E-mail Spam Detection and Phishing link Detection using Machine Learning

[1]Keerthika J, [2]Adisvara A, [3]Akash S, [4]Jayanesh B, and [5]and Arul Prakash T
[1,2,3,4,5]Department of Computer Science and Engineering,Sri Eshwar College of Engineering,Coimbatore,India.
[1]keerthika.jcse@sece.ac.in, [2]adishdmc5@gmail.com, [3]akash.s2019cse@sece.ac.in, [4]jayanesh.b2019cse@sece.ac.in, [5]arulprakash.t2019cse@sece.ac.in

Article Info
A. Haldorai et al. (eds.),*2nd International Conference on Materials Science and Sustainable Manufacturing Technology,*
Advances in Computational Intelligence in Materials Science.
Doi:https://doi.org/10.53759/acims/978-9914-9946-9-8_9

Abstract-Phishing, which tricks individuals into revealing delicatedata like login credentials and financial details, is the most widespread type of cybercrime. Attackers typically use electronic mail, prompt messaging, and telephone calls to initiate these attacks. Despite ongoing efforts to prevent phishing attacks, current measures are not entirely effective, as the amount of phishing emails has enlarged significantly in current years. While numerous methods have been developed to filter out phishing emails, there is still a need for a comprehensive solution. This survey is the first of its kind to examine the use of N-L-P and ML methods for identifying phishing electronic mail. The analyzesof state_of_the_art N-L-P approaches that are presently being used to detect phishing electronic mail at different periods of the outbreak, with a focus on M-L methods. These methods are compared and evaluated in-depth.

Keywords– Spam, E-mail, Phishing, Machine Learning

I. INTRODUCTION

Phishing

The rapid advancement of internet technology has significantly transformed workinghandler interactions while also generatingextraunadornedsafety issues. The current developingpressures not individual affect the worker'sCPU but can also snip their uniqueness and currency. Phishing terrorizations not only rely on the development of technology but also utilize communal engineering to obtain the victim's individuality and account data. It is crucial to reduce the risk and illegal activities related with phishing. According to the Q3 2020 report unconfined by the Anti-Phishing_Working_Group (APWG), the amount of detected phishing emails increased significantly, from 44,497 in Q2 2020 and 44,008 in Q1 2020 to 128,926. Phishing bouts using the topic of coronavirus sickness 2019 (COVID-19) have been occurring later mid-September of the subsequenttime. The phishing emails often use subject lines such as net and safetyknowledges or COVID-19 information to lure their targets. As per the data, phishing has increased exponentially, causing significant harm.

Description of Phishing

Phishing is a well-known term that has received widespread attention from scientific journals, news outlets, and various organizations, such as banks and law enforcement agencies. The question that arises, however, is what phishing exactly entails. Some publications explicitly define the term, while others use examples or assume that readers are already familiar with the concept. As a result, there are various definitions proposed by scholars in the scientific literature, leading to a wide range of interpretations of phishing. Due to its broad nature, the prose does not deliver a precise explanation of phishing bouts, which can encompass different scenarios. For instance, PhishTankdescribes phishing as "a fakeeffort, typicallycompletedoverelectronic mail, to snip your personal info", which covers most phishing attacks, although not all of them involve the theft of personal data. Other definitions proposed by APWG, Xiang et al., and Ramesh et al. centre on the effort to obtain privateindividualparticulars, such as login-credentials. Another description defines phishing as "any web-page that, deprived ofconsent, contends to turn on behalf of a 3rdgathering to cloudspectators into executing an act with which the spectator would individualfaith a factual".

Spam

Email spam is the practice of using email to send unsolicited messages, including advertisements, to a hugeamount of receivers without their consent. This has become a major problem on the internet as it wastes storage space and slows down the speed of messages. While automatic email filtering is a highly effective method of detecting spam, spammers are now able to easily block these filters. In the past, most spam was blocked manually by identifying certain email

addresses. However, machine learning approaches are now being secondhand for unsolicited maildiscovery, with NB being ace of the most commonly useful algorithms[1]. NB is a supervised ML algorithm that is cast-off to solve cataloguing problems, making it possible to build fast ml representations that can make rapidforecasts. Unsolicited mail emails are typically sent as bulk messages, often containing malicious links. If the sender is unknown, the email may be classified as spam. Users may inadvertently sign up for spam when they takeunrestrictedfacilities or update software. On the other hand, "normal" refers to email that is not spam. mlmethods are highly effective, and focus on emergingcomputer_programs and procedures that can admissioninformation[1]. A set of training information, consisting of pre-classified emails, is used to train the algorithms for email filtering. There are many algorithms available for machine learning approaches in email filtering, with Naïve Bayes being one of the most effective, producing the highest accuracy.[2]

