PHISHING EMAIL DETECTION USING CNN

Dr. PRABHAKARA R UYYALA
Executive Delivery Manager & Architect ǀ IBM USAǀ & ǀ Research Scholar ǀ COMPUTER SCIENCE AND ENGINEERING ǀ JS UNIVERSITY, UP ǀINDIA

ABSTRACT: The phishing email is one of the significant threats in the world today and has caused tremendous financial losses. Although the methods of confrontation are continually being updated, the results of those methods are not very satisfactory at present. Moreover, phishing emails are growing at an alarming rate in recent years. Therefore, more effective phishing detection technology is needed to curb the threat of phishing emails. In this paper, we first analyzed the email structure. Then based on an improved Convolutional Neural Networks (CNN) model with multilevel vectors and attention mechanisms, we proposed a new phishing email detection model named, which is used to model emails at the email header, the email body, the character level, and the word level simultaneously. To evaluate the effectiveness of, we use an unbalanced dataset that has realistic ratios of phishing and legitimate emails. Experimental results show that the. Meanwhile, this ensures that the filter can identify phishing emails with high probability and filter out legitimate emails as little as possible. This promising result is superior to the existing detection methods and verifies the effectiveness of detecting phishing emails.

I.INTRODUCTION: Phishing is a lucrative form of fraud, in which fraudulent recipients cheat and receive confidential information under false pretenses. Fisher emails will guide users click on the link or link on the website where you need to provide confidential information, such as passwords, credit card information. Recent advances in web technology have attracted most business and technology companies, including banks, to offer their services online. As people rely on Internet services to do their transactions, cyber fraud can become a major threat to people's privacy and security. Disclose personal information that the fraudster uses to gain unauthorized access to the user's account. For example, fraudulent email sent to a user may contain malware, in the form of malware plug-ins or email attachments. If this user downloads a link to the PC, the malware will install itself on the desktop and it will transfer money to the fraudster's bank account whenever the user tries to make an online transaction.

Phishing attacks use email messages and websites to be similar to legitimate company emails and websites and to force users to disclose their personal or financial information. An attacker may use sensitive user information for good. Users may be tricked into publishing their information-sensitive information through a web form, responding to fraudulent emails, or downloading and installing Trojans that search users 'computers or track users' online activities.

Fraudulent activities are increasing daily, and victims of the past are now seeking ways to protect themselves from being attacked again. To achieve this, they need to further protect their security mechanism, which implies that the existing security system is greatly improved. The system can identify fraudulent activities and prevent them from happening.

Phishing is a fraudulent attempt to disguise itself as a trusted company and obtain sensitive information. These strategies for conflict are constantly being updated, and the results of those strategies are currently not very good. We use the Naive Bayes theorem to analyze emails and find a solution to this problem. The rapid progression of Internet innovation has greatly changed the understandingofon-linecustomers.Thecurrentsituationisthatthent hreatsareaimedatnotonlycausing serious damage to customers' computers but also stealing their money and their identity.

II.EXISTING SYSTEM: Various techniques for detecting phishing emails are mentioned in the literature. In the entire technology development process, there are mainly three types of technical methods including blacklist mechanisms, classification algorithms based on machine learning and based on deep learning. From previous work, the existing detection methods based on the blacklist mechanism mainly rely on people’s identification and reporting of phishing links requiring a large amount of manpower and time. However, applying artificial intelligence to the detection method based on a machine learning classification algorithm requires feature engineering to manually find representative features that are not conducive to the migration of application scenarios. Moreover, the current detection method based on deep learning is limited to word embedding in the content representation of the email. These methods directly transferred natural language processing (NLP) and deep learning technology, ignoring the specificity of phishing email detection so that the results were not ideal Given the methods mentioned above and the corresponding problems, we set to study phishing email detection systematically

In the vast landscape of cybersecurity, phishing remains one of the most prevalent and insidious threats, targeting individuals, businesses, and organizations alike. To combat this ever-evolving menace, researchers and practitioners have explored a myriad of techniques for identifying and thwarting phishing attempts.

