Abstract

Malware detection has always been a hot issue and a priority task in cyber crimes. Despite a lot of work in the past, malware detection in Microsoft Word remains a significant challenge for researchers and other practitioners.

Recent targeted attacks on organisations have made use of the new Microsoft Word documents (.docx). Anti-virus software is incapable of detecting new unknown malicious files, including malicious docx files. This reaserch examines the malware and detects the malicious files with the help of structural path features and lexical based features of extracted URL from unzipped XML files of Microsoft Word.

This research carried out three experiments and finally reached to a goal with 0.97 accuracy with a highest true positive rate of 0.98 and lowest false positive rate of 0.012. Our method showed a significant improvement in detecting malicious docx files using ensemble learning, as compared to random forest and SVM margin alone, thus giving a better solution for detection model.