International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Phishing Email Detection Using AI

Shaik Armaan Shoaib[1], Sanke Sathish[2], Ms. K. Shirisha[3], Mr. P Satya Shekar Varma[4]

[1,2]Student, Department of Computer Science Engineering, MGIT
[3,4]Assistant Professor, Department of Computer Science Engineering, MGIT, Hyderabad, India

Abstract

Phishing emails continue to pose a significant threat, causing financial losses and security breaches. This solution addresses limitations in existing applications, such as reliance on proprietary datasets and lack of real-world application, by proposing a high-performance machine learning model for email classification. Phishing email detection using Artificial Intelligence (AI) is an innovative approach against the ever-growing threat of email-based cyberattacks. This project uses advanced machine learning algorithms through AI to classify emails into categories such as phishing and safe.

Keywords: Phishing Detection, Artificial Intelligence, Machine Learning, Email Security, Application Programming Interface (API)

1. Introduction

Phishing attack is a cybercrime to deceive users into stealing their information, such as personal identity, financial details, etc. Masquerading as legitimate sources, attackers can reach victims by sending fraudulent messages using emails (such as Gmail, Outlook, etc.). Users become vulnerable if they input their information or download attachment files.[2]

Email phishing is one of the widespread and continually increasing threats in the digital world targeting individuals and organizations to obtain confidential information or break into their security systems. Such scams usually involve seemingly legitimate emails whose intention is to entice people to provide passwords, financial information, or other types of personal identification. Phishing schemes are no longer just a cause for financial loss but also a threat to privacy, trust, and potentially the entire operations of an organization [3]. The advancement in modern phishing emails that frequently appear as trusted brands or are crafted with advanced social engineering techniques makes them increasingly tough to detect using traditional approaches.

Modern attackers employ dynamic and adaptive strategies, such as embedding malicious content in seemingly harmless attachments or links. The sheer volume of emails sent daily, combined with evolving attack patterns, overwhelms existing security systems that rely on static filters or rule-based algorithms. Traditional spam filters, while effective against generic spam, often struggle to detect sophisticated phishing emails, leading to high false-positive or false-negative rates. This not only exposes the users to possible attacks but also destroys the confidence of users with email systems as a secure medium for communication.[7]

The current solutions for phishing detection are mainly based on manual review, rule-based mechanisms, and static heuristics. These systems often make analyses of email metadata, sender addresses, and known malicious patterns to classify the emails as phishing or legitimate. However, they have huge drawbacks [9]. The methods adopted are manual, cumbersome, and slow, which makes it an unsuitable option for large scale deployment. Static filters cannot cope up with the new emerging phishing techniques and hence provide poor resistance against zero-day attacks [1][6].

2. Literature Survey

Sanjay Ramdaus et al. (2024) aims to highlight the similarities between phishing email detection using machine learning and big data analytics. Thanks to machine learning big-data analytics solutions, a new research avenue is opened to improve cybersecurity based on the latest threats associated with criminal opportunities, terrorist activities, political or personal issues, illicit drugs, actors, or different transactions. The remainder of the paper is organized as follows: Section 2 presents the literature survey. Section 3 provides the early detection of phishing emails using data mining-based machine learning. The big data analytics used in cybersecurity are provided in Section 4. [1].

Nishant, A et al. (2024) proposed ensemble techniques where various machine learning algorithms are combined to improve the accuracy and strength of spam detection systems. Using different algorithms, it tries to create an appropriate systematic behavior to increase the detection rates and reduce the number of misclassification cases. In this research, four machine learning algorithms were selected to build the meta-learning model; these algorithms have been chosen based on their proven effectiveness in spam detection systems, such as Naive 6 Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K Nearest Neighbors (KNN).[2]

Eman et al. (2024) shows the effectiveness of an ensemble approach for email spam detection by aggregating multiple weak machine learning algorithms to produce a strong machine learning model. The purpose of this research is to enhance the accuracy and robustness of the predictive model to detect spam emails. As a result, the proposed approach produced a better performance with 95.8% accuracy.[3]

