

CONVERSATIONAL AI ASSISTANT FOR PHISHING AND MALICIOUS URL DETECTION

B Rajani¹, B Poojitha², G Rahul Reddy³, J Laxman⁴, D Ashok⁵

¹Asst.Professor Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

²³⁴⁵Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

Abstract - Phishing attacks are being carried out in a sophisticated manner using various techniques, including social engineering attacks to extract critical information from users. The conventional approach to identify phishing emails uses rules, which are not sufficient to handle newly generated and semantically sophisticated phishing emails. The objective of the proposed project is to create a conversational AI service using Large Language Models (LLMs) to analyze emails based on semantic characteristics such as urgency, impersonation, authority, deception, etc.

The system can effectively classify in an intelligent manner, thus performing better for the identification of potential risks. Lastly, it can easily be incorporated into online applications, with privacy guaranteed at the time of analysis, ensuring effective phishing awareness and decision-making, as evident from the user assessment, while at the same time demonstrating the efficiency of implementing the system with interaction, as evident from the evaluation performed.

Key Words: Phishing Detection, Cybersecurity, Large Language Models, Conversational AI, Email Security, Semantic Analysis, Usable Security.

1. INTRODUCTION

It has turned out to be one of the most prevalent as well as detrimental varieties of cybercrime in recent times. According to the Anti-Phishing Working Group, there were over 1.2 million reported cases of phishing in the second quarter of 2023 alone. It is estimated that around 3.4 billion phishing emails are transmitted every day across the world, which translates into more than a trillion emails every year. Phishing makes up for 91% of all cyberattacks and triggers 36% of data breaches; therefore, it is one of the most severe cybersecurity concerns.

This gives rise to social engineering as the basis for phishing attacks. Phishing attacks basically exploit human psychology more than anything else. The attackers play on emotions like urgency, power, fear, or curiosity to trick the users into clicking on certain web pages or giving away sensitive information. Due to these techniques based on psychology, phishing attacks are basically cases of human error. Human error gives rise to 95% of all successful cyberattacks. Lack of knowledge, tension, cognitive overload, or security training

are some factors that minimize a user's ability to a great extent.

Although there is an existence of automated phishing detection tools, the emphasis of most of these tools is mainly technical features; nonetheless, there is a lack of explanation which is easily understandable, especially for non-technical people, who are the target of most attackers.

In order to address these challenges, it has been proposed to develop a system that can assist the user in the more efficient processing of phishing emails with the help of Artificial Intelligence. To achieve this objective, there is a plan to incorporate Large Language Models to identify not only textual phishing attacks but also semantic and psychological attacks. Additionally, sentence-level insights have been proposed to offer detailed explanation to the user on why it has marked the email as suspicious.

Moreover, there is the ability for the user to query as well as provide explanations in a conversational manner. This will enhance the awareness of the system, thus eliminating the need to have knowledge of the technology, which can enhance the effectiveness of the solution by incorporating the intelligent analysis, the ease of use, as well as the privacy-preserving ability of local processing.

1.1 Background of Phishing Attacks

Phishing is one of the techniques of cyberattacks where hackers try to fool a victim into revealing their sensitive information to the cyber attackers by pretending to be a trusted source. Phishing attacks differ from other kinds of attacks, such as exploitation attacks, which are successful based on the system's vulnerabilities, but in the case of a phishing attack, all its efforts are focused on soliciting a victim's psychological responses with the help of factors such as "urgency," "fear," "authority," and "curiosity."

With several volumes of phishing messages reaching users annually, it has now grown into one of the biggest and most widespread cybersecurity challenges worldwide. Nowadays, many phishing attacks are adequately intelligent; they use texts and brand names that bypass old-fashioned methods of detection. Moreover, the introduction of QR code phishing and multi-channel phishing attacks is a further reminder of the importance of developing intelligent, adaptive, and user-centric solutions.

1.2 Motivation and Problem Overview

The phishing attacks are becoming highly sophisticated, targeting the human behavior rather than targeting system flaws. Most of the traditional phishing solutions, i.e., blacklist-based filtering and rule-based heuristics, have limitations as they are only based on previous experiences. Since the attackers are dynamically changing URLs, emails, and social engineering campaigns, the conventional phishing solutions are not capable of detecting new phishing attacks and attacks based on semantic attacks. Most of the solutions available in the traditional approach have no feedback, thus users are not able to trust these solutions.

