EXPLORING NATURAL LANGUAGE PROCESSING CHATBOTS AND PHISHING WEBSITE DETECTION: A LITERATURE PERSPECTIVE

By

ROSHANI TALMALE * HARSHITA WANKHEDE ** PRANAV LOKHANDE ***

PRANAY THAKRE **** VRUNDA MISHRA *****

PRIYA DHOLE ****** BHAGYSHRI BALPANDE *******

*-******* *Department of Computer Science and Engineering, S. B. Jain Institute of Technology, Management and Research, Nagpur, India.*

https://doi.org/10.26634/jdf.2.2.21248

Date Received: 08/10/2024 *Date Revised: 14/10/2024* *Date Accepted: 20/10/2024*

ABSTRACT

The paper proposes a novel approach that integrates machine learning and NLP in order to identify phishing sites and create multilingual interaction by providing better user engagement through the chatbot. It will utilize the XG-Boost algorithm in order to show the phishing detection system with more than 90% accuracy rate in the identification and classification of websites as legitimate or phishing for a set of 10,000 websites. A major contribution of this work is the embedding of a multilingual chatbot, developed on Dialog- flow, with support for English, Hindi, and Marathi, thereby broadening the possible community of users for the system. This paper describes the architecture of the system at its different layers- feature extraction, model training, and its integration with the chatbot. The proposed work fills the gaps in earlier literature as it provides a user interface accompanied by robust detection. This system will also be extended by the provision of language support by adding more languages in the near future and also by increasing the detection accuracy using deep learning models. The results demonstrate that combining machine learning with user-centric design may improve the detection of phishing sites considerably and enhance user engagement.

Keywords: Phishing, Dialog-flow, XG-Boost, Convolutional Neural Network, Chatbot, Cyber Crime, Multilingual.

INTRODUCTION

In the ever-expanding realm of online services, phishing attacks have emerged as a significant threat in modern-day cybercrime. Cybercriminals exploit trust to steal sensitive information, including passwords, financial data, and private credentials. Traditional blacklists and heuristic methods struggle to keep pace with the evolving tactics of these attackers. As phishing techniques become increasingly sophisticated, distinguishing legitimate websites from fraudulent ones becomes a challenge for average users. This necessitates the development and deployment of advanced, automated detection systems.

Recent studies on phishing detection using machine learning algorithms show significant potential. However, many solutions currently available in the market fail to address crucial aspects such as real-time detection capabilities and support for non-English languages. This paper fills in those gaps by proposing a dual-component system: the XG-Boost-based phishing detection model and a multilingual chatbot that would help support users in three languages, namely, English, Hindi, and Marathi (Butnaru et al., 2021). The phishing detection model reaches an accuracy of more than 90%, thereby ensuring that the developed tool would provide a dependable way of identifying malicious websites. Multilingual support to the chatbot further enhances user engagement since it would support real- time support and educational content in multiple languages.

The novelty of this work is that it has combined a phishing detection with a user-friendly multilingual chatbot.

1. Literature Review

Alswailem et al. (2019) highlighted the alarming growth in phishing attacks, where human weaknesses, rather than software vulnerabilities, are exploited to steal sensitive information such as usernames, passwords, and financial data from users. They suggest an intelligent phishing detection system based on machine learning, utilizing the RF algorithm. The system is designed as an automatic browser extension that notifies users whenever they attempt to navigate to a potentially fraudulent site. The research methodology extracts features such as the website's URL, page content, and page rank from primary sources. Initially, thirty-six features are considered, but the system narrows down to optimize performance through feature selection. The Random Forest algorithm was chosen for its proven higher accuracy in classification. These features were thoroughly tested, and 26 were found to effectively enhance the system's accuracy to 98.8 percent.

Korkmaz et al. (2020) focus on phishing website detection by analyzing the structure of URLs using machine learning techniques. Phishing websites typically mimic legitimate ones to deceive users into providing sensitive information, including passwords and credit card numbers. This study uses several machine learning algorithms to classify URLs as either legitimate or phishing based on 58 different features extracted from the URLs. Key elements such as URL length, special characters, and specific tokens are considered. The experiment uses three datasets sourced from PhishTank and Alexa, totaling over 126,000 URLs. The aim is to develop a fast and accurate phishing detection system that does not rely on third-party services.

