Research Article

Detecting Phishing Attacks using NLP

Mr.T.H.Feiroz Khan[a], and Mr. Anwar Basha H[b]

[A] Department of Computer Science & Engineering, SRM Institute of Science & Technology, Ramapuram, Chennai, India.
[b]Department of Computer Science & Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: In present world phishing attacks are the most common and easily targeted attacks. In order to analyse texts and detect improper statements that show the phishing attacks, we have come with an idea which will use (NLP) Natural language processing techniques. Compared to previous work, our approach is different because the emphasis is on the text data found in the attack, which is semanticized in order to identify malicious intent. We have evaluated it using a huge phishing e-mail test data to illustrate the efficacy of our strategy.

Keywords: Phishing attacks, Machine learning, Natural Language Processing

1. Introduction

The detection of finding the faults in the fabrics and the classification of the fabrics are being done by the Computer Vision. With the help of Computer Vision we are able to detect the defects in the fabrics such as oil stains identification costs time, cash and buyer fulfillment. In this way, early and exact texture imperfection location is a significant period of value control. Computerized texture deformity review framework has been pulling in broad consideration of the specialists of numerous nations for quite a long time. Fraction includes two testing issues, to be specific imperfection recognition and imperfection arrangement.

2. Literature Survey

Many methods look at the URLs in the message [2]. Three methods for detecting phishing websites were explored in this paper suggested by Vaibhav Patil. The third solution is based on a visual appearances review to validate the validity of the site[2]. The downside to this technique is the identification of certain small false positives and false negative tests. In comparison, a third method includes evaluating the various features of URL.

G. Jaspher Willsie Kathrine evaluated several different detection approaches such as Heuristic based, and the technique used was decision tree algorithm and by his he helped in choosing the best approach to detect with least failure rate [1]. It consumes a lot of time as we need to check for each and every algorithm then come to final result.

Previous work of Ebubekir BUBER [4] has also employed syntactic parsing to infer malicious intent. It is seen as an algorithm which is better compared to previous ones and gives success rate of 89.9%. This Success rate can be improved further with a more efficient algorithm.

Muhammet Baykara [5] developed an Anti-phishing simulator is developed to detect the phishing attacks. It works on a simple logic on detecting attacks by comparing the mails and checking the URLs. [5] Major drawback of this system was that it works only to detect the URL induced mail type phishing attacks.

Prasanta Kumar Sahoo [3] used Data mining algorithms to detect the fake E-mail using Naïve Bayesian classification. It was efficient because of its methodology in terms of complexity and overhead to detect phishing attacks. The drawback of this system was its incompatibility to work on different system and error rate.

In order to produce the best performance depending on the Sharma A. K criterion [6], a comparative study was performed of various spam filtering policies focused on specific roles in order to optimize the efficiency of the spam detection algorithms or various current data mining algorithms in its job.

Li et al [7] suggested an online learning approach to website identification. This article discusses website functionalities like site graphics and model database objects to refine the characteristics obtained from the site used as an evolutionary algorithm dependent on quantities. To define the website as valid or phishy, the configured functions are transferred through transductive help vector machine.

3. Proposed System

Our method carries out a semantical analysis of the attacker's text to validate the adequacy of each paragraph. Our method decides whether the sentence is a question or a order depending on the functions played by each word in the sentence. The potential topics of questions and commands will be collected by pairs. Every pair is then checked to see if it is included in a registry of malicious pairs. The program reads a text file one sentence at a time and returns true if a social manipulation attack is present in the record.

Identification of malicious questions and commands depends on the presence of a pair blacklist object (verb-direct object) which implies a malicious purpose to include in the application or order. We use machine learning to create a blacklist of the subject, constructing a decision tree that is designed for multiple distributed results. We also used Multinomial NB (the method implementing this algorithm) as the Scikit-Learn Python Library[17]. This algorithm produces a pre-setting label with any single pair that generates a predictive rating of trust. The size of the confidence levels is 0 and 1, with a score of 1 indicating certainty.