However, the main issue lies in the incapability of these systems to grasp the semantic meaning as well as the underlying psychological manipulation used in phishing messages. Modern phishing messages look technically correct but, in reality, try to convey hidden meanings of urgency, impersonation, or consequences. A new dimension of phishing risks has also appeared, as QR-based phishing provides customers no opportunity to look at any embedded URL before opening it. Therefore, it is important that a user-centric, intelligent, and transparent solution is devised, not only capable of tackling phishing risks in emails or QR codes, but also potentially supportive in increasing user understanding.

2. PROPOSED SYSTEM

The proposed system is an intelligent system in the form of an AI-based conversational assistant that is expected to detect and analyze phishing threats from emails and QR codes in a way that is easily understandable. The system brings together under one roof the facilities of Large Language Models using the Google Gemini API with a Django-based web application to facilitate intelligent phishing detection.

In the architecture proposed, three main modules are foreseen to be implemented, notably Email Phishing Detection, QR Code Malicious URL Detection, and finally Conversational AI Assistant. From there, after the input of the mail message by the user, it is expected that the system will make use of the Gemini tool in carrying out semantic analysis. In order to detect malicious features within the QR code, it will go through image scanning before being subjected to potential malicious features using an OpenCV library.

Unlike traditional systems that output only two classification results, such as Phishing or Safe, with no explanation, the proposed method enables users to obtain contextual explanation of their decisions. It does so through the ability to interact with the interface, asking questions and receiving responses in natural language.

It is designed with a modular and secure architecture, using the authentication framework for user management given by

the Django framework, for role-based access. It is designed in such a way that all components are made to work together in order to provide accurate detection, usability, and cybersecurity awareness.

2.1 System Architecture

The system architecture is built as a modular form of a Django-based web application, which incorporates AI-based phishing detection for email, analysis of QR codes, and a conversational assistant, all under a secure authentication setup. It consists of four components: User Interface, Application Layer, AI Processing Layer, and Data Storage Layer.

Both users and administrators interact through the web interface, wherein users can register, log in, and input email content, upload images of QR codes, and retrieve results from the AI module. The Django application layer is used to handle all the request handling and authentication of users. The email content is sent to the Gemini API through LangChain for analysis of content, and the QR code is scanned using OpenCV to detect malicious URLs. The user details and results of the analysis are stored in the database to maintain system reliability.

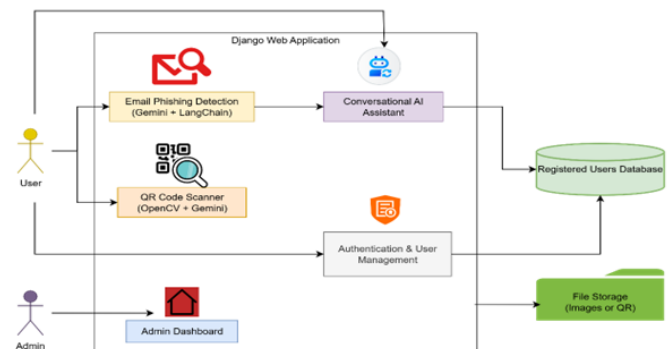


Fig -1: System Architecture

2.2 Workflow

The workflow of the system would start with the registration of the user, administrative approval, and then an authenticated login. Once authenticated, the user would be presented with two choices: one for analyzing emails and the other for QR code analysis.

Email analysis path: The system reads the contents from the email submission and performs phishing detection with the Gemini LLM through Lang Chain. It classifies the content as phishing or legitimate and displays the result along with the contextual reasoning. Then, the conversational assistant gives advice on cybersecurity to enhance user awareness.

In the QR code analysis pathway, a user is allowed to upload an image of a QR code. First, by means of OpenCV, the embedded URL is decoded. After extracting the URL, it's fed

to the Gemini model for classifying it as malicious or safe. Finally, the classification result will be shown to the user.

The process in both cases ends with educating the user by giving explanatory feedback that will eventually enhance better cybersecurity awareness to assure secure user protection.

