IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

Suspicious Link Detection using AI

Prof. Amar Palwankar[1], Afiya Borkar[2], Pranali Shingare[3], Rifah Solkar[4], Shreya Khedaskar[5]
Assistant Professor, Department of Information Technology[1]
Students, Department of Information Technology[2,3,4,5]
Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Abstract: *With the increase in internet usage, cybersecurity has become a major concern for computer systems, as malicious URLs can release different forms of malware and attempt to collect user data. The global lockdown in 2020 led to a significant rise in the use of internet services for business, which in turn resulted in a surge of cybercrimes committed by cybercriminals and significant data losses for businesses. To prevent such attacks, it is important to identify malicious URLs and understand the types of threats they pose. Signature-based approaches are often used to find such websites, and security tools are deployed to impose access restrictions on them. This chapter proposes using the linguistic aspects of related URLs to enhance the effectiveness of classifiers for identifying dangerous websites through Machine Learning algorithms such as Logistic Regression and Random Forest Technique. The study shows that being able to identify spam URLs solely based on URLs and categorizing them without relying on page content can lead to significant resource savings and a safer browsing experience for users.*

Keywords: Suspicious URL Detection, Machine Learning, Supervised Learning, Logistic Regression, Random Forest ,Cybersecurity.

I. INTRODUCTION

The internet has become an integral part of our lives, connecting people and businesses across the globe. However, with the convenience and accessibility of the internet comes the risk of cyber attacks. Phishing and malware attacks are some of the most common cyber threats, and can cause significant harm to individuals and businesses. In response to this threat, various tools and technologies have been developed to help detect and prevent these attacks. One such tool is Suspicious Link Detection, a web application designed to detect and classify suspicious links.

Suspicious Link Detection is a web application that uses machine learning models to classify URLs as safe or suspicious. The application leverages two trained models, Logistic Regression and Random Forest, to classify URLs based on various features. Logistic Regression is a binary classifier that predicts whether a URL is safe or suspicious, while Random Forest is a decision tree-based classifier that can handle non-linear and complex data. Both models have been trained on a large dataset of URLs, and are designed to improve over time as more data is added.

In addition to the machine learning models, Suspicious Link Detection also uses the VirusTotal API to further verify the safety of URLs. VirusTotal is a free online service that analyzes URLs and files for malware and viruses. Suspicious Link Detection sends the URLs to VirusTotal for analysis and receives a report on the safety of the URL. This report is then used to provide the user with a more accurate assessment of the safety of the URL.

Furthermore, Suspicious Link Detection also uses the GNews API to provide users with up-to-date information on malware and phishing attacks. This API provides access to news articles and reports on the latest cyber threats, which can help users stay informed and take proactive measures to protect themselves.

Overall, Suspicious Link Detection provides a comprehensive solution for suspicious link detection, combining the power of machine learning models with external APIs to provide users with accurate and up-to-date information on the safety of URLs. This paper presents the design and implementation of Suspicious Link Detection, as well as the methodologies used to train and evaluate the machine learning models. We also discuss the challenges and limitations of current approaches to detecting suspicious links, and how Suspicious Link Detection overcomes some of these limitations. Finally, we evaluate the effectiveness of Suspicious Link Detection in detecting suspicious links, using real-world examples and user feedback.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

II. PROJECT SCOPE

This project aims to develop a machine learning model that can detect potentially malicious links present in online content. This project is crucial to prevent users from falling victim to cyber attacks, such as phishing and malware, that can compromise their online security and privacy.

The scope of this project includes developing and evaluating the performance of the machine learning models, building a web application for link scanning, and conducting experiments to compare the effectiveness of logistic regression and random forest. The project does not include developing a comprehensive cybersecurity solution or addressing all potential security threats.

III. PROBLEM STATEMENT

The problem that Suspicious Link Detection is addressing is the rise of online phishing and malware attacks. These types of attacks can cause significant harm to individuals, organizations, and even entire networks. Phishing attacks involve fraudulent attempts to obtain sensitive information, such as passwords and credit card numbers, by masquerading as a trustworthy entity in electronic communication. Malware attacks, on the other hand, involve the installation of malicious software on a user's device, which can steal sensitive data or take control of the device.