II. PROPOSED METHODOLOGY

Spam recognition

The approach employed to filter e-mailingunsolicited mail is founded on the NB algorithm.

Data Pre-processing

Data pre-processing is a method employed to convert raw information into a structured &refined dataset, which can be analysed with ease. When information is composed from numerousbases, it is in its raw form and is unsuitable for analysis. Pre-processing involves several steps that prepare the data for further analysis.

Tokenization is a process of breaking down a large amount of transcript into lessermasses, also known as Signs. These Signs help to identify designs& are separated by whitespace typescripts such as line breaks, spaces or by punctuation marks.[2]

Dropping Values

Dropping is a frequently used technique to handle missing values in a dataset. This involves removing rows or entire columns that contain lost values in order to prevent mistakes during data study.[2]

Stop Words

Break words refer to common English words that do not contribute much to the meaning of a sentence. These words can be securelyoverlookeddeprived of affecting the completelogic of the verdict.

BOW

A BOW is a way of representing text that shows how often different words appear in a document. This method is commonly used to extract important attributes from the document.[3]

This Procedureencompasses the subsequentphases:

Stage 1: Contemplate a casualelectronic mail from the unsolicited mail dataset for implementation.

Stage 2: To carry out the process of attribute extraction/selection and classification, the email needs to be pre-processed first, as the email under consideration is in a basic form.

Stage 3: To start, the email should be tokenized into separate keywords. Tokenization will break down each individual word. If there are any copies in the dataset, they will be removed. Stop words will also be removed from the tokens. The collection of typescript will be transformed into a medium of symbolictotals. Finally, the data will be split into training and testing data.

Step 4: By assessing the prototypical on the exercise and testing data it forecasts the accurateness of the prototype.

NB Classifier

NB is a ml procedure that makes predictions grounded on the likelihood of an entity. It is commonly cast-off in typoscriptsorting, particularly in identifying spam emails. The algorithm relies on the likelihood of the occurrence of certain words in unsolicited mail and normal emails to classify them. It has become a popular and effective technique for detecting spam. NB analyses the likelihood of each period and chooses the supremelikelihood as the yield. This algorithm is known for providing accurate results. The formulation for NB procedure is characterized as trails.[4]

$P(A|B) = P(B|A) * P(A) / P(B)$

Phishing detection

Preprocessing and data assortment

Collecting data is a significant task in creating a ml prototypical. It involves meetingchore-specific data based on predefined factors to generate useful outcomes. However, some data may contain inaccuracies, such as imprecise, incomplete, or erroneous values. Therefore, data processing is necessary before analyzing and drawing conclusions. Data pre-processing comprises activities such as data cleansing, data transformation, and data selection.[5]

Data Cleansing and alteration
Data spring-cleaning is the procedure of eradicating any irrelevant or incomplete data from the dataset. The data may contain missing values, which are considered noisy and must be eliminated. The entire dataset is thoroughly examined to ensure that only clean data is used for the subsequent analysis. Transformations such as smoothing, aggregation, generalization, and data transformation can be used to improve the excellence of the data.[5]

Datasets Assortment
Data assortment involves selecting the most valuable data for our system through a series of procedures or functions. Once we have determined the best algorithm for discovering phishing websites, we can input data into the algorithm and calculate the output based on the results. A set refers to a collection of data that is used for various purposes, such as training in machine learning. The data set is fed into our machine learning algorithm to train the model, and it is the data type cast-off to afford an impartialassessment of the concluding product that is fitted on the exercise data set.[5]

Algorithm
Regression: This is a prediction model that utilizes probability to anticipate the relationship between two variables. The variable being forecasted is referred to as the dependent variable, such as predicting the sales figure[5]