Amith Khandakar et al. (2024) addresses limitations in existing research, such as reliance on proprietary datasets and lack of real-world application, by proposing a high-performance machine learning model for email classification. Utilizing a comprehensive and largest available public dataset, the model achieves a f1 score of 0.99 and is designed for deployment within relevant applications. Additionally, Explainable AI (XAI) is integrated to enhance user trust. This research offers a practical and highly accurate solution, contributing to the fight against phishing by empowering 7 users with a real-time web-based application for phishing email detection.[4]

Alohaly, M et al. (2023) have proposed a new model capable of improving the results of SMS classification and spam detection. They first employed a recent text embedding technique based on the GPT-3 Transformer to represent the text message in a dense numerical vector. Then, they gathered four classifiers (SVM, KNN, CNN and LightGBM) in an Ensemble module to classify the vector representations obtained from the previous module and to make the right decision regarding the input instance, they created a weighted voting algorithm that collected the results of the four classifiers and calculated the most suitable output.[5]

Thakur, K et al. (2023) has identified several limitations and future directions for phishing email detections using deep learning. While deep learning has shown promising results in detecting phishing emails, there

is room for improvement. Some of the key limitations include the focus on a limited dataset without clear explanations on the dataset used, misclassification still existing, minimal time consumption, being limited to one language, and the lack of tools and resources required for research. To overcome these limitations, future research can focus on privacy preservation in phishing email detection, increasing the dataset size, optimizing feature selection mechanism using 8 evolutionary algorithms, and expanding the research to other languages and types of crises.[6]

Dinesh, K et al. (2023) highlights the importance of taking proactive measures to prevent phishing attacks. It also provides valuable insights into developing and testing a phishing detection algorithm using Natural Language Processing and Python. While the algorithm demonstrates high accuracy in detecting phishing emails, there is still room for improvement, particularly in considering attachments and subject lines. Future research could focus on integrating these tools with popular email services and developing real time alert systems for users. Overall, this paper contributes to the ongoing efforts to improve cybersecurity and protect against phishing attacks.[7]

Rehman, S et al (2023) implemented multiple classifiers of ML and a classifier of deep learning (DL) were applied to the SMS and e-mail dataset for spam detection with higher accuracy. After conducting experiments on the real dataset, the researchers concluded that the proposed system performed better and more accurately than previously existing models. Specifically, the support vector machine (SVM) classifier outperformed all others. These results suggest that SVM is the optimal choice for classification purposes.[8]

Abdul Basit et al. (2021) did an extensive review of AI-based methods for detecting phishing attacks, covering machine learning, deep learning, hybrid, and scenario-based approaches. It evaluates the strengths and weaknesses of each technique, highlighting challenges like high false positives and evolving phishing tactics. The authors suggest future research directions for developing more robust and scalable detection systems.[9]

Priyanka Verma et al. (2020) presented in four sections. An introduction about phishing its types, its history, statistics, life cycle, motivation for phishers and working of email phishing have been discussed in the first section. The second section covers various technologies of phishing- email phishing and also description of evaluation metrics. An overview of the various proposed solutions and work done by researchers in this field in form of literature review has been presented in the third section. The solution approach and the obtained results have been defined in the fourth section giving a detailed description about NLP concepts and working procedure.

Table 2.1 Literature survey on Phishing Email Detection

S. No.	Title	Authors	YOP	Algorithms	Remarks
1.	AI-Driven Phishing Email Detection - Leveraging Big Data Analytics for Enhanced Cybersecurity	Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Janardhana Rao Sunkara,	2024	(AI)-based technique for online phishing email optics (OPE) called "ABC-PEM-HPSO-RFO algorithm",high-	Provides valuable insights using AI and big data analytics for phishing email detection.

