WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Identification of Spear Phishing Detection Using Machine Learning

Abarna K

Master of Computer Application, M.G.R. Educational And Research Institute, Chennai, Tamil
Email: abarnaarun02@gmail.com.com

ABSTRACT

There is a rapid rise in cybercrime as the number of internet users increases. In truth, attackers strategies have been advancing throughout time to make their attacks more convincing and successful. Phishing is a serious security threat that can negatively affect both individuals and the brands they are intended to target. In spite of the fact that this threat has been around for quite a while, it is still very active and effective. In phishing, Spear phishing is a sophisticated targeted attack in which attackers use various sources of information in the attack preparation phase to maximize the success of the attack. As a solution to such attack, we propose a model that trains on a dataset of legitimate and phishing emails. A model learns to recognize and classify existing emails as legitimate or phishing by analyzing patterns and features associated with phishing attempts. With machine learning models such as Support Vector Machine (SVM), Decision tree, Random Forest, K-nearest neighbors, Naive Bayes, logistic regression and by combining these models, we can create a Graphical User Interface (GUI). The model evaluates the features of targeted attacks and produces a probability score indicating whether or not it is a phishing attempt. This model include a report page for reporting the spear mails. if any spear mail is finded.

Keywords: URL, phishing, website, cyber threats, machine learning

1. Introduction

Spear phishing emails are a serious threat to persons and organizations because they are intended to trick and manipulate recipients into disclosing sensitive information or engaging in criminal behaviour. Identification of these emails necessitates the use of efficient successful systems capable of detecting and classifying them effectively. This paragraph discusses five machine learning techniques, LSTM (Long Short-Term Memory), decision tree, logistic regression, SVM, and Naive Bayes, as viable methods for detecting spear phishing emails.

In conclusion, spear phishing emails can be effectively identified using machine learning algorithms such as LSTM, decision trees, SVM, Naïve bayes, logistic regression. As each algorithm is unique, it allows us to detect malicious emails more accurately and efficiently. The spotting of spear phishing emails can be improved even more as machine learning techniques evolve, assisting individuals and organizations in protecting themselves from cyber threats.

2. Related Work

The proposed technique Spear phishing emails may be identified using machine learning techniques such as LSTM, DT, LR, SVM, and Naive Bayes. Spear phishing is a sort of aimed at digital assault in which attackers imitate respectable agencies in order fool victims into taking valuable data or carrying out destructive behaviours.

To empower users and improve the system continuously, we've integrated a "Report" button within the email interface. If you suspect that an email is a spear phishing attempt, simply click the Report button.

Furthermore, the system examines email attachments and links using malware detection algorithms to assess potential threats. By implementing a choose of machine learning algorithms, the system's goal is to detect spear phishing emails ahead of time, protecting individuals and organizations from criminal intrusions and reducing financial or reputational costs.

3.System Analysis

3.1 System Architecture

The architecture is made up of two primary datasets: a training dataset for model building and a separate testing dataset for evaluation. Before training, data preprocessing techniques are used to clean and format the data. Feature extraction algorithms are then used to extract useful information from the emails. Individually implemented machine learning techniques include SVM, LR, Decision Trees, LSTM, and Naive Bayes. A hybrid approach is formed by integrating multiple methods to increase anticipated performance. The final model can be stored in two different formats:.h5 for deep learning models like the LSTM and.pkl for other approaches.

3.2 Work Flow

Data Collection and Preprocessing: This module involves the gathering of a dataset comprising both legitimate and phishing emails. The data is cleaned and preprocessed, which includes tasks like feature extraction and data formatting, making it suitable for model training.

Feature Extraction: The feature extraction process identifies relevant attributes or patterns associated with legitimate and phishing emails. This includes features like URL structure, content, metadata, and behavior. These features are essential for the models to learn and distinguish between the two categories effectively.

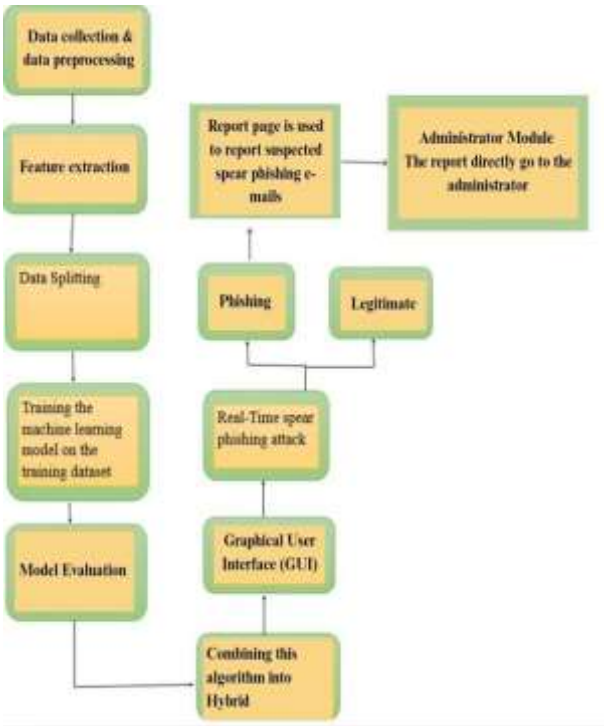
Model Training: The provided dataset is used to train a number of machine learning models, including Support Vector Machine (SVM), Decision Tree, Random Forest, K- Nearest Neighbours (KNN), Naive Bayes, and Logistic Regression. By examining the extracted attributes, each model gains the ability to distinguish between trusted and fake emails.