Logistic Regression
The outcome of logistic regression is expressed in 0's and 1's, and these values are used to forecast the output of the provided data,such as High (or) Low. Normally, the range of linear regression is 0-∞ and -∞ to +∞.Because the sigmoid function is used in logistic regression, the linear line must be trimmed to 0 and 1. The dawnrate represents the likelihood of 0 or 1, or good or bad, via sigmoid with this. To derived the logistic regression range 0→∞ to 0→1, these equations are obtained from the straight-line equation y = m1x1+ c (-∞ →+∞). So, we use equation: - If y=0, the equation is equal to 0. As a result, the y value varies between y=1→∞ and you may use the log final logistic regression equation to convert it further to get between -∞ →+∞.[6] in **Fig 1.**

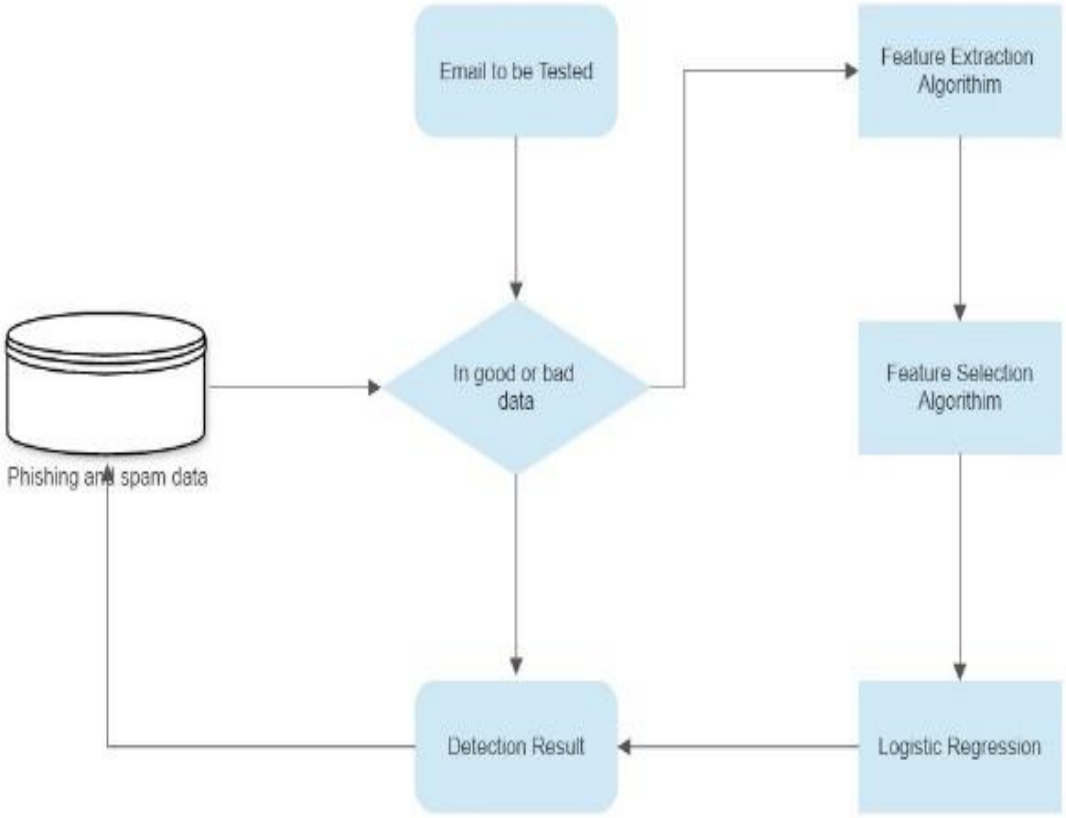

Fig 1. Logistic Regression

III. IMPLEMENTATION

Implementation for spam

The implementation of this model is done using the Visual Studio Code platform, with a training dataset obtained from the Kaggle website. To ensure effective machine execution, the data is check for copies and null standards before being dividedinto 2 datasets- the training data and the test data, in an 80:20 ratio. Text processing is performed on both datasets, removing stop words and punctuation symbols to obtain clean words. The model is then evaluated using both datasets to obtain a confusion matrix, calculated using accuracy, memory, and F1-score. Accuracy and recollectionassistance to determine optimisticexamples in the model, while the F1-grade provides a biasedavgof both accuracy and recollection. The mix-up matrix includes 4unlikemixtures of forecast and definitestandards. The accurateness of the model is dependent on the ratio of preciseforecasts for the test data.[7]