		Hemanth Kumar Gollangi		performance parallel swarm operators (HPSO)	Emphasizes the need for robust AI systems optimized for large-scale security operations.
2.	AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI	Kommisetty, P. D. N. K., & Nishanth, A	2024	Generative Adversarial Networks (GANs), Deep Learning (DL), Bidirectional Encoder Representations from Transformers (BERT)	Proposed ensemble techniques where various machine-learning algorithms are combined to improve the accuracy and strength of spam detection systems.
3.	Enhancing email spam detection through ensemble machine learning: a comprehensive evaluation of model integration and performance.	Al-shanableh, Najah; Alzyoud, Mazen S.; and Nashnush, Eman	2024	Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN)	The selected algorithms were applied individually on different datasets. As a result, the proposed approach produced a better performance with 95.8% accuracy.
4	Novel interpretable and robust web-based AI platform for phishing email detection	Abdulla Al-Subaiey, Mohammed Al-Thani, Naser Abdullah Alam, Kaniz Fatema Antora, Amith Khandakar, SM Ashfaq Uz Zaman	2024	Random Forest, Support Vector Machine, Logistic Regression, AdaBoost	XAI complexity: Interpretability tools can be difficult for non-technical users to understand.
5.	An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large	Suhaima Jamal, Hayden Wimmer	2024	Bidirectional Encoder Representations from Transformers (BERT)	Fine-tuned BERT-based model tailored for classifying phishing, spamemails.

	Language Model Approach				
6.	SpearBot: Leveraging Large Language Models in a Generative-Critique Framework	Qinglin Qi, Yun Luo, Yijia Xu, Wenbo Guo, Yong Fang	2024	Language Learning Model (LLM)	When a phishing email is identified by the critic, SpearBot refines the generated email based on the critique feedback.
7.	Phishing Email Detection Through Amygdala Hijack Threats	Chemudupati, Soumik; Valecha, Rohit,	2024	Support Vector Machine, Logistic Regression	Machine learning model to classify phishing emails based on the presence of threat cues.
8.	An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach.	Mohammad Amaz Uddin, Iqbal H. Sarker	2024	Explainable-AI (XAI), Local Interpretable Model-Agnostic Explanations (LIME)	Model effectively achieves high accuracy, demonstrating its capability to perform well.
9.	Unveiling the Efficacy of AI-based Algorithms in Phishing Attack Detection	Shahzad, T., & Aman, K	2024	14 AI algorithms such as RF, CNN, NB, KNN, LSTM, LR, ANN, AdaBoost, SVM	Among all the algorithms CNN, Multi-layer perceptron and AdaBoost achieved more than 90% accuracy.
10.	ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection	Koide, T., Fukushi, N., Nakano, H., & Chiba, D.	2024	Large language models (LLMs)	The system provides a highly accurate determination of whether an email is phishing or not.
11.	Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning.	Abdallah Ghourabi; Manar Alohaly	2023	Support Vector Machines, K-Nearest Neighbors, Light Gradient Boosting Machine (LightGBM)	Accuracies of the 10 classifiers using the GPT-3 embedding were all significantly superior to those of the classical embedding.

12.	A Systematic Review on Deep-Learning-Based Phishing Email Detection	Thakur, K.; Ali, M.L.; Obaidat, M.A.; A. Kamruzzaman	2023	Deep learning, Long Short-Term Memory (LSTM)	The literature review has identified several limitations and future directions for phishing email detections using deep learning.
13.	Phishing detection implementation using data bricks and artificial Intelligence	Kalla Dinesh, Fnu Samaah, Sivaraju Kuraku, and Nathan Smith	2023	Natural Language Processing (NLP), Natural Language Toolkit (NLTK).	Provides valuable insights into developing and testing a phishing detection algorithm using NLP and Python.
14.	An Intelligent Framework Based on Deep Learning for SMS and e-mail Spam Detection	Maqsood, U.; Ur Rehman, S.; Ali, T.; Mahmood, K.; Alsaedi, T.; & Kundi, M.	2023	Deep learning (DL), Support vector machine (SVM), Random Forest	On e-mail and SMS datasets, the researchers used NB, SVM, RF, and CNN. In both datasets, the SVM outperforms ML classifiers
15.	Improving Phishing Email Detection Using the Hybrid Machine Learning Approach	Palanichamy, Naveen & Murti, Yoga	2023	Term frequency-inverse document frequency (TF-IDF), Feature extraction technique (FET)	The research findings indicate that the hybrid model utilizing TF-IDF achieved superior performance, with an accuracy of 87.5%.
16.	Optimal Deep Belief Network Enabled Cybersecurity Phishing Email Classification	Dutta, A. K., et al	2023	Deep belief network (DBN), Term frequency-inverse document frequency (TF-IDF)	(DBN) model is used for the email classification and its efficacy can be boosted by the BBO based hyperparameter tuning process.

