Intelligent deep machine learning cyber phishing URL detection based on BERT features extraction

ABSTRACT

Recently, phishing attacks have been a crucial threat to cyberspace security. Phishing is a form of fraud that attracts people and businesses to access malicious uniform resource locators (URLs) and submit their sensitive information such as passwords, credit card ids, and personal information. Enormous intelligent attacks are launched dynamically with the aim of tricking users into thinking they are accessing a reliable website or online application to acquire account information. Researchers in cyberspace are motivated to create intelligent models and offer secure services on the web as phishing grows more intelligent and malicious every day. In this paper, a novel URL phishing detection technique based on BERT feature extraction and a deep learning method is introduced. BERT was used to extract the URLs' text from the Phishing Site Predict dataset. Then, the natural language processing (NLP) algorithm was applied to the unique data column and extracted a huge number of useful data features in terms of meaningful text information. Next, a deep convolutional neural network method was utilised to detect phishing URLs. It was used to constitute words or n-grams in order to extract higher-level features. Then, the data were classified into legitimate and phishing URLs. To evaluate the proposed method, a famous public phishing website URLs dataset was used, with a total of 549,346 entries. However, three scenarios were developed to compare the outcomes of the proposed method by using similar datasets. The feature extraction process depends on natural language processing techniques. The experiments showed that the proposed method had achieved 96.66% accuracy in the results, and then the obtained results were compared to other literature review works. The results showed that the proposed method was efficient and valid in detecting phishing websites' URLs.