Suspicious Link Detection's purpose is to provide a reliable and efficient solution to detect and prevent these types of attacks. The web application utilizes advanced machine learning models, such as logistic regression and random forest, to analyze URLs and provide accurate assessments of their authenticity. By analyzing various features of the URL, such as its domain name, path, and metadata, the machine learning models can determine whether the URL is suspicious or not.

IV. LITERATURE SURVEY

In order to have more information about Malicious Link Detection Systems which are already used to detect suspicious domains, IP's and URL, we referred Research papers based on Malicious Link Detection System using Machine Learning. It gave us information about different techniques used to detect malwares and other breaches with their advantages and disadvantages.

4.1 Overview of Earlier Research Work Done

[1] Mr. Mohammed Alsaedi student of CSE Engineering dept, Mr. Fuad A. Ghaleb a faculty of University Technology Malaysia , Mr. Faisal Saeed student from Birmingham City University and Mr. Jawad Ahmad from Edinburgh Napier University, named "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning" had put forth their focus and published their International article in (Sensors 2022, 22, 3373. https://doi.org/10.3390/s22093373). This paper describes that a malicious website detection model was designed and developed with a hypothesis stating that cyber threat intelligence is an effective and safer alternative to improve the detection accuracy of malicious websites.

[2] Mr. Shantanu & Mr. Janet B. from Department of Computer Application National Institute of Technology Tiruchirappalli, India. and Mr. Joshua Arul Kumar, Department of ECE MAM College of Engineering Tiruchirappalli, India had studied on the concept of detection of malicious URLs as a binary classification problem and evaluated the performance of several well-known machine learning classifiers entitled "Malicious URL Detection" which was published in (International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/ ©2021 IEEE).

[3] Mr. Zhiqiang Wang, Mr. Xiaorui Ren, Shuhao Li, Bingyan Wang, JianyiZhang and Tao Yang from Beijing Electronic Science and Technology Institute researched on malicious URL detection model based on deep learning such that the system model uses word embedding method based on character embedding by combining it, titled "A Malicious URL Detection Model Based on Convolutional Neural Network" published in research paper (Hindawi Security and Communication Networks Volume 2021, Article ID 5518528, https://doi.org/10.1155/2021/5518528).

[4] Mr. Jino S Ganesh, Mr. Niranjan Swarup.V, Mr. Madhan Kumar.R, Mr. Harinisree.Astudents of P.G under the guidance of Prof. Dr. Giri Raj.M of Mechanical Engineering, Vellore Institute of Technology, Tamil Nadu, India, had worked and made a system by using four different machine learning algorithms, namely logistic regression,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

decision tree, random forest, multilayer perceptron neural networks to detect malwares and phishing sites entitled "Machine Learning based Malicious Website Detection" published in (International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020).

[5] Mr. Doyen Sahoo, Mr. Chenghao Liu, Mr and Mr. Steven C.H. Hoi from School of Information Systems, Singapore Management University described that Malicious URL detection plays a critical role for many cybersecurity applications by categorizing them into Blacklist or Heuristic Approach and also used ML approach to classify different spams and malwares named "Malicious URL Detection using Machine Learning: A Survey", published in International article (Vol. 1 August 2019, https://doi.org/10.1145/nnnnnnn.nnnnnnn).

4.2 Summary

So basically, after studying the research and review papers of various authors we found that various authors have created a URL/website which is system eco-friendly to Analyse suspicious domains, IPs and URLs to detect malware and other breaches. There are many online websites available for detection of spam and phishing URLs which can be done by entering the link in their system. Therefore, our system will scan and analyse the URL based on the ML approach and update the result, whether the URL is malicious or not on a single click either it may be from social site or email.

V. METHODOLOGY

The Suspicious Link Detection project is a web application developed using Python programming language and Flask web framework, that aims to detect suspicious links using machine learning algorithms and API integration. The project follows a methodology that includes several stages such as data collection, data pre-processing, feature extraction, model training, model evaluation, and integration and deployment. The web application uses various machine learning algorithms to detect suspicious links, and the process involves collecting data, processing it to extract relevant features, and then training a model to recognize patterns in the data that indicate suspicious links. The model is then evaluated and integrated into the web application, which can be deployed to provide a real-time detection service.

Overall, the Suspicious Link Detection project is an innovative approach to improving online security by leveraging the power of machine learning to detect and protect against potentially harmful links.