17.	An Enhanced Deep Learning-based Phishing Detection Mechanism to Effectively Identify Malicious URLs	Manoj Kumar Prabakaran, Parvathy Meenakshi Sundaram, Abinaya Devi Chandrasekar	2023	Variational Autoencoders (VAE), Deep neural networks (DNN)	The experimental results suggested that the proposed model has reached a maximum accuracy of 97.45%
18.	Phishing email detection using deep learning algorithms.	Singh, M. C., Sumanth, P., Sathyanarayana, S. B., & Rithika, G.	2022	Long Short-Term Neural Network (LSTM), Convolutional Neural Networks	The total accuracy of the experiment achieves a high percentage.
19.	A comprehensive survey of AI-enabled phishing attacks detection techniques	Abdul Basit & Maham Zafar & Xuan Liu & Abdul Rehman Javed & Zunera Jalil & Kashif Kifayat,	2021	Support Vector Machine, Random Forest, Artificial Neural Networks, Decision Trees, AdaBoost	Less Dynamic phishing tactics: Techniques may struggle to keep up with fast-evolving attack methods
20.	Email phishing: text classification using natural language processing	Verma, Priyanka & Goyal, Anjali & Gigras, Yogita	2020	SVC (support vector classifier), K Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier	Gives a detailed description on the classification of phishing emails using the natural language processing concepts

3. PROPOSED METHODOLOGY

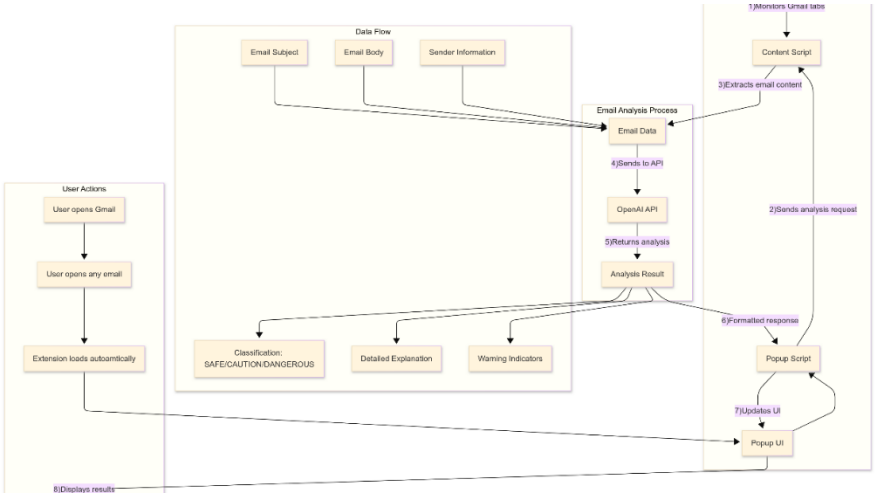

Figure 3.1 Design of the phishing detection extension

Figure 3.1 shows the main design and workflow of the extension. The system operates as a Chrome extension integrated with a phishing detection engine powered by OpenAI's API. The following components work in coordination to analyze and classify emails:

1. User Actions

- The process begins when the user opens Gmail and subsequently opens an individual email.
- The extension automatically runs, which triggers the analysis workflow.

2. Chrome Extension

- The **background script** continuously monitors browser tabs to detect Gmail usage.
- Once Gmail is detected, a **content script** is injected into the email view to extract the necessary information:
  - Email subject
  - Email body
  - Sender information

3. Data Flow

- The extracted data is organized into structured email data comprising:
  - Email Subject
  - Email Body
  - Sender Information
- This structured data is passed into the email analysis module.

