Phishing Mail Detection

Submitted in the partial fulfillment of the requirements

for the degree of B.Tech in Computer Engineering

by

Ashutosh Kulkarni (22CE1139)

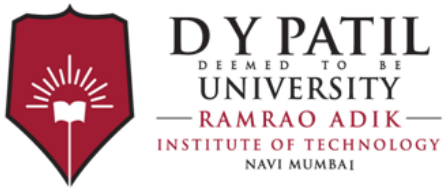
Aman Kumbhar (22CE1284)

Shreya Gupta (22CE1270)

Shaili Kakde (22CE1236)

Supervisor

Ms. Vaishali Jadhav



Department of Computer Engineering

Ramrao Adik Institute of Technology

Sector 7, Nerul, Navi Mumbai

(Under the ambit of D. Y. Patil Deemed to be University)

November 2024

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Institute of Technology

(Under the ambit of D. Y. Patil Deemed to be University)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706

CERTIFICATE

This is to certify that, the Mini Project-III report entitled

Phishing Mail Detection

is a bonafide work done by

Ashutosh Kulkarni (22CE1139)

Aman Kumbhar (22CE1284)

Shreya Gupta (22CE1270)

Shaili Kakde (22CE1236)

and is submitted in the partial fulfillment of the requirement for the degree of

B.Tech in Computer Engineering

to the

D. Y. Patil Deemed to be University

Supervisor	Project Co-ordinator
(Ms. Vaishali Jadhav)	**(Ms.Shweta Ashtekar)**
Head of Department	Principal
(Dr. Amarsinh V. Vidhate)	**(Dr. Mukesh D. Patil)**

akeapproval

Mini Project Report III Approval

This is to certify that the Mini Project -III Entitled ***"Phishing Mail Detection"*** is a bonafide work done by ***Ashutosh Kulkarni(22CE1139) Aman Kumbhar(22CE1284)***,***Shreya Gupta(22CE1270)***,and ***Shaili Kakde(22CE1236)*** under the supervision of ***Ms.Vaishali Jadhav***.This minirpoject is approved in the partial fulfillment of the requiremnt for the degree of ***BTech Computer Engineering***

Bachelor Degree In Computer Engineering,University of Mumbai

Mini Project Report - III Approval

This is to certify that the Mini Project - III entitled ***" Phishing Mail Detection"*** is a bonafide work done by ***AshutoshKulkarni (22CE1139)***, ***Aman Kumbhar(22CE1284)***, ***Shreya Gupta (22CE1270)***, and ***Shaili Kakde (22CE1236)*** under the supervision of ***Ms. Vaishali Jadhav***. This Mini Project is approved in the partial fulfillment of the requirement for the degree of ***B.tech in Computer Engineering***

Internal Examiner :

1. .....

2. .....

External Examiners :

1. .....

2. .....

Date : .../.../.....

Place : .....

DECLARATION

I declare that this written submission represents my ideas and does not invovle plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words have been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: _____

Ashutosh Kulkarni (22CE1139)

Aman Kumbhar (22CE1284)

Shreya Gupta (22CE1270)

Shaili Kakde (22CE1236)

Abstract

Phishing attacks are threats based on fraudulent communication, typically via email, in which cybercriminals attempt to entice and persuade a target by posing as a reliable individual or entity. Phishing detection techniques that extract highly representative features from the emails' text are an effective way to combat these threats because machine learning algorithms can be trained using these features to create models that can distinguish between phishing and authentic emails.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Phishing remains a prevalent and evolving cyber threat that targets sensitive user information through deceptive means. Despite advancements in detection and prevention techniques, the sophistication of phishing attacks continues to grow. Effective countermeasures require a combination of technological solutions and user education to mitigate the risks associated with phishing.

Our strategy is concentrated in a holistic procedure based on lemmatization, bag of words, latent dirichlet allocation, and powerful classification algorithms. After passing the e-mails to an array structure, a preprocessing step over the e-mail texts is executed. Next, it is conducted a lemmatization using the WordNet lexical database as a dictionary to obtain a semantic-based reduction. Then, it is extracted a document-term matrix, from the BoW model, that is used in two different fronts: directly as the classification algorithms features attributes (Method 1), and as the input for Latent Dirichlet Allocation, whose obtained topics are used to express new reduced features in function of their proportions in each message (Method 2). Finally, it feed the same algorithms with these two different sets of features, concluding each method

