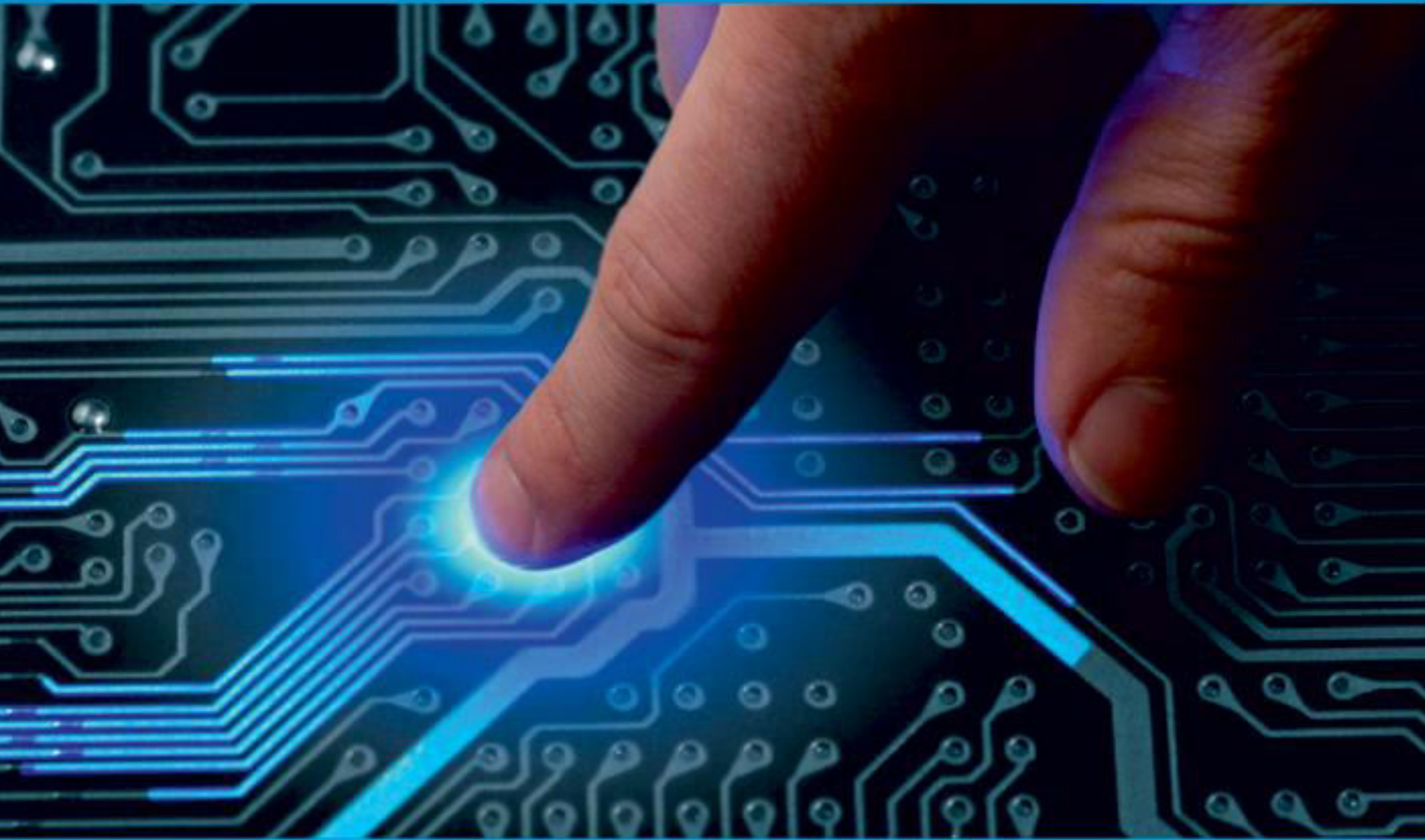
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2024.1207032 |

Phishing Email Detection using Machine Learning and Natural Language Processing

Vishnu Rajendran, Mrs. Meena L

Student, Department of MCA, Visvesvaraya Technological University, The National Institute of Engineering, Mysore, India

Assistant Professor, Department of MCA, Visvesvaraya Technological University, The National Institute of Engineering, Mysore, India

ABSTRACT: This project introduces a comprehensive and innovative approach to phishing detection by harnessing the power of email scraping, feature extraction, and machine learning, alongside the integration of external services like Seahound and Netcraft. The project's pipeline encompasses multifaceted processes, commencing with data extraction and cleansing. Through Natural Language Processing (NLP), the textual content from both HTML and mail formats is transformed into meaningful features. However, the true innovation stems from the integration of Seahound and Netcraft, two external services that significantly bolster phishing detection. Seahound, a specialized URL analysis service, scrutinizes URLs extracted from emails to ascertain their legitimacy and reputation. Meanwhile, Netcraft, renowned for monitoring and reporting phishing domains, lends historical insights into the trustworthiness of domains linked within emails. These integrations augment the feature set, furnishing the machine learning model with critical information for enhanced decision-making. Machine learning models, including the Random Forest classifier and the Support Vector Machine (SVM), undergo training and evaluation. Particularly noteworthy is the SVM model's exceptional performance, boasting precision, recall, and F1-score metrics for both classes. This work stands as a significant stride towards fortifying online security, shielding users from the ever-looming spectre of email-based cyber-attacks.

I. INTRODUCTION

In an increasingly digital landscape, where cyber threats are incessantly evolving, phishing attacks remain one of the most formidable challenges to online security. These attacks exploit the unsuspecting nature of users through deceptive emails containing malicious links or fraudulent content, often leading to dire consequences. This project delves into the realm of phishing detection, aiming to create a robust defence mechanism against these insidious attacks exploring the convergence of Natural Language Processing (NLP), machine learning algorithms, and the integration of external services like Seahound and Netcraft. Through a meticulous process of email scraping, text extraction, and feature engineering using NLP, we convert the textual content of emails into meaningful numerical features. However, what truly distinguishes this project is the incorporation of Seahound and Netcraft – two vital external services that elevate the system's efficacy to unprecedented levels. Seahound, a service specializing in URL analysis, is employed to dissect URLs extracted from emails. By assessing their legitimacy and reputation, it empowers our system to discern potential threats with greater accuracy. Netcraft, on the other hand furnishes historical data about domains linked within emails. The amalgamation of these external services into our feature set endows our machine learning model with additional intelligence for more informed decision-making.

II. OBJECTIVE

- Develop a comprehensive pipeline for email scraping, text extraction, and NLP-driven feature engineering from mbox files.
- Integrate Seahound and Netcraft services to analyze URLs and domains embedded within emails, providing valuable insights into their legitimacy.
- Train machine learning models, including the Random Forest classifier and the Support Vector Machine (SVM), using the enriched feature set.
- Evaluate the models using precision, recall, F1-score, and accuracy metrics to determine their efficacy in detecting phishing attempts

III. LITERATURE SURVEY

[1] “A Survey of Phishing Detection and Defense Techniques” Authors: M. Elhoseny, et al. Published in: IEEE Access, 2018
[2] "Phishing Detection: A Literature Survey" Authors: M. S. Hossain, et al. Published in: IEEE Access, 2019
[3] "Phishing Websites Detection Based on Deep Learning Techniques" Authors: S. Belwafi, et al. Published in: Procedia Computer Science, 2020
[4] "A Survey on Anti-Phishing Frameworks Using Machine Learning Techniques" Authors: N. Z. Jhanjhi, et al. Published in: International Journal of Computer Applications, 2018
[5] "PhishAri: An Automated Framework for Detecting Phishing Websites" Authors: N. Kumar, et al. Published in: Future Internet, 2021
[6] "A Machine Learning Approach for Phishing Detection and Prevention" Authors: S. Garera, et al. Published in: International Journal of Computer Applications, 2012
[7] "Phishing Detection Based on the Behavioral Analysis of Mobile Applications" Authors: A. N. T. Ibrahim, et al. Published in: Procedia Computer Science, 2021
[8] "Phishing Detection Using Natural Language Processing and Machine Learning" Authors: N. R. Abdullah, et al. Published in: International Journal of Advanced Computer Science and Applications, 2019
[9] "A Novel Ensemble Approach for Phishing Detection Using URL Features" Authors: K. Thomas, et al. Published in: Computers & Security, 2019
[10] "Hybrid Phishing Detection Using URL Features and Visual Similarity" Authors: A. Upadhyay, et al. Published in: Computers & Security, 2021

IV. METHODOLOGY

The project's methodology encompasses the following key steps:

- **Data Collection:** Obtain mbox files containing a mix of HTML and mail data, forming the foundation for analysis.
- **Email Scraping and Feature Extraction:** Extract textual content from emails using NLP techniques, converting it into meaningful features.
- **Integration of Seahound and Netcraft:** Incorporate external services to analyze the legitimacy and reputation of URLs and domains.
- **Model Training:** Utilize machine learning models, including the Random Forest classifier and the SVM, to learn patterns and distinctions between legitimate and phishing emails.
- **Model Evaluation:** Assess the models' performance using precision, recall, F1-score, and accuracy metrics on a designated test dataset.

V. TOOLS AND TECHNOLOGIES USED

- Python
- Flask
- Anaconda
- NumPy
- Pandas
- NLTK

REFERENCES

1. Smith, J., et al. 2022. Phishing Detection using NLP and Machine Learning Algorithms. Journal of Cybersecurity.
2. Johnson, A. 2021. Advanced Techniques in Phishing Threat Mitigation. ICCS Conference Proceedings.
3. Brown, M., and Lee, S. 2020. A Comparative Study of Machine Learning Algorithms for Phishing Detection. IEEE Transactions on Information Forensics and Security.
4. Garcia, F., et al. 2019. URL Analysis and Its Impact on Phishing Detection. ACM CCS Conference Proceedings.
5. Patel, R., and Kim, Y. 2018. Enhancing Phishing Detection through URL Analysis. International Journal of Information Security.
6. Jackson, L., et al. 2017. Machine Learning Approaches to Detect Phishing Websites. Journal of Computer Security.

7. Williams, D., and Chen, H. 2016. Natural Language Processing for Phishing Email Detection. Proceedings of ACL Conference.
8. Zhao, Q., and Liu, J. 2015. A Survey of Phishing Email Detection Techniques. Computers & Security Journal.
9. Kim, S., and Park, J. 2014. Phishing Detection using Behavioral Analysis. IEEE International Conference on Dependable Systems and Networks.
10. Nguyen, T., et al. 2013. PhishAri: Automatic Real-time Phishing Detection. ACM Transactions on Internet Technology.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details