A comprehensive review of the existing literature reveals a multitude of strategies deployed in phishing email detection. Amidst the technological advancements, three primary technical methodologies have emerged as prominent contenders: blacklist mechanisms, classification algorithms rooted in machine learning principles, and those leveraging the intricate architecture of deep learning.

Blacklist mechanisms, a traditional yet effective approach, rely heavily on human intervention for the identification and reporting of suspected phishing links. However, this method presents inherent challenges, demanding significant manpower and time to maintain and update the blacklist database, rendering it somewhat labor-intensive and time-consuming.

In contrast, the integration of artificial intelligence (AI) into machine learning-based classification algorithms offers a promising avenue for automated phishing detection. Yet, this approach requires meticulous feature engineering to identify relevant characteristics within phishing emails, limiting its scalability and adaptability across diverse application scenarios.

Meanwhile, the emergence of deep learning has ushered in a new era of cybersecurity, with its ability to extract intricate patterns and features from complex datasets. However, current deep learning-based detection methods often focus solely on word embedding techniques within email content representation, overlooking the nuanced attributes specific to phishing emails. This oversight compromises the efficacy of the detection process, resulting in suboptimal outcomes.

In light of the inherent limitations and challenges posed by existing methodologies, our research endeavors to undertake a systematic exploration of phishing email detection. Through a comprehensive and interdisciplinary approach, we aim to bridge the gap between theoretical insights and practical implementations, thereby enhancing the efficacy and resilience of phishing detection mechanisms in the digital age of cyber security. based on deep learning. Specifically, this paper makes the following contributions.

Disadvantages:

With respect to the particularity of the email text, we analyze the email structure, and mine the text features from four more detailed parts: the email header, the email body, the word-level, and the char- level.

Noise is introduced as little as possible, and the context information of the email can be better captured.

III.PROPOSED SYSTEM: With the emergence of email, the convenience of communication has led to the problem of massive spam, especially phishing attacks through email, which studied the effectiveness of phishing blacklists. Blacklists mainly include sender blacklists and link blacklists. This detection method extracts the sender's address and link address in the message and checks whether it is in the blacklist to distinguish whether the email is a phishing email. The update of a blacklist is usually reported by users, and whether it is a phishing website or not is manually identified. To some extent, the perfection of the blacklist determines the effectiveness of this method based on the blacklist mechanism for phishing email Detection. The current situation is that new threats may not only cause severe damage to customers' computers but also aim to steal their money and identity. Among these threats, phishing is a noteworthy one and is a criminal activity that uses social engineering and technology to steal a victim's identity data and account information. According to a report from Anti-Phishing Working compared with the fourth quarter of According to the striking data, it is clear that phishing has shown an apparent upward trend in recent years. Similarly, the harm caused by phishing can be imagined as well.

Advantages:

Phishing email refers to an attacker using a fake email to trick the recipient into returning information such as an account password to a designated recipient.

Additionally, it may be used to trick recipients into entering special web pages, which are usually disguised as real web pages, such as a bank's web page, to convince users to enter sensitive information such as a credit card or bank card number and password. Although the attack of phishing email seems simple, its harm is immense.

