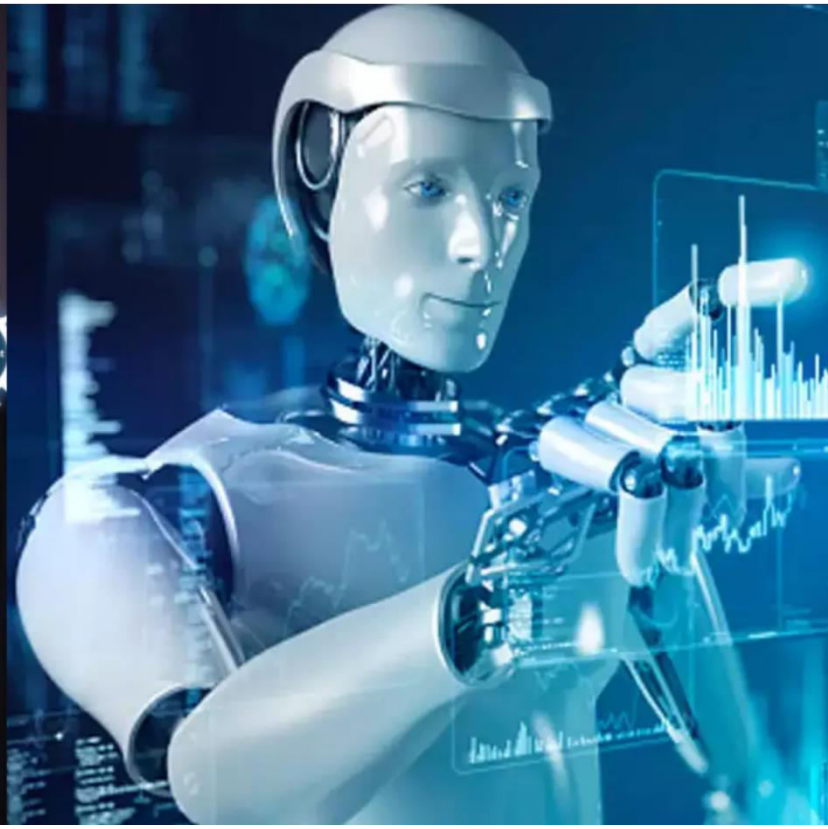
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

To Implement a System for Phishing Website Prediction using SVM Algorithm

Thaniga Chalam M[1], Vimala M[2]

Associate Professor, Department of Electronics and Communication Engineering, Er. Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India

PG Student, Department of Applied Electronics, Er. Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India

ABSTRACT: This paper presents that the phishing is a technique used by cyber-criminals to impersonate legitimate websites in order to obtain personal information. This paper presents a novel lightweight phishing detection approach completely based on the URL (uniform resource locator). The mentioned system produces a very satisfying recognition rate this system, is an SVM (support vector machine) tested phishing URLs records. In the literature, several works tackled the phishing attack. How ever those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage. The proposed system uses only six URL features to perform the recognition. The mentioned features are the URL size, the number of hyphens, the number of dots, the number of numeric characters plus a discrete variable that correspond to the presence of an IP address in the URL and finally the similarity index. Proven by the results of this study the similarity index, the feature we introduce for the first time as input to the phishing detection systems improves the overall prediction rate.

KEYWORDS: Phishing URL's, Python frontend, Database, non phishing URL's, SVM algorithm, IDE VS code, dataset.

I. INTRODUCTION

Over the most recent couple of years because of the persistent development of utilization of web we utilize the mail benefits to be specific the mass conveyance of undesirable messages, principally of business sends, yet in addition with damaging substance or with false objectives, has turned into the primary issue of the email benefit for Internet specialist co-ops (ISP), corporate and private clients. Ongoing overviews detailed that more than 60% of all email traffic is Phishing. Phishing causes email frameworks to encounter over-burdens in transmission capacity and server stockpiling limit, with an expansion in yearly expense for organizations of more than several billions of dollars. What's more, Phishing messages are a significant issues for the security of clients, since they endeavor to get the data from them to surrender their own data like stick number and record numbers, using parody messages which are taken on the appearance of originating from trustworthy online organizations, for example, financial foundations. Messages can be of Phishing compose or non-Phishing compose.