4. Email Analysis Process

- The email data is sent to the **OpenAI API** with a specially crafted prompt designed to assess business legitimacy and risk indicators.
- The OpenAI model returns a detailed analysis result which includes:
  - A classification of the email as **SAFE** or **SUSPICIOUS** or **DANGEROUS**
  - A detailed explanation of the reasoning
  - A list of warning indicators if phishing is suspected

5. Popup UI and Result Display

The popup script formats the analysis response and updates the popup UI. The **results** are then shown to the user structured analysis with:

- Specific risk factors identified
- Overall confidence score (0-100%)
- Explanation of findings
- Recommended actions
- Flags emails with:
  - High Risk: ≥90% phishing probability
  - Medium Risk: 70-89% probability
  - Low Risk: <70% probability
- Analyzes key phishing indicators:
  - Sender Legitimacy and history
  - Communication tone and urgency
  - Credential/information request
  - Link/attachment patterns
  - Language and formatting

- Business context alignment

6. Data Flow

- The data of the email is collected like Sender & Receiver address, email subject, email body
- This data is sent as a request to the OpenAI API for analysis
- The response contains the classification, detailed explanation and warning indicators
- This response is formatted and displayed to the user using popup script

4. IMPLEMENTATION

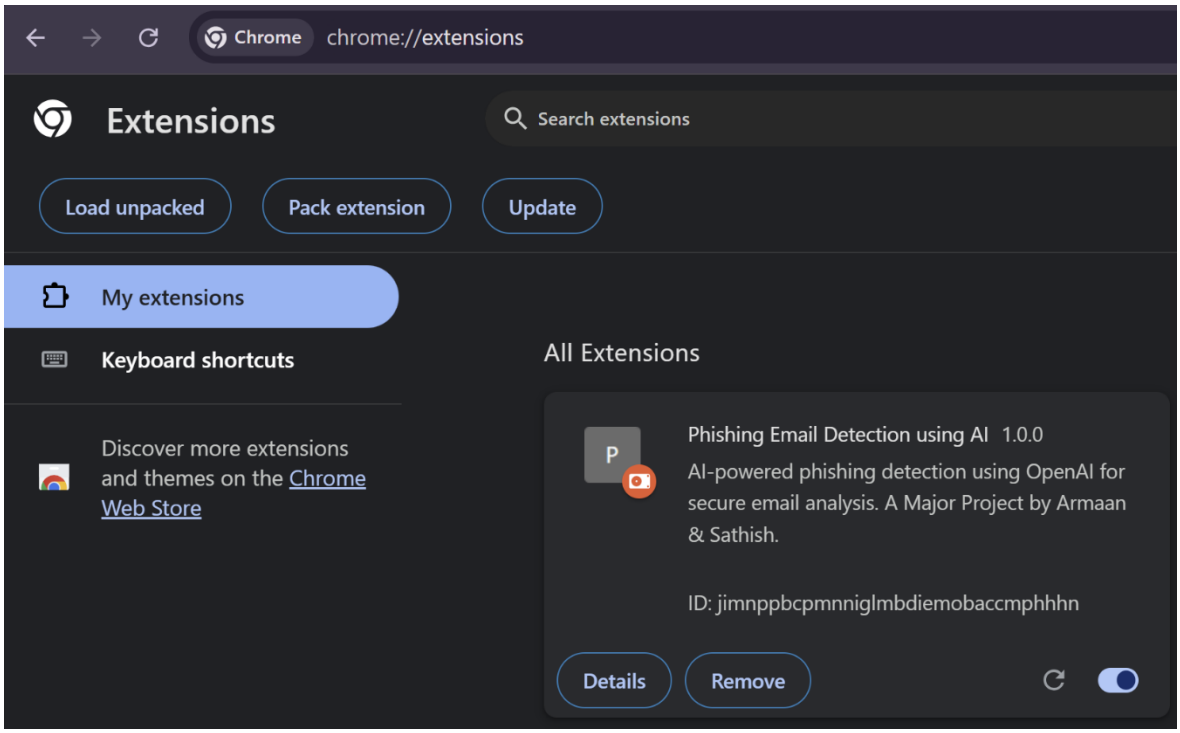

Figure 4.1 Loading extension in browser

We need to load the extension in the browser and to do that we have to perform the following steps:

1. Open your desired browser
2. Click “Manage Extensions” in settings
3. Turn on developer mode option
4. Click “Load Unpacked” button. It opens a window where we have to select the extension folder
5. Select “dist” folder from the project
6. The extension will be loaded automatically in the browser
7. Optionally, we can pin the extension for quick access

5. RESULTS AND DISCUSSION

Running the extension on safe email

The extension is automatically loaded on the Gmail website and there is no need for clicking on the extension manually. The user has to simply open Gmail website in preferred browser and open any email.