Kumar et al. (2020) proposed a machine learning-based methodology for detecting phishing websites by analyzing various URL-based features. In the current digital world, phishing attacks have become a significant issue, as they target victims through fake websites masquerading as legitimate ones to steal usernames, passwords, and, most importantly, financial data. These attacks primarily occur in the banking and e-commerce industries, where users unknowingly share their details on phishing sites. The work presented in this study provides a machine learning framework to classify URLs as either phishing or legitimate based on key features extracted from the URLs. The paper begins with a description of phishing attacks, stating that attackers replicate trusted sites to deceive victims into disclosing sensitive information. Most of these attacks are carried out using fake URLs, which are nearly identical to legitimate web pages, causing users to believe they are interacting with trusted services. Traditional phishing attack detection methods, such as blacklists, whitelists, and heuristic approaches, are identified as insufficient, particularly for new phishing sites or zero-day attacks.

Chatterjee and Namin (2019) propose a new approach to detect phishing websites using a Deep Reinforcement Learning (DRL) model. Phishing is one of the simplest and most effective forms of cybercrime, baiting individuals into providing sensitive information such as personal data, banking details, or login credentials. Traditional methods, including blacklisting and heuristic analysis, are static in nature and do not always guarantee effectiveness against newly created or highly dynamic phishing websites. A DRL-based method is proposed to adapt to changes in phishing websites while learning their associated features for phishing URL detection. A deep neural network is used to describe a model that identifies whether a URL is phishing or benign based on 14 lexical features of the URL. Extensive experimentation was conducted on the Ebbu2017 Phishing Dataset, and the proposed method proves to be dynamic and self-adaptive, showing promising performance metrics.

Abedin et al. (2020) explored the use of machine learning methods to distinguish between phishing and legitimate websites. Phishing attacks are mechanisms of user deception, wherein sensitive information is extracted from users by luring them to fake sites that resemble legitimate ones. Three supervised machine learning algorithms were used in this experiment: K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest algorithms, applied to a dataset of URL-based features. The results indicate that Random Forest may perform best, with a precision rate of 97%, a recall of 99%, and an F1 score of 98%. The paper concludes that the Random Forest classifier is effective for phishing detection and may be further improved by adding more features or using deep learning models.

Alkawaz et al. (2021) provide a comprehensive review of how machine learning techniques are applied to the identification and detection of phishing websites. Phishing represents a cybercrime in which an attacker assumes the identity of a legitimate website to steal users' private information. Blacklists prove ineffective against dy

2. Proposed System

The proposed work will focus on the development of a complete system, integrating machine learning into a multilingual chatbot for phishing detection and promoting user interaction. The primary components of this system are data collection, feature extraction, model training for machine learning, and integration with a chatbot (Sahingoz et al., 2019). Figure 1 shows the system architecture, showcasing the flow of these components. The system is designed to enhance the accuracy of phishing detection and offer real-time multilingual support to users through a chatbot interface. Detailed elaborations on the major steps in the process are described below:

2.1 Phishing and Legitimate Dataset Collection

This involves the collection of phishing and legitimate datasets, which are used to train and test the model.