Phishing mail is additionally called garbage mail or undesirable mail though non-Phishing messages are veritable in nature and implied for a particular individual and reason. Data recovery offers the devices and calculations to deal with content records in their information vector frame. The Statistics of Phishing are expanding in number There are extreme issues from the Phishing messages, viz., wastage of system assets (data transfer capacity), wastage of time, harm to the PCs due to infections and the moral issues, for example, the Phishing messages publicizing obscene locales which are hurtful to the youthful ages.

1.1 Objective:

Phishing detection techniques have been the focus of considerable research. Typical phishing detection techniques include the blacklist-based detection method and the heuristic-based technique. The blacklist-based technique maintains a uniform resource locator (URL) list of sites that are classified as phishing sites; if a page requested by a user is present in that list, the connection is blocked. This technique is commonly used and has a low false-positive rate; however, its accuracy is determined by the quality of the list that is maintained. Consequently, it has the disadvantage

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

of being unable to detect temporary phishing sites. The heuristic-based detection technique analyzes and extracts phishing site features and detects phishing sites using that information. To propose a new heuristic-based phishing detection technique that resolves the limitation of the blacklist-based technique. We implemented the proposed technique and conducted an experimental performance evaluation. The proposed technique extracts features in URLs of user-requested pages and applies those features to determine whether a requested site is a phishing site. This technique can detect phishing sites that cannot be detected by blacklist-based techniques; therefore, it can help reduce damage caused by phishing attacks.

II. LITERATURE SURVEY

1. Survey of review Phishing detection using machine learning techniques.

Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services. Because of their impact, manufacturers and retailers are highly concerned with customer feedback and reviews. Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. This practice is known as Opinion (Review) Phishing, where Phishingmer manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain. Since not all online reviews are truthful and trustworthy, it is important to develop techniques for detecting review Phishing. By extracting meaningful features from the text using Natural Language Processing (NLP), it is possible to conduct review Phishing detection using various machine learning techniques. Additionally, reviewer information, apart from the text itself, can be used to aid in this process. In this paper, we survey the prominent machine learning techniques that have been proposed to solve the problem of review Phishing detection and the performance of different approaches for classification and detection of review Phishing. The majority of current research has focused on supervised learning methods, which require labeled data, a scarcity when it comes to online review Phishing. Research on methods for Big Data are of interest, since there are millions of online reviews, with many more being generated daily. To date, we have not found any papers that study the effects of Big Data analytics for review Phishing detection. The primary goal of this paper is to provide a strong and comprehensive comparative study of current research on detecting review Phishing using various machine learning techniques and to devise methodology for conducting further investigation.

2. Fast and effective clustering of Phishing URL' s based on structural similarity.

Phishing URL' s yearly impose extremely heavy costs in terms of time, storage space and money to both private users and companies. Finding and persecuting Phishingmer and eventual Phishing URL's stakeholders should allow to directly tackle the root of the problem. To facilitate such a difficult analysis, which should be performed on large amounts of unclassified raw URL's, in this paper we propose a framework to fast and effectively divide large amount of Phishing URL' s into homogeneous campaigns through structural similarity. The framework exploits a set of 21 features representative of the email structure and a novel categorical clustering algorithm named Categorical Clustering Tree (CCTree). The methodology is evaluated and validated through standard tests performed on three dataset accounting to more than 200k real recent Phishing URL's.

3. Cosdes: A collaborative Phishing detection system with a novel e-mail abstraction scheme.

E-mail communication is indispensable nowadays, but the e-mail Phishing problem continues growing drastically. In recent years, the notion of collaborative Phishing filtering with near-duplicate similarity matching scheme has been widely discussed. The primary idea of the similarity matching scheme for Phishing detection is to maintain a known Phishing database, formed by user feedback, to block subsequent near-duplicate Phishing On purpose of achieving efficient similarity matching and reducing storage utilization, prior works mainly represent each e-mail by a succinct abstraction derived from e-mail content text. However, these abstractions of e-mails cannot fully catch the evolving nature of Phishing, and are thus not effective enough in near-duplicate detection. In this paper, we propose a novel e-mail abstraction scheme, which considers e-mail layout structure to represent e-mails. We present a procedure to generate the e-mail abstraction using HTML content in e-mail, and this newly devised abstraction can more effectively capture the near-duplicate phenomenon of Phishing Moreover, we design a complete Phishing detection system Cosdes (standing for Collaborative Phishing Detection System), which possesses an efficient near-duplicate matching scheme and a progressive update scheme. The progressive update scheme enables system Cosdes to keep the most up-to-date information for near-duplicate detection. We evaluate Cosdes on a live data set collected from a real e-mail server and show that our system outperforms the prior approaches in detection results and is applicable to the real world

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

4. Apache Mahout: Scalable machine learning and data mining.
Mahout's goal is to build scalable machine learning libraries. With scalable we mean: Scalable to reasonably large data sets. Our core algorithms for clustering, classification and batch based collaborative filtering are implemented on top of Apache Hadoop using the map/reduce paradigm. How ever we do not restrict contributions to Hadoop based implementations: Contributions that run on a single node or on a non-Hadoop cluster are welcome as well. The core libraries are highly optimized to allow for good performance also for non-distributed algorithms. Scalable to support your business case. Mahout is distributed under a commercially friendly Apache Software license. * Scalable community. The goal of Mahout is to build a vibrant, responsive, diverse community to facilitate discussions not only on the project itself but also on potential use cases. Come to the mailing lists to find out more.
Currently Mahout supports mainly four use cases: Recommendation mining takes users' behavior and from that tries to find items users might like. Clustering takes e.g. text documents and groups them into groups of topically related documents. Classification learns from existing categorized documents what documents of a specific category look like and is able to assign unlabeled documents to the (hopefully) correct category. Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.

5. Comparative study on email Phishing classifier using data mining techniques.
In this e-world, most of the transactions and business is taking place through e-mails. Nowadays, email becomes a powerful tool for communication as it saves a lot of time and cost. But, due to social networks and advertisers, most of the URL's contain unwanted information called Phishing. Even though lot of algorithms has been developed for email Phishing classification, still none of the algorithms produces 100% accuracy in classifying Phishing URL's. In this paper, Phishing dataset is analyzed using TANAGRA data mining tool to explore the efficient classifier for email Phishing classification. Initially, feature construction and feature selection is done to extract the relevant features. Then various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. Finally, best classifier for email Phishing is identified based on the error rate, precision and recall.

III. SYTEM ANALYSIS

3.2 PROPOSED METHODOLOGY:

In the handling of electronic Phishing, it is a tougher job to segregate a huge burden of URL's in a recipient's inbox and preventing from the attack of Phishing URL's. It depends on the taste acceptance and the approach towards utilizing email conversations by an individual recipient. A Phishing for an ordinary person could be a ham for an authority or official who used to take actions against it. Some mails also may be sent by the control authorities or in a noble cause to aware people from Phishing could be classified as Phishing because the only reason it uses such Phishing words often.
In order to avoid these kinds of misclassification and also strictly prevent from attack of Phishing with less requirement of training the proposed methodology is derived. This methodology will utilize the probability of occurrence of several independent words in an email and their probability of Phishing and make conclusions out of it like whether the mail is Phishing or ham. Proposed methodology uses SVM classifier for classification purpose to make accurate decisions on a mail to be Phishing or ham. SVM works mainly to accomplish two purposes; one is to classify mails precisely into ham and Phishing URL's; second is to classify a mail according to the relative occurrence of words to specify ham or Phishing with the approach to make sure that none of the healthy mails for recipient should not specify as Phishing.
In general SVM classifier classifies set of objects based on training to identify what kind of data belongs to a certain category. If it finds similar while testing phase, then it will mark it up to that corresponding category. The basic work function of such NB classifier is described as follows in order to understand the fundamental classification.

HARDWARE/SOFTWARE DESCRIPTION

HARDWARE REQUIREMENTS:

- Processor : Pentium IV
- Hard Disk : 80 GB
- CPU : 512 MHz
- RAM : 1 GB

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

SOFTWARE REQUIREMENTS:

- OPERATING SYSTEM : WINDOWS XP
- FRONT END : JAVA

IV. SOFTWARE OVERVIEW

4. SYSTEM SPECIFICATION:

4.1 HARDWARE REQUIREMENTS:

- System : Intel 6.0.
- Hard Disk : 250 GB.
- RAM : 2 GB.
- Monitor : 14‟ Color Monitor.
- Mouse : Optical Mouse.

