World Academy of Science, Engineering and Technology
International Journal of Social and Business Sciences
Vol:17, No:08, 2023

Towards an Adversary-Aware ML-Based Detector of Spam on Twitter Hashtags

Authors : Niddal Imam, Vassilios G. Vassilakis

Abstract : After analysing messages posted by health-related spam campaigns in Twitter Arabic hashtags, we found that these campaigns use unique hijacked accounts (we call them adversarial hijacked accounts) as adversarial examples to fool deployed ML-based spam detectors. Existing ML-based models build a behaviour profile for each user to detect hijacked accounts. This approach is not applicable for detecting spam in Twitter hashtags since they are computationally expensive. Hence, we propose an adversary-aware ML-based detector, which includes a newly designed feature (avg posts) to improve the detection of spam tweets posted by the adversarial hijacked accounts at a tweet-level in trending hashtags. The proposed detector was designed considering three key points: robustness, adaptability, and interpretability. The new feature leverages the account's temporal patterns (i.e., account age and number of posts). It is faster to compute compared to features discussed in the literature and improves the accuracy of detecting the identified hijacked accounts by 73%.

Keywords : Twitter spam detection, adversarial examples, evasion attack, adversarial concept drift, account hijacking, trending hashtag
Conference Title : ICFNSM 2023 : International Conference on Fake News and Social Media
Conference Location : London, United Kingdom
Conference Dates : August 17-18, 2023