JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2268 - 2273
https://publishoa.com
ISSN: 1309-3452

Phishing Detection in Email Using Machine Learning

Alfiya sheikh[1], Dr. Shubhangi Neware[2]

Computer Science & Engineering,Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
Email: sheikhas2_@rknec.edu[1]
Assistant Professor Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
Email: newares@rknec.edu[2]

ABSTRACT

Nowadays everyone uses an email for personal or professional use and the information is used over email is confidential and professional uses. such as credential atm card details and their card details and this is malicious information used by attackers t use your credential and access your login and other platforms

To use ur details phishing is strategies of attack that will accumulate sensitive information from a user to pretending and a trusted organization.

In phished mail attacker will send u a phishing mail to collect the information precedence. phishing detection is a type of classification problem which sort the ham spam classification as legitimate or phish email .in this paper we are discussing and classifying the phished email and also prevent from that we are researching on whole phish ing types contained and also provide a measure.

Keywords*:* support vector machine, phishing,Decision tree.

I. I NTRODUCTION

Phishing is a strategy of collecting sensitive pieces of information from the users and pretending as a trustworthy organization or site. phishing is a type of social engineering attack act of collecting ar payoff the information like credit card, password details username as a trustworthy entity and collecting it from a trustworthy site. prurporting for social media and website and auction sites, online payment.
Mainly phishing starts with email and other communication sites which are mainly designed for attacking the victims. and directly click to the website ar attach documents and the required from us to edit the credentials details in it. Phishers send a message to the thousands of users and in which only a small percentage of users trap in it also gain a high a profit for the users.

The process is going to be followed as :

1. An attacker will send a phishing email to the users.
2. The victim will click on that site and go `through a phishing website
3. An attacker will collect all the credentials from users on its site
4. The attacker will collect all this and access the victim a social media and personal site credits as well; as in Fig 1.

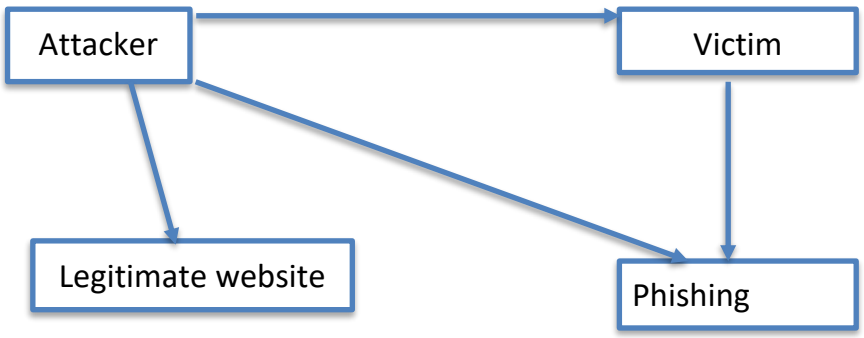


Fig 1: phishing attack

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2268 - 2273
https://publishoa.com
ISSN: 1309-3452

This is the following going to happen for the process going to be held in our system and will get phished.
That message is coming from the trusted sender and if the victim gets fooled and if they provide the credential to a spam website
And also the phishing is target to our system ar target to computers as the ever-increasing use of email and downloading some phish documents and attachments growth of technologies and valuable information to the fraudster to be increasing day by day this paper is focused on the phishing mail and spam detection using machine learning.

II. LITERATURE WORK

[1]This paper discusses the history of phishing attacks and the motivation behind this research by performing this attack Detection of phishing attacks and required highest accuracy for that has always issue withy previous development and design techniques. Have led to required new technologies so there are several ways to implement which one solution is not adequate for that problem.

[2]in this paper they define and research on deceptive phishing is discussed as the most common type of phishing
attack deceptive phishing is occur when recognized source email to comprise information this email refers to your account request on ur email verify account information uses your details.

[3]in this paper, deceptive phishing and prevention in social networking sites using data mining and word net autology using deceptive phishing the major problem is instant messages but detection of phishing through voice chatting is not done yet but good to work on it. Using g data speech recognition system

[4] in this paper author Ma, Yearwood et al. [19] luster phishing emails automatically they are using the mechanism of 13 orthographic features, produce the objective function values, clustering email using algorithms K-means clustering algorithm

[5] C. Emilin Shyni, S. Sarju, and S. Swaminathan are developing a model A MultiClassifier Based Prediction Model for Phishing Emails Detection Using the Topic Modelling, Named Entity Recognition and Image Processing, SciRes 2016 .