In the first phase of the Suspicious Link Detection project, the team implemented the logistic regression algorithm to convert URLs into feature vectors. Logistic regression is a popular supervised learning algorithm that's often used for binary classification problems, like determining whether a given URL is safe or not.To train the model, the team used a dataset containing both safe and malicious URLs. They extracted features from the URLs and used the CountVectorizer method from the scikit-learn library to vectorize the URLs. The model was then trained on this data to predict whether a given URL is safe or not based on its features.

In the second phase of the Suspicious Link Detection project, the Random Forest algorithm was used to extract features of URLs using various functions. The Random Forest algorithm is an ensemble learning algorithm that is widely used for classification and regression problems. Similar to the logistic regression algorithm used in the first phase, the Random Forest model was trained using a dataset containing safe and malicious URLs.

However, in this phase, the features of the URLs were extracted using various functions. These functions included extracting the length of the URL, presence of keywords in the URL, and domain information. These features were then sent to the Random Forest model, which predicted whether the given URL is safe or not.

After training the model, it was saved in h5 format for future use. The combined use of both the logistic regression and Random Forest algorithms in the Suspicious Link Detection project resulted in a more accurate and reliable system for detecting suspicious links.

Model evaluation was done using metrics such as accuracy, precision, recall, and F1-score. Cross-validation was performed to ensure that the models generalized well to new data.

API integration was done by using the VirusTotal API to verify the URL's safety and the GNews API to display the latest news regarding malware and phishing. The system displayed the output of both machine learning models to the user to provide a more accurate assessment of the URL's safety.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

Finally, the trained models were integrated into a web application called Suspicious Link Detection, which provides a user-friendly interface for users to check the safety of URLs. The application also contained a prevention webpage that provided users with tips and tricks for safe browsing and general cybersecurity measures.

Overall, the methodology used in the Suspicious Link Detection project involved the use of multiple machine learning algorithms, APIs, and datasets to create a comprehensive system for detecting suspicious URLs and providing users with cybersecurity guidance.

VI. SYSTEM ARCHITECTURE

This is the DFD that how the system will proceed and simulate its process for detection of malicious link.