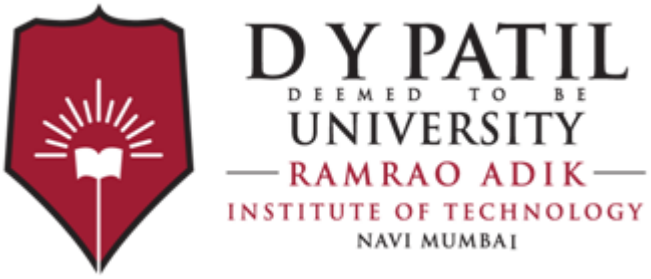


Figure 1.1: DY Patil Deemed University logo

The following Table 1 shows the Name of the Students.

Table 1.1: Name of Students

Name	**Roll No.**
Ashutosh Yogesh Kulkarni	22CE1139
Aman Kumbhar	22CE1284
Shreya Gupta	22CE1270
Shaili Kakde	22CE1236

1.1 Motivation

Addressing a gap in the field of spam email identification that has grown over time is the driving force behind this research project. Most of the existing solutions are falling behind. the innovative ideas that spammers consistently introduce, which strongly supports the development of machine learning-based anti-spam solutions. A number of these relatively recent solutions are critically assessed in this review article, which also offers suggestions on how to improve them further. In order to draw attention to the shortcomings and current situation, the article also examines several non-machine learning based frameworks that are now in use. The rise of machine learning-based anti-spam solutions is strongly justified by the fact that the existing solutions are largely falling behind the new ideas that spammers are continuously introducing.

1.2 Problem Statement and Objectives

The goal of the project is to develop a machine learning-based phishing email detection system that will enable users to promptly and precisely recognize phishing attempts. This tool determines whether an email is safe or potentially dangerous by looking for patterns in the sender information, links, and email content.

Collect and examine a range of emails to identify phishing-related trends.Build a machine learning model that can distinguish between ”safe” and ”phishing” emails.Create a web application that enables users to obtain a phishing risk report by uploading an email.Evaluate the model’s performance in terms of precision and dependability.

1.3 Organization of the report

The report is organised as follows: The Chapter 2 reviews the literature. Chapter 3 focuses on defining the system's issue. That includes problem categorization, proposed technologies, device architecture, and hardware/software requirements. On the other hand, Chapter 5 describes the inference and future work on the technique to be utilized as a more improved model.

Chapter 2

Literature Survey

1. A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques SAID SALLOUM 1,TAREK GABER 1,2, SUNIL VADERA 1,AND-KHALEDSHAALAN

This research paper provides a comprehensive overview of the existing research on phishing email detection methods using NLP and machine learning techniques. The purpose and scope The increasing threat of phishing is a significant issue with online security-losses run into billions of dollars annually. Phishing is more exclusively executed through e-mails; for the most part, it utilizes advanced techniques to create a false sense of security in making users divulge otherwise sensitive information. Based on this research, the authors have carried out a systematic literature review of articles published between 2006 and 2022 and have extracted and analyzed 100 research articles mainly in four areas of core research areas with different algorithms, text features with datasets, and evaluation criteria.Methodology: The authors performed an SLR on research articles published between 2006 and 2022 and found 100 relevant papers. The paper highlights the key areas of research, algorithms, text features, datasets, and criteria for the evaluation of phishing email detection using NLP. Common Techniques: Feature Extraction and classification are highly research able techniques along with tools like Support Vector Machines and NLP techniques such as TF-IDF and word embeddings.Datasets: Known datasets for benchmarking phishing detection methods are phishing corpus and the Spam Assassin Public Corpus. Challenges: There is limited research on detecting phishing emails in non-English languages, particularly Arabic. Implications: Research gaps emerge, including Arabic phishing detection and a development need toward sophisticated NLP techniques to keep abreast of the ever-changing phishing strategy.

