# FUZZY CLASSIFICATION AND DECISION TREE BASED ESTIMATION OF PHISHING MESSAGES AND URLS

**Abhishek Chunawale[1]*, Sumedha Sirsikar[2], Aniket Ingavale[3], Pratik Kedar[4]**

**Abstract—**
The growth of internet users and the overall evolution of the internet platform have been highly innovative and useful. With the increase in number of users, there has also been an increase in the number of services that are offered by the internet platform. Due to these factors, a very popular and widely used social engineering method for Cyber-attack called phishing is being used. Phishing is the process by which the cyber-attack utilizes a fake web page to trick the unknowing user into revealing their confidential information or authentication credentials to the attacker. This type of attack is highly difficult to identify and ameliorate due to the variable nature of phishing attacks and the deception from the phishing attack coordinator. Therefore numerous researches on phishing attacks have been analyzed effectively to propose our approach which is based on the implementation of natural language processing along with fuzzy classification and decision tree. The proposed approach is effectively evaluated for its errors to achieve the performance metrics indicating a satisfactory performance of the approach.

**Keywords—**Fuzzy Classification, Natural Language Processing, Decision Tree, Phishing

[1]*School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University Pune, India, E-mail:- abhishek.chunawale@mitwpu.edu.in
[2]School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University Pune, India, E-mail:- sumedha.sirsikar@mitwpu.edu.in
[3]School of Computer Engineering and Technology Dr. Vishwanath Karad MIT World Peace University Pune, India, E-mail:- aniket.ingavale@mitwpu.edu.in
[4]Department of Information Technology, MIT Pune, India, E-mail:- thepratikkedar@gmail.com

***Corresponding Author:** Abhishek Chunawale
*School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University Pune, India, E-mail:- abhishek.chunawale@mitwpu.edu.in

**DOI:** 10.48047/ecb/2023.12.si5a.0552

## I. INTRODUCTION

Before the advent of the internet platform and large-scale use of information systems across the globe, the main mode of information transfer and retrieval was in the physical format through the use of written material such as a book or the use of word of mouth. These techniques were gradually improved over time while the books are on large scale used even today there are a lot more techniques for the storage and retrieval of data electronically. These techniques are highly useful as they can easily store large amounts of data in a small space and can be retrieved effectively without any errors, unlike the traditional and conventional non-electronic storages.

With the introduction of electronic storage and the utilization of these approaches for the purpose of storing data, large amounts of data can be stored easily by any individual. With the increase in the requirement of an effective communication technique, the various researchers and inventors over the years improve the networks to create a large worldwide connection called the internet. The internet has been effective in enabling useful communication over large-scale distances to reduce the cost and increase the pace of innovation and research significantly.

The improvements in the electronic space have led to increasing affordability of various electronic devices and smartphones that are capable of connecting to the internet platform. This has in turn significantly increased the number of users on the internet platform and a large number of services have also been created to facilitate the large influx of users. This has been significant in the overall development and the evolution of the internet platform with social media websites and e-commerce portals being created and used by a large number of individuals and organizations.

This had an undue on the overall security of the users and organizations on this platform. With the increasing number of Cyber-attacks and the ever-evolving modus operandi of the attackers, it has become highly difficult to secure the online space. The paradigm of social engineering has been significant in this field out of which the phishing attacks are one of the most popular and widely used online mechanisms. Phishing attacks effectively deceive the user by providing a fake URL of a legitimate website and asking these individuals or organizations to enter their valid credentials for authentication. As soon as the user enters the relevant information it directly gets into the hands of the attacker who can utilize it to perform data leaks and other criminal activities.

Therefore it is imperative to stop these phishing attacks and reduce the number of individuals that fall for these tactics and other techniques to reveal their information and credentials. There are many techniques that have been designed to achieve this goal and these approaches have been effectively studied in this literature survey paper thoroughly. These techniques have been useful for the purpose of designing our methodology.

This research study devotes section II to an overview of previous work as a literature survey; section III describes the proposed methodology; section IV shows analysis and the system's performance; and lastly, section V finishes the article with conclusion and hints for future scope.

## II. LITERATURE SURVEY

P. Legg et al. explain that with the increasing number of websites and web portals on the online platform there is also an increase in phishing based attacks on the internet. The increased number of users on the internet platform also contributed to the large-scale improvement in this paradigm and the subsequent increase of these websites [1]. The phishing attack effectively performs a highly sophisticated approach for the purpose of stealing the personal data of the user including the login credentials. This is a breach of privacy and needs to be eliminated effectively. There are a number of techniques and tools that are utilized to improve awareness based on cyber situation and reduce phishing attacks effectively which are elaborated in this research paper.