IV. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

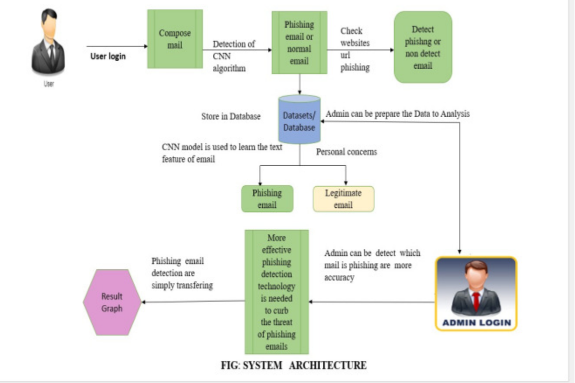

User
User login
Compose mail
Detection of CNN algorithm
Phishing email or normal email
Check websites url phishing
Detect phishng or non detect email
Store in Database
Datasets/ Database
Admin can be prepare the Data to Analysis
CNN model is used to learn the text feature of email
Personal concerns
Phishing email
Legitimate email
More effective phishing detection technology is needed to curb the threat of phishing emails
Admin can be detect which mail is phishing are more accuracy
Phishing email detection are simply transfering
Result Graph
ADMIN LOGIN
FIG: SYSTEM ARCHITECTURE

fig 4.1 System Architecture

4.2 DATA FLOW DIAGRAM

The DFD is also called a bubble chart. It is as simple graphical formalism that can be used to represent a system in terms of input data to the system various process in carried out on this data, and the output data is generated by this system

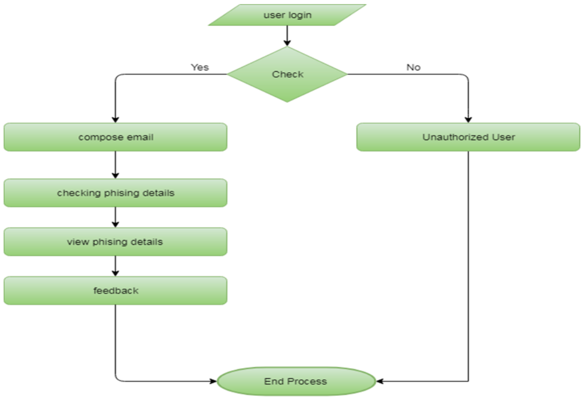


Fig 4.2 Data Flow diagram :(User)

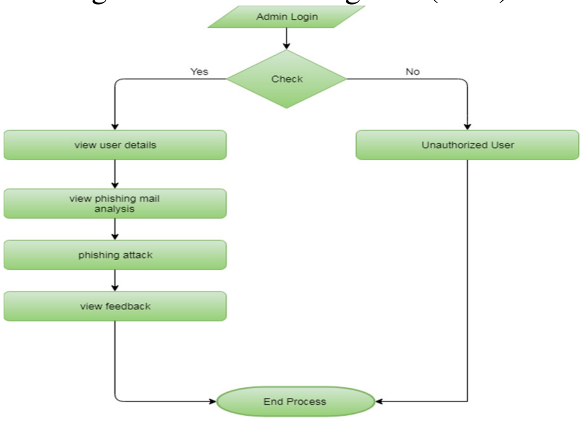


Fig 4.3 Data Flow diagram :(Admin)

4.3 ER DIAGRAM

ER Diagram stands for Entity Relationship Diagram also known as ERD is a diagram that displays the relationship of entity sets stored in a database. …ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes and diamond shapes to represent relationships An ER MODEL is usually the result of systematic analysis to define and describe what is important to processes in an area of business. It does not define the business processes: it only presents a business data scheme in graphical form. It is usually drawn in graphical form as boxes that are connected by lines which express the associations and dependencies between entities. an er model can also be expressed in a verbal form for example: one building. Entities may be characterized not only by relationships, but also by additional properties, which include identifiers called "Primary Keys".

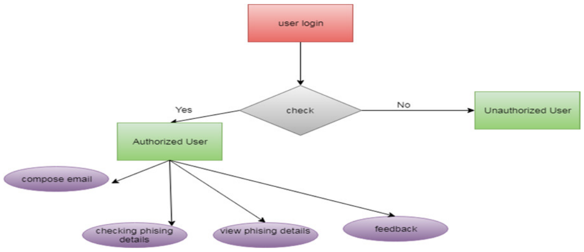


Fig 4.4 ER diagram :(User)