III. METHODOLOGY WORK

We are deciding to proposed work our project is to classify the ham and spam messages and authorize it and using our dataset in which we are using the most prominent and less number of features to classify and obtain the highest accuracy so our model process going to follow as :

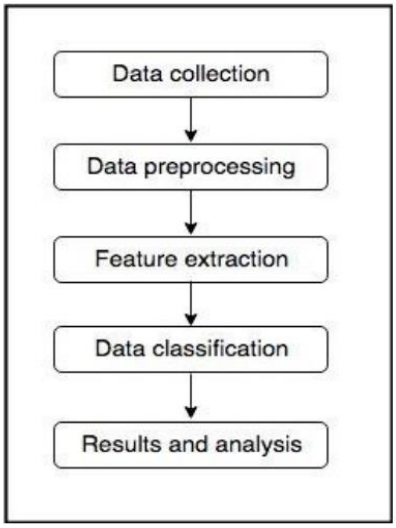


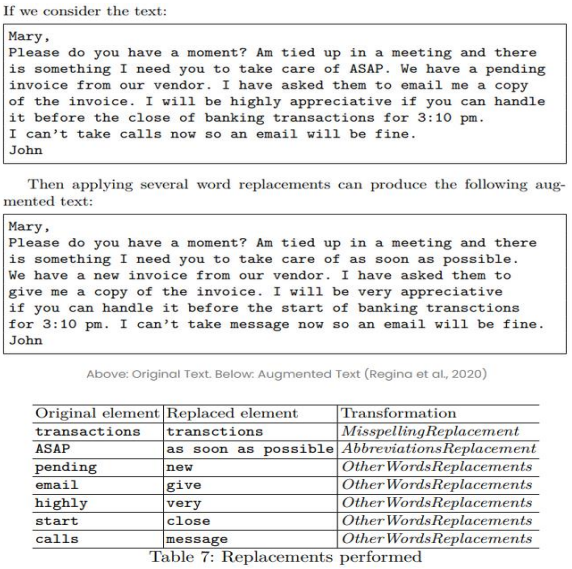
Fig2: Process Flow

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2268 - 2273
https://publishoa.com
ISSN: 1309-3452

FEATURES

So here feature we are using the same concept of data augmentation's in phishing using the text argumentation for a spear phish text augmentation techniques for email phish detection and also text data augmentation meaning the creation of new data from existing data

Indeed for an account of language complexity while relevant to the natural language processing and (machine learning for translation text classification.) So the example of text data augmentation implementation is as follows which is shown in fig [3].


If we consider the text:

```
Mary,
Please do you have a moment? Am tied up in a meeting and there
is something I need you to take care of ASAP. We have a pending
invoice from our vendor. I have asked them to email me a copy
of the invoice. I will be highly appreciative if you can handle
it before the close of banking transactions for 3:10 pm.
I can't take calls now so an email will be fine.
John
```

Then applying several word replacements can produce the following augmented text:

```
Mary,
Please do you have a moment? Am tied up in a meeting and there
is something I need you to take care of as soon as possible.
We have a new invoice from our vendor. I have asked them to
give me a copy of the invoice. I will be very appreciative
if you can handle it before the start of banking transctions
for 3:10 pm. I can't take message now so an email will be fine.
John
```

Above: Original Text. Below: Augmented Text (Regina et al., 2020)

Original element	Replaced element	Transformation
transactions	transctions	*MisspellingReplacement*
ASAP	as soon as possible	*AbbreviationsReplacement*
pending	new	*OtherWordsReplacements*
email	give	*OtherWordsReplacements*
highly	very	*OtherWordsReplacements*
start	close	*OtherWordsReplacements*
calls	message	*OtherWordsReplacements*

Table 7: Replacements performed

Fig 3:Replacement performs

Natural language processing :

NLP application involves context translation, automatic text correction, sing language to text correction use of chat-bots on websites, speech recognition also classification of email phishing ham spam detection.

