МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики
Кафедра математического моделирования и анализа данных

Дорофеев Герман Сергеевич

ОБНАРУЖЕНИЕ ВРЕДОНОСНЫХ DOCX-ФАЙЛОВ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация к дипломной работе

Научный руководитель:
кандидат физико-математических наук,
М.С. Абрамович

Минск, 2022

РЕФЕРАТ

Дипломная работа: 50 стр., 19 рис., 4 таблицы, 12 источников, 3 приложения.

Ключевыеслова: docx, malware, macro, machine learning, classification.

Объект исследования: docx-файлы.

Цель работы: исследовать применение методов машинного обучения в обнаружении вредоносных docx-файлов, разработать приложение.

Методы исследования: методы машинного обучения.

Результат: изучено применение методов машинного бучения в обнаружении вредоносных docx-файлов, разработано приложение.

Область применения: сфера информационной безопасности.

РЭФЕРАТ

Дыпломная работа: 50 стр., 19рыс., 4 табл, 12 крыніц, 3 прыкл.

Ключавыясловы: docx, malware, macro, machinelearning, classification.

Аб’ектдаследавання: docx-файлы.

Мэтыпрацы: даследавацьужываннеметадаўмашыннаганавучанняўвыяўленнішкоднасныхdocx-файлаў, распрацавацьпраграму.

Метадыдаследавання: метадымашыннаганавучання.

Вынікі: вывучанаўжыванеметадаўмашыннаганавучання ў выяўленнішкоднасныхdocx-файлаў, распрацаванапраграма.

Вобласцьужывання: абаронаінфармацыі.

ABSTRACT

Diploma thesis: 50 pages, 19 figures, 4 tables, 12 sources, 3 attachments.

Keywords: docx, malware, macro, machine learning, classification.

Object of research: docx-files.

Purpose: studythe application of machine learