OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

DETECTION OF PHISHING EMAIL BASED ON NLP AND ML TECHNIQUES

SHAIKH SAYEMA ANWER[1], ASST.PROF. V. KARWANDE[2]

ME Student, Department of Computer Science & Engineering, EESGOI , India[1]
HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI,India. [2]

Abstract: Spam email has become a major concern for customers and Internet providers alike. One of the main obstacles to its removal is that the proposed remedies need to have a very low, practical, false-positive rate. Natural Language Processing is a major subject of research in a range of areas. Text Classification is part of the NLP in which a range of approaches are used to transform text into a machine-readable format. Methods for categorizing texts include tokenization, speech-part tagging, stemming, and chunking. Using Scikit-Learn Classifiers to train the model to recognize spam and ham communications following the implementation of these data processes, offers us a categorized data set in which the model is trained to detect spam and ham messages. By exploring and comparing the relative capabilities of several machine learning approaches, we constructed a model for the problem of spam or ham transmissions. We are looking at the problem of classification in the context of naive Bayes, one of the most utilized machine learning models in the field of spam filtering. The performance metrics of the algorithms we employed in this study are comparable logically by Naive Bayes (NB) and Multinomial Naive Bays (MNB). The approach we offered in our 'Spam Mail Collection' data set showed an average accuracy of 98.49% utilizing the Naive Bayes (NB) model. The suggested system would compare the contents of the message with the contents of the spam keyword database, and the mail will be categorized as spam, if the contents match the contents of the database.

Keywords*: Naive Bayes (NB); Multinomial Naive Bayes (MNB); part-of-speech (POS);Natural Language Processing(NLP); Confusion Matrix(CM).*

I INTRODUCTION

The rapid progress of Internet technology has dramatically modified the experience of online users, while security issues have grown increasingly overpowering. Under the current conditions, new assaults can not only significantly impair the systems of customers, but also try to steal their money and identity. Phishing is a criminal activity that uses social engineering and technology to obtain identity information and account information of the victim. Spam is unwanted mass communications that are saved over their desires in user inboxes. The most prevalent senders of spam are advertisers, tricksters and scammers.

It is easy to assess whether or not a communication is spam. Our objective is to identify spam in order to find out the best suited approach by evaluating the accuracy of many algorithms. The typical strategies of employing the whitelists and blacklists do not work, because they can only prohibit the delivery of messages from a whole server (source) (false positives). Many people also receive hundreds of messages from unknown sources and spam blocked in inboxes. Spam refers to unwanted messages, whereas ham mails are crucial communications. We will develop a model that categorizes e-mails as spam or ham. In order to do this, data from messages must be gathered first, then techniques for natural language processing must subsequently be used. Spam filtering helps the mobile user to more clearly view their mailbox. Unnecessary communications are classed as spam, so that consumers don't waste time reading them. In this research we propose to classify data as either spam (unwanted) or ham in communications (wanted).

Phishing is a sort of comprehensive fraud occurring when a malicious website is acting like a legitimate one, such as passwords, account points or MasterCard numbers.

Although there are a number of contradictions to the programming of phishing and methods for distinguishing potential phishing activities on messages and the identification of phishing substances on sites, phishers are considering new and half-race strategies to address accessible programming and systems. Phishing is a deceit technique which employs a mixture of social design and innovation to compile sensitive and individual data, such as passwords and the charging of subtle aspects by incorporating into an electronic correspondence the look of a reliable person or firm. Phishing uses spoof communications that are created genuine and indicated to originate from honest to reputable sources such as money-related foundations, e-commerce destinations, etc. to entice customers to visit false websites by providing phishing e-mails.

II LITERATURE SURVEY

In this research by looking at the email's structure. Then, using an upgraded Recurrent Convolutional Neural Networks (RCNN) model with multilevel vectors and an attention mechanism, they introduced THEMIS, a novel phishing email detection model that models emails at the email header, email content, character level, and word level all at the same time. We utilise an unbalanced dataset with actual phishing and legitimate email ratios to assess the effectiveness of THEMIS. THEMIS' total accuracy is 99.848 percent, according to experimental results. The false positive rate (FPR) is currently 0.043 percent. The filter's excellent accuracy and low FPR ensure that phishing emails are detected with high probability while real emails are filtered out as little as possible. This promising finding outperforms existing detection approaches and validates THEMIS's ability to detect phishing emails [1].

In this research Phishing emails are phishing emails that contain illegal links that direct users to spoofed pages of real websites, or pages where real HTML has been inserted with malicious HTML code, in order to steal personal information such as bank or credit card account numbers, email account numbers, and passwords. The most vulnerable aspect of security is people. Human weaknesses are exploited in phishing emails. This article outlines how the persuasion principle is used in phishing emails, and it suggests a phishing email detection method based on the persuasion principle based on existing approaches. The persuading principle is to count how many times the feature's related word appears in the mail. An information gain approach is used to choose the feature, and then 25 features are chosen for detection. Finally, the accuracy rate was tested and found to be 99.6% [2].

In this research, With the increased use of mobile devices in recent years, there has been a growing trend to shift nearly all real-world functions to the cyberworld. Although this makes our daily lives easier, it also leads to numerous security breaches due to the Internet's anonymous structure. Most attacks may be avoided with the use of antivirus software and firewall systems. Experienced attackers, on the other hand, try to exploit computer users' vulnerabilities by phishing them with fake websites. These pages impersonate popular banking, social media, e-commerce, and other websites in order to steal personal data such as user IDs, passwords, bank account, and credit card numbers, among other things. Phishing detection is a difficult problem, and many alternative methods, such as a blacklist, rule-based detection, anomaly-based detection, and so on, have been suggested in the market. Due to its dynamic structure, recent works tend to apply machine learning-based anomaly detection, notably for spotting "zero-day" attacks, according to the literature. We suggested a machine learning-based phishing detection system in this paper, which used eight different algorithms to evaluate URLs and three distinct datasets to compare the results to previous work. The experimental results show that the proposed models operate exceptionally well, with a high success rate [3].

In this research, In recent years, phishing websites have become a major cybersecurity issue. Spam, malware, ransomware, drive-by vulnerabilities, and other malicious software can be found on phishing websites. Many times, a phishing website will look exactly like a well-known website, luring an unsuspecting visitor into the trap. The victim of the fraud suffers a monetary loss, as well as the loss of personal information and reputation. As a result, it's critical to identify a solution that can quickly neutralise such security vulnerabilities. Blacklists have traditionally been used to detect phishing websites. Many notable websites, such as PhisTank, host a list of blacklisted domains. The blacklisting strategy has two flaws: it may not be comprehensive, and it does not detect a freshly created phishing website. Machine learning techniques have been utilised to classify and detect phishing websites in recent years. In this study, we examined multiple machine learning techniques for phishing URL categorization and found that the Nave Bayes Classifier with precision=1, recall=.95, and F1-Score=.97 achieved the maximum accuracy of 98 percent. [4].

In this research, Phishing assaults, which are focused on social engineering and malware, are a type of cybercrime in today's world. It is one of the most dangerous risks that every person and organisation has to deal with. Users find information on the internet using URLs, which are also known as web links. The review raises readers' knowledge of phishing assaults, helps them spot them, and encourages them to practise phishing prevention. In phishing, phishers use email or text messages as a weapon to fool individuals or organisations by sending URL links to them. Companies and individuals are unable to detect all phishing emails or messages due to the large number of them received every day. Various reviews are presented here for detecting phishing attacks using machine learning. It is used to determine whether the web links are phishing or real [6].

In this research, Phishing is a type of cyber-attack that aims to get personal information, account information, credit card information, organisational information, or a user's password in order to execute transactions. Phishing websites seemed to prefer legitimate ones, making it difficult to distinguish between them. The goal of that study is to perform ELM based on 30 primary components that are categorised utilising a machine learning methodology. To avoid being discovered, most phishing URLs use HTTPS. Phishing on the internet can be detected in three ways. The first strategy assesses different elements of the URL, the second technique analyses the authority of a website and calculates whether the website is introduced or not, as well as who is supervising it, and the third technique verifies the website's authenticity [9].

In this research, Phishing is a technique in which someone pretends to be a trustworthy entity in order to collect personal and sensitive information such as login ids, passwords, and credit/debit card information for nefarious purposes. Many websites that appear to be trustworthy to us can be phishing sites and be the source of different internet scams. Phishing websites may try to collect our personal information in a variety of ways, including phone calls, text messages, and pop-up windows.As a result, the need of the hour is to safeguard information communicated over the internet, and one specific approach to do so is to combat phishing assaults. The focus of this study is on various Machine Learning techniques for detecting if a website is phishing or not. Machine learning solutions are favoured because they can detect zero-hour phishing assaults and are better at handling new forms of phishing attempts. In our testing, we were able to predict if a website was phishing or authentic with an accuracy of 98.4 percent [11].

