Adversarial Phishing Attacks & Defences

Archana Sreekumar

4 September 2019

Abstract

Phishing detection models are becoming a target for adversarial attacks. How to defend against these adversarial attacks is a problem which is still not completely solved. These attacks can evade the classifier and fail it to protect the users from a phishing attack. Even though there exist some defence mechanisms against these adversarial attacks, they cannot guarantee complete protection[1]. Since millions of users are using phishing classifiers, such an attack could lead to exposure of their sensitive information[2]. In this paper I am going to do a study on the existing methods used for adversarial attacks and defences on phishing detection models and expecting to propose a new defense method with improved robustness.

Keywords

Phishing detection, adversarial attacks, defences

References

[1] Anirban Chakraborty et al. “Adversarial attacks and defences: A survey”. In: *arXiv preprint arXiv:1810.00069* (2018).

[2] Bin Liang et al. “Cracking classifiers for evasion: a case study on the google’s phishing pages filter”. In: *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, pp. 345–356.