2 **From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection**

EDER S. GUALBERTO THIAGO P. DE B. VIEIRA 1,RAFAEL T. DE SOUSA, JR. 1,(Senior Member, IEEE), 1,JOÃOPAULO C. L. DA COSTA AND CLÁUDIO G. DUQUE

This paper considers the following all literature review topics when it comes to phishing detection techniques as based on natural language processing and machine learning. The key points from the literature are mentioned below: Traditional Designs: Early phishing detection relied on metadata, blacklisting, and rule-based filtering. Traditional designs were found to be insufficient for advanced attacks. They were not very effective because invasion techniques can easily bypass them, and static rule-based systems are inherently shallow. Feature Engineering and Machine Learning: Recent approaches focus on enriching the distinguishing features of an email, including metadata and body text, to attain better detection rates. There is also promise with algorithms like Support Vector Machines and Random Forests while combining structural features with traditional NLP techniques such as TF-IDF and word embeddings to achieve high accuracy.Topic Modeling and Text Representations: It uses topic modeling (LDA, Latent Dirichlet Allocation) over the text sparsity/dimensionality issues. Effective feature representation was achieved by combining Bag of Words (BoW), TF-IDF with LSA, Latent Semantic Analysis. Such representations, machine learning classifiers, were really accurate in terms of both precision and recall, at a computational cost. The use of standardized datasets like Phishing Corpus and SpamAssassin to benchmark should be encouraged for making consistent comparisons among studies. Nonetheless, limiting dataset and possible biases remain a challenge; diverse and comprehensive datasets are called for.

3.Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism

Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism" covers the evolution of phishing detection methods and their limitations, leading to the development of more advanced techniques like deep learning. Key points include:Traditional Phishing Detection Techniques: Blacklist Mechanism: Early detection methods relied on blacklists, which required constant updates based on manually reported phishing sites. This approach was limited by high maintenance demands and ineffectiveness against new phishing attacks. **Advancements in Contextual and Semantic Analysis**

a.**RCNNs and Attention Mechanisms**: The RCNN model combines the advantages of CNNs (for feature extraction) and RNNs (for contextual understanding), which has shown to improve text classification performance. The attention mechanism further enhances model focus on critical parts of the email, such as unusual headers or language patterns that indicate phishing attempts.**Hierarchical and Multilevel Embedding**: The THEMIS model builds on these advancements, introducing multilevel embeddings at both the character and word levels to capture intricate details in email headers and bodies. This hierarchical approach better addresses the differences between legitimate and phishing emails, resulting in more accurate detection.

4.A Comprehensive Survey for Intelligent Spam Email Detection

ASIF KARIM , SAMI AZAM , BHARANIDHARAN SHANMUGAM , KRISHNAN KANNOORPATTI , AND MAMOUN ALAZAB

The paper **"Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism"** presents THEMIS, a model designed to improve phishing email detection by leveraging an enhanced Recurrent Convolutional Neural Network (RCNN) that integrates multilevel embeddings and an attention mechanism. This approach addresses limitations in traditional methods by focusing on complex semantic and contextual cues within email headers and bodies.

Summary of the Model's Features and Evaluation:

1. **Multilevel Embeddings**: THEMIS captures email structure at both the character and word levels, aiding in detecting phishing patterns in headers and content.

2. **RCNN with Bi-LSTM**: By using bidirectional LSTM layers, the model improves contextual understanding and reduces noise.

3. **Attention Mechanism**: THEMIS applies attention to the most relevant parts of emails, effectively enhancing classification accuracy.

Results: The model was tested on realistic datasets, achieving high accuracy (99.848%) and a low false positive rate (0.043%), outperforming standard deep learning models.

In summary, THEMIS demonstrates a robust approach to phishing detection with high precision, making it a valuable tool for improving email security.

2.1 Limitations of Existing System or Research Gap

Spurred by this research project was the fact that, time and again, there seemed to be a widening gap to be filled in identifying spam e-mail. Solutions developed until now are constantly lagging behind innovative ideas spammers constantly introduce-factual support for developing machine-learning-based anti-spam solutions.

That the new machine learning-based anti-spam solutions are well-supported by the fact that available solutions are very much lagging behind with new ideas spammers introduce for pursuit of their nefarious cause.

Chapter 3

Proposed System

Methodology/Techniques... The purpose of this research endeavor has guided the selection of the papers for review. We have examined a number of papers that were chosen based on the index terms listed. We have carefully examined the method that has been presented, whether machine learning principles have been used effectively, how reliable and significant the suggested solution is, and, lastly, the extent of modification needed to address any potential drawbacks. Only the works that demonstrate intelligent automation and a significant influence have been chosen for more study because they were thought to be promising. The purpose of the additional two parts, which cover a limited number of static and bio-inspired techniques, is primarily to illustrate the current status of email spam frameworks and the variety of research avenues.

