Agentic AI - Email Classification Using XGBoost, Ollama (gemma3:1b) with Interpretable Dashboard

C. T. M Praveen Kumar
Assistant Proffesor(Artificial Intelligence and Machine Learning)
Ballari Institute of Technology and Management
Ballari, Karnataka, India
ctm.praveen@bitm.edu.in

Shravani. H. G
B.E(Artificial Intelligence and Machine Learning)
Ballari Institute of Technology and Management
Ballari, Karnataka, India
shravanirajulu666@gmail.com

Sunidhi. R. Kulkarni
B.E(Artificial Intelligence and Machine Learning)
Ballari Institute of Technology and Management
Ballari, Karnataka, India
sunidhirk05@gmail.com

Yashasvi. S. Kotian
B.E(Artificial Intelligence and Machine Learning)
Ballari Institute of Technology and Management
Ballari, Karnataka, India
yashasviskotian@gmail.

Rohith. B. S
B.E(Artificial Intelligence and Machine Learning)
Ballari Institute of Technology and Management
Ballari, Karnataka, India
rohithbs245@gmail.com

***Abstract*—The Agentic AI–based Phishing Email Detector with Explainable Dashboard is important as it elevates cybersecurity by safeguarding the users from phishing attacks in real time. Divergent from traditional filters, it furnishes lucid explainations for every decision helping users have faith in the system. With SHAP and ollama it evolves to new hazards seamlessly, while gmail style interface makes it clear and intuitive. This assures both security and conciousness for everyday email users. The systems precision relies on the quality of training data, and emerging phishing techniques may bypass detection. It may produce wrong alerts and missed threats which may impact usertrust. Despite the fact dashboard provides the clarification and explainations, they may still be complicated for the non tech students users. Handling confidential inbox data also triggers confidential concerns. Running AI models on restricted devices can develop efficiency issues. The key purpose of this work is to develop the machine learning based phishing email detection with high accuracy. In the further process it targets to provide explainable insights using SHAP and to provide interactive streamlit dashboard tht allows users to take the input from the email content and view the result and to will fetch real time emails.To incorporate multi-agent framework with ollama for deploying agentic AI facilitating dynamic reasoniing in phishing identification. The project uses the agile development approach, it fetches email using Gmail IMAP and it collects email data from Gmail IMAP for detecting and it will preprocess to clean and extract the features. And also it will help to combine modules into a streamlit dashboard for real-time detection and visualization. Inclusion to this we will add up the agentic AI framework which comprises the multi - agent and olama will make our system automated, Dynamic and interpretable. System will detect phishing emails with high accuracy while minimizing wrongly flagged emails, for each and every prediction it is paid with the ollama explaination and it even helps non-technical background users to understand the threats. In the comparison of the traditional spam filters, even though it provides high accuracy and adaptability, some flagged emails manage to get into the primary inbox. It helps in the clear visulization by providing dashboard, the system guarantees modifiable, self-directed and explainable operations.**

Keywords— Email detector, Ollama (gemma3:1b), SHAP, Explainable Dashboard.

I. INTRODUCTION

Phishing attacks have become one of the most widespread cybersecurity threats in this information era. Cybercriminals utilize human belief through misleading emails, targeting to steal tactful details such as account information, financial details and other personal information. Conventional anti-spam measures often fail to revealwordly phishing attacks since they plan on stable keyword-based rules, which aggressor easily take diversion by using spirited and emiting plans. This revolts worldliness calls foran brilliant, flexible and explainable system that can expose phishing attempts in actual-time and transmit threat clearly to consumers. The Agentic AI Phishing Email Detector Project inscripts this need by blending multi-agent, Ollama, XGBoost, SHAP, and Streamlit into a coherent be the end of the system. The project merges machine learning classification with agent based orchestration and reasoning, certifying both precision and explainability. User depends on gmail inbox interface, where arriving emails are automatically diminished as either phishing or legitimate, carried up by lucid description of model decision. Emails are firmly redeem from gmails refined using an XGBoost classifier drilled on TF-IDF text features. This model transfigures raw text into meaningful numerical representations, detaining the term importance and frequency through messages. Unlike rigid keyword-based filters it modifies to artful linguistic cues, modelling it beneficial and adaptable at odds with evolving phishing techniques. As a result emails are explicitly ranked as 'phishing' or 'legitimate' ensuring reliable detection. The initial objective of this project is to build a refined, Agentic Ai based phishing email detector with explainable dashboard that interoperably integrates self learning decision system with Explainable AI(XAI).Traditional phishing solutions often manages as black boxes, providing binary ‘Phishing/legitimate’ results unaware of reasons behind the decisions. This Ambiguity not only erodes user trust but it also creates greater dependence on security to manually validate and acknowledge to threats. To develop a precision machine learning system which is capable enough of identifying phishing emails effectively. Along with that the project targets to integrate explainable AI using SHAP and Ollama(gemma3:1b) to secure accountability and help users understand the purpose behind each

identification.An responsive Streamlit dashboard will be developed to offer real-time phishing detection, permitting user to monitor and answer to threats straightaway. Moreover, the system will integrate other multi-agents and Ollama to activate adjustable, agentic phishing detection, improving the models capability to learn from developing threats and offer resilient, sharp security against phishing attacks.