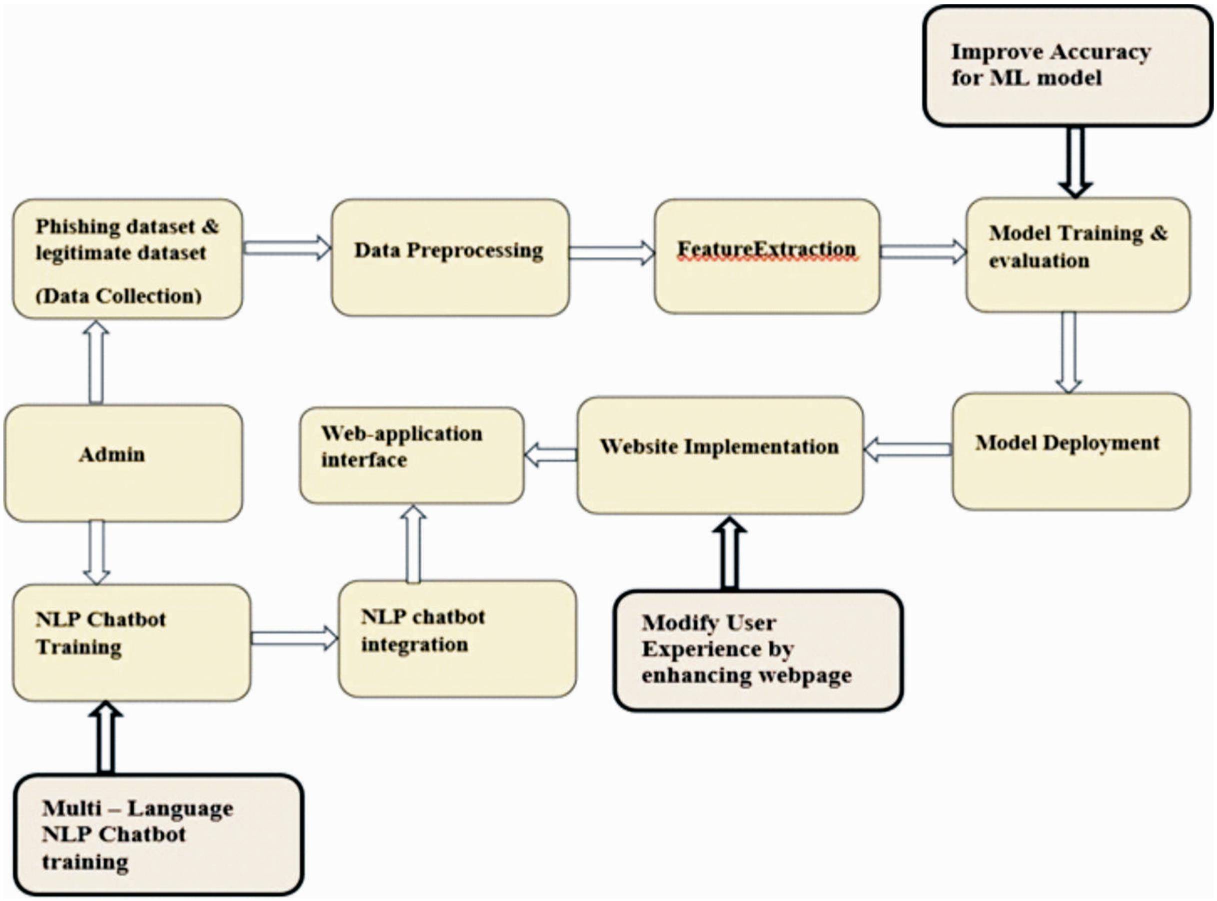
Improve Accuracy for ML model
Phishing dataset & legitimate dataset (Data Collection)
Data Preprocessing
FeatureExtraction
Model Training & evaluation
Admin
Web-application interface
Website Implementation
Model Deployment
NLP Chatbot Training
NLP chatbot integration
Modify User Experience by enhancing webpage
Multi – Language NLP Chatbot training

Figure 1. System Architecture

2.2 Preprocessing of Data

The preprocessing of data involves cleaning raw data by removing duplicate entries, eliminating missing values, and normalizing the information. These steps are crucial for enhancing the accuracy of the model.

2.3 Feature Extraction

The given model identifies a phishing website from a legitimate one because it focuses on some critical features such as the characteristics of the URL, ages of domains, and actions of the website.

2.4 Model Training and Evaluation

The preprocessed data used to train machine learning models. Here, it is evident that XG-Boost is the best performer. Metrics used were accuracy, precision, and recall for model evaluation.

2.5 Model Deployment

Model deployment involves deploying the trained and fine-tuned model to analyze websites in real time and begin detecting phishing threats.

2.6 Enhancing the Accuracy of the ML Model

Enhancing the accuracy of the ML model involves the use of optimization techniques, such as hyperparameter tuning, to continually improve its performance.

2.7 Admin Role

Admin manages the entire system by managing both models to work smoothly without any issues.

2.8 Training the NLP Chatbot

The chatbot is trained on the platform using Dialog-flow in various languages to enable its users who ask them questions in the English, Hindi, as well as in the Marathi languages. The Chatbot with multilingual NLP training is learned to generate highly smooth multi-lingual responses, so that users would be able to interact with it in their native language (Kaushik & Rahul, 2023).

2.9 NLP Chatbot Integration

The chatbot is integrated into the system in order to respond to user queries and to assist the user during the phishing detection process.

2.10 Web Application Interface

The system is exposed through a web application, and users can submit URLs and interact with chatbot for phishing detection.