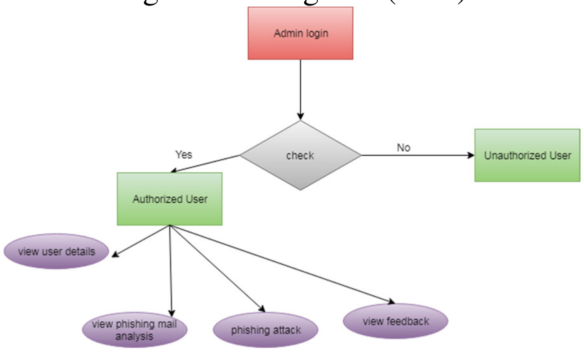


Fig 4.5 ER diagram :(Admin)

Input Design:

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining privacy. Input Design considered the following things:

What data should be given as input?

How should the data be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when errors occur.

Objectives:

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

It is achieved by creating user-friendly screens for the data entry to handle large volumes of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data can be performed. It also provides record viewing facilities.

When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize instant. Thus the objective of input design is to create an input layout that is easy to follow

Output Design:

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other systems through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source of information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

Select methods for presenting information.

Create documents, reports, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

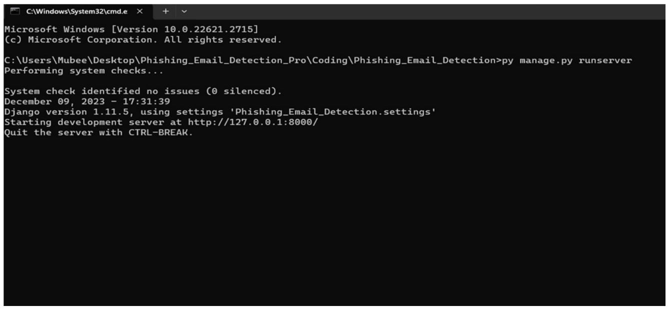
Convey information about past activities, current status or projections of the Future.

Signal important events, opportunities, problems, or warnings.

Trigger an action.

Confirm an action.