Most using an algorithm in natural language processing is as included: in which we will briefly focus on the BERT technology which is specifically used for mail ham spam classification .so included algorithm we can take it one by one:

Bag of the word: a bag of words also is used to vectorize the information from text. Is that the way it checked the occurrences of words and count the words?

TF-IDF: This is the algorithm that will take the account the occurrences and count the frequency in which words appear in the text. some of the words having positive weightage are some having the negative weight

Steaming: stemming and calculating the upcoming with using that words and finding that will goes on
And while using it goes to

Lemmatization: Lemmatisation is a technique used to convert a word into its basic lemma to group the different forms of the same term.
Its also used for the text normalization

BERT: BERT is a high-performance algorithm that is used to understand and analyze text based on the context of a word.

BERT (Bidirectional Encoder Representations from Transformers) is an algorithm in the NLP field that can analyze and learn relationships between words based on a context. In NLP, this mechanism is called **attention.**

Another great advantage of BERT over other language models is that it was designed to analyze texts in both directions. That is, from right to left, and from left to right so it will easy to recog. This mechanism is called **bidirectionally.**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2268 - 2273
https://publishoa.com
ISSN: 1309-3452

The combination of **attention and bidirectionally** mechanisms allows some systems based on BERT to be extremely efficient in identifying and classifying texts. And this evolution in detecting malicious emails comes in. This way, our system can analyze and understand the message's context and then define with precision whether it is a legitimate or malicious one, such as a spam or phishing campaign.

IV. RESULT AND CONCLUSION

A. Dataset:the dataset is comprised of emails containing all ham and spam(phished) massages with real-time data also we added a dataset available on the public site Kaggle and we used it together. First of all, we analyze our data and perform Eda on it. Preposed our data including Normalisation, lemmatization, and steaming, then after obtaining the transformed text from the original dataset and then used it for model building training and testing using a classifies and algorithm including svc, etc,decision tree, naive byes,

B. Accuracy Prediction:In this Accuracy prediction simple we use four Algorithms to find accuracy between Tweets and results. Here are those four Algorithms..

1)Logistic Regression: Logistic regression is a considerable Machine learning Technique that uses Supervised Learning. It operated a predicting unconditional dependent value by using two independent sets. Logistic regression predicts the accuracy uses of dependent values.

2) Decision Tree:In that Decision tree, it is also a Supervised Machine Learning where data is split between nodes and leaves.

3) Random Forests:Random forest is one of the Supervised Machine Learning techniques which is use Classification and Regression problems. It makes a Decision tree on another sample set and takes the majority value for classification and regression using the average case.

4) Support Vector Machine (SVM):SVM also uses a supervised technique. This algorithm uses both regression and classification.

5) in this Fig (4)we can see that our analysis according to Algorithms. Our best Accuracy is 0.9990 for Testing data set which is given by Logistic Regression, Random Forests and Support Vector Machines.