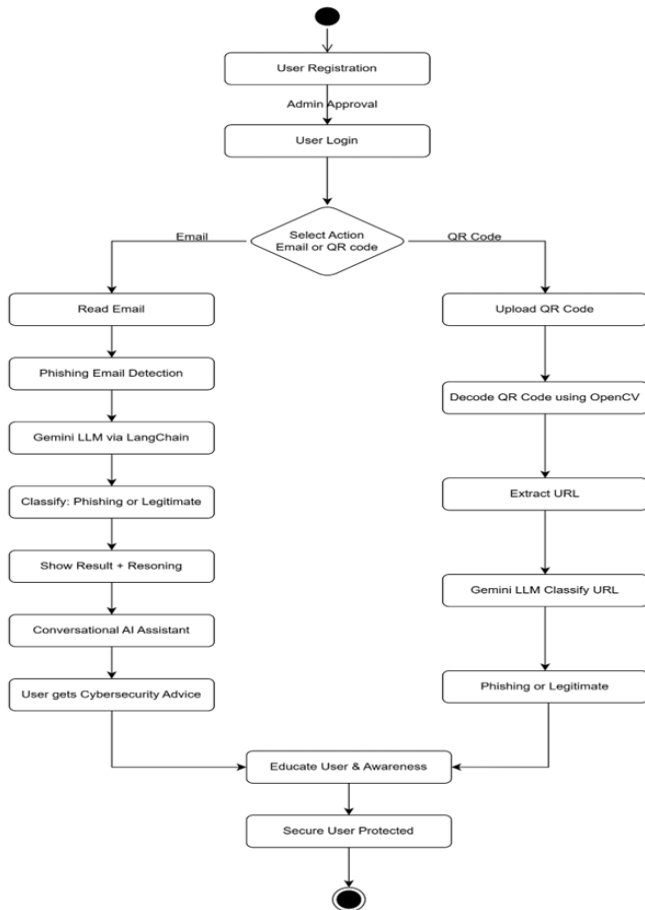


Fig -2: Activity Diagram

3. IMPLEMENTATION

The proposed system is achieved by using a secure and modular web-based application, and as such, the base platform will be Django. Also, the layered architecture style is adopted, as it meets the needs of being scalable, maintainable, and incorporates AI-based phishing mechanisms.

3.1 Backend Development (Django Framework)

The backend framework of the application is Django, and it offers the Model View Template (MVT) pattern. In Django, the processing of the request, along with session handling and validation, is done on the backend. Django is a high-level framework and offers a clean abstraction of business logic, view, and data model.

The user management system is based on the user authentication system provided by Django. The authentication of users can be achieved by registering users, approving accounts by administrators, logging in, and session usage. Role-based access control is used for restricting access of certain phishing detection tools by the user. The user can be managed by administrators using the dashboard. Hashing of passwords and protection of CSRF are enabled to avoid unauthorized access of any user.

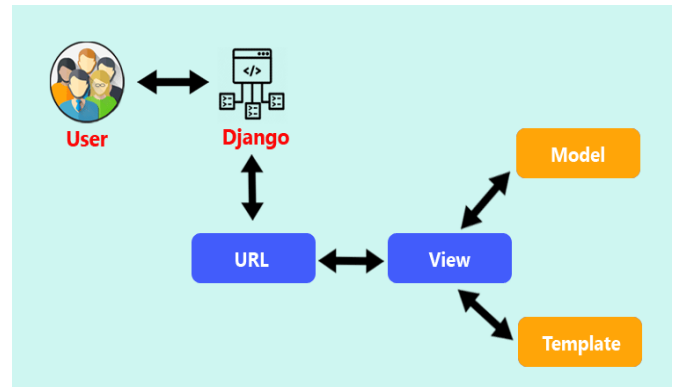


Fig -3: Django Model-View-Template (MVT) Architecture

3.2 Gemini API Integration via Lang Chain

The mechanism of phishing detection is now integrated through Lang Chain with the Gemini API of Google. Lang Chain provides an orchestration layer to normalize prompt structuring and standardizes managing model interactions, thereby processing AI responses in a consistent format.

This system receives user-supplied content in standardized form and requests Gemini LLM to perform semantic analysis. The model searches for markers, such as urgency, impersonation, authority claims, suspicious requests, and misleading tone. This system classifies the content as phishing or legitimate and provides human-readable reasoning for the same; hence, it builds transparency and user trust.

These QR code extracted URLs are then fed into the Gemini model, which scores their malicious intent. The AI model further analyzes the structure of the URL, domain patterns, and other contextual threat indicators before the final result of the classification is returned.

3.3 QR Code Processing using OpenCV

The extraction of the QR code image is performed through the library called OpenCV. This means that any QR code image entered into the system as part of the image analysis process allows for the extraction of the data within the image. This implies that if the extracted data was a URL, then the URL is used as the classifier for the URL through the AI system.

This integration will help to identify malicious links that are embedded within a QR code, thus attending to some of the emerging attacks, such as QR phishing or “quishing”. As previously discussed, the pipeline shall ensure speed in decoding and integration with the proposed AI analysis module.

3.4 Database and Data Management

SQLite serves as the backend database to store credentials of the users, authentication data, and activities. Here, Object Relational Mapping of Django is helpful in data management in an organized and secured manner.

The images of uploaded QR codes are stored by utilizing the file handling facility offered by Django for a temporary period of time. Additionally, the sensitive data is not leaked to outsiders, but the AI component is processed within a specific environment for maintaining user privacy.