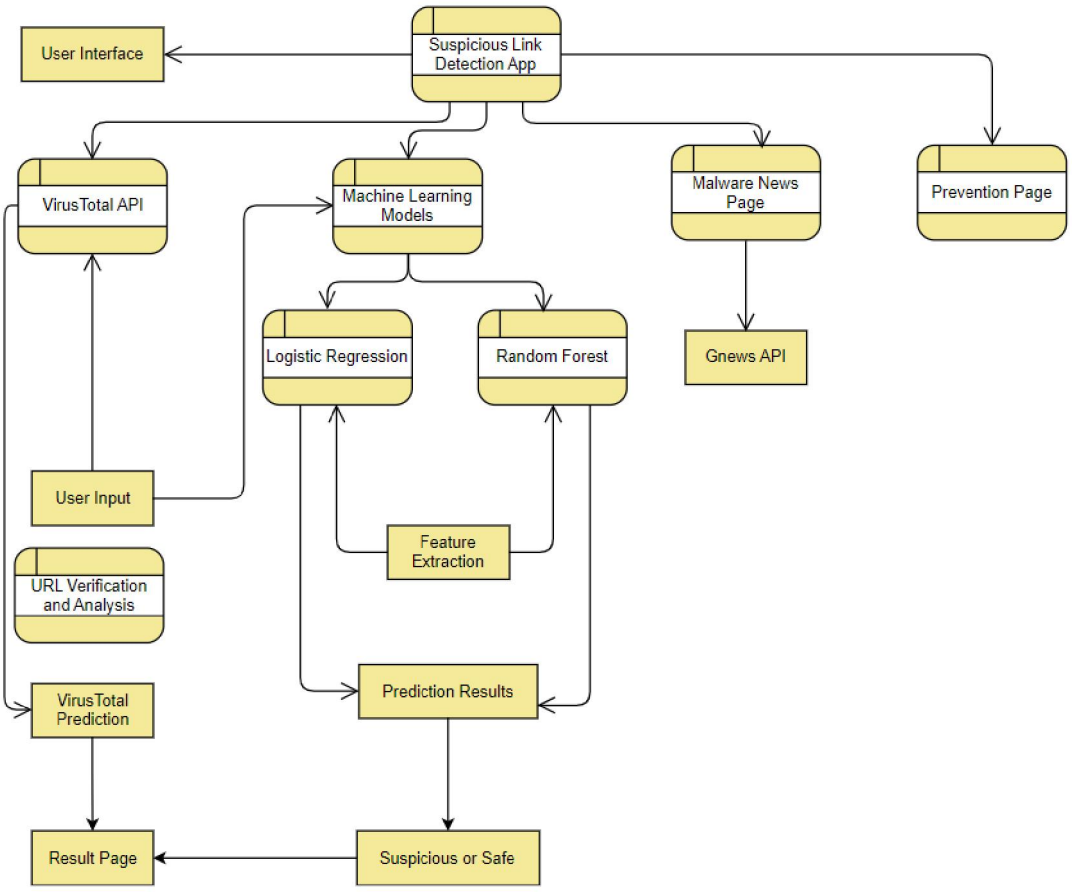

Fig1. Level 0 DFD

The high-level diagram shows the flow of data between different components of the system. The user interacts with the Suspicious Link Detection web application through a user interface. The application then uses two primary components: the VirusTotal API and the machine learning models to detect whether the URL is suspicious or not.

In the Level 1 DFD, more details are added to the URL verification process. The URL entered by the user is verified against various checks, including VirusTotal API and Machine Learning models (Logistic Regression and Random Forest). The result of this verification process is then assessed to determine if the URL is suspicious or safe. If the URL is found to be suspicious, the user is redirected to a Prevention Web Page where they can learn more about the potential risks associated with the URL and take necessary actions to protect themselves.

Overall, the Suspicious Link Detection project uses machine learning algorithms and API integration to detect suspicious links and provide users with up-to-date information about malware and phishing attacks. The DFD diagrams provide a detailed view of how data flows through the system, and the various components interact with each other to ensure the safety of the user.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

VII. RESULTS AND OUTPUTS

Following are the screenshots of our results obtained from our system.