C. N. Gutierrez et al. express that there has been an increase in the number of electronic devices and internet-capable smartphones across the world. This is a highly useful situation as it allows a large number of users to offset the large costs related to running the internet and keeping the servers online. This led to a significant increase in internet platforms as the number of connected devices increases exponentially every single day [2]. These large-scale increases have also been the main reason behind a large number of cyber attackers attacking the users and stealing their data through phishing. The authors propose the use of a safe PC which stands for semi-automated feature generation for phishing classification for effective identification of new phishing techniques.

K. S. C. Yong et al. introduced the concept of phishing attacks as one of the problematic occurrences that are widespread on the internet nowadays. These attacks are highly focused on deceiving the user into giving out their credentials to a compromised website [3]. Phishing attacks are highly difficult to detect and even more difficult to counteract as the attackers utilize new and ingenious ways to perform these activities.

Therefore the authors in this paper proposed the utilization of an effective countermeasure to reduce phishing attacks on websites through the use of the quick response code.

V. Lyashenko et al. elaborate on the large scale improvements in the info-communication systems that are gaining popularity all across the globe. These information and communication systems provide effective communication and transfer of information through the use of widespread networks such as the internet [4]. These systems are highly reliant on various functions that improve reliability through the protection of information and effective monitoring. The paradigm of phishing attacks is highly detrimental to this kind of system as it can allow the attackers to effectively gain access to these systems and wreak havoc. The researchers in this publication investigate the tools that are predominantly used for the purpose of understanding the dynamics of phishing attacks.

M. Zabihimayvan et al. narrate that the approaches of social engineering have been effective in the advent of phishing attacks for the purpose of extracting confidential and personal information of the user. Phishing attacks have been gaining popularity with that actors all over the world as it requires very less effort and can be used effectively over multiple layers of security without any detection [5]. The phishing process involves the use of a fake website or a webpage that is used to steal the user's confidential information or login credentials effectively. To reduce this effect the authors propose the use of an enhanced phishing attack detection technique that utilizes feature selection through the fuzzy rough set formation.

M. Baykara et al. state that there has been an increase in cybercrime performed online due to the increasing number of users or potential victims that can be attacked by the attackers. The rise of the social media approaches has also accelerated this process significantly [6]. The process of phishing is highly complex and difficult to detect by traditional approaches. The authors in this publication have proposed the utilization of a Bayesian algorithm for effective classification and evaluation of spam and phishing emails.

S. Patil et al. elaborate on the various techniques that are being used to perform attacks on the online environment nowadays. One of the most popular and most used techniques for the purpose of attempting cyber-attacks is the utilization of phishing. The phishing attacks are highly complicated social engineering methodologies that deceive the user into revealing their sensitive information or provide access to confidential data easily. This is highly problematic as it can lead to data leakage scenario of confidential data that can be detrimental to the individual or an organization [7]. In this research, the authors have proposed the construction of a framework for anti-phishing through performing an overview on phishing identification methodically.

T. Nathezhtha et al. express that there has been a significant increase in the number of phishers or attackers that utilize phishing attacks to steal the information of the users such as password and username to gain unauthorized access on their behalf. This process can lead to decreased reliability of the system as well as it can lead to a data leak scenario [8]. This operation should be eliminated significantly to achieve effective security for the users online. The researchers in this publication have proposed the use of a phishing attack detector that utilizes web crawling for the purpose of phishing detection.

Athulya A. A. et al. explain that there are various techniques that are utilized by social engineering to gain an effective intrusion into any system. One of the most popular social engineering approaches is the utilization of phishing attacks to achieve effective gathering of sensitive data that can be useful in the intrusion. These attacks are getting increasingly popular due to the versatility and the lack of detection techniques for phishing attacks with high accuracy [9]. Therefore to ameliorate this effect the authors have proposed the utilization of a hybrid technique for the detection of phishing attacks in a quick and extremely accurate manner.

G. J. W. Kathrine et al. introduce the concept of cyber-attacks and the various techniques that are utilized by the attackers to perform these kinds of attacks. The cyber-attacks are mainly performed to extract confidential information or the financial traditional sketches bank account details or other card details [10]. These attacks are performed by highly qualified cybercriminals that utilize a plethora of different techniques where phishing is one of the most primary and high success rate technique. The authors of this publication have effectively elaborated on the various machine learning techniques and the other variants of phishing attacks.

