provides a thorough explanation of how phishing emails are categorised using principles from natural language processing. The prediction matrix and classification report are very helpful and the accuracy rates of the classifiers that are provided in this paper will perform very well. Therefore we can predict very precisely whether the mail is legitimate one or not. So, the handling of phishing email will be done by recognizing them in spam folder. The research potential in this field is enormous. Future work will involve classifying and grouping large, raw, unstructured datasets.

REFERENCES

[1] G. Mujtaba, L. Shuib, RP. Verma, A. Goyal, and Y. Gigras, ‘‘Email phishing: Text classification using natural language processing,’’ Comput. Sci. Inf. Technol., vol. 1, no. 1, pp. 1–12, May 2020, doi: 10.11591/csit.v1i1.p1-12.

[2] S. Salloum, T. Gaber, S. Vadera and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," in *IEEE Access*, vol. 10, pp. 65703-65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

[3] G. Raj, N. Majeed and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," in *IEEE Access*, vol. 5, pp. 9044-9064, 2017, doi: 10.1109/ACCESS.2017.2702187.

[4] Borg, A., Boldt, M., Rosander, O. *et al.* E-mail classification with machine learning and word embeddings for improved customer support. *Neural Comput & Applic* **33,** 1881–1902 (2021). https://doi.org/10.1007/s00521-020-05058-4

[5] Srinivasarao, U., & Sharaff, A. (2021). Sentiment analysis from email pattern using feature selection algorithm. Expert Systems, e12867. https://doi.org/10.1111/exsy.12867

[6] Nikhil Govil, Kunal Agarwal, Ashi Bansal, Astha Varshney. “A Machine Learning based Spam Detection Mechanism”, IEEE Xplore Part Number: CFP20K25-ART; ISBN:978-1-7281-4889-2.

[7] Chode Abhinav, K Jayachandra, Kommu Pranith Kumar, V Sowmya “Spam Mail Detection using Machine Learning”, International journal for research in applied science & engineering technology.

[8] Manoj Sethi, Sumesha Chandra, Vinayak Chaudhary, Yash, “Email Spam Detection using Machine Learning”, International Research Journal of Engineering and Technology (IRJET).

[9] Nikhil Kumar, Sanket Sonowal, Nishant, “Email Spam Detection Using Machine Learning Algorithms”, IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.

[10] S. Suryawanshi, A. Goswami and P. Patil, "Email Spam Detection : An Empirical Comparative Study of Different ML and Ensemble Classifiers," 2019 IEEE 9th International Conference on Advanced Computing (IACC), 2019, pp. 69-74, doi: 10.1109/IACC48062.2019.8971582..