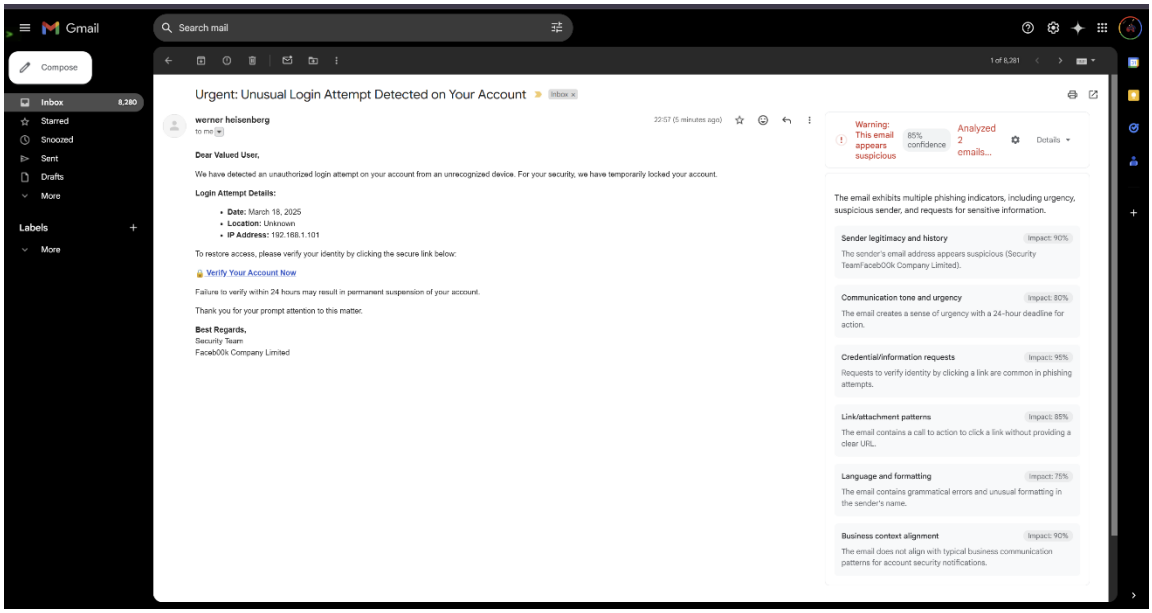

Figure 5.1 Running the extension on a safe email

We can see the detailed analysis of the email in **Figure 5.1**. It displays that the email is from "Kotak Bank" and the email is considered to be "SAFE" with 85% confidence. This confidence rating is calculated based on the following factors:

- Sender Legitimacy and history – 10% impact
- Communication tone and urgency - 5% impact
- Credential/information request - 5% impact
- Link/attachment patterns - 10% impact
- Language and formatting - 5% impact
- Business context alignment - 10% impact