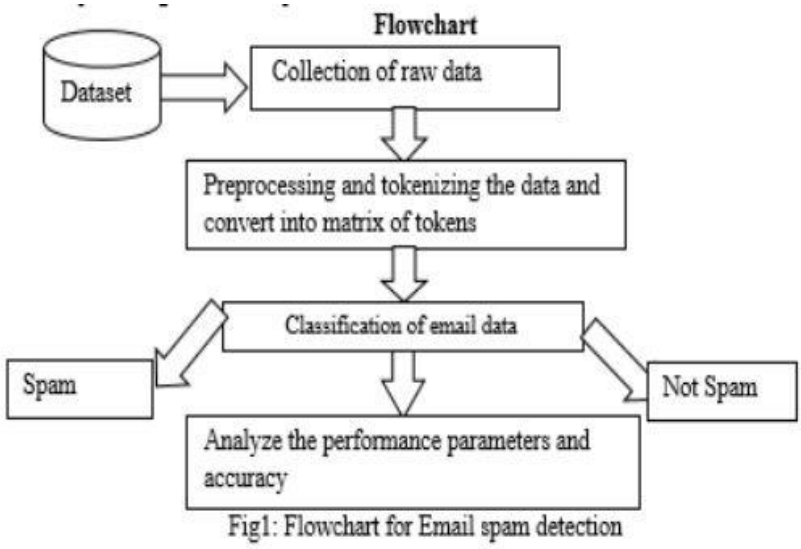

Fig 2. E-Mail Spamming

Implementation for Phishing

To extract the link's features, you'll need a means to enter the link. A model generated with Python code will be used tocarry out the task. After clicking the link, the input block redirects to python code, which uses machine learning methods toextract the features. Phishing sites are usually masked as authentic. By using this method, an invader can trick a victim into clicking on their fake websites that appear to be genuine. These phishing URLs have few distinguishing characteristics. These characteristics are divided into four types, as seen in **Fig 2** above[8]. **Fig 3** shows Implementation for Phishing.

- Feature extraction
- Address bar
- Abnormal Features
- Domain Features

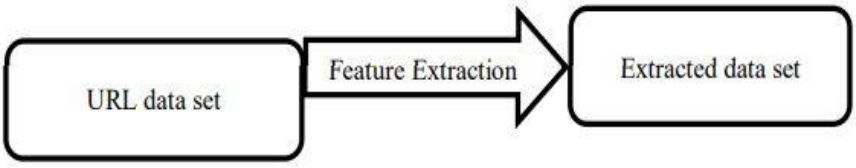

Fig 3. Implementation for Phishing

Use of Regular Expressions

Address bar features: To correctly identify phishing attempts, special characters such as @,"“, /, _, -, and the length of the input URLs must be given in this IP address. It also has a URL length, as well as the availability of Short URLs. The length of a domain registration is also examined. These features are reviewed and confirmed to see if the supplied link includes all of them. If the requirements are met, the link will be forwarded to the next one.

Abnormal Features
The intreated links tags like <META>, <Script>, <Link>, and the SFH status data may beused to check for abnormal characteristics. I'm getting feedback from the supervisor. Examine whether the web page contains a link that would send the data submitted to an email address. Finally, specify if it is an Abnormal URL or not.[9]

Domain features
Because phishing websites have such a brief lifespan, this piece of the code examines the domain's age. It examinesthe DNS highest, web stream of traffic, and page vigorous. Because phishing pages have relatively few visits, the index will be 0 or 1 at most. It alsoexamines the Google catalogue for that, as well as the quantity of pages connecting to that page. All of the feature data is kept inthe Arithmeticalaccount.Logistic regression technique is used to estimate the consequencebuilt on the data provided[9]