T. Peng et al. narrate the various cyber-attacks and other security threats that are being monitored carefully and big researched extensively to build countermeasures and detection approaches for the same. The authors have evaluated the various cyber-attacks and found that there is very little attention being paid towards the prevention and reduction of phishing attacks [11]. For this

purpose, the authors have prompted the use of machine learning along with natural language processing for the highly accurate detection of phishing emails.

K. M. Zubair Hasan et al. state that with the rapid evolution all across the world there is a speed increase focus on the World Wide Web and the development of the Internet Services. This is due to the increased convenience offered by the internet platform that attracted a large number of users and services to this paradigm. This is also been the cause of an increased number of cyber-attacks especially phishing attacks on various businesses and individuals [12]. To counteract this process, the authors have proposed the utilization of machine learning algorithms such as a deep convolutional neural network for the implementation of an automatic prediction approach for phishing websites.

## III.PROPOSED METHODOLOGY

The proposed technique for phishing URL detection has been illustrated in the system overview as shown in the Figure 1. The presented approach is executed in a stepwise manner to achieve the outcomes which have been described below.

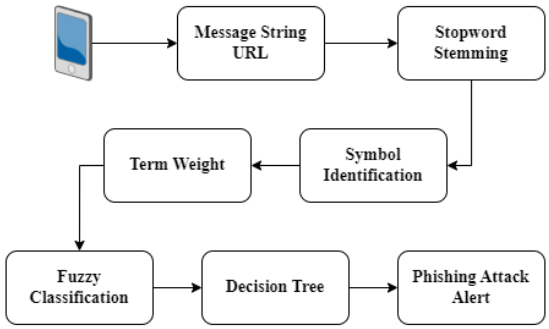


**Fig. 1.** Mobile Phishing Detection System Overview

### *A. Mobile Phishing Detection System*

#### *Step 1: Message Collection and URL Handling*

The presented approach for the detection of the phishing attacks has been developed on a mobile application. The suspected phishing message is provided to the system as an input through the graphical user interface provided in the application. This message must be entered by the user into the text field as an input to the application. This message is crucial for realization of the phishing attacks and which can be identified using the fuzzy classification and decision tree.

The input message is first searched for the presence of 'http' string. This is achieved by splitting the words from the input message and detecting the URL in the message. After the detection of the URL, the length of the URL is calculated. This length is compared with the length of the message. If the length of the message is equal to the length of the URL then the message contains only the URL, otherwise the message contains a text as well as the URL. As an output of comparison binary value 0 or 1 is obtained where the value '1' is considered for the input containing the message as well as the URL. On the other hand, the value '0' is considered for the input message containing only the URL.

The system is built for three different scenarios. The input message provided to the system will either contain both the message and the URL or only the message or only the URL. The proposed approach considers such scenarios for phishing detection.

The URL is acting as a subject to identify the presence of phishing elements in the input message. The extensive processing of the URL is achieved by reading the URL against various conditions such as length, position and number of symbols etc. Prediction of safe or unsafe URL can be done by considering presence of following special characters such as:

1. The number of dots appearing / present in the URL should not be more than or equal to 4
2. The length of the URL should remain less than 25 characters
3. The different symbols and hyphens should not be more than 2
4. Certain words are checked for their presence in the URL, such as PayPal, confirmed and suspend, and if these words are encountered more than 2 times, the flag is turned to true.

All of these parameters are used to facilitate the detection and are subjected to analysis of the flag values for each of the conditions. If the resultant achieves more than 2 attributes to be true, then the URL is confirmed as a phishing. This flag output is then provided to the system for further evaluation.

#### *Step 2: Preprocessing and Term Weight*

This process is the first logical step in the methodology where the extracted input message received from the previous step which was provided as an input to the system. After the elimination of the URL from the message, the text is first taken into a string format and split into words, this is then preprocessed by removal of the stop words.

Stop words are the words that are used in the English language to connect two phrases or to integrate two separate portions of the same

sentence. Stop word removal method is the part of preprocessing which receives the input message to remove stop words. Current methodology does not require this, and the removal of the stop words imposes no penalty in terms of comprehension of the sentence. Therefore, the words such as 'and', 'is', 'the' are eliminated from the message to achieve the preprocessed text which is provided to the next step for further processing.

The message text with the stop words removed is subjected to the term weight calculation. Term Weight is nothing but the frequency of the words that are appearing in the message. This process determines the value of the words and their importance in the message. The words and their weights are calculated and then subjected to the Bag of Words approach for the further evaluation.