2.11 Website Deployment

Website deployment involves deploying all phishing detection models and other chatbot components as the final version of the system on a functional website.

2.12 The User Experience Enhanced

The website interface is regularly improved so that the website becomes better accessible and convenient for all users.

3. Findings

The focus has primarily been on improving the accuracy of phishing detection using machine learning models such as Random Forest and Decision Tree. However, few techniques provide real-time user interaction or multi-language support. Table 1 shows how the proposed work addresses this gap by utilizing XG-Boost, achieving more than 90% accuracy, while also incorporating a multilingual chatbot, enhancing accessibility for users in English, Hindi, and Marathi. The system offers real-time interaction in multiple languages and provides immediate support by educating users about phishing threats. This contrasts with traditional techniques, which lack such user-centric features.

4. Application

The application of this research will have a far-reaching impact on cybersecurity by offering an accessible and effective system to detect phishing websites and interact with users through a multilingual chatbot. It enhances user security by providing real-time analysis of website authenticity, thereby preventing phishing attacks. The multilingual chatbot ensures that a wider audience can use the system without language barriers, making cybersecurity tools accessible to diverse linguistic communities. This system is also of immense worth for educational institutions, corporate enterprises, financial institutions, and government agencies where phishing attacks are prevalent and frequently turn out to be serious data breaches or financial losses. This system enhances

Ref No	Author	Methods	Advantages	Challenges	Accuracy
[1]	Alswailem et al. (2019)	Random Forest Classifier, Dataset of 16,000 URLs and 36 features were considered.	High accuracy, Efficient feature selection and Real-time detection.	False positives might still occur depending on feature selection and Potential hardware limitations impact processing time.	Achieved 98.8% accuracy using Random Forest.
[2]	Korkmaz et al. (2020	Random Forest, XGBoost, Decision Tree are used and 58 URL features were initially extracted.	High Accuracy, Efficient Processing and Comprehensive Feature Set.	Require significant computational resources and longer training times. and Dependency on URL Structure.	The Random Forest (RF) algorithm achieved the highest accuracy, with 94.59%.
[3]	Kumar et al. (2020)	K-Nearest Neighbors, Decision Tree, Logistic Regression are used.	High Accuracy and Effective Detection.	Limited Dataset and require significant computational power.	The Random Forest algorithm performed the best, with a accuracy of 91.4%.
[4]	Chatterjee and Namin (2019)	Deep Reinforcement Learning Model	Adaptiveness, Dynamic Learning and Feature-Based Classification.	Feature Selection Limitations and not yet optimized for real-world deployment.	Precision: 86.7% Recall: 88% Accuracy: 90.1% F-measure: 87.30%
[5]	Abedin et al. (2020)	Random Forest and 32 attributes, such as UsingIP, LongURL, HTTPS, SubDomains are used.	High Precision and Recall and Scalability.	Feature Selection and Zero-Day Attack Detection.	Precision: 97% Recall: 99% F1 Score: 98% ,AUC Score:1.0
[6]	Alkawaz et al. (2021)	Decision Tree and Random Forest are used.	High Accuracy, Dynamic Learning and Reduced False Positives.	Computational Complexity and Handling Zero-Day Attacks.	Random Forest classifier performed the best, achieving an accuracy of 97.14%
[7]	Anil et al. (2020)	Random Forest and Support Vector Machine are used.	High Accuracy and Automated Feature Extraction.	Handling Evolving Phishing Technique and Computational Complexity.	Accuracy: 97.10%
[8]	Kumar et al. (2020)	Logistic Regression, Random Forest, Decision Tree, K- Nearest Neighbor are used.	Comparative Analysis and High Accuracy.	Class Imbalance and Generalization.	The Random Forest and Decision Tree classifiers both achieved an accuracy of 98%.
[9]	Chawla (2022)	30 parameters are used along with Max Vote Classifier algorithm.	Robust Model and Comparative Analysis.	Data Imbalance and Scalability	The Max Vote Classifier achieved the highest accuracy of 97.73%
[10]	Teja et al. (2020)	Random Forest, Support Vector Machine, Gradient Boosting Tree, Artificial Neural Network.	High accuracy and robust performance across all metrics.	Hyperparameter tuning in models.	The Random Forest classifier achieved the highest accuracy of 97%
[11]	Sangeetha et al. (2021)	Speech Recognition and Intent Classification and Entity Extraction	Customization, Hands-free Operation and Efficiency	Accent Recognition Issues and Training Data Limitations.	The Multinomial Naive Bayes model achieved an accuracy of 75.43%
[12]	Mulik et al. (2021)	The chatbot uses a Keras sequential model and employs ReLU and softmax activation functions	User-Friendly, Spell Check and Context Maintenance and Low Computational Cost	Rigid Responses and Accent and Input Issues.	The model was tested with five different optimizers, with Adagrad achieving the highest accuracy at 99.08%
[13]	Aleedy et al. (2019)	The study uses three models: LSTM, GRU, and CNN, trained on 700,000 query-response pairs	Accurate Response Generation and Efficient Handling of Large Datasets.	Informative Queries and longer training times.	LSTM achieved the best results with a BLEU score of 0.36.
[14]	Lalwani et al. (2018)	Personal Query Response, AIML Response System and Query Analysis and Response System.	Efficient Information Retrieval, User- Friendly and Continuous Improvement.	Limited Flexibility and Accuracy Dependence on Keywords.	Responses are generated with a confidence threshold of 0.5.
[15]	Safi and Singh (2023)	List-Based Techniques, Visual Similarity Techniques and Deep Learning are used.	High Accuracy and Adaptability	Data Requirements and Evasion Tactics	CNNs achieved the highest accuracy in phishing detection at 99.98%
[16]	Zuraiq and Alkasassbeh (2019)	Content-Based Approach, Heuristic-Based Approach and Fuzzy Rule-Based Approach are used.	Effective at identifying phishing sites and can handle ambiguity.	Content-Based Approach and Requires large amounts of data.	Content-based approach: 99.14% heuristic-based method: 98.23%