V.RESULTS:



```
Microsoft Windows [Version 10.0.22621.2715]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Mubee\Desktop\Phishing_Email_Detection_Pro\Coding\Phishing_Email_Detection>py manage.py runserver
Performing system checks...

System check identified no issues (0 silenced).
December 09, 2023 - 17:31:39
Django version 1.11.5, using settings 'Phishing_Email_Detection.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

Fig:5.1 Code Executing Page

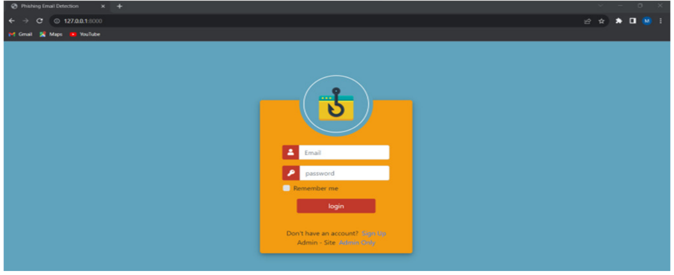


Fig:5.2 Home Page

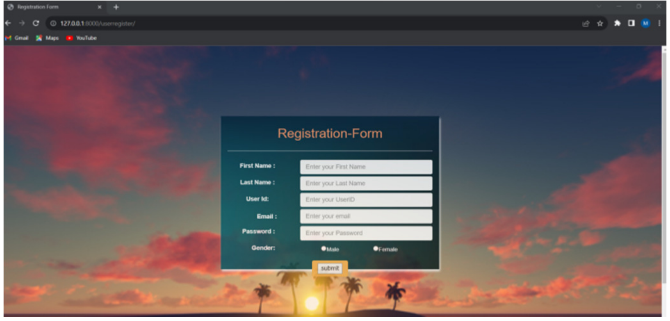


Fig: 5.3 User Registration Page

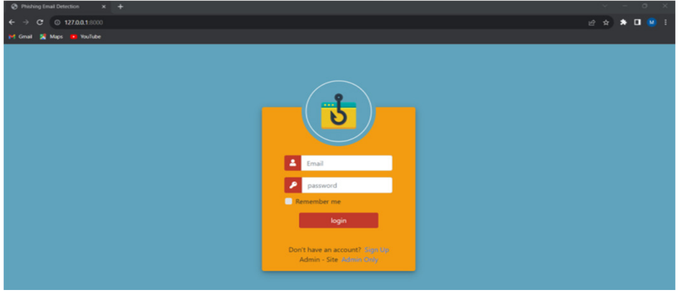


Fig: 5.4 User Sign Up Page

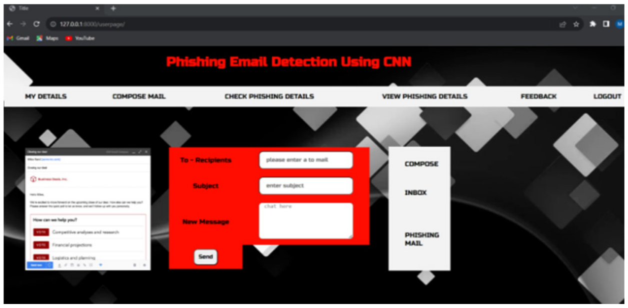
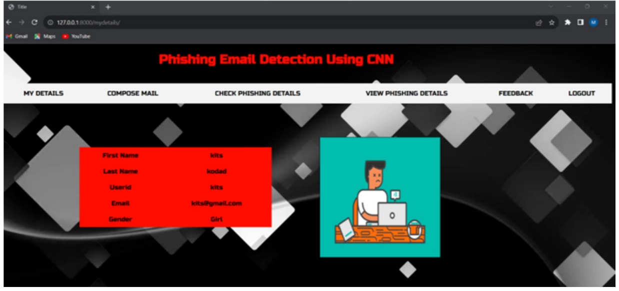


Fig: 5.5 User Page



First Name	kits
Last Name	kodad
Userid	kits
Email	kits@gmail.com
Gender	Girl

Fig: 5.6 User Details

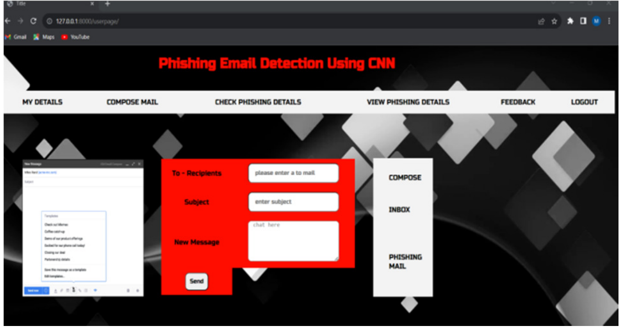


Fig: 5.7 Compose mail Details

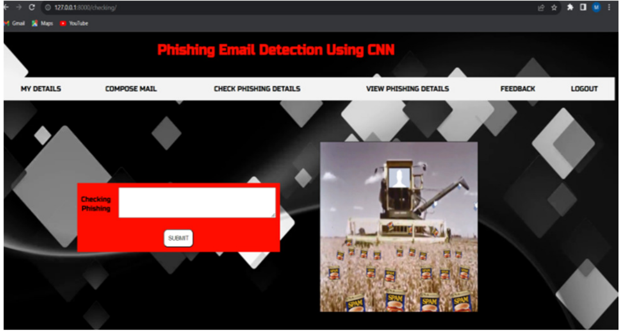
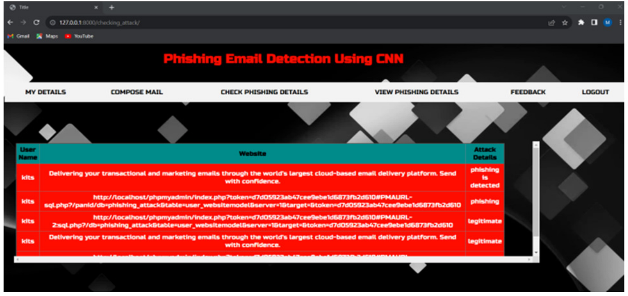


Fig:5.8 Check Phishing Details



User Name	Website	Attack Details
kits	Delivering your transactional and marketing emails through the world's largest cloud-based email delivery platform. Send with confidence.	phishing is detected
kits	http://localhost/phpmyadmin/index.php?token=d7d05923ab47cee9ebe1d6873fb2d610#PMAURL-sql.php?/panid/db=phishing_attack&table=user_websitemodel&server=1&target=&token=d7d05923ab47cee9ebe1d6873fb2d610	phishing
kits	http://localhost/phpmyadmin/index.php?token=d7d05923ab47cee9ebe1d6873fb2d610#PMAURL-2:sql.php?/db=phishing_attack&table=user_websitemodel&server=1&target=&token=d7d05923ab47cee9ebe1d6873fb2d610	legitimate