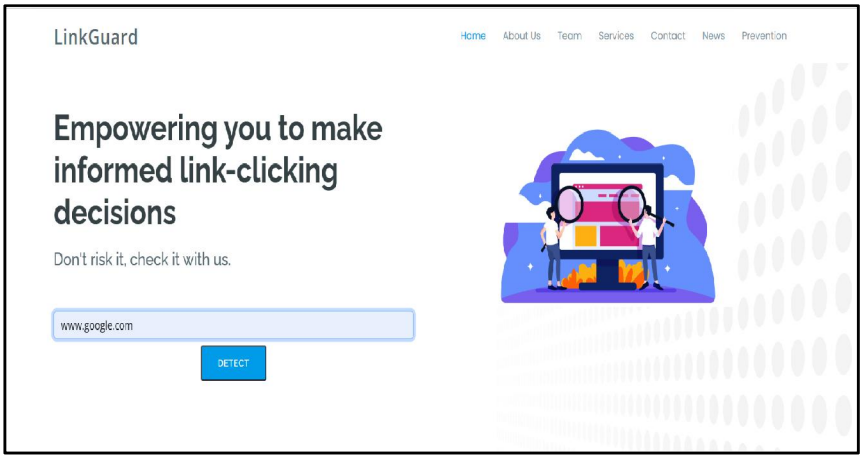

Fig 2. Front Page of Our website

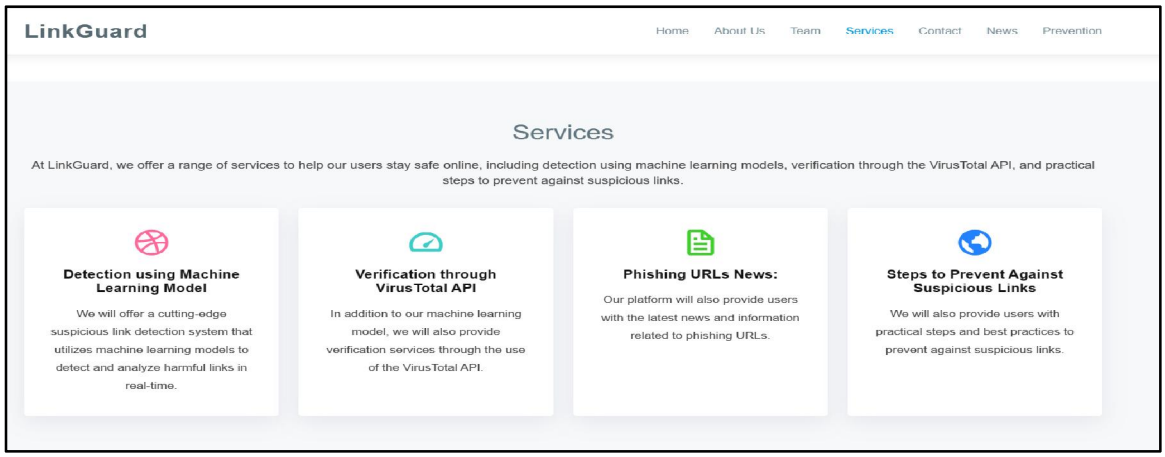

Fig 3. Front Page of Our website

Fig 2,3. Shows the front page of our website, you will find an input textbox where you can enter the URL that you want to check for potential threats. Our powerful detection system will then analyse the URL to determine whether it is safe or malicious.

In addition to the input textbox, you will also see various navigation tabs at the top of the page. By clicking on these tabs, you can explore the different sections of our website and learn more about our system.