(Contd...)

(...Contd.)

Ref No	Author	Methods	Advantages	Challenges	Accuracy
[17]	Garje et al. (2021)	K-Nearest Neighbors, Naive Bayes and Decision Tree	High Accuracy	Feature Complexity and Data Imbalance.	The Decision Tree model performed the best, achieving an F1 score of 0.94
[18]	Flayh (2023	URL Analysis, Content Analysis and Security Indicator Analysis.	Effective Feature Extraction, Automation and High Accuracy.	Resource-Intensive and Evolving Tactics	Random Forest, can achieve an accuracy of up to 99%
[19]	Mahajan & Siddavatam (2018)	Decision Tree, Random Forest and Support Vector Machine	Scalability and Low False Positive Rate	Data Dependence	Random Forest: 97.14% Decision Tree: 97.11% Support Vector Machine: 96.51%
[20]	Dutta (2021)	Feature Extraction and RNN-LSTM	Effective for Real- Time Detection, Improved Learning Rate and High Accuracy	Complex Model and Feature Dependence	Accuracy: 97.4% F1-Score: 96.4%
[21]	Theja and Krishnaveni (2013)	MD5 Hashing Algorithm and Session Key Authentication	Enhanced Security and Mobile Integration	Mobile Vulnerabilities and Limited Scope	-
[22]	Adebowale et al. (2019)	35 features are extracted from websites, including text-based, frame- based and image- based.	High Accuracy and Real-Time Detection	Complexity of Feature Extraction and Dependence on Dataset Quality.	Achieved an accuracy of 98.55% for text features, 98.06% for frame features, 97.2% for image features
[23]	El Aassal et al. (2020)	PhishBench integrates 226 features and machine learning algorithms like Random Forest, SVM, AutoML	Comprehensive Benchmarking, Customizable and Extensible and High Performance.	Computational Resources and Dataset Diversity	Random Forest and AutoML were among the top- performing classifiers, achieving F1- scores of over 98%.
[24]	Aljofey et al. (2020)	The model uses character-level embedding and one-hot encoding.	Independence from Third-Party Services and Zero-Day Attack Detection.	Training Time and Potential Misclassifications	The model achieved an accuracy of 95.02%
[29]	Abdulla et al. (2022)	Artificial Intelligence Markup Language and Natural Language Processing	Scalability and 24/7 Availability	Scalability and 24/7 Availability	-
[30]	Baby et al. (2017)	The chatbot is built using NLP.	24/7 Availability and Improved User Interaction	Limited Conversational Scope and less efficient	-
[32]	Singh et al. (2023)	Multilingual BERT Models and Fixed-Response and Context-Based QA	Multilingual Support, Efficient Query Processing and Contextual Answering.	Representation of Languages and Scalability	Top-1 Accuracy: 60.73% Top-2 Accuracy: 70.73% Top-3 Accuracy: 76.34%

Table 1. Overview of Literature

security with the integration of machine learning and natural language processing in addition to raising awareness among users about safe browsing (Sharma, 2012).

4.1 Key Application

4.1.1 Detection of Phishing Website for End-User

The system provides real-time detection of a phishing website so that the users can validate the authenticity of the website in real-time and thus avoid an impending cyber- attack.

4.1.2 Multilingual User Support

This system's chatbot engages in conversation in English, Hindi, and Marathi, thereby making cybersecurity tools accessible to end-users belonging to any linguistic background.

4.1.3 Corporate and Enterprise Usage

The tool can be integrated into the company security infrastructures so that employees can evaluate the website before engaging with them, thus reducing risks of phishing scams on the office workplace.

4.1.4 Advance Cyber Security for Financial Institutions

Protects the financial industries by giving customers a reliable method to verify whether a site is authentic before they can do business with it, thus reducing fraudulent crimes.

4.1.5 Government and Public Sector Adoption

Government websites are provided with enhanced security against phishing attacks, ensuring that citizens cannot access malicious services and can communicate safely through official portals.

5. Advantages

- *High Accuracy in Phishing Detection:* It is achieved by this system with the help of the XG-Boost algorithm, resulting in more than 90% accuracy and a drastic reduction of false positives and false negatives. Reliable and precise results are provided to users, ensuring the maximum level of security is attained.
- *User Accessibility Multilingual Chatbot:* Effective communication is provided by a chatbot integration based on Dialogflow in English, Hindi, and Marathi. The usability of the tools is enhanced by this multilingual feature, bringing it within reach of more user groups, regardless of their language proficiency.
- *Real-Time URL Analysis:* The system can actually analyze URLs in real time, thus immediately informing the users as regards the legitimacy of the website. The real-time ability is much needed because it helps prevent the users from falling prey to phishing attacks in real-time applications.
- *Ease of Interaction:* The smooth interface of this system makes interaction easy, making it accessible even to nontechnical users. The increased level of user interaction with the system makes the phishing detection process more approachable and exciting.

6. Limitations

- *Dependency on Quality and Diversity of Dataset:* Effective functioning depends on the quality and diversity of the dataset. Failure in representing the newer phishing techniques may limit its effectiveness to detect the changing trends in phishing activities.
- *Limited Language Support Beyond Three Languages:* Although three languages, English, Hindi, and Marathi, are supported, the language support is limited in terms of diversity. Communication in local and international languages may be hindered, especially in the globalized world.
- *Complexity in Integration with Organizational Cybersecurity Infrastructure:* Challenges may arise in integrating the system with existing frameworks in organizations due to variations in technical requirements and potential compatibility issues with different infrastructures.

7. Future Scope

Future research would provide extensive applicability through the scope of phishing detection systems. As threats in cyberspace keep on changing, upgrading the system with more robust machine learning models, such as deep learning techniques, may be possible for detecting the most advanced phishing attacks. The continuous upgrading of the dataset with new phishing tactics would be important to support high levels of detection accuracy. This would further enhance the multilingual capabilities of the chatbot by adding more languages, thereby increasing its accessibility to global users and making the system more versatile.

A direction for future work is implementing a system with real-time monitoring and proactive threat detection, where the system automatically flags potentially malicious sites before users interact with them. This adds an additional layer of security by preventing any initial interaction with phishing sites. Integration with browser extensions or mobile applications would provide a seamless user experience, allowing for easy adoption and use. Furthermore, the system could be personalized for specific industries such as finance, healthcare, and e-commerce, which are frequently targeted by phishing attacks. Tailoring the system to specific sectors might enhance security against targeted phishing campaigns. Establishing shared databases or threat intelligence with cybersecurity agencies and organizations could also contribute to more robust global efforts in phishing attack detection.

Conclusion

The work presents an integrated approach to address the serious problem of phishing by using machine learning and natural language processing. A detection system will be implemented using the XG-Boost algorithm, which results in an accuracy of more than 90% for identifying phishing websites from a dataset of 10,000 samples. This effectiveness demonstrates that it is indeed possible to protect users from advanced cyber threats with the proposed approach. Furthermore, the multilingual chatbot provides instant help and guidance in English, Hindi, and Marathi, making the system accessible to a wider audience.