Additionally, we can see the safe email is in green colour as a visual indicator

Running the extension on Phishing Email

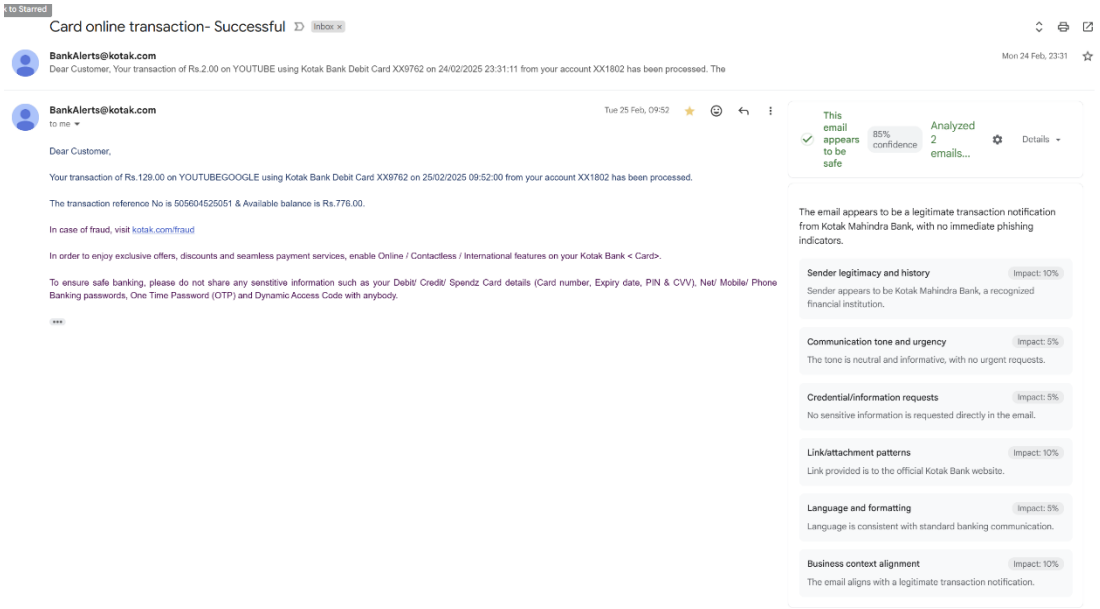

Figure 5.2 Running extension on Phishing email

We can see the detailed analysis of the email in **Figure 5.2**. It displays that the email is considered to be "Phishing" with 85% confidence. This confidence rating is calculated based on the following factors:

- Sender Legitimacy and history – 90% impact
- Communication tone and urgency - 80% impact
- Credential/information request - 95% impact
- Link/attachment patterns - 85% impact
- Language and formatting - 75% impact
- Business context alignment - 90% impact

We can see the suspicious email is in red color as a visual indicator.

6. FUTURE IMPROVEMENTS

Although the current system delivers promising results, the field of phishing detection continues to evolve, offering numerous avenues for further research and enhancement. As phishing techniques become more advanced and multilingual, it becomes essential to improve the model's sophistication and adaptability to stay ahead of attackers.

The following could be improved:

- **Multilingual Support**: Expanding the AI model's capabilities to detect phishing emails in multiple languages beyond English. This will significantly increase the system's utility in non-English-speaking regions like Hindi, Telugu etc. offering a more global solution to phishing threats.
- **Performance Optimization**: We can compress the model for faster interface and if possible, GPU Acceleration for faster classification of the email.
- **Custom Rule Framework**: Introducing a feature that allows cybersecurity professionals or end-users to define their own detection rules. This hybrid approach—combining rule-based and AI-based detection—can increase flexibility and cater to niche use cases or organizational requirements.
- **Internet Independency**: Since the application requires internet to classify the emails, there is a possibility to create a local LLM (Language Learning Model) that runs on the user personal computer which doesn't require internet.

7. CONCLUSION

The *Phishing Email Detection using AI* project leverages the power of artificial intelligence (AI), this project demonstrates a practical and scalable approach to detecting phishing emails with high accuracy and speed. The implementation validates the theoretical design through real-world scenarios, showcasing its applicability in protecting users from fraudulent messages that aim to extract sensitive information or manipulate user behaviour.

The integration of the AI model into a web-based application elevates the project's utility by offering an intuitive and interactive interface for users. This web extension empowers users to receive instant feedback on potential phishing threats directly within their email client, making the technology not only accessible but also user-friendly. This real-time detection mechanism significantly enhances the responsiveness and relevance of phishing prevention tools, providing a proactive defence against increasingly sophisticated phishing attacks.