SOFTWARE REQUIREMENTS:

- Operating System
- Front end
- Database
- Window 8(64 bits)
- Python
- Dataset

4.2SOFTWARE DESCRIPTION:

Introduction:

Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and Unix systems and has also been ported to Java and .NET virtual machines. Python is a fairly old language created by Guido Van Rossum.

The design began in the late 1980s and was first released in February 1991. Python is considered a scripting language, like Ruby or Perl and is often used for creating Web applications and dynamic Web content. It is also supported by a number of 2D and 3D imaging programs, enabling users to create custom plug-ins and extensions with Python. Examples of applications that support a Python API include GIMP, Inkscape, Blender, and Autodesk Maya. Scripts written in Python (.PY files) can be parsed and run immediately.

They can also be saved as a compiled programs (.PYC files), which are often used as programming modules that can be referenced by other Python programs. In late 1980s, Guido Van Rossum was working on the Amoeba distributed operating system group.

Key Strengths

Python's clean object-oriented design and extensive support libraries offer two to tenfold the programmer productivity seen with languages like C, C++, C#, Java, VB, and Perl.

V. SYSTEM ARCHITECTURE

Phishing are more hostile for ordinary clients and unsafe additionally they cause the less efficiency, diminishing the transfer speed of system and costs organizations as far as part of cash. Hence, every business organization proprietor

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

who utilizes email must process keeping in mind the end goal to square Phishing from getting data by utilizing their email frameworks. Despite the fact that it might difficult to obstruct all Phishing sends, simply hindering a some of it will diminish the effect of its unsafe impacts. Keeping in mind the end goal to successfully sift through Phishing and garbage mail, the proposed framework can recognize Phishing from genuine messages and to do this it needs to distinguish run of the mill Phishing attributes and practices. These practices are known once to client, best standards and estimations can be utilized to hinder these messages.

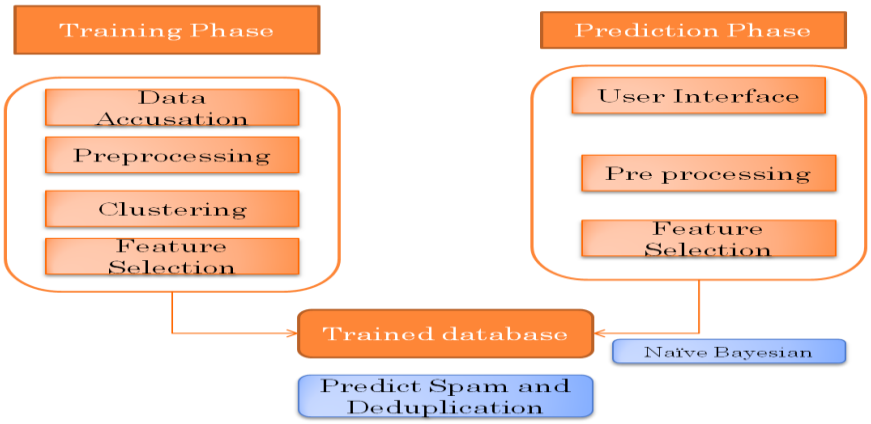


Figure 1

SYSTEM DESIGN

5.1 DATA FLOW DIAGRAM

A two-dimensional diagram that explains how data is processed and transferred in a system. The graphical depiction identifies each source of data and how it interacts with other data sources to reach a common output. Individuals seeking to draft a data flow diagram must identify external inputs and outputs, determine how the inputs and outputs relate to each other, and explain with graphics how these connections relate and what they result in. This type of diagram helps business development and design teams visualize how data is processed and identify .