4. Module Description

Data assessment

At Data Assessment, the data is analyzed one by one. Each variable in the data set is analyzed and then it is made as a proper data set for training purposes. This increases the accuracy and reliability of the data.

Pre-processing

At Pre-Processing, an extensive evaluation is performed on standard benchmarks from text categorization and semantic analysis. This involves the steps of data cleaning, transformation and reduction.

Feature selection:

At Feature selection, Decision Tree algorithm is performed and the emails are being classified into either 0 or 1 category based on the features of the text mail.

Prediction:

At Prediction level, the classified mail is being checked with the listed dataset for the comparison and finally determined if the mail is malicious or not based on the threshold value percentage.

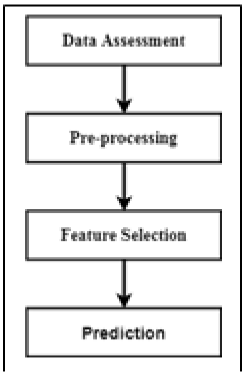

Fig.1 Module diagram for detecting phishing attacks in emails

5. Architecture Diagram

At the first level (Input Raw E-mails), raw e-mails are read by the program and e-mail messages delete text. A semantic analysis is used for extracting word line by line for testing in the second level. In the third level, an SVM classification model will be equipped. In the fourth stage values are classified into 0,1. 1 state true and 0 state false. After getting the values in the fifth stage, a check is done see the text with the blacklisted texts in the database and finally determine if the email is malicious or not based on the threshold percentage value.

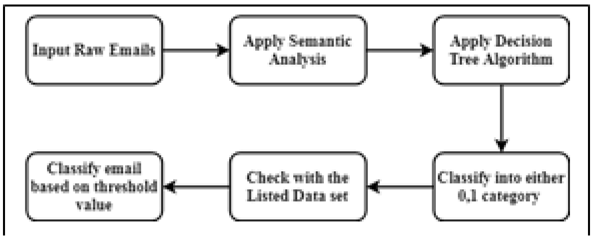
Input Raw Emails
Apply Semantic Analysis
Apply Decision Tree Algorithm
Classify into either 0,1 category
Check with the Listed Data set
Classify email based on threshold value

Fig.2 Architecture diagram for detecting phishing attacks in emails

6. Implementation

We used an email data collection to accurately test our approach to detect the positives and negatives. We have assembled a phishing email package that is publicly accessible. For the legal e-mail URL corpus we used Enron Corpus [8]. Some of the e-mails contained only pictures outside the images without text. We have missed phishing e-mails only with pictures and all 1780 remaining e-mails. We have evaluated the test corpus for comparison to the NLP algorithm, which only shows phish URL connections. In Python scripts our algorithm has been implemented and graphs are generated using R tools.