List of References

1. Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Janardhana Rao Sunkara, Hemanth Kumar Gollangi (2024): AI-Driven Phishing Email Detection - Leveraging Big Data Analytics for Enhanced Cybersecurity. Library Progress International, 44(3), 7211-7224.
2. Kommisetty, P. D. N. K., & Nishanth, A. (2024). AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI. In IARJSET (Vol. 9, Issue 10)
3. Al-shanableh, Najah; Alzyoud, Mazen S.; and Nashnush, Eman (2024) "ENHANCING EMAIL SPAM DETECTION THROUGH ENSEMBLE MACHINE LEARNING: A COMPREHENSIVE EVALUATION OF MODEL INTEGRATION AND PERFORMANCE," Communications of the IIMA: Vol. 22: Iss. 1, Article 2.
4. Abdulla Al-Subaiey, Mohammed Al-Thani, Naser Abdullah Alam, Kaniz Fatema Antora, Amith Khandakar, SM Ashfaq Uz Zaman, Novel interpretable and robust web-based AI platform for phishing email detection, Computers and Electrical Engineering, Volume 120, Part A, 2024, 109625, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2024.109625.
5. Suhaima Jamal & Hayden Wimmer (2024) An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach.
6. Qinglin Qi, Yun Luo, Yijia Xu, Wenbo Guo, Yong Fang SpearBot: Leveraging Large Language Models in a Generative-Critique Framework for Spear-Phishing Email Generation (2024)
7. Chemudupati, Soumik and Valecha, Rohit, "Phishing Email Detection Through Amygdala Hijack Threats" (2024). NEAIS 2024 Proceedings.
8. Mohammad Amaz Uddin, Iqbal H. Sarker (2024) An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach. arXiv:2402.13871
9. Shahzad, T., & Aman, K. (2024). Unveiling the Efficacy of AI-based Algorithms in Phishing Attack Detection. Journal of Informatics and Web Engineering, 3(2), 116–133.
10. Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2024) ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection. arXiv:2402.18093
11. Ghourabi, A.; Alohaly, M. Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning. Sensors 2023, 23,3861.
12. Thakur, K.; Ali, M.L.; Obaidat, M.A.; Kamruzzaman, A. A Systematic Review on Deep-Learning-Based Phishing Email Detection. Electronics (2023), 12,4545.
13. Kalla, Dinesh, Fnu Samaah, Sivaraju Kuraku, and Nathan Smith. "Phishing detection implementation using databricks and artificial Intelligence." International Journal of Computer Applications 185, no. 11 (2023): 1-11.
14. Maqsood, U., Ur Rehman, S., Ali, T., Mahmood, K., Alsaedi, T., & Kundi, M. (2023). An Intelligent Framework Based on Deep Learning for SMS and e-mail Spam Detection. Applied Computational Intelligence and Soft Computing, 2023(1), 6648970.
15. Palanichamy, Naveen & Murti, Yoga. (2023). Improving Phishing Email Detection Using the Hybrid Machine Learning Approach. Journal of Telecommunications and the Digital Economy. 11. 120-142. 10.18080/jtde. v11n3.778
16. Dutta, A. K., et al. (2023) Optimal Deep Belief Network Enabled Cybersecurity Phishing Email Classification DOI:10.32604/csse.2023.028984
17. Prabakaran et al. (2023) An Enhanced Deep Learning-based Phishing Detection Mechanism to Effectively Identify Malicious URLs

18. Singh, M. C., Sumanth, P., Sathyanarayana, S. B., & Rithika, G. (2022). Phishing email detection using deep learning algorithms. International Journal of Health Sciences, 6(S3), 8130–8139.
19. Abdul Basit & Maham Zafar & Xuan Liu & Abdul Rehman Javed & Zunera Jalil & Kashif Kifayat, 2021. "A comprehensive survey of AI-enabled phishing attacks detection techniques," Telecommunication Systems: Modelling, Analysis, Design and Management, Springer, vol. 76(1), pages 139-154, January.
20. Verma, Priyanka & Goyal, Anjali & Gigras, Yogita. (2020). Email phishing: text classification using natural language processing. Computer Science and Information Technologies. 1. 1-12. 10.11591/csit.v1i1.p1-12.