4. RESULTS AND PERFORMANCE ANALYSIS

The performance of a proposed Conversational AI Assistant will be analyzed in the context of detecting phishing emails and malicious QR-based URLs. Detection was carried out using a structured dataset and user input to test the performance of the system regarding correctness of classification, dependability, and ease of use.

4.1 Experimental Setup

The system has been tested against different types of phishing emails, legitimate emails, phishing URLs, and legitimate URLs in the form of QR codes. There are different types of phishing schemes in the system, including urgency-based phishing schemes, impersonating schemes, and URL-based phishing schemes.

The input will then be processed by passing it through the Django application, followed by analysis using the Gemini language model. In addition, the URLs provided in the QR code will be decoded using OpenCV. This will also involve an assessment of the malicious intention. This will be achieved through consideration of the accuracy, stability, and quality of explanation by the conversational AI.

4.2 Performance Analysis

The performance of the proposed model has been validated using a balanced data set of 420 phishing emails and another set of 420 legitimate emails. The performance of the classification has also been validated based on Accuracy, Precision, Recall, and F1 Score metrics. Overall, the correctness of the model is quite high in classifying phishing and legitimate emails, as indicated by an accuracy of 95.24%. In addition, the precision level, which reflects the ratio of correct classification of legitimate emails as phishing emails,

is as high as 96.8%. This shows that most of the emails that are given higher priority as phishing emails are actually such, thereby reducing false positives. Similarly, the recall level, which reflects the capability of the model in terms of reducing false negatives, is as high as 93.56%, as illustrated by its high value. This is further validated by its score of 95.15%, thereby affirming the robustness of the model as a correct balance between precision and recall.

Other than this, the reliability and consistency of the QR-based classification of malicious URLs approach were also ensured through the use of the actual correct QR code decoding achieved through the application of OpenCV and the semantic evaluation of the identified URLs through the application of the Large Language Model. The semantic evaluation of the URLs through the application of AI methods has shown improvements in terms of the enhancement of the actual accuracy of detection, compared to the traditional methods of detection, which allow the best of them to come under the category of psychological attacks.

4.3 System Reliability

The system reliability has also been validated by undergoing extensive functional testing on all the integrated modules in the system. This has validated the consistency of function interaction between the Django backend and Gemini API for the purpose of carrying out semantic analysis, devoid of failed responses. The operation of the QR code module has successfully decoded the uploaded image using OpenCV, ensuring effective acquisition of the details of the URLs to be processed. Also, the implementation of the authentication function has shown to be effective in ensuring user login, session operation, and admin functionality, devoid of unauthorized operation or session conflict. Further, the system has successfully operated in the process of invalid, incomplete, and empty inputs, devoid of crashes and instabilities. This has therefore validated the smooth navigation to the entire modules in the system, ensuring effective operation under normal circumstances.

4.4 Security and User Experience Analysis

The proposed system utilizes various security factors that include consideration of security factors related to access control and data security. For example, the Django authentication module provides clients with access to protection of password hashing, support of sessions, and protection against a variety of web-based attack patterns, such as Cross-Site Request Forgery Attacks. Role-based access control of phishing detection modules is restricted to authorized users, whereas inappropriate admin access is restricted to authorized personnel only. Inappropriate inputs, such as emails and QR code entries, are controlled using the system to avoid superfluous data storage, which can compromise these risks. Moreover, the presented system will be more effective at phishing detection using the help of AI, bypassing the need for rule-based detection.

From the point of view of usability, it is seen that the system has been developed in such a way that it can be used by any user, whether technical or nontechnical, due to its simple and organized interface. There is also clarity in terms of classification results and explanations provided by the system to make better decisions. A conversational assistant has also been integrated, which would be extremely beneficial to promote further user cyber awareness.