3.1 Problem Statement

The goal of the project is to develop a machine learning-based phishing email detection system that will enable users to promptly and precisely recognize phishing attempts. This tool determines whether an email is safe or potentially dangerous by looking for patterns in the sender information, links, and email content

3.2 Proposed Methodology/Techniques

This survey report is organized so that the background information required for the research under analysis is covered first. The components of an email are described in Section II, along with how spammers use these components to create different kinds of email spam assaults

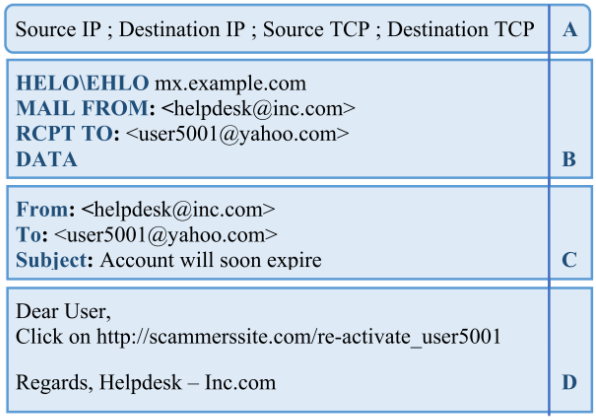


FIGURE 1. Email data parts.

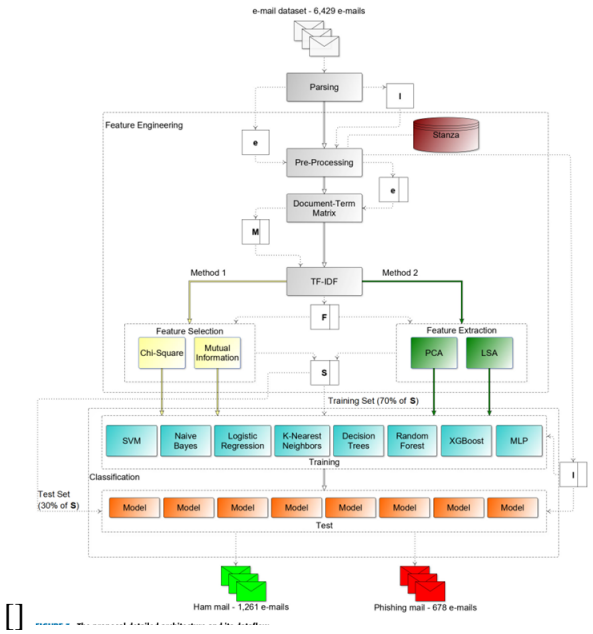
Figure 3.1: Email Data Part

on users. Section III will cover a number of general-purpose, non-AI-based spam detection systems and frameworks that do not rely on machine learning principles, even though the main goal of this review study is to assess machine learning-based solutions targeted at phishing and spoofing attacks. It is crucial to investigate such ideas in order to acquire a better understanding of our current position with spam emails and the need to integrate automated intelligence into both new and existing procedures.

3.3 System Design

The types of spam attacks and the anatomy of emails The annoyance of spam will eventually infiltrate practically every type of digital communication platform in use today. Among these, sending unsolicited emails has consistently been one of the most often used methods by scammers. This section shows the email's overall structure in detail as well as the several tactics the scammers used to attack it. Email data components are made up of many blocks, as seen in

2.Spear Phishing Spear phishing is an advanced targeted impersonation attack where all attacker impersonates a trusted one over email,that could be an employee or an external partner,often to extort money or funds ,deploy malware or steal credential. Four key thing about spear fishing is that **a.The target**:Someone entity that target their trusted employers,research their background and personal information **b.The Intent**:The attackers want the prey to do something which include wiring money to them,sharing credential or sensitive company files. **c.Sender Identity** they appear as one of the important person from the company or enterprise



[]

Figure 3.2: Email Data part explanation

so that their prey can fall into their trap.

Machine Intelligent platform are able to learn and adopt to threats by focusing on people relationships with employee and their work history

3.Whaling Whaling attack is a method used by cyber criminal to masquarade as the CEO or a senior executive or member with the aim of stealing money and sensitive information .These criminals use tricks like they craft mail as a trusted partner or senior member which often lead to malware installation and information leaked. We can reduce this by educating key individual of company or enterprise to ensure they are routinely on guard about possiblity of being targeted.They should be trained to look out for the tell-tale sign of an attack,such as spoofed email addresses and names . Executives should also learn to take special care when posting and sharing information online on social media sites,details about birthdays hobbies can all be use by cyber criminals to craft more sophisticated emails. Anti Phishing Software a.URL Screening , b.Link Validation Key to avoid phishing attack is by a.Educating people about whaling attack b.Robust verification process like two factor authentication c.Regular software update

ii NON-AI BASED CURRENT ANTI-SPAM SYSTEMS The majority of the popular anti-spam frameworks listed below are offered on many platforms, including standalone software applications and web-based solutions.