kits	Delivering your transactional and marketing emails through the world's largest cloud-based email delivery platform. Send with confidence.	legitimate

Fig: 5.9 View Phishing Details

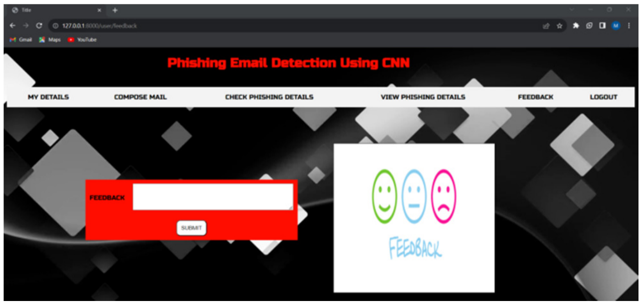


Fig: 5.10 User Feed Back Page

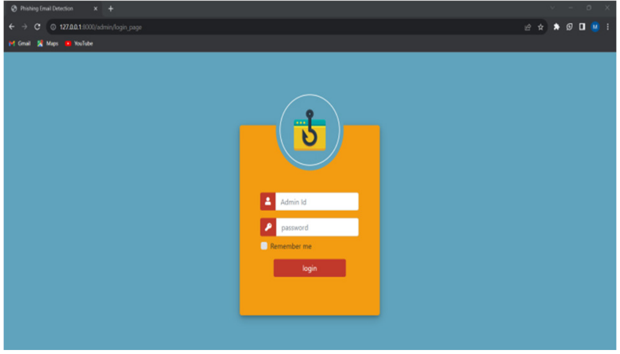
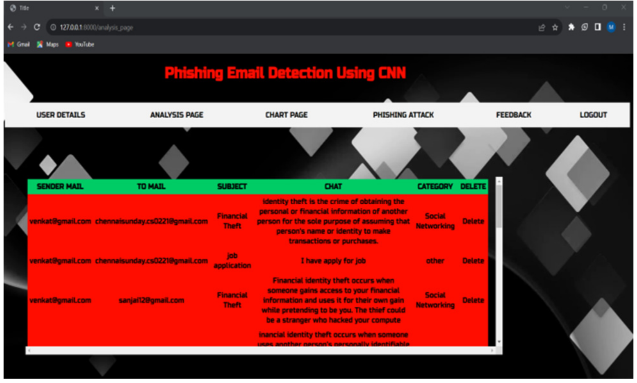


Fig: 5.11 Admin Page



SENDER MAIL	TO MAIL	SUBJECT	CHAT	CATEGORY	DELETE
venkat@gmail.com	chennaisunday.cs0221@gmail.com	Financial Theft	identity theft is the crime of obtaining the personal or financial information of another person for the sole purpose of assuming that person's name or identity to make transactions or purchases.	Social Networking	Delete
venkat@gmail.com	chennaisunday.cs0221@gmail.com	job application	I have apply for job	other	Delete
venkat@gmail.com	sanjai12@gmail.com	Financial Theft	Financial identity theft occurs when someone gains access to your financial information and uses it for their own gain while pretending to be you. The thief could be a stranger who hacked your compute	Social Networking	Delete
			inancial identity theft occurs when someone uses another person's personally identifiable		

Fig: 5.12 Analysis Page

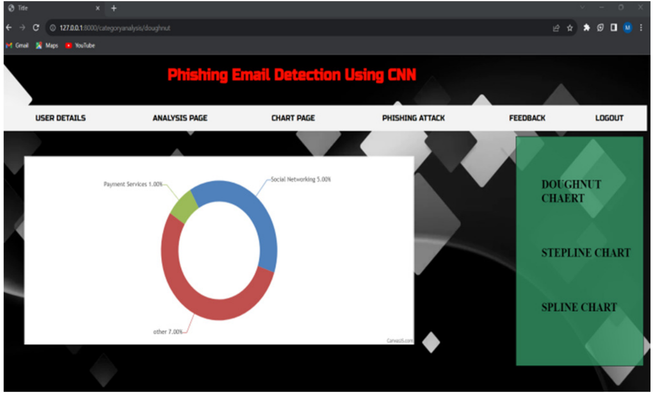


Fig: 5.13 Dough Chart Page

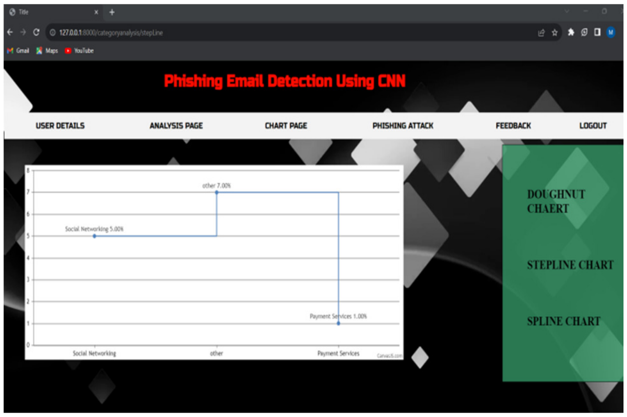


Fig: 5.14 Step line Chart

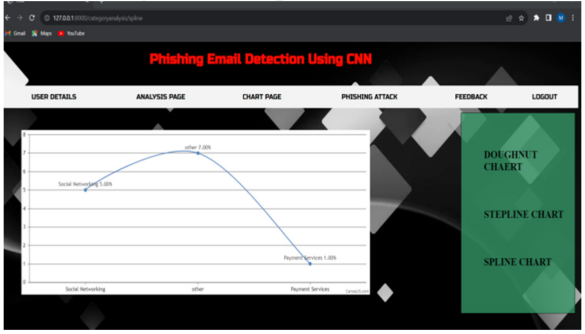
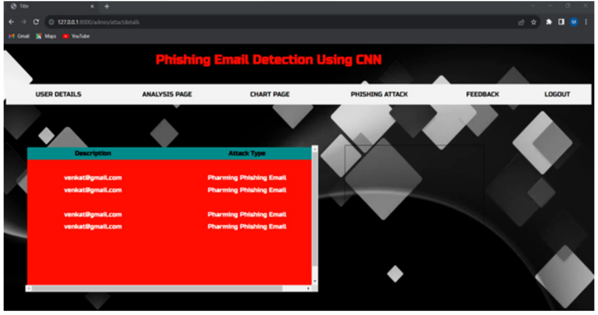
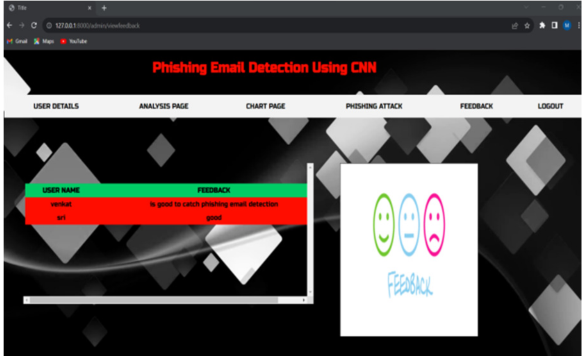


Fig:5.15 Spline Chart


127.0.0.1:8000/admin/attackdetails
Phishing Email Detection Using CNN