Accuracy Score
The Naïve Bayes model is utilized in this endeavour to achieve the highest level of accurateness, and the classifier will deliver its estimated outcomes to the worker. The "spam.csv" dataset obtained in the "Kaggle" website is cast-off for training purposes. The training and testing datasets are evaluated by comparing their ability to correctly distinguish between unsolicited mail and normal messages. The confusion matrix method is used to determine the amount of instances of each class within the dataset under consideration[10]. For FP,FN, TP, and TN, the avgof dataset as follows:

FP: The amount of unsolicited mail that were classified incorrectly is eight in total.
FN: Only one spam email was classified incorrectly in total.
TP: The total amount of unsolicited mail that were accurately sorted as spam is 747.
TN: The entireamount of normal emails that were accurately sorted as normal is 4825.
The model's exactness was evaluated on both the training and testing datasets, and the result was a 99% accuracy rate, as depicted in the **Table 1** below.

Table 1. Training and Testing Datasets

	Exactness	Recall	F1-Score	Support
0	1.00	0.99	0.99	4825
1	0.97	1.00	0.98	747
Correctness			0.99	5572
Macro Average	0.98	0.99	0.99	5572
Weighted Average	0.99	0.99	0.99	5572

Table 2. Confusion Matrix and accurateness consequences for the test data.

Confusion-Matrix:
[[4825 8]
[1 747]]
Accuracy: 0.99209

The test dataset consisted of 5572 text messages. Out of 747 non-spam messages in the test dataset, 746 were accurately sorted as normal, although 1 was erroneously sorted as spam. Out of 4825 spam messages in the test dataset, 4817 were preciselysorted as unsolicited mail, and 8 were incorrectly sorted as normal.**Table 2** displays the confusion_matrix and accuratenessconsequences for the test data.

IV. RESULT

After Creating the model (completion of the project) we can predict whether the comment is SPAM or HAM or its a Phishing link We are able to categorize the messages as unsolicited mail or normal or a phishing site. After implementing algorithm (model) on dataset, after testing our model ,the efficiency of our model will be up to 97%.**Fig 4** shows Implementing Algorithm (Model) On Dataset.

```
In [137]: sample1 = """Pre-Approved Top Up Loan is from Kotak Mahindra Prime Limited, a subsidiary of Kotak Mahindra Bank Limited,
          loan disbursement in 2 days will be subject to complete documentation and credit approval at sole discretion of
          Kotak Mahindra Prime Ltd and subject to guidelines issued by RBI from time to time. Pre-approved Top Up loan is
          offered on the basis of existing car loan repayment. Available at select locations. Mandatory and Stamp duty
          charges applicable.
          click the link to learn more https://www.kotak.com/en/home.html"""

          sample2 = """Hi arul , how are you ? Here is my the  new website I developed can you please login to this and
          let me know your suggestions http://0ed0609.wcomhost.com/app/seco80665452/ , https://www.bgenearl.com/login?rid=WGRbgZx"""

In [140]: predict(sample1)

          This is a spam message

          https://www.kotak.com/en/home.html  is a  good URL

In [141]: predict(sample2)

          This is a normal message

          http://0ed0609.wcomhost.com/app/seco80665452/  is a  bad URL
          https://www.bgenearl.com/login?rid=WGRbgZx  is a  bad URL
```

```
In [153]: sample1 = """Pre-Approved Top Up Loan is from Kotak Mahindra Prime Limited, a subsidiary of Kotak Mahindra Bank Limited,
          loan disbursement in 2 days will be subject to complete documentation and credit approval at sole discretion of
          Kotak Mahindra Prime Ltd and subject to guidelines issued by RBI from time to time. Pre-approved Top Up loan is
          offered on the basis of existing car loan repayment. Available at select locations. Mandatory and Stamp duty
          charges applicable.
          click the link to learn more https://www.kotak.com/en/home.html"""

          sample2 = """Hi arul , how are you ? Here is my the  new website I developed can you please login to this and
          let me know your suggestions http://0ed0609.wcomhost.com/app/seco80665452/ .
          Also follow me on linkedIn for updates https://in.linkedin.com/"""

In [154]: predict(sample1)

          This is a spam mail

          The URL(s) from this mail :
          https://www.kotak.com/en/home.html  is a  good URL

In [155]: predict(sample2)

          This is a normal mail

          The URL(s) from this mail :
          http://0ed0609.wcomhost.com/app/seco80665452/  is a  bad URL
          https://in.linkedin.com/  is a  good URL
```