CLASS DIAGRAM :

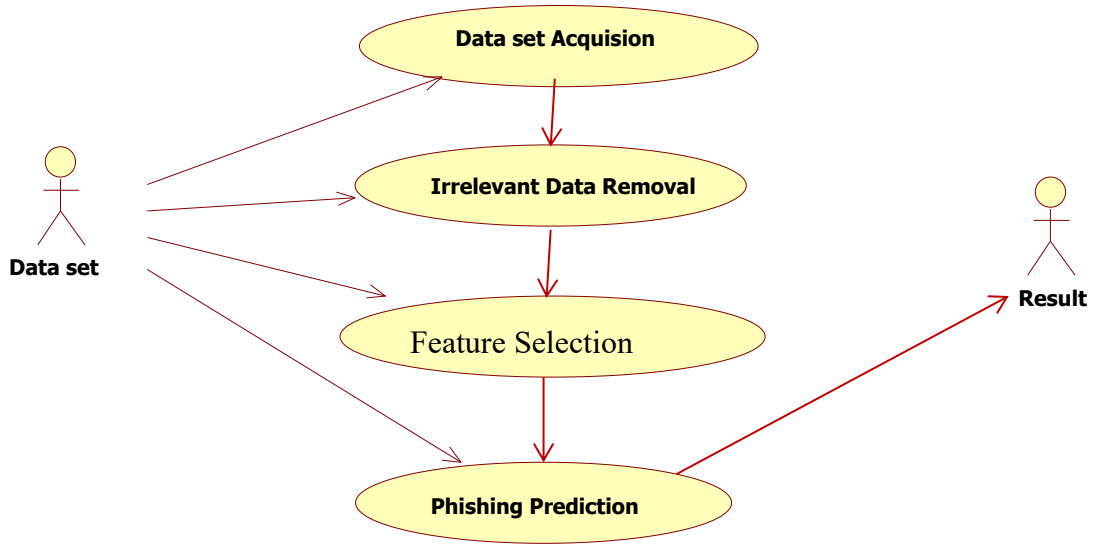


Figure 2

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

Sequence Diagram:

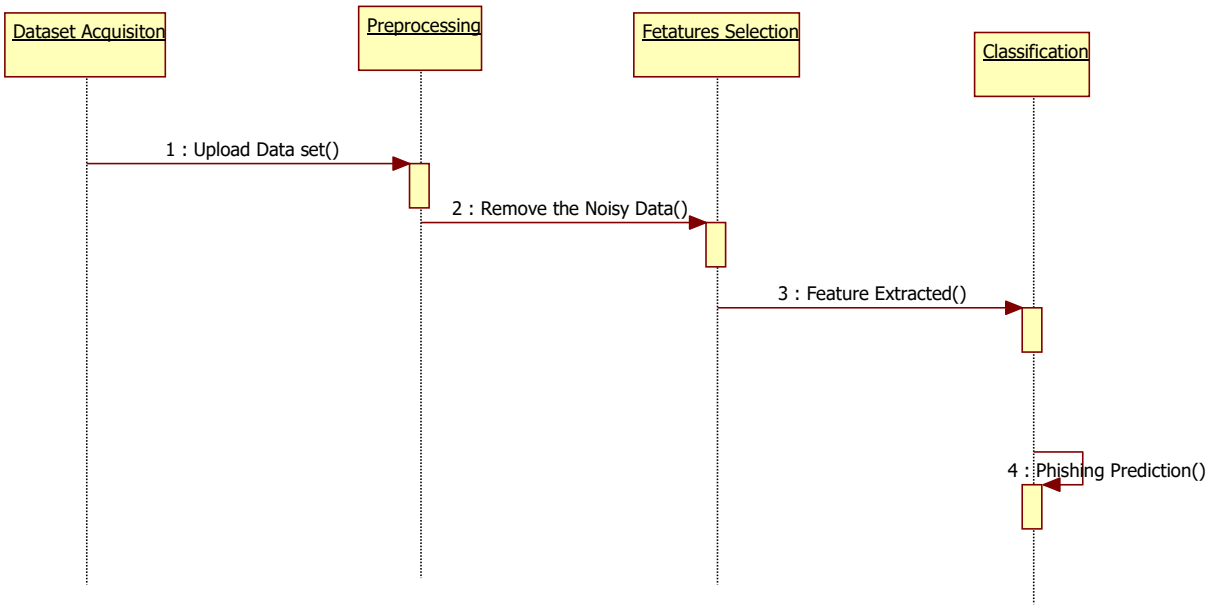


Figure 4

TO RUN THE WEB APPLICATION:

1. Open a web browser in any system connected to the local network to which the server is connected.
2. If the server's IP address is 192.168.1.121 or hostname is local host, then type the following in the URL of the browser:

http://localhost/Novel-Duplicate-Page/page_detect.php

3. The browser opens a web page containing the home page of the application with which the user can proceed.

VI. RESULT

DATA SET

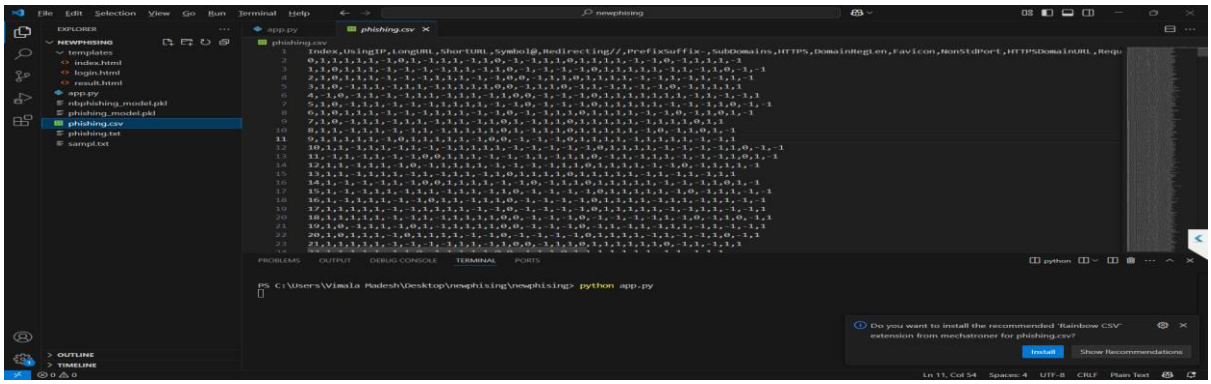


Figure 5

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

Home

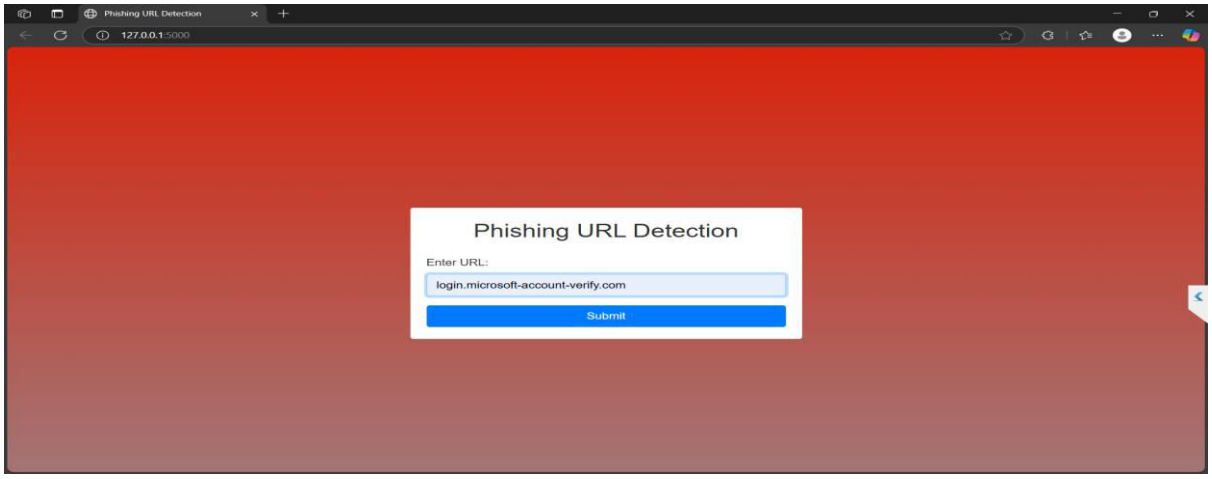


Figure 6

RESULT

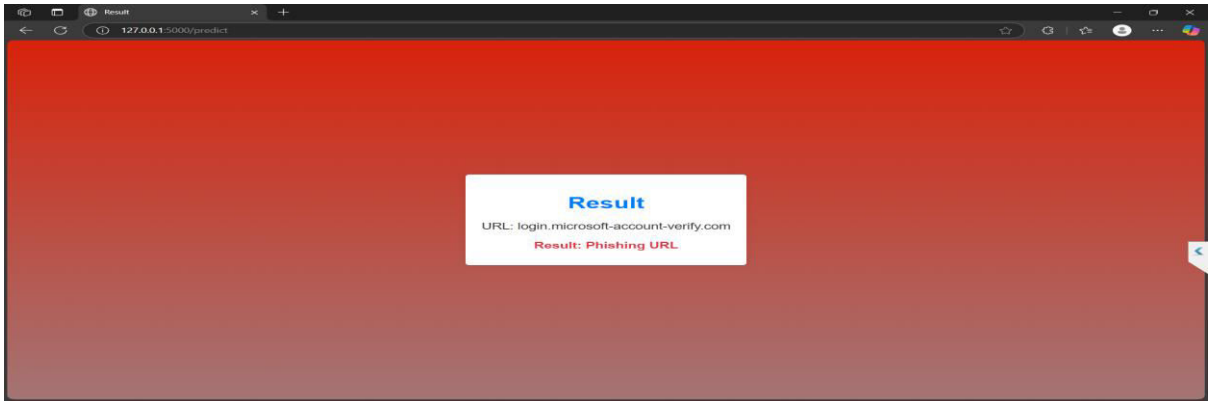


Figure 7

VII. CONCLUSION

SVM is an Phishing classifier which can capable of classifying with an average of 99.5% accuracy. Moreover, it requires a lesser amount of data for training and to give its standard performance with a very low training time of 3.5 seconds. So far from this study, it is inferred that SVM is a fast and reliable classifier because of its nature of relating independent probabilities of words in the content of an email. SVM gives out a new ethical approach of email classification with combining independent probabilities of consecutive words. In future, by improving the method for classifying unidentified or new words from a test email efficiently SVM can be more accurate in classification of phishing. And also by decreasing the total number of mails in dataset and maintaining the same accuracy will also help to reduce the build time of training dataset.

VIII. FUTURE ENHANCEMENT

Achieving accurate classification, with zero percent (0%) misclassification of Ham E-mail as Phishing and Phishing E-mail as Ham. The efforts would be applied to block Phishing E-mails, which carries the phishing attacks and now-days

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404440|