USER DETAILS
ANALYSIS PAGE
CHART PAGE
PHISHING ATTACK
FEEDBACK
LOGOUT

Description	Attack Type
venkat@gmail.com	Pharming Phishing Email
venkat@gmail.com	Pharming Phishing Email
venkat@gmail.com	Pharming Phishing Email
venkat@gmail.com	Pharming Phishing Email

Fig: 5.16 Phishing Attack Type



USER NAME	FEEDBACK
venkat	is good to catch phishing email detection
sri	good

Fig: 5.17Admin Feedback page

VI.CONCLUSION: we use a new deep learning model name dto detect phishing emails. The model employs an improved CNN to model the email header and the email body at both the character level and the word level. Therefore, the noise is introduced into the model minimally. In the model, we use the attention mechanism in the header and the body, making the model pays more attention to the more valuable information between them. We use the unbalanced dataset closer to the real-world situation to conduct experiments and evaluate the model. The model obtains a promising result. Several experiments are performed to demonstrate the benefits of the proposed model. For future work, we will focus on how to improve our model for detecting phishing emails with no email header and only an email body.

REFERENCES:

[1] Anti-Phishing Working Group. (2018). *Phishing Activity Trends Report 1st Quarter 2018*. [Online]. Available: http://docs.apwg.org/Preports/apwg_trends_report_q1_2018.pdf PhishLabs. (2018). *2018 Phish Trends & Intelligence Report*. [Online]. Available: https://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend%20Report_2018-digital.pdf