The phishing detection mechanism might improve interaction with the corresponding chatbot but allows a user to be informed about possible risks, encouraging proactive measures toward cybersecurity. Detection algorithms must, similarly, be updated concerning cybercrime tactics, and the dataset size should be increased for constant effectiveness. Thus, the conclusion of this study emphasizes combining advanced technological solutions with user-centered interfaces to further develop a safer and more informed online environment.

References

[1]. Abdulla, H., Eltahir, A. M., Alwahaishi, S., Saghair, K., Platos, J., & Snasel, V. (2022, July). Chatbots development using natural language processing: A review. In 2022 *26th International Conference on Circuits, Systems, Communications and Computers (CSCC)* (pp. 122-128). IEEE.

https://doi.org/10.1109/CSCC55931.2022.00030

[2]. Abedin, N. F., Bawm, R., Sarwar, T., Saifuddin, M., Rahman, M. A., & Hossain, S. (2020, December). Phishing attack detection using machine learning classification techniques. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1125-1130). IEEE.

https://doi.org/10.1109/ICISS49785.2020.9315895

[3]. Adebowale, M. A., Lwin, K. T., Sanchez, E., & Hossain, M. A. (2019). Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. *Expert Systems with Applications,* 115, 300-313.

https://doi.org/10.1016/j.eswa.2018.07.067

[4]. Aleedy, M., Shaiba, H., & Bezbradica, M. (2019). Generating and analyzing chatbot responses using natural language processing. *International Journal of Advanced Computer Science and Applications,* 10(9), 60-68.

[5]. Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics,* 9(9), 1514.

https://doi.org/10.3390/electronics9091514

[6]. Alkawaz, M. H., Steven, S. J., Hajamydeen, A. I., & Ramli, R. (2021, April). A comprehensive survey on identification and analysis of phishing website based on machine learning methods. In *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 82-87). IEEE.

https://doi.org/10.1109/ISCAIE51753.2021.9431794

[7]. Alswailem, A., Alabdullah, B., Alrumayh, N., & Alsedrani, A. (2019, May). Detecting phishing websites using machine learning. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-6). IEEE.

https://doi.org/10.1109/CAIS.2019.8769571

[8]. Anil, G. N., Prakash, G. O., Manoj, K. H., Lokesh, M., & Madhusudhan, K. M. (2020). Detection of phishing websites based on feature extraction using machine learning. *International Research Journal of Engineering and Technology (IRJET),* 7 (7), 476-481.

[9]. Baby, C. J., Khan, F. A., & Swathi, J. N. (2017, April). Home automation using IoT and a chatbot using natural language processing. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-6). IEEE.

[10]. Butnaru, A., Mylonas, A., & Pitropakis, N. (2021). Towards lightweight URL-based phishing detection. *Future Internet,* 13(6), 154.

https://doi.org/10.3390/fi13060154

[11]. Chatterjee, M., & Namin, A. S. (2019, July). Detecting phishing websites through deep reinforcement learning. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC),* 2, 227-232.

IEEE.

https://doi.org/10.1109/COMPSAC.2019.10211

[12]. Chawla, A. (2022). Phishing website analysis and detection using Machine Learning. *International Journal of Intelligent Systems and Applications in Engineering,* 10(1), 10-16.

[13]. Dutta, A. K. (2021). Detecting phishing websites using machine learning technique. *PloS one,* 16(10), e0258361.

https://doi.org/10.1371/journal.pone.0258361

[14]. El Aassal, A., Baki, S., Das, A., & Verma, R. M. (2020). An in-depth benchmarking and evaluation of phishing detection research for security needs. *IEEE Access,* 8, 22170-22192.

https://doi.org/10.1109/ACCESS.2020.2969780

[15]. Flayh, N. A. (2023). Phishing website detection using machine learning: A review. *Wasit Journal for Pure Sciences,* 2(2), 270-281.

[16]. Garje, A., Tanwani, N., Kandale, S., Zope, T., & Gore, S. (2021). Detecting phishing websites using machine learning. *International Journal of Advances in Engineering and Management (IJAEM),* 3 (4), 496-503.