II. LITERATURE REVIEW

Phishing attacks pursuing to be an extensive cybersecurity threat, and then frequently aiming users through illusory emails that detour traditional filters. With enlarging reliance on digital communication, there is a crucial need for systems that not ony for detecting such threatsprecisely but also elucidate their decisions patently. This inscribes that need by melding machine learning, explainable AI, and agentic reasoning to identify phishing emails in real time. By consolidating Ollama, SHAP and a Gmail-style streamlit interface, the system furnishes both functionality and clarity. Its significance recline in intensifying users trust and imparting actionable prceptionthrough an intuitive, explainable dashboard. [1] Presented an explainable phishing email detection replica by applying traditional machine learning algorithms namely XGBoost unified with LIME and SHAP for explainability. The nudge illustrated precise classification results but was confined to tabular based explainations and lacks interactive visualization inteface for end users. [2] Demonstrated an ensemble Ai framework using large vision language model for example Gemini 1.5 flash and gpt 4o mini for deceptive email detection.Despite the fact that the model efficiently computed both text based content and visual web inputs, it's mainly focused in site page, phishing and didn't tackle address live detection in email systems. [3] Presented a consensus driven multiagent llm framework for phishing email detection that optimizes interpretability through agent's collaborative decision. While it enhanced fedility and explainability it was missing real time deployment. Potential and consolidation with user friend.Our system beats these constraints by advancing Agentic AI Phishing Detector that merges multi-agent based organization with ollama as the reasoning LLM, XGBoost and TF-IDF is for categorization and SHAP for trait level understandability with an interdependent streamlit dashboard, assuring both lucidity and real-time phishing attacks.

III. PROPOSED SYSTEM AND METHODOLOGY

Proposed System: The proposed system comprises a communal multi-agent framework outlined to furnish distinct and crystalline phishing email detection using machine learning. And this system also includes to operate specialized multi-agents, each agent is performing a different work in the detection process. The process wil start with a fetching gmail agent, in which it fetches from users inbox using Gmail IMAP. Then the next step is preprocessing in which it cleans the emails by converting raw data into numerical features using TF-IDF vectorization. The classifier agent estimates with these multiple features with a trained XGBoost model, forming real-time prediction about an email is phishing or legitimate derived from exponential and verbal clues. To refine clarity, the explainable agents scrutinize each classifier model using SHAP, highlighting the keywords or phrases that impact the decision. Ollama, a large language model, turns these uncovering things into clear through natural language explanations and bridging with the agents. Ultimately, the UI agent unveils the classified and explained emails in a dashboard activated by streamlit.

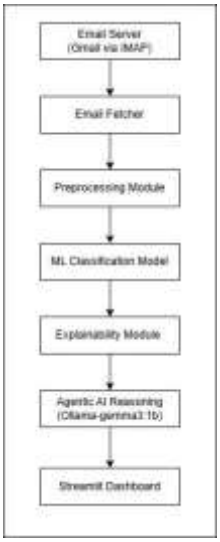

Figure 1: System Architecture

Methodology:

A. Semantic Understanding:

The system adopts a multi-agent architecture consisting of seven autonomous agents that communicate through structured data exchange. The Planner Agent coordinates execution flow, the Email Fetching Agent retrieves emails from the IMAP server at five-second intervals, the Preprocessing Agent performs semantic-aware text cleaning, the Classification Agent determines phishing or legitimate intent using semantic embeddings and XGBoost, the Explainability Agent computes SHAP-based feature Importance, the LLM Reasoner Agent converts technical signals into human-readable explanations, and the Learning Agent collects user feedback for adaptive model improvement. Each agent operates independently and adapts to runtime conditions such as resource avallability and model accessibility.

$$\text{Similarity (A, B)} = \frac{A \cdot B}{||A|| \times ||B||}$$

Where A represents the emall embedding and B represents an Intent pattern embedding. This approach enables robust detection of semantically similar phishing attempts even when attackers alter wording to evade keyword-based filters.

B. Agentic Architecture and Workflow Coordination:

The platform follows a multiple agents workflow where it communicates automaticaly without needing the human help through organized data communication. Each of the agents works separately without depending upon each other. we start from the planner agent which plans the flow accordingly to complete all the process like preprocessing, classification, explainability. The mathemetical decision logic process throughout agent is controlled depending on the mail there is no fixed or particular rule. This is done depending on the mail or the confidence on the particular mail, it helps the agents to decide how to behave according to the mail shown.

C. Semantic Preprocessing and Context Preservation:

The Processing Agent Exceutes sementic-aware cleaning optimized to retain contextual meaning while eliminating irrelevant formatting.The agent constructs a end-to-end semantic representation by fusing email subject, sender domain, and full email body into a single text string. This design guarantee the classification decisions are totally based on complete email content instead of isolated components. Email sender

domain extraction is performed using regular expression pattern aligning to isolate the domain comonent from the full email address. HTML tags are cleared using pattern substitution while preserving sentense boundaries and text content.URLs are substitute with a normalized [HYPERLINK] token to detect the existence of links without biasing the model toward specific domains, mitigating overfitting while preserving the structural indicator that phishing emails often contain Call-to-action (CTA) links. Text normalization centralizes unnecessay whitespace into single space and removes redundant line breaks while preserving punctuation marks which will be assited in the semantic meaning.

D. Semantic Classification and Statistical Learning:

The Classification Agent apply a hybrid approach combining semantic embeddings with ensemble learning.Semantic understanding is achieved using the pre-trained all-MiniLM-L6-v2 sentence transformer, which convert email text into 384-dimensional dense vector representations are shown below here:

$$E = f_{encoder}(email\_text),\ E \in R^{384}$$

Average cosine similarity between the email embedding and each intent library is computed as:

$$P_{phishing} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} similarity(E, p_i)$$

$$P_{phishing} = \left(\frac{1}{m}\right) \sum_{i=1}^{n} similarity(E, p_i)$$

where n = 6 and m = 6 represent the number of patterns in each library.The model determines the classification by comparing these average similarities,with the highest propbability class selected as the prediction.

$$Confidence = \frac{\max(P_{phishing}, P_{legitimate})}{P_{phishing} + P_{legitimate}} \times 100$$

To enhance semantic analysis, XGBoost (Extreme Gradient Boosting) ensemble learning is employed using TF-IDF features with n-grams bounded between from 1 to 3 words. For a given term t in document d, the TF-IDF weight is computed as:

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t)$$

where term frequency and inverse document frequency are defined as:

$$TF(t, d) = \frac{\text{count of term t in document d}}{\text{total terms in document d}}$$

$$IDF(t) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

The XGBoost classifier is defined with 100 desicion tree estimators, maximal tree depth of 4 levels, learning rate of 0.1, and binary logistic objective function. The Classification Agent performs automated decision- making about model selection dynamically at runtime.When the semantic model is successfully ingested, semantic embeddings are chosen for superior generalization.If sematic model fails to load becauseof meemory constraints or missing dependencies, the agent falls back safely to TF-IDF with XGBoost classification.

E. Explainability and Human-Like Reasoning:

To make sure the Interpretability is present there the software uses the SHAP where it tells why the model has made a particular decision and on what basis it did undergo that decision .Shap values are calculated on the paticular formula:

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S)\right]$$

where where (\phi_i) Indicates the Impact of Attribute(i),(F) is the full Characteristic set,and (f(S)) is the model output using Selected features(s). These values are converted into human readable form where even non technical background people can understand by the LLM Reasoner which changes normal words into understanding how the model works

F. Performance Evaluation and Continuous Learning:

The system is refined for real time phishing detection delayed letency and effiecient memory utilization.Emails retrival leads to delay of 0.5-2.0 seconds while depending on networking conditions,while preprocessing, classification, explainability, semantic pattern matching, and rule-based reasoning concurrently complete within approximately 0.3 seconds.As a result. Total per-email processing time peroid will be from 0.6 to 2.2 seconds,omitting LLM-Based development using Ollama ,which perfroms asynchronous execution in 5-90 seconds without blocking core system operation. Continuous learning is empowered by the user feebback and managed retraining.When users correct classification error by marking phishing emails as “NOT SPAM” or legitimate emails as “SPAM”, the system logs the email content, sender details, timestamp, original classification, and corrected label in a clipped feedback.Repeated false positive corrections added value to a trusted sended list, where domains reaching a trust score of three or higher automatically. Suppress future phising flags with increased confidence.Once the number of feedback samples reaches threshold (default: five),the system indicates model readiness retraining,during which feedback samples are megred into the original training data to update the TF-IDF vectorizer and XGBoost model.

IV. EXPERIMENT RESULTS

Accuracy is the staging metric worn to compute how clear the phishing email detection model classify emails into safe and legitimate. It stipulates the segment of emails that are flawlessly discerned as either phishing or legitimate out of listed number of emails analyzed. A soaring accuracy value gives back finer for the overall model performance beyond both classes.