[2] M. Nguyen, T. Nguyen, and T. H. Nguyen. (2018). ''A deep learning model with hierarchical LSTMs and supervised attention for anti-phishing.'' [Online]. Available: https://arxiv.org/abs/1805.01554

[3] Anti-Phishing Working Group. (2016). *Phishing Activity Trends Report 4th Quarter 2016*. [Online]. Available: http://docs.apwg.org/ reports/apwg_trends_reportq4_2016.pdf

[4] Anti-Phishing Working Group. (2015). *Phishing ActivityTrends Report 1st-3rd Quarter 2015*. [Online]. Available: http://docs.apwg.org/Preports/apwg_trends_report_q1-q3_2015.pdf

[5] L. M. Form, K. L. Chiew, S. N. Sze, and W. K. Tiong, ''Phishing email detection technique by using hybrid features,'' in

Proc. 9th Int. Conf. IT Asia (CITA), Aug. 2015, pp. 1–5.

[6] M. Hiransha, N. A. Unnithan, R. Vinayakumar, and K. Soman, ‘‘Deep learning based phishing email detection,’’in *Proc.1st Anti Phishing Shared Pilot 4th AC MInt.Workshop Secur.Privacy Anal.(IWSPA)*,A. D. R. Verma, Ed. Tempe, AZ, USA, Mar. 2018.

[7] C. Coyotes, V. S. Mohan, J. Naveen, R. Vinayakumar, and K. P. Soman, ‘‘ARES: Automatic rogue email spotter,’’ in *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur.Privacy Anal. (IWSPA)*,A. D. R. Verma, Ed. Tempe, AZ, USA, Mar. 2018.

[8] Uyyala, Prabhakara. "Delegated Authorization Framework for EHR Services using Attribute Based Encryption." The International journal of analytical and experimental modal analysis 13, no. 3 (2021): 2447-2451.

[9] Uyyala, Prabhakara. “ The advanced proprietary AI/ML solution as Anti-fraudTensorlink4cheque (AFTL4C) for Cheque fraud detection .” The International journal of analytical and experimental modal analysis 15, no. 4(2022):1914 - 1921.