4.5 Analysis of Results



Fig -4: Home Page

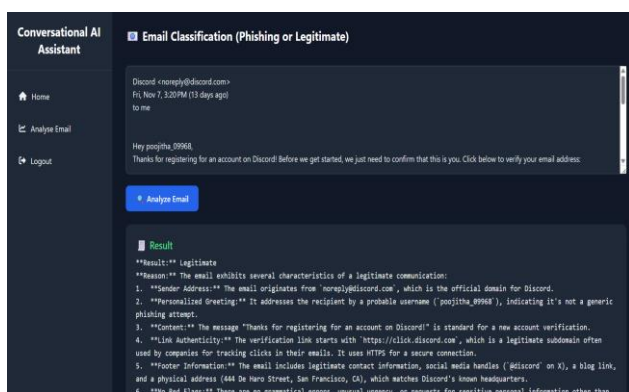


Fig -5: Email Classification

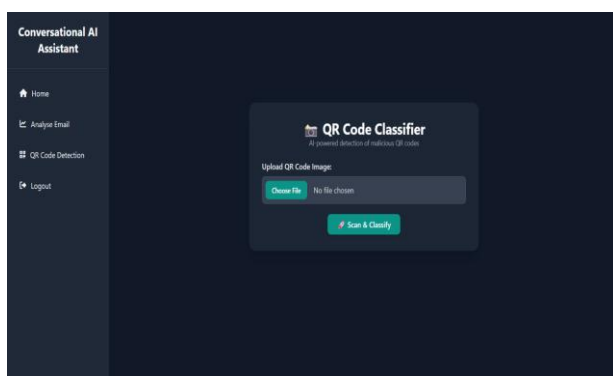


Fig -6: QR Code Classification

The high precision value also implies that the system is not likely to classify legitimate emails as phishing emails, which is quite useful. Furthermore, the high value for the recall test

shows the system is able to detect almost all phishing attempts, thus being effective.

With the assistance of semantics analysis, the detection can be improved over traditional rule-based systems, especially when trying to filter more complex phishing scenarios that are based on psychological manipulation.

Also, the application of the conversational interface promotes usability since users can obtain explanations for better understanding of the rationale behind classification. This way, there is an improvement in cybersecurity awareness.

5. CONCLUSION

This paper has proposed an AI-based phishing detection system in order to enhance user awareness and security against modern phishing attacks. The proposed phishing detection system is composed of several different functional elements, including safe user and administrator authentications as well as phishing email analysis, dialogue, and QR code detection for malicious activities. This proposed system, based on the Google Gemini API and LangChain and computer vision technology using OpenCV, is capable of classification and easy understanding as well as safety tips.

This ensures controlled access to the system, authentication, and sufficient administration, further enhancing reliability and governance. Overall functional testing confirmed that the system was successful in identifying phishing content, analyzing sender content, handling invalid content, and generating relevant context.

In contrast, when considering conventional phishing detection system development, the primary focus is normally on developing a mechanism for such identification at the binary level. However, more focus is placed on explainability and user interaction using the proposed method, which is extremely useful for developing cybersecurity awareness and informed decisions. The system appears to have promising features to develop an effective mechanism against phishing emails and malicious QR code attacks using relevant attributes of AI-driven semantic analysis.

6. FUTURE WORK

Even though the proposed system features efficient phishing detection and user navigation, there are still improvements to be made to enrich the scope of this system. Some of these improvements for the proposed system might include additional support for the detection system, which can be used for various phishing attacks, including SMS phishing, voice phishing, and social media phishing.

In addition to that, having support for multi-language phishing detection can make the application more palatable to different categories of users. The application can also be extended to incorporate intelligence feeds from different sources, such as Phish Tank and Open Phish, which can enhance its detection capabilities using updated phishing data. The application also has the ability to be used by mobile users via a mobile application.

Further enhancements could include the integration of automated incident response with Security Management Systems (SIEM/SOAR) as well as the implementation of interactive cybersecurity training modules to reduce the susceptibility of humans to phishing attempts.

REFERENCES

- [1] S. Hossain, D. Sarma, and R. J. Chakma, "Machine Learning-Based Phishing Attack Detection," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2020. [Online]. Available: www.ijacsa.thesai.org
- [2] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. E. Ulfath, and S. Hossain, "Phishing attacks detection using machine learning approach," in *Proc. 3rd Int. Conf. Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Aug. 2020, pp. 1173–1179. doi: 10.1109/ICSSIT48917.2020.9214225
- [3] N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-side defense against web-based identity theft." [Online]. Available: www.ebaymode.com
- [4] D. Miyamoto, H. Hazeyama, and Y. Kadobayashi, "An evaluation of machine learning-based methods for detection of phishing sites," *Proc. IEEE International Symposium on Applications and the Internet (SAINT)*, 2008.
- [5] Anti-Phishing Working Group (APWG), "Phishing E-mail Reports and Phishing Site Trends," *Phishing Activity Trends Report*, 2022. [Online]. Available: <http://www.apwg.org>
- [6] D. Stuttard and M. Pinto, *The Web Application Hacker's Handbook: Finding and Exploiting Security Flaws*. Hoboken, NJ, USA: Wiley, 2011.
- [7] Fatima Salahdine, Zakaria El Mrabet, Naima Kaabouch, "Phishing Attacks Detection A Machine Learning-Based Approach", *Proc. IEEE*, Dec. 2021. doi: 10.1109/UEMCON53757.2021.9666627