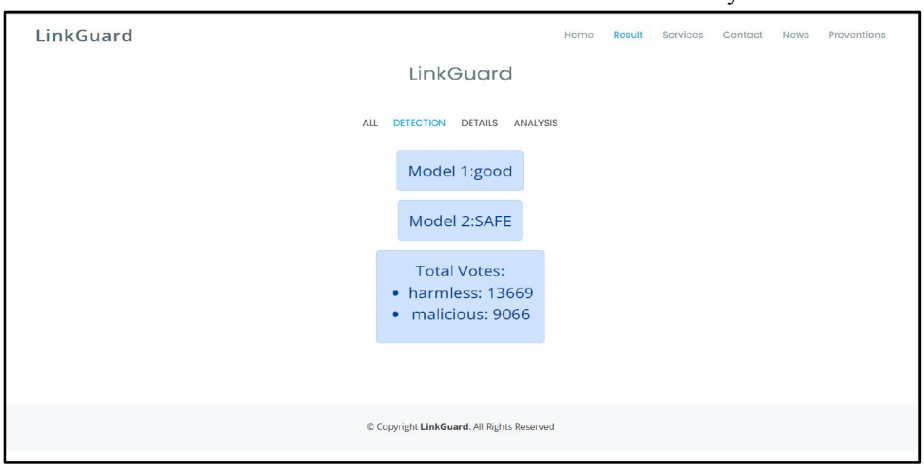

Fig 4. Result Page

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

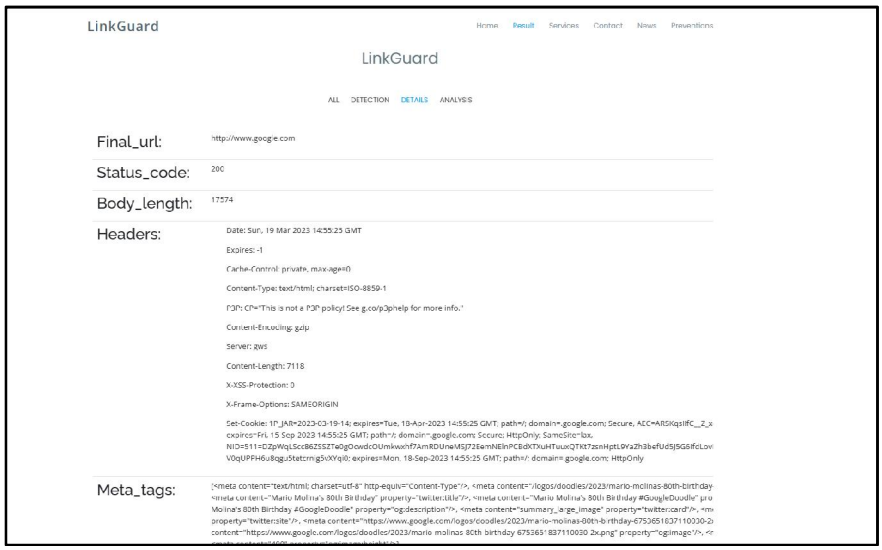

Fig 5. Result Page

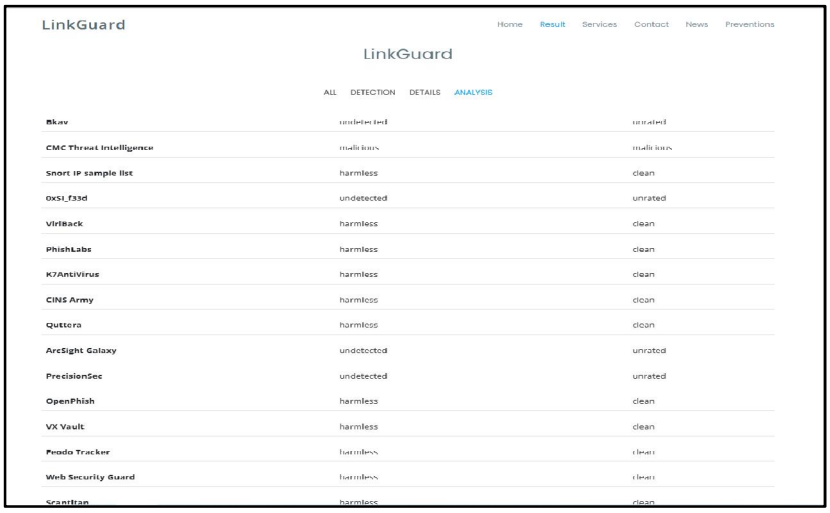

Fig 6. Result Page

Fig 4,5,6 displays the predicted results of the Logistic Regression model, Random Forest model, and the VirusTotal API result.

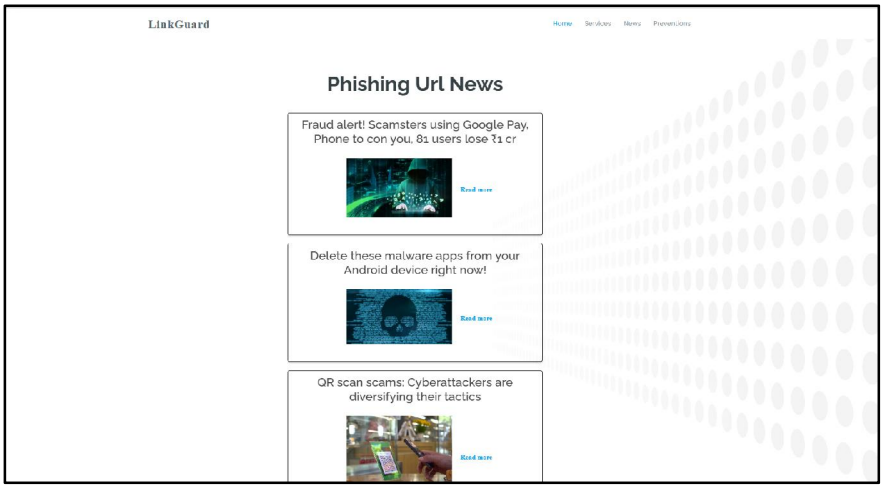

Fig 7. News Page

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

Figure 7 displays the output of the Malware/Phishing News API integrated into the Suspicious Link Detection system. The news articles related to malware and phishing attacks are fetched by the GNews API and displayed to the user. This feature provides users with up-to-date information about current threats and helps them stay informed about potential risks.

Fig 8. Prevention Page

Fig 8 illustrates the Prevention Page, which is displayed to users. The Prevention Page provides detailed information about the potential risks associated with the URL and offers recommendations for users on how to protect themselves.

VIII. FUTURE SCOPE

The Suspicious Link Detection project has significant potential for further research and development. One area of research could be to improve the accuracy of the machine learning models used in the project. This could be achieved by collecting more data and refining the feature extraction process. Additionally, incorporating other machine learning techniques, such as deep learning, could also improve the accuracy of the models.

Another potential area of research is to extend the project to detect other types of cyber threats, such as malware and phishing attacks. This could involve developing additional machine learning models and integrating new APIs to provide a more comprehensive solution for detecting potential cyber threats.

Furthermore, the Suspicious Link Detection project can be extended to integrate with other security solutions, such as antivirus software and firewalls. This could provide users with an even more comprehensive security solution, which can protect against various types of cyber threats.

Additionally, the system could be integrated with other cybersecurity tools and services, such as firewalls or intrusion detection systems, to provide a more comprehensive security solution.

Another potential future scope could be to improve the accuracy of the machine learning models by incorporating more advanced techniques, such as deep learning, and by continuously updating the models with new data to adapt to evolving threats.