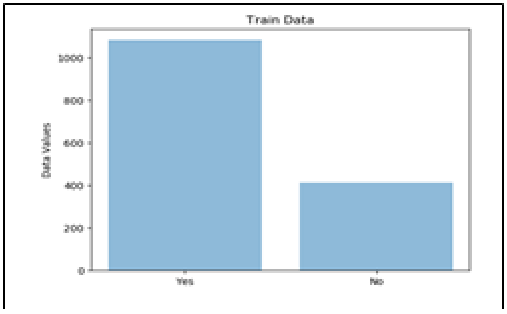

Fig.3 Matplot of training data

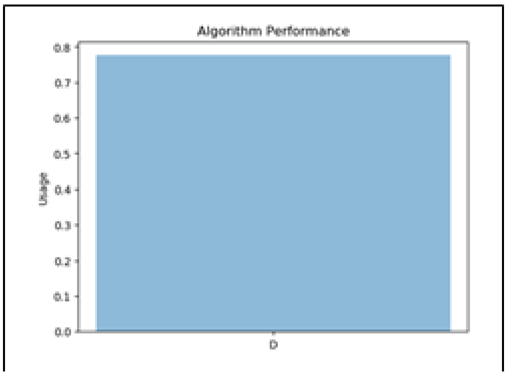

Fig 4.Matplot of algorithm performance

7. Conclusion

The new framework helps internet users to browse safely and securely. This allows users to save valuable information which should not be leaked. It is much easier to remove our system if our proposed system is supplied in form of an extension to users. These findings demonstrate the efficiency that heuristic characteristics, visual characteristics and blacklist and white-list approach can achieve with a hybrid solution. We propose a method to detect phishing attacks on targeted emails. We are not dependent on metadata linked to emails but on text analysis. Our method is also successful in the detection of text-only phishing e-mails. Our findings on phishing e-mails provide substantially stronger warning that semantic knowledge is a good social manipulation predictor.

References

1. G.JaspherWillsieKathrine, A. Amrutha Rose Department, EligiousKalaivani. C Department "Variants of phishing attacks and detection techniques"3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019.
2. VaibhavPatil, TusharBhat, PriteshThakkar, Chirag Shah “Detection and Prevention of Phishing Websites using Machine Learning Approach”2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
3. Prasanta Kumar Sahoo “Data mining a way to solve Phishing Attacks” Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India
4. Ebubekir BUBER, Banu D “Detecting Phishing Attacks from URL by using NLP Techniques”2nd International conference on computer science and Engineering-2017
5. Muhammet Baykara, ZahitZiyaGürel “Detecting of phishing attacks" 2018 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, pp. 1-7, 2018.
6. K. Sharma and R. Yadav, "Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, pp. 1089-1093, 2015.
7. Yuancheng Li, Rui Xiao, JingangFeng, Liujun Zhao, “A semi-supervised learning approach for detection of phishing webpages,” Optik-International Journal for Light and Electron Optics, vol.124, Issue 23, December 2013.
8. KLIMT, B., AND YANG, Y. The Enron Corpus: A New Dataset for Email Classification Research.2004
9. Mahmoud Khonji, Youssef Iraqi, Andrew Jones, “Literature Review on Phishing Detection”, Institute of Electrical and Electronics Engineers Communication Surveys and Tutorials, Volume.15 (Issue 04), 2013.

10. Kiruthika, U., Somasundaram, T.S. & Raja, S.K.S. Lifecycle Model of a Negotiation Agent: A Survey of Automated Negotiation Techniques. Group Decis Negot (2020). https://doi.org/10.1007/s10726-020-09704-z
11. Ms. Pallavi D. Dudhe, Prof. P. L. Ramteke, "A Review on Phishing Detection approaches", International Journal of Computer Science and Mobile Computing, Volume. 4 (Issue 02), page no. 166-170, February 2015.
12. K. Jain and B. B. Gupta, "Comparative analysis of features based machine learning approaches for phishing detection," in Proceedings of the 10th INDIA-COM, New Delhi, India, 2016.
13. Andronicus A. Akinyelu and Aderemi O. Adewumi,"Classification of Phishing Email Using Random Forest Machine Learning Technique", Hindawi Publishing Corporation Journal of Applied Mathematics, 2014.
14. Marchal, S., François, J., State, R., & Engel, T. (2012). Proactive discovery of phishing related domain names. Research in Attacks, Intrusions, and Defenses, 190-209
15. Godwin Caruana and Maozhen Li. A survey of emerging approaches to spam filtering. ACM Computing Surveys (CSUR), 44(2):9, 2012.
16. R Kishore Kumar, G Poonkuzhali, and P Sudhakar. Comparative study on email spam classifier using data mining techniques. In Proceedings of the International Multi Conference of Engineers and Computer Scientists, volume 1, 2012.
17. Sampathkumar, A., Murugan, S., Rastogi, R., Mishra, M.K., Malathy, S. and Manikandan, R., 2020. Energy Efficient ACPI and JEHDO Mechanism for IoT Device Energy Management in Healthcare. In Internet of Things in Smart Technologies for Sustainable Urban Development (pp. 131-140). Springer, Cham.