1.DOMAINKEYS IDENTIFIED MAIL Domain Key Identified mail that signs all the

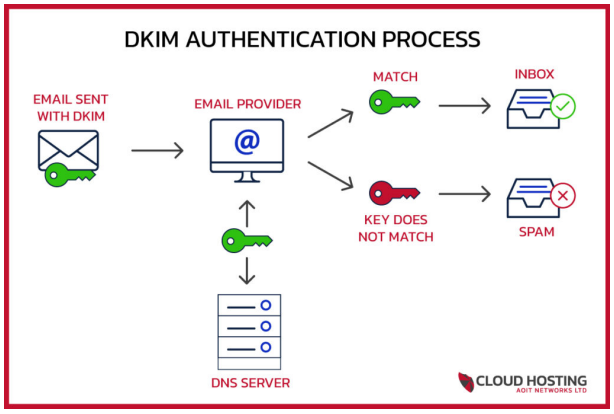


Figure 3.3: flow chart of Domain Key

outbound messages from a domain with a specific key which we will name as private key and then the corresponding public key is published to dns to record and then the receiving server can compare the two keys.

When an mail is sent it's going to take a list of fields and create a hash with content of those the email server then signs that hash with a private key which only the email server has this is important because that helps validate authenticity of email. When an email server receives an email it pulls down the public key from the dns entry and validate the hash ,if it matched it is legitimate.

2. **SENDER POLICY FRAMEWORK**

Sender Policy framework is a mechanism that tells servers receiving email what servers alowed to send mail on behalf of our domains.It uses an SPF record published in dns to do that .So when we are setting up spf we will publish our SPF Record to dns and then send this to the sending organisation server.When email was sent out from our domain the receiving server receives it and perform an spf record lookup by querying dns for the spf record in ascending domain to verify whether it came from authorised sending server and an Ip address. SPF usually contain list of ip address or dns that we autorised to have email sent on our behalf. For the record if there is no spf as specifies that the message came from source was unautorized then this domain rejects the mail.

3.4 Details of Hardware/Software Requirement

Anti phishing Software

a. Modulus Cloud: the generic IP filtering engines: It attempts to find look-alike domains

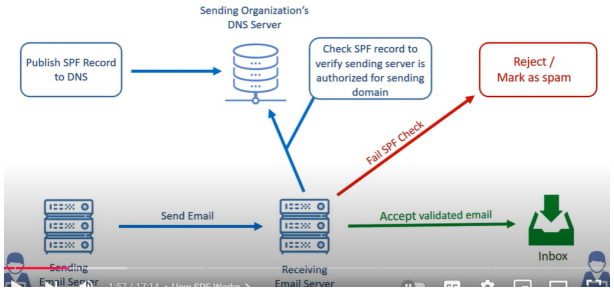


Figure 3.4: SPF

and emails targeting the user, imitating high-level corporate and executive names. For example, a domain with the letter 'l', typically represented by a small letter L, may imitate it by replacing the small letter with an I, the capital letter version. This is extended further by the software to give a "direct Quarantine" add-on which only Microsoft Outlook is able to use; this allows the user to view his quarantined files point-blank; which can be easily deleted with just one click of the mouse by the user. Using the software, users can directly forward emails to Vircom and instantly reclassify them. Emails that previously had got wrongly quarantined would be reclassified on the spot.

b Sword phish-Swordfish is actually a ReST-based API by EASYSOLUTIONS uniquely allowing the insertion of intel into already existing anti-fraud systems providing classifiers that make distinctions between good and bad domains. In principle, this technology existed for years irrespective of whether it's part of spam firewall, proxies, filters, etc. but such tools have always been behind. That is, it has learned from attacks that have already been done and is focused on blacklist. In place of this, Swordphish relies on predictive technology that utilizes its three unique classifiers programmed to understand and flag phishing and even domain generating malware. Without getting too technical, Sword phish is able to extract features from millions of domains, distinguishing between good and bad with zero further inspection of the domain in the DNS or requiring support in an external environment. Sword phish is extremely fast with an execution time of 10 milliseconds per search and a measured accuracy of 95