In this research, Phishing is one of the most dangerous social engineering tactics for enslaving end users and gaining access to crucial information systems. The use of e-mail communication with an embedded hyperlink is a frequent phishing technique. Due to the intricacy of today's phishing attempts, detecting and mitigating phishing attacks is a major difficulty. In terms of detection and mitigation time, existing solutions are frequently too time consuming to be applied in the actual world. Similarly, due to the dynamic nature of phishing attempts, they use static detection methods that are ineffective in the actual world. We offer Phish Limiter, a new detection and mitigation methodology in which we offer a novel deep packet inspection (DPI) technology and then combine it with software-defined networking (SDN) to detect phishing activities via e-mail and web-based communication. Phishing signature categorization and real-time DPI are two components of the suggested DPI methodology. We use an artificial neural network model to classify phishing attack signatures and design the real-time DPI based on the programmability of SDN to develop the store and forward mode and the forward and inspect mode to direct network traffic, so that Phish Limiter can flexibly address the dynamics of phishing attacks in the real world. Because it offers a global view of a network through SDN, PhishLimiter also provides superior network traffic control for containing phishing attempts. Furthermore, we put PhishLimiter to the test in a real-world testbed setting with real-world email containing embedded URLs. PhishLimiter delivers an effective and efficient way to deter harmful activity, according to our lengthy experimental study [13].

III. SYSTEMS ARCHITECTURE

The system model architecture presented includes Natural Language Processing is a major subject of research in a range of areas. Text Classification is part of the NLP in which a range of approaches are used to transform text into a machine-readable format. Methods for categorising texts include tokenization, speech-part tagging, stemming, and chunking. Using Scikit-Learn Classifiers to train the model to recognise spam and ham communications following the implementation of these data processes, offers us a categorised data set in which the model is trained to detect spam and ham messages. By exploring and comparing the relative capabilities of several machine learning approaches, we constructed a model for the problem of spam or ham transmissions. We are looking at the problem of classification in the context of naive Bayes, one of the most utilised machine learning models in the field of spam filtering. The performance metrics of the algorithms we employed in this study are comparable logically by Naive Bayes (NB) and Multinomial Naive Bays (MNB).

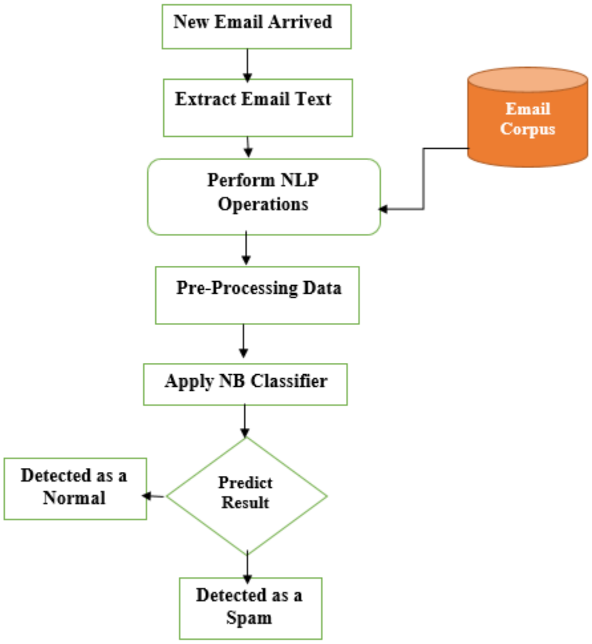

Figure No 3.1: System Architecture

The architecture above was The methodology we supplied with the Spam Mail Collection data set averaged 98.49 percent utilizing the Naive Bayes (NB) model. The system will examine the text of the message using the spam keyword database, and the mail will be categorized as spam if the content of the email corresponds to the content in the data base.

IV EXPERIMENTAL RESULTS

The We balanced the dataset before using the spam model such that train and test data sets contain class labels that are evenly distributed. We divide the data into 70% for the training dataset and 30% for the test dataset. During the model training, a validation dataset was randomly picked as 10% of the training data set. Three baseline classification models were compared to the suggested approach, Naive Bayes. We assessed the performance of separate models to categories each data set and also defined parameters in order to present its optimum performance. NB model with regularization parameter 1 and a linear kernel was reported. The Naive Bayes model with a smoothing parameter has been published. Fig displays.

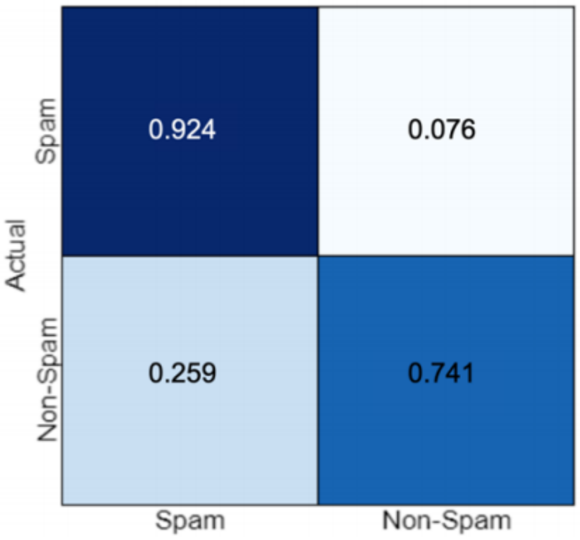

Figure No 4.1: Performance of the Independent Model Confusion Matrix for Spam Mail.