[17]. Kaushik, S., & Rahul. (2023). Chatbot using Natural Language Processing (NLP) techniques. *Journal of Emerging Technologies and Innovative Research (JETIR),* 10 (9), 1-17.

[18]. Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, July). Detection of phishing websites by using machine learning-based URL analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

https://doi.org/10.1109/ICCCNT49239.2020.9225561

[19]. Kumar, D. N., Hemanth, N. S. R., Premnath, S., Kumar, V. N., & Uma, S. (2020). Detection of phishing websites using an efficient machine learning framework. *International Journal of Engineering Research and Technology,* 9(5), 1282-1286.

[20]. Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B. S. (2020, January). Phishing website classification and detection using machine learning. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). IEEE.

https://doi.org/10.1109/ICCCI48352.2020.9104161

[21]. Lalwani, T., Bhalotia, S., Pal, A., Bisen, S., & Rathod, V. (2018). Implementation of a Chatbot system using AI and NLP. *International Journal of Innovative Research in Computer Science & Technology,* 6(3), 26-30.

[22]. Mahajan, R., & Siddavatam, I. (2018). Phishing website detection using machine learning algorithms. *International Journal of Computer Applications,* 181(23), 45-47.

[23]. Mulik, D. S., Sawant, P., & Bhosale, V. (2021). Application of NLP: Design of Chatbot for new research scholars. *Turkish Online Journal of Qualitative Inquiry,* 12(8), 2817-2823.

[24]. Patra, B., & Kumar, M. (2020). Natural language processing in Chatbots: A review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* 11(3), 2890-2894.

[25]. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University-Computer and Information Sciences,* 35(2), 590-611.

https://doi.org/10.1016/j.jksuci.2023.01.004

[26]. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications,* 117, 345-357.

[27]. Sangeetha, S., Sahithya, C., Rasiga, M. R., & Shalini, N. (2021). Chatbot for personal assistant using natural language processing. *International Journal of Research in Engineering, Science and Management,* 4(3), 96-97.

[28]. Sharma, R. (2012). An analysis of an intelligent chatbot using natural language processing. *International Journal of Food and Nutritional Sciences (IJFANS),* 11 (6), 936-942.

[29]. Singh, U., Vora, N., Lohia, P., Sharma, Y., Bhatia, A., & Tiwari, K. (2023, July). Multilingual chatbot for Indian languages. In *2023 14th International Conference on Computing Communication and Networking*

Technologies (ICCCNT) (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCNT56998.2023.10307978

[30]. Teja, C. S. B., Sasank, T., & Reddy, Y. (2020). Phishing website detection using different machine learning techniques. *International Research Journal of Engineering and Technology (IRJET)*, 7 (10), 607-610.

[31]. Theja, Y. R., & Krishnaveni, R. (2013). Security based phishing website detection. *International Journal of Computer Science and Mobile Computing*, 2 (4), 523-527.

[32]. Zuraiq, A. A., & Alkasassbeh, M. (2019, October). Phishing detection approaches. In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICTCS.2019.8923069

ABOUT THE AUTHORS

Dr. Roshani Talmale is working as an Assistant Professor in the Computer Science Department at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. She completed her Ph.D. degree at Vignan's University, Andhra Pradesh, India. She received her B.Tech degree in Computer Science from S.N.D.T. University, Mumbai, in 2002 and her M.E. degree from RTM Nagpur University, Nagpur, Maharashtra, India, in 2014. She has over 15 years of academic experience in various institutes.

Harshita Wankhede is working as an Assistant Professor in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. She has over 5 years of academic experience in various institutes.

Pranav Lokhande is a Student in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With a keen interest in Web Development, Artificial Intelligence, and IoT, he is proficient in programming languages such as Python, HTML, and C.

Pranay Thakre is a Student in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With a keen interest in Data Science, Data Analytics and Full-Stack Development, he is proficient in programming languages such as C, Python, Java, SQL, and PHP.

Vrunda Mishra is a Student in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With a keen interest in Data Science, Machine Learning and NLP, she is proficient in programming languages such as Python, HTML and Java.

Priya Dhole is a Student in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With a keen interest in Data Analytics and Visualization, she is proficient in programming languages such as Python, HTML, SQL and C.

Bhagyshri Balpande is a Student in the Department of Computer Science and Engineering at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With a keen interest in Artificial Intelligence and Data Science, she is proficient in programming languages such as Python, C, and Java.