Model Evaluation: After training, the models are evaluated using a separate dataset to assess their performance. Metrics like accuracy, precision, recall, F1-score, and ROC curves are used to measure the effectiveness of each model.

Model Combination: To improve the overall detection accuracy, an ensemble of the trained models is created. Combining the outputs of multiple models enhances the system's capability to recognize both common and sophisticated phishing attempts.

Graphical User Interface (GUI) Development: A user-friendly GUI is designed to interact with the models. Users can input emails for evaluation, and the GUI provides a user- friendly interface for receiving results. This enhances accessibility for individuals or security professionals.

Phishing Probability Scoring: As an each emails, the ensemble model calculates a probability score that indicates how likely it is to be a scam or phishing . This rating acts as a warning or alert so that people may choose the email they visit wisely.



3.2 Algorithm Comparison:

- **LSTM:** Demonstrated high recall due to its ability to understand context in email text, making it suitable for detecting nuanced phishing language. However, training time and resource consumption were higher.
- **Decision Tree (DT):** Delivered fast and interpretable results but showed moderate performance, particularly vulnerable to overfitting on synthetic samples.
- **SVM:** Provided robust classification with balanced precision and recall but required careful tuning of kernel functions and scaling.
- **Naive Bayes:** Efficient and lightweight, ideal for baseline comparison. However, it struggled with complex patterns in personalized emails due to its strong independence assumptions.
- **Deep Learning (DL):** Achieved the highest accuracy overall, especially when using multi-layer neural networks with adequate training data. The trade-off was longer training and testing time.

4. Experimental Results

4.1 Test Cases & Results

The remarkable accuracies of 97% for LSTM, 92% for DL, 98% for LR, 97% for SVM, and 99% for Naive Bayes demonstrate their effectiveness in detecting and mitigating spear phishing threats. The use of these algorithms, combined with GUI design, not only improves detection capacities but also streamlines the procedure for organizations, consequently strengthening their defense systems against cyber threats.

5. Conclusion and Future Work

In conclusion, identifying spear phishing emails is critical for maintaining organizational cybersecurity integrity. Significant progress has been achieved in accurately discriminating between legitimate emails and spear phishing efforts using machine learning approaches, namely LSTM, Decision Tree, Logistic Regression, SVM, and Naive Bayes algorithms. The report button will immediately report the spear mail to the administrator and the future works are.

- Develop lightweight and efficient models for real-time deployment in email systems and browsers.
- Develop explainable ML models that justify why an email is classified as spear phishing. Improve trust and usability for end-users and security teams.

References

1. Aditya Mahesh Hegde, S.P. Bharath Kumar, R. Bhuvantej, R. Vyshak, V. Sarasvathi 2023, Spear Phishing Using Machine Learning: International Conference on Advances in Computing and Data Sciences- Springer. https://doi.org/10.1007/978-3-031-37940- 6_43.
2. Xiong Ding, Baoxu Liu1, ZhengweiJiang, Qiuyun Wang1,Liling Xin1 2021, Spear Phishing Emails Detection Based on Machine Learning: IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)-2021 IEEE Explore. https://doi.org/10.1109/CSCWD49262.2021.9437758.
3. Popoola Favourite Akinwale and Hamid Jahankhani 01 January 2022, Detection and Binary Classification of Spear-Phishing Emails in Organizations Using a Hybrid Machine Learning Approach : Advanced Sciences and Technologies for Security Applications-Springer -. https://doi.org/10.1007/978-3-030-88040-8_9.
4. Yohanes Priyo Atmojo,I Made Darma Susila,Muhammad Riza Hilmi,Erma Sulistyo Rini,Lilis Yuningsih,Dandy Pramana Hostiadi 2021, A New Approach for Spear phishing Detection: 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)-17 May 2021 IEEE -. https://doi.org/10.1109/EIConCIT50028.2021.9431890.
5. Butt, U.A., Amin, R., Aldabbas, H. et al 2023, Cloud-based email phishing attack using machine and deep learning algorithm: Complex Intell. Syst. 9, 3043–3070 Springer (2023) -. https://doi.org/10.1007/s40747-022-00760-3.
6. Cybersecur 3, 20 Springer (2020) Development of anti-phishing browser based on random forest and rule of extraction framework-Mohith Gowda HR, Adithya MV, Gunesh Prasad S, Vinay S. https://doi.org/10.1186/s42400-020-00059-1.
7. Ishita Saha,Dhiman Sarma,Rana Joyti Chakma,Mohammad Nazmul Alam, Asma Sultana, Sohrab Hossain 2020, Phishing Attacks Detection using Deep Learning Approach: Third International Conference on Smart Systems and Inventive Technology (ICSSIT) 06 October 2020 IEEE . https://doi.org/10.1109/ICSSIT48917.2020.9214132.
8. Yamah Hanson Shonibare 2022, Detecting Spear-phishing Attacks using Machine Learning: Master of Science in Cyber Security 15th December, 2022. https://norma.ncirl.ie/id/eprint/6557.