Fig 4. Implementing Algorithm (Model) On Dataset,

V. CONCLUSION

In the world of the internet, phishing is an unfathomable danger. In this assault, the average person enters personal information onto a phoney website that appears to be a legitimate website. Based on current research, a survey of phishing strategies is conducted.This offered a thorough knowledge of the assault as well as a number of potential responses. In this study, several methodologies for detecting phishing have been discussed; nevertheless, most of the algorithms still have limitations such as accuracy, failure to differentiate objects, and so on. However, logistic regression approach is used to determine accuracy in this study because it is an exposed problem in the phishing industry. The correctness gained is 95%, and based on the research, it may grow further. This project is designed to efficiently detect spam emails based on their content. One approach to detectionunsolicited mails is by verifying the domain names of the sender. Categorizing emails as unsolicited mail or normal is a critical task for effective email management. The NB algorithm is a suitable baseline technique for spam detection, as it deliverssmallincorrectoptimistic rates that are usuallysatisfactory to users. By additional optimizing the limits of the NB algorithm, the accuracy of the spam detection process can be increased. In the future, other optimization algorithms could be used in conjunction with the Naive Bayes algorithm. The examination of the testing is based on f1-score, accuracy, exactness, and recollection. By analysing the outcomes, we can conclude that the integrated approach leads to higher exactness and exactness than the separate NB method.

References

[1]. D. Wu, L. Fan, C. Zhang, H. Wang, and R. Wang, “Dynamical Credibility Assessment of Privacy-Preserving Strategy for Opportunistic Mobile Crowd Sensing,” IEEE Access, vol. 6, pp. 37430–37443, 2018, doi: 10.1109/access.2018.2847251.

[2]. K. R. Jansi and S. V. Kasmir Raja, “A survey on Privacy Preserving Data Aggregation Schemes in People Centric Sensing Systems and Wireless Domains,” Indian Journal of Science and Technology, vol. 9, no. 37, Oct. 2016, doi: 10.17485/ijst/2016/v9i37/102072.

[3]. K. R. Jansi and S. V. Kasmir Raja, “Design Perspectives of People Centric Sensing Systems,” Indian Journal of Science and Technology, vol. 9, no. 37, Oct. 2016, doi: 10.17485/ijst/2016/v9i37/102076.

[4]. N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, “Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets,” IOP Conference Series: Materials Science and Engineering, vol. 226, p. 012091, Aug. 2017, doi: 10.1088/1757-899x/226/1/012091.

[5]. D. Sen, C. Das, and S. Chakraborty, “A New Machine Learning based Approach for Text Spam Filtering Technique,” Communications on Applied Electronics, vol. 6, no. 10, pp. 28–34, Apr. 2017, doi: 10.5120/cae2017652572.

[6]. K. R. Jansi, S. V. Kasmir Raja, and G. K. Sandhia, “Efficient privacy-preserving fault tolerance aggregation for people-centric sensing system,” Service Oriented Computing and Applications, vol. 12, no. 3–4, pp. 305–315, Sep. 2018, doi: 10.1007/s11761-018-0241-5.

[7]. T.-C. Chen, S. Dick, and J. Miller, “Detecting visually similar Web pages,” ACM Transactions on Internet Technology, vol. 10, no. 2, pp. 1–38, May 2010, doi: 10.1145/1754393.1754394.

[8]. M. R. Wijaya, R. Saptono, and A. Doewes, “The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naive Bayes Classifier for The Classification of the Ratio of Inpatients,” Scientific Journal of Informatics, vol. 3, no. 2, pp. 139–148, Nov. 2016, doi: 10.15294/sji.v3i2.7910.

[9]. Anjana Kumari, ”Study on Naive Bayesian Classifier and its relation to Information Gain*,” International Journal on Recent and Innovation Trends in Computing and Communication*, Volume: 2, Issue- 3, pp.601 – 603,March 2014.

[10]. R. Shams and R. E. Mercer, “Classifying Spam Emails Using Text and Readability Features,” 2013 IEEE 13th International Conference on Data Mining, Dec. 2013, doi: 10.1109/icdm.2013.131