V CONCLUSION

The proposed system Natural Language Processing is an essential branch of research with applications in numerous fields. The previously collected Spam Mails Collection is utilized as a dataset and a class is predicted and provided for each input in the collection as output. They must first be appropriately labelled before using algorithms in the mail messages. Using different classifiers, we can decide which approaches are best and worst for a certain job. The Naive Bayes(NB) classification algorithm was more effective and successful compared to other methods. As a result, it is often used in spam filters with success rates of up to 98 percent. This shows that the classifier of Naive Bayes (NB) are the best successful approaches for spam detection in the inbox.

REFERENCES

1. Yong Fang, Cheng Zhang, Cheng Huang, Liang Liu, And Yue Yang,"Phishing Email Detection Using Improved RCNN Model with Multilevel Vectors and Attention Mechanism",IEEE Access,2018.

2.Xue Li, Dongmei Zhang, Bin Wu,"Detection method of phishing email based on persuasion principle",4th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC),IEEE,2020.

3.Mehmet Korkmaz,Ozgur Koray Sahingoz,Banu Diri,"Detection of Phishing Websites by Using Machine Learning-Based URL Analysis", IEEE Xplore,2020.

4.Jitendra Kumar,A. Santhanavijayan,B. Janet,Balaji Rajendran and Bindhumadhava BS,"Phishing Website Classification and Detection Using Machine Learning",International Conference on Computer Communication and Informatics (ICCCI),Jan. 22-24, 2020.

5.Weiheng Bai,"Phishing Website Detection Based on Machine Learning Algorithm", International Conference on Computing and Data Science (CDS),IEEE,2020.

6.Charu Singh,Smt.Meenu,"Phishing Website Detection Based on Machine Learning: A Survey", 6th International Conference on Advanced Computing & Communication Systems (ICACCS),IEEE,2020.

7.Smita Sindhu,Sunil Parameshwar Patil,Arya Sreevalsan,Ms. Saritha A. N. and Faiz Rahman,"Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation",IEEE,2020.

8.Amani Alswailem,Norah Alrumayh,Bashayr Alabdullah and Dr.Aram Alsedrani,"Detecting Phishing Websites Using Machine Learning",IEEE,2019.

9.Mahajan M. V.,Kakade P. G.,Sawant P. J. and Pawar Shila,"Detection of Phishing Website Using Machine Learning Approach",4th International Conference on Electrical,Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT),IEEE,2019.

10. Mishra Alekh K., Tripathy A. K. and Satyabrata Swain,"Analysis and Prevention of Phishing Attacks in Cyber Space", First International Conference on Secure Cyber Computing and Communication (ICSCCC),IEEE,2018.

11.IshantTyagi, Jatin Shad, Shubham Sharma,IshantTyagi, Jatin Shad, Shubham Sharma,"A Novel Machine Learning Approach to Detect Phishing Websites",5th International Conference on Signal Processing and Integrated Networks (SPIN),IEEE,2018.

12.Vaibhav Patil,Pritesh Thakkar,Chirag Shah,Tushar Bhat and Prof.S. P. Godse,"Detection and Prevention of Phishing Websites using Machine Learning Approach",IEEE,2018.

13.Tommy Chin, Kaiqi Xiong , And Chengbin Hu,"Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking",IEEE Access,2018.

14.Ebubekir Buber, Onder Demir,Ozgur Koray Sahingoz,"Feature Selections for the Machine Learning based Detection of Phishing Websites", IEEE,2018.

15.Neda Abdelhamid,Fadi Thabtah and Hussein Abdel jaber ,"Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features",IEEE,2017.

16.Dea Delvia Arifin,Shaufiah and Moch. Arif Bijaksana,"Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier",IEEE,2016.

17.Priyanka Singh,Yogendra P.S. Maravi, Sanjeev Sharma,"Phishing Websites Detection through Supervised Learning Networks",IEEE,2015.

18.Yanhui Du ,Fu Xue,"Research of the Anti-Phishing Technology Based on E-mail Extraction and Analysis",International Conference on Information Science and Cloud Computing Companion,IEEE,2014.

19.Joby James,Sandhya L.,Ciza Thomas,"Detection of Phishing URLs Using Machine Learning Techniques", International Conference on Control Communication and Computing (ICCC),IEEE,2013.