Furthermore, the system could be adapted to work with different languages and character sets, as many phishing attacks and malware threats are becoming more sophisticated and using non-English characters or languages.

IX. CONCLUSION

As technology and internet usage continues to grow, users share sensitive information online without always being aware of the risks associated with it. Malicious URLs pose a significant threat to cybersecurity and detecting them is crucial in protecting users. Machine learning techniques offer promising solutions in identifying and classifying

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

malicious URLs. As a result, it is crucial to have effective methods for detecting and protecting against these threats, with machine learning techniques showing promise in this area.

The Suspicious Link Detection project presents a comprehensive solution for detecting potentially malicious URLs and protecting users from cyber threats. The system uses machine learning techniques to classify URLs as either safe or suspicious and provides users with up-to-date information about potential threats. With the growing number of people conducting their personal and professional lives online, there is an increasing need for cybersecurity solutions like Suspicious Link Detection.

The evaluation of the system demonstrated its high degree of accuracy in classifying URLs, indicating its potential as a reliable tool for detecting malicious URLs. The integration of APIs like VirusTotal, Malware/Phishing News, and GNews also enhances the system's ability to provide users with real-time updates about emerging threats.

Furthermore, the project has significant potential for further research and development. Future work could focus on improving the accuracy of machine learning models, extending the system to detect other types of cyber threats, and integrating it with other cybersecurity tools and services. With these potential improvements, the Suspicious Link Detection project can continue to play a critical role in protecting individuals and organizations from potential cyber threats.

X. ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the individuals and organizations who have supported and contributed to the successful completion of this project "Suspicious Link Dtection". We would like to articulate our profound gratitude and indebtedness to those people who helped us in completion of our project.

First and foremost, our deepest thanks to Prof. Amar Palwankar Guide of our project for guiding and correcting us with attention and care and for her constant motivation and valuable suggestions. She has taken pain to go through the project and make necessary corrections as and when needed. We also want to express our sincere gratitude to Prof. Vinayak Bharadi, Head of Department, Information Technology for putting constant efforts and being enthusiastic for successful completion of the project. We are very grateful to Prof. Priyanka Bandagale, Project Coordinator for her support and motivation. We truly appreciate all our faculty members for providing a solid background for helping a lot to properly shape the problem and providing insights to the solution. Last but not the least, we thank all those friends who helped us in the course of this entire week. We extend our heartfelt thanks to our family and well-wishers whose blessings made us reach this stage.

REFERENCES

[1]. Mohammed Alsaedi, Fuad A. Ghaleb, Faisal Saeed, Jawad Ahmad_(2022) Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning. International article in (Sensors 2022, 22, 3373.https://doi.org/10.3390/s22093373).

[2]. Shantanu, Janet B, Joshua Arul Kumar R_(2021) Malicious URL Detection. (International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281- 9537-7/20/ ©2021 IEEE).

[3]. Zhiqiang Wang, Xiaorui Ren, Shuhao Li, Bingyan Wang, Jianyi Zhang, Tao Yang_(2021) A Malicious URL Detection Model Based on Convolutional Neural Network. (Hindawi Security and Communication Networks Volume 2021, Article ID 5518528, https://doi.org/10.1155/2021/5518528).

[4]. Jino S Ganesh, Niranjan Swarup.V, Madhan Kumar.R, Harinisree.A and Dr. Giri Raj.M_(2020) Machine Learning based Malicious Website Detection. (International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020).

[5]. Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi_(2019) Malicious URL Detection using Machine Learning: A Survey International article (Vol. 1 August 2019, https://doi.org/10.1145/nnnnnnn.nnnnnnn).

[6]. Ayon Gupta, Sanghamitra Giri, R. Naresh_(2020) Malicious URL Detection System using combined SVM and Logistic Regression Model. (International Journal of Advanced Research in Engineering and Technology, JARET Volume 11, Issue 4, April 2020).

[7]. Cho Do Xuan, Hoa Dinh Nguyen_(2020) Malicious URL Detection based on Machine Learning. (IJACSA International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Impact Factor: 7.301
Volume 3, Issue 3, April 2023

[8]. Mr. Ferhat Ozgur Catak professor of University of Stavenger, Ms. KevserSahinbas from Istanbul Medipol University and Mr. Volka Dortkardes from Turkey_(2020) (USA by IGI Global Engineering Science Reference).