Anti phishing Hardware

a.**Web Titan Gateway** is a hardware filtering device, which, in addition to the usual phishing and fraud protection, comes with spyware and malware protection, an anti-virus, and inspection of your SSL/HTTPS. More importantly, it can be integrated into your corporate network either as a hardware appliance or software. Web Titan was developed by TitanHQ-a company that specializes in both software and hardware security-having been formed over twenty years

ago with the goal of responding to the needs of organizations with thousands of users that would seamlessly fit into the existing network. Amongst its options it offers are proxy cache, integration with your current directory, translucent proxy options, and automated updating and backups.

b.Protector P 500 is the other anti-phishing solution. Sec point developed it, an appliance that is automatically updated through several times a day to inhibit passing on phishing attacks. It provides different security modules dependent on your needs, from hardware for 50 users, considered suitable for a small network, 50-500 users in a middle-sized network, and more than 1000 users for corporate networking. Its anti-phishing system warns users in real time, and the appliance also includes spyware protection features. It further features a quick wizard setup to make installation very fast in any network.

Chapter 4

Results and Discussion

As previously announced, 47,107 are the amount of features employed in both methods before performing the feature selection or feature extraction techniques to dimensionality reduction, which is the same number of output terms from the lemmatization process, explained in subsection III-E

Method 1: Selection of Features: A Viewpoint Based on Mutual Information Measure The Mutual Information measure is employed as a dimensionality reduction strategy for this Method 1 viewpoint. You may find its prediction assessment values. A desired number of features is chosen based on this measure from the original features in DTM columns that have been weighted using TF-IDF. displays the outcomes obtained using this viewpoint and twenty-five attributes. The best result utilizing the Mutual Information measure in Method 1 was achieved in this configuration, with accuracy, precision, recall, F1 score, and speci city rates of 99.90%. The Random Forest ML classification algorithm is used to do this, and the entropy function is used to gauge the caliber of The Random Forest machine learning classification algorithm is used to

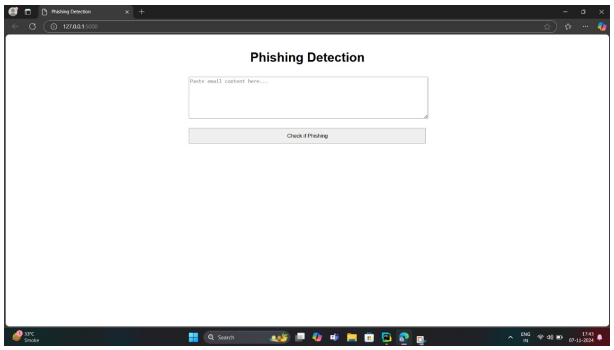


Figure 4.1: output

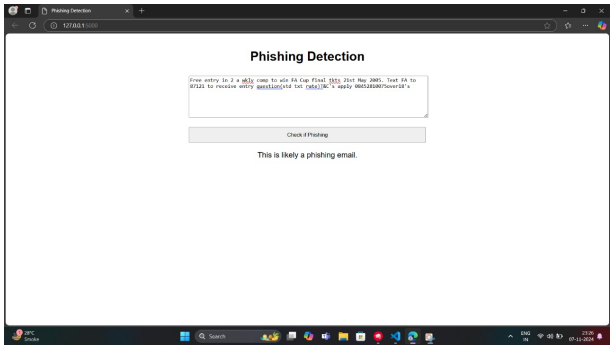


Figure 4.2: output2

accomplish this, with the entropy function used to gauge the quality of a split, five features to take into account when determining the optimal split, two samples as the bare minimum needed to split an internal node, and the remaining parameters set to default.

4.1 Implementation Details

Their equations are expressed below, in function of true positive (tp), false positive (fp), false negative (fn), and true negative (tn) rates. Accuracy (a): tp+tn/ tp+fp+tn+fn

Accuracy Precision p=tp/(tp+fp)

4.2 Result Analysis

FEATURE EXTRACTION-The main objective of this techniques is to extract new features from the original features set. It is expected these new features bring more distinctive information about the texts, with less noise. Following two techniques are used.