***Step 3: Bag of Words***

The preprocessed message along with the term weight is subjected to the bag of words as an input. The Bag of Words approach is a personalized dictionary that contains suspected words for the phishing process. The message is analyzed for their presence in the Bag of Words to achieve detection of these terms in the message. This is done by comparing each of the words to achieve their presence in the form of a Boolean value. This is then used to count the number of words that are corresponding to the Bag of Words and this count and the respective words are provided to the next step for the purpose of classification. The term frequency of bag of words is extracted as mentioned in the algorithm below.

ALGORITHM: Bag of Word Term Frequency Estimation

//Input=Bag of Words: $BOW_{LST}$
//Input=Preprocessed Message List: $PM_{LST}$
//Term Frequency List: $TF_{LST}$
Term Frequency List($BOW_{LST,}$ $PM_{LST}$)
1: Start
2: $TF_{LST}$ =∅
3: ***for*** i=0 to Size of $PM_{LST}$
4: $TMP_{LST}$ =∅
5: $P_{Word}$= $PM_{LST\,[i]}$
6: count=0
7: ***for*** j=0 to Size of $BOW_{LST}$
8: $B_{Word}$ = $BOW_{LST\,[j]}$
9: ***if*** ($P_{Word}$= $B_{Word}$)***, then***
10: count++
11: ***end for***
12: $TMP_{LST[0]}$ = $P_{Word}$
13:$TMP_{LST[1]}$ = count
14: $TF_{LST}$ = $TF_{LST}$ + $TMP_{LST}$
15: ***end for***
16: return $TF_{LST}$
17: Stop

***Step 4: Fuzzy Classification***

The output from the Bag of Words approach is provided as an input to this step for the purpose of classification. The Fuzzy classification is one of the most powerful classification processes that have been effective in realization of the complete classification of the input. This is achieved through the use of the fuzzy crisp values that are in a range which provides better segregation of the input. In this system, the fuzzy categories are created into 3 labels as medium, high and none. The medium label classifies any word with the Bag of Words score in the range of 1 to 3, the high label accepts a Bag of Words score between 4 and 5, and the rest of the scores are labeled as none. The non-zero words are the words with non-zero Bag of Words score which are provided as an input to execute the fuzzy classification process.

***Step 5: Decision Tree***

This is the final step of the procedure where the output from the fuzzy classification process as well as the flag value from the URL handling process is provided as an input. If-then rules make it easier to segregate the output into the separate parts needed for the analysis of the phishing content in the provided input string.

The output from the URL handling returns a Boolean value for the presence of phishing contents or not. The other output from the fuzzy classification is a label in the form of medium, high or none. These values are provided as an input to the decision tree which results into following status of vulnerability:

- Medium Vulnerability: If the flag is false and the label is medium
- High Vulnerability: If the flag is false and the label is high or the flag is true and the label is medium
- Very High Vulnerability: If the flag is true, and the label is high
- Low Vulnerability: If the flag is true and the label is none
- No Vulnerability: If the label is false and the label is none

## IV. RESULTS AND DISCUSSIONS

The results obtained for phishing attacks detection is achieved through the use of fuzzy classification and decision tree. It is necessary to determine the effective error achieved by the system to understand if the implementation of fuzzy classification and decision tree approaches has been done correctly. Any inconsistency in the implementation of these approaches would lead to

massive errors that will be indicative of deployment error.

The precision of the detections achieved by the system for the phishing attacks must be evaluated for the presence of errors as part of the evaluation method. The Root Mean Square Error (RMSE) is the performance metric being used for this purpose. This is due to the fact that the RMSE is capable of extracting the true error. This error evaluation is an effective tool that can lead to the determination of the execution preciseness of the phishing detection methodology. This method is used to calculate the effective rate of error between two continuous and correlated variables.

The continuous and connected things that our system has produced are the expected phishing attack detections and the obtained phishing attack detections. The value of RMSE can be realized as per Equation 1 given below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \quad (1)$$

Where,

$(x_1 – x_2)^2$: Differences Squared for the summation in between the expected phishing attack detections and the obtained phishing attack detections.

n: Number of samples or Trails.

Experimental results achieved through RMSE are as shown in the Table I below.

**TABLE I: MEASUREMENT OF MEAN SQUARE ERROR**

| Experiment No. | Number of expected phishing attack detections | Number of obtained phishing attack detections | Mean Square Error (MSE) |
|---|---|---|---|
| 1 | 10 | 5 | 25 |
| 2 | 10 | 7 | 9 |
| 3 | 10 | 3 | 49 |
| 4 | 10 | 6 | 16 |
| 5 | 10 | 4 | 36 |

Figure 2 shows Comparison of MSE between Number of expected phishing attack detections v/s Number of obtained phishing attack detections.