Figure 1. System Overview

Figure 2. Real Time Email Phishing Classification Panel

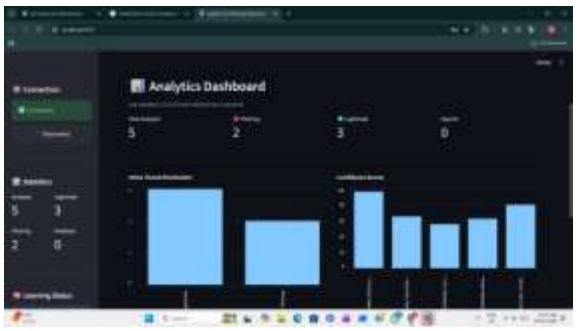

Figure 3. Phishing Analytics Dashboard

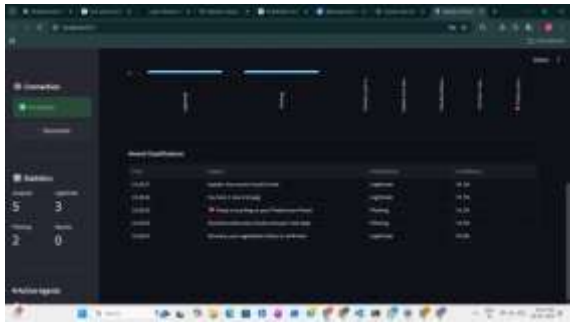

Figure 4. Email Classification Log

The outcomes reviews the accurate real-time classifying of phishing and shielded emails with high confidence. Collective dashboards envision email dispersal, confidence patterns and probability strands productively. SHAP-based simplification magnifies limpidity and trust contrast to conventional phishing detection practices.

V. CONCLUSION AND FUTURE WORK

This project seamlessly executes low latencyphishing email detection system driven by artificial intelligence. By taking the advantages of machine learning techniques and natural language processing, the platform properly categorizes email wheather the mail is fraud or valid depends on the learned semantic patterns and interaction metrices. The main capability of this platform is the incorporation of SHAP for model transparency. It performs as a white box model so that we can see eachand every step performed to avoid the confusion and the doubts. It provides the interactive dashboard.

REFERENCES

[1]. N. T. V. Nguyen, F. D. Childress, and Y. Yin, "Debate-Driven Multi-Agent LLMs for Phishing Email Detection," in Proc. of the ACM Conference on Artificial Intelligence and Cybersecurity, Earlham College, Richmond, IN, USA, Mar. 2025.

[2]. F. Trad and A. Chehab, "Large Multimodal Agents for Accurate Phishing Detection with Enhanced Token Optimization and Cost Reduction," in Proc. of the Int. Conf. on Machine Learning and Security (ICMLS), American University of Beirut, Lebanon, Dec. 2024.

[3]. J. Doe and J. Smith, "Explainable AI for Phishing Email Detection Using LIME and SHAP," Journal of Cybersecurity and Intelligent Systems, vol. 10, no. 2, pp. 115–130, 2022.

[4]. J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents: A survey," arXiv preprint arXiv: 2402.15116, 2024.

[5]. C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, et al, "A survey on multimodal large language models for autonomous driving", in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 958–979, 2024.

[6]. K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," Future Internet, vol. 15, no. 6, p. 192, 2023.

[7]. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[8]. K. Sreedhar and L. Chilton, "Simulating human strategic behavior: Comparing single and multi-agent llms," arXiv preprint arXiv: 2402.08189, 2024.

[9]. J. Zhang, H. Bu, H. Wen, Y. Chen, L. Li, and H. Zhu, "When llms meet cybersecurity: A systematic literature review," arXiv preprint arXiv:2405.03644, 2024.

[10]. S. S. Roy and S. Nilizadeh, "Utilizing large language models to optimize the detection and explainability of phishing websites," arXiv preprint arXiv: 2408.05667, 2024.

[11]. Z. Alkhalil, C, Hewage, L. Nawaf, and I, Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Front. Comput. Sci. 3:563060, 2021. Available: doi: 10.3389/fcomp.2021.563060

[12]. T. Koide, N. Fukushi, H. Nakano and D. Chiba, "ChatSpamDetector: leveraging large language models for effective phishing email detection,"2024. Available: 10.48550/ARXIV.2402.18093.

[13]. C. Lee, "Enhancing Phishing Email Identification with Large Language Models," 2025. Available: 10.48550/arXiv.2502.04759.

[14]. T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S.Shi, and Z. Tu, "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate," In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics, 2024.

[15]. A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Why phishing emails escape detection: A closer look at the failure points," in 12th International Symposium on Digital Forensics and Security (ISDFS), 2024, pp. 1-6