which is more matter of concern. Also, the work can be extended to keep away the Denial of Service attack (DoS) which has now, emerged in Distributed fashion called as Distributed Denial of Service Attack (DDoS).

REFERENCES

[1] I. Idris, and A. Selamat, "Improved email Phishing detection model with negative selection algorithm and particle swarm optimization," Applied Soft Computing, vol. 22, pp. 11-27, 2014.
[2] F. Gillani, E. Al-Shaer, and B. As Sadhan, "Economic metric to improve Phishing detectors," Journal of Network and Computer Applications, vol.65, pp. 131-143, 2016.
[3] M. Luckner, M. Gad, and P Sobkowiak, "Stable web Phishing detection using features based on lexical items," Computers & Security, vol. 46, pp. 79-93, 2014.
[4]. Jain, A., Gupta, P., Saran, H. K., Parmar, D. S., Bhati, J. P., & Rawat, D. (2024, November). Forecasting Future Sales Using Linear Regression Approach. In 2024 International Conference on Cybernation and Computation (CYBERCOM) (pp. 269-272). IEEE.
[5] S. Maldonado, and G. L'Huillier, "SVM-based feature selection and classification for email filtering," Pattern Recognition-Applications and Methods, Springer Berlin Heidelberg, pp.135-148, 2013.
[6]. Muniraju Hullurappa, Sudheer Panyaram, "Quantum Computing for Equitable Green Innovation Unlocking Sustainable Solutions," in Advancing Social Equity Through Accessible Green Innovation, IGI Global, USA, pp. 387-402, 2025.
[7] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three-way email Phishing filtering," Journal of Intelligent Information Systems, vol. 42, pp. 19- 45, 2014.
[8] M. Mohamad, and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in Phishing email classification," in International Conference of Computer, IEEE Communications, and Control Technology (I4CT), 2015, pp. 227-231.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details