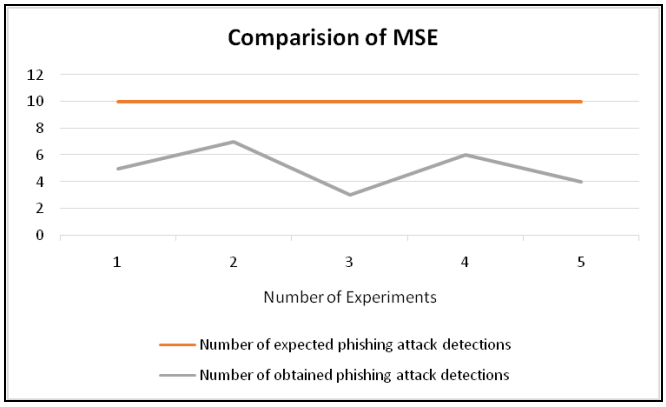


**Fig. 2.** Comparison of MSE between Number of expected phishing attack detections v/s Number of obtained phishing attack detections

A set of five tests were used to analyze the error calculated between the values of expected phishing attack detections and the obtained phishing attack detections. Each of the tests is subjected to ten trials. The obtained values for the MSE and RMSE are 27 and 5.19 respectively. The error rate of the system is within the expected range with a highly satisfactory performance of the initial implementation. The obtained RMSE value dictates a correct implementation of the proposed methodology using fuzzy classification and decision tree.

## V. CONCLUSION

The paradigm of phishing attacks and the various techniques for its detection and evaluation has been discussed in detail in this research paper. There has been an immense growth in the number of these attacks been performed on the Internet platform. These attacks have been effective in various data breaches which have led to a lot of

undesirable situations for various individuals and organizations. Therefore there is an urgent need for purpose of identifying phishing attacks on the online platform. Our methodology implementation takes the suspected text as an input which is then preprocessed and the message and URL are isolated to check the presence of the phishing. This is achieved through the use of Bag of Words and Natural Language Processing. The output form the URL handler and the fuzzy classified messages are provided as an input to decision tree which detects the presence of phishing attack through the use of if-then rules. The experimental evaluation achieves an effective score that is adequate for the prescribed methodology.

The future research prospects can be focused on achieving this phishing attack detection approach into to an API for universal application and integration.

## REFERENCES


1. P. Legg and T. Blackman, "Tools and Techniques for Improving Cyber Situational Awareness of Targeted Phishing Attacks," 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), Oxford, United Kingdom, 2019, pp. 1-4, doi: 10.1109/CyberSA.2019.8899406.
2. C. N. Gutierrez et al., "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks," in IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 6, pp. 988-1001, 1 Nov.-Dec. 2018, doi: 10.1109/TDSC.2018.2864993.
3. K. S. C. Yong, K. L. Chiew and C. L. Tan, "A survey of the QR code phishing: the current attacks and countermeasures," 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843688.
4. Lyashenko, O. Kobylin and M. Minenko, "Tools for Investigating the Phishing Attacks Dynamics," 2018 International Scientific - Practical Conference Problems of Info communications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 2018, pp. 43-46,doi:10.1109/INFOCOMMST.2018.8632100.
5. M. Zabihimayvan and D. Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2019.8858884.
6. M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5, doi: 10.1109/ISDFS.2018.8355389.
7. S. Patil and S. Dhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 588-593, doi: 10.1109/ICACCS.2019.8728356.
8. T. Nathezhtha, D. Sangeetha and V. Vaidehi, "WC-PAD: Web Crawling based Phishing Attack Detection," 2019 International Carnahan Conference on Security Technology (ICCST), CHENNAI, India, 2019, pp. 1-6, doi: 10.1109/CCST.2019.8888416.
9. A.A. and P. K., "Towards the Detection of Phishing Attacks," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 337-343, doi: 10.1109/ICOEI48184.2020.9142967.
10. G. J. W. Kathrine, P. M. Praise, A. A. Rose and E. C. Kalaivani, "Variants of phishing attacks and their detection techniques," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 255-259, doi: 10.1109/ICOEI.2019.8862697.
11. T. Peng, I. Harris and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, 2018, pp. 300-301, doi: 10.1109/ICSC.2018.00056.
12. K. M. Zubair Hasan, M. Z. Hasan and N. Zahan, "Automated Prediction of Phishing Websites Using Deep Convolutional Neural Network," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036647.