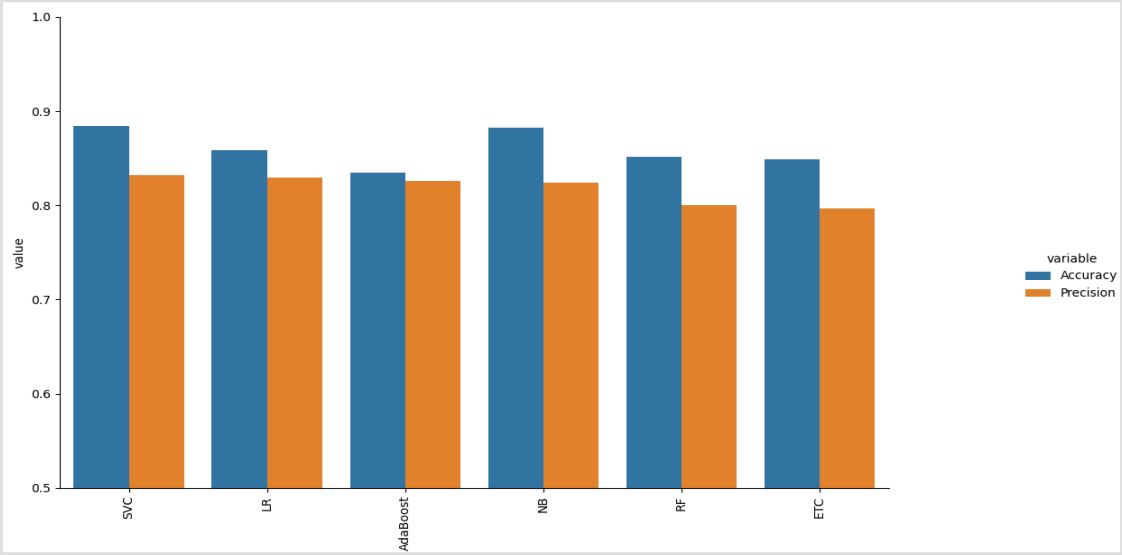


Fig4:Accuracy diagram

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2268 - 2273
https://publishoa.com
ISSN: 1309-3452

Model	**accuracy**
Linear Regression	0.85
Support vector classifier	0.88
Adaboost classifier	0.83
Naive byes	0.88
Random forest	0.85

Fig 5: Comparison of Accuracy table

6) CompareAfter implementing the algorithm it will give all precision and recall values as shown below diagram, pragmatic.

	Algorithm	Accuracy	Precision
0	SVC	0.884266	0.831643
1	LR	0.858919	0.829135
3	AdaBoost	0.835007	0.826036
5	NB	0.882353	0.823705
2	RF	0.851267	0.800617
4	ETC	0.848876	0.796524

	Algorithm	variable	value
0	SVC	Accuracy	0.884266
1	LR	Accuracy	0.858919
2	AdaBoost	Accuracy	0.835007
3	NB	Accuracy	0.882353
4	RF	Accuracy	0.851267
5	ETC	Accuracy	0.848876
6	SVC	Precision	0.831643
7	LR	Precision	0.829135
8	AdaBoost	Precision	0.826036
9	NB	Precision	0.823705
10	RF	Precision	0.800617
11	ETC	Precision	0.796524

Fig6: Precision and Accuracy Analysis

7)The results show our model produces high accuracy in detecting phished emails. By using the most relevant features, the number of features has been reduced as compared to other works but at the same time, accuracy is improved .

V. CONCLUSION

Phishing is an n attack that is used to steal your money or your identity by using and getting you to reveal your personal information such as bank information including credit card numberer password on a website that pretends to be legitimate from this study or analysis From this Research paper discuss the phishing how it goes on how it will frauds you so that purpose is to aware about phishing attack and protect the users from this phishing attack so we are creating a model in which we an including the classifications of mails into phished ham and spam messages by using the classifier and algorithm .before of that we need preposed and train our dataset by using a classifier and preprocessing step and extract the most relevant features from it .and then fed model into random forest model, SVM, Naive byes, linear regression, and highest accuracy is found from random forest model, Svc, so we clear that svc gives you the highest accuracy. So we again improved it after agin involving some other algorithms so right now our research is up to this model which we created to avoid the user being a fraud by the phishing attack.

VI. REFERENCES

1. Noor Ghazi M. Jameel, Loay E. George. They study on Detection of Phishing Emails using Feed Forward Neural Network, in International Journal of Computer Applications 2013.

2. Ian Fette, Norman Sadeh, Anthony Tomasi, they Learning to Detect Phishing Emails, and processing . And they Proceedings research of the International World Wide Web Conference (WWW), 2006 department.

3. Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf .

4. Gori Mohamed .J, M. Mohammed Mohideen, Mrs. Shahira Banu. Worked on the Email Phishing - An open threat to everyone, in the publication of International Journal of Scientific Research Publications, 2014 .

5. C. Emilin Shyni, S. Sarju, S. Swaminathan found A Multi- Classifier Based Prediction Model for Phishing on Emails Detection Using Topic Modelling, and the Named Entity Recognition and Image Processing, SciRes 2016 .

6. Noor Ghazi M. Jamee , Loay E. George (2014), working on the “Detection Phishing Emails Using Features Decisive and the machine learning model Values”,257-259.

7. Rakesh M. Verma and Nirmala Rai. Are worked and uses Phish-IDetector: Message-Id Based Automatic Phishing Detection, in the International Joint Conference on e-Business and Telecommunications 2015 .