1. PCA (Principal Component Analysis) is a dimensionality reduction technique used to simplify data by converting a large set of features into a smaller set while retaining most of the original information. Important Key Points are as under. Finds new axes (principal components) that capture the most variance in the data. Reduces the number of features, making models faster and less prone to overfitting.Useful for visualizing high-dimensional data. PCA helps transform data into fewer dimensions while preserving as much of the underlying patterns as possible.

2. LSA (Latent Semantic Analysis) is a technique in NLP used to find hidden relationships between words and documents. It starts by creating a term-document matrix and uses

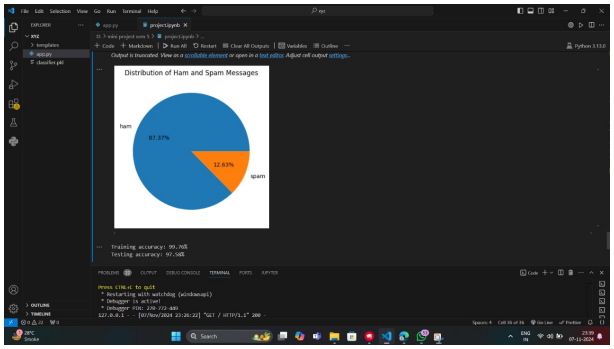


Figure 4.3: accuracy

Singular Value Decomposition (SVD) to reduce the matrix to a lower-dimensional space, capturing important concepts and patterns. Important Key Points are as under.

Helps identify synonyms and related terms.

Reduces the complexity of text data.

Improves information retrieval by understanding context. LSA is useful for tasks like document similarity and topic modelling but can be computationally intensive and sometimes lacks interpretability.

Chapter 5

Conclusion and Further Work

The types of spam emails and their effects on contemporary society and business were covered in the survey work reported in this research. A wide range of spam detection frameworks, including both normal non-automated and machine learning-based ones, have been rigorously examined to provide a comprehensive view of the field's present state and future prospects. It is anticipated that sufficient research will soon expand into the less-explored field of machine learning-based spam detecting techniques. The reviews make it very evident that the frameworks that are now being developed, despite their use of automated machine learning-based solutions, are frequently ill-prepared to handle the various ways that an email spam threat might propagate to create antispam software that, taking into account the many assault angles mentioned above, can combat different email spam kinds at once with only one software installation.

Following a comprehensive investigation, the study yields a number of distinct findings, particularly in the area of propositions based on machine learning. It is evident that supervised techniques are widely used, and the improved consistency of the model's performance is the primary cause of this. Additionally, it has been noted that several algorithms, including SVM and Naïve Bayes, are highly sought after. We also concluded that although single-algorithm anti-spam systems are rather prevalent, research into hybrid and multi-algorithm systems has a lot of promise. Additionally, there needs to be a significant increase in study on email header elements that do not include the "subject" field, URLs in the email body, and sender domain information.

Addressing "Concept Drift" is another crucial issue that requires more focus because it would undoubtedly enable a system to function at its best when spamming tactics and moti-

vations are gradually changed. Furthermore, the existing method of handling phishing-related spam emails is not as effective as stated; therefore, a more creative strategy that considers the various aspects of the issue is needed. It is concerning that the governments of several of the world's most powerful nations have failed to enact laws that will effectively address this problem in the long run, despite several warnings from various organizations. However, efforts to improve cybersecurity have gained more attention recently, leading to more research and more efficient

References

Appendices

Appendix A

Weekly Progress Report

Figure A.1: Weekly Progress Report

Appendix B

Plagiarism Report

Appendix C

Publication Details / Copyright / Project Competitions

References

References

4]S. Salloum and T. Gaber and S. Vadera and Khaled Sharan,IEEE Access,2022

Acknowledgments

We take this oppurtunity to express my profound gratitude and deep regards to my guide Ms.Vaishali Jadhav for her exemplary guidance ,monitoring and constant encouragement throughout the completion of this report.We are truly grateful to her efforts to improve our understanding towards various concepts and technical skills required in our project.The blessing,help and guidance give by her time to time shall carry us a long way in the journey of life on which were we are about embark .We take this privilage to express my sincere thanks to Dr.Mukesh.D.Patil,Principal RAIT DyPatil Deemed to be University .Much necessary facilities ,we are also thankful to Dr AV Vidhate ,HOD OF Computer engineering and Mrs Shweta ashtekar Miniproject co-ordinator.Last but not least we would also like to thank all those who have directly and indirectly helped us in completion of this project

Date: _____