An Empirical Approach on Detection and Prevention of E-mail Phishing using Machine Learning Techniques

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

INFORMATION TECHNOLOGY

by

SHWETA SANKHWAR

Supervised by

Dr. DHIRENDRA PANDEY
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Co-Supervised by

PROF. R. A. KHAN
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Submitted to

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW

DECEMBER-2018

ABSTRACT

Internet is becoming a significant resource to people in the modern society. To cope-up with the increasing demand for robust and novel web applications, programmers are developing new tricks to add a unique flavour to their web sites. Though these advanced features are based on mature languages and standards, new security problems are often unearthed with each new trick. These new security problems are not only due to technological nuances of new programming techniques, but are also dependent on the way people interact with the web sites. This research work focuses on highlighting one such problem named Phishing, which has become a serious problem for all enterprises utilizing e-communication and online e-commerce tools.

The word phishing is a variation of the word 'fishing'. The concept of 'phishing' can be best understood by observing a typical 'fishing' activity, whereby a fisherman offers a tempting bait to the fishes and waits for them to bite onto it hungrily. In the cyber world, 'phishers' follow the same methodology; only in this case the bait turns out to be a seemingly harmless email or a text message that demands the receiver's personal information in lieu of providing him with some 'too good to be true' offers. Many people fall prey to such phishing attacks by believing that the messages and emails have originated from some legitimate source and naively reply back with all their sensitive information and credentials. In general terms, these people are said to have been 'phished' over the internet.

Phishing websites are cleverly designed replicas of original and trusted websites. Most spoofed websites would seem indistinguishable from their genuine counterparts at first glance. Such websites are dedicated to

fraudulent activities and a phishing email is most likely to contain a link to such a website. When visited, these sites would prompt the user to provide their valuable credentials like passwords or credit card numbers assuring them of huge monetary gains or lucrative offers. Needless to say, this information is utilized by the phishers operating that site for their own benefits. This is precisely why phishing attacks are synonymously termed as 'identity theft' attempts.

The problem of phishing attacks in enterprise is next issue rising in wide scale and complexity, as phishers use email phishing via obfuscated, malicious or phished URLs and continuously adapt or innovate their strategies to lure victims. To gain trust and confidence of victim's phishers have started using visceral factors and Familiarity cues. Although in most cases a phisher's clear motive is to commit identity theft in order to benefit from it financially; it is wrong to assume that phishing is always money centric. A phisher can also rob an internet user of his goodwill and character. By phishing the login credentials of an employee or a student, a phisher can trespass their accounts and steal their personal information. There are no limits to what a phisher can do in such a scenario. Earning a bad name for oneself in a professional or academic arena can prove much more traumatic than being embarrassed at a social networking site. It is a challenging task to address this issue.

Therefore, in this thesis novel approaches for E-mail Phishing Detection using Machine Learning techniques are proposed. Firstly, A heuristic based model is designed to detect email phishing via URLs (obfuscated, malicious or phished URLs) using Naïve Bayes and Support Vector Machine classifiers. This model includes a URL detection algorithm which efficiently detects phishing and legitimate URLs. It is evident through extensive literature review that single filter approaches are insufficient to detect different categories of phishing attempts in enterprise

environ. Therefore, a novel anti-phishing model for enterprise using artificial neural network is proposed. Basically, it is s heuristic-based approach with ANN. This proposed phishing detection model incorporates an anti-phishing multi-filter for detection phishing/malicious/obfuscated/spoofed URLs. This anti-phishing multi-filter has ability to scan the effective list of URL features with social features or social human factors. In addition, this model effectively identifies whether the phishing email is known phishing or unknown phishing to reduce the trust and familiarity-based email phishing enterprise environ. The Feed-Forward Backpropagation and Levenberg- Marquart methods of ANN are adopted to enhance the URL classification process and with Fuzzy Inference System to get result with imprecise data of social features. The proposed model can effectively handle zero-day phishing and also accurately classify the known and unknown email phishing via URLs. Thirdly, a novel anti-phishing effectiveness evaluator is developed with a formula to calculate the effectiveness of anti-phish armatures or mechanisms. This evaluator calculates the effectiveness based on email structure vulnerability.

Lastly, guidelines for email phishing prevention are chalked out to avoid the phishing attack. The major key point is listed for naive user to recognize the phishing email and to differentiate between phished and legitimate email via URL. Dissemination of cyber awareness and digital hygiene is recommended for the society. Some suggestive measures are also discussed with case studies for prevention and protection of user from cyber-crime. As it would keep them secured in cyber, social and monitory aspects and also would help them spread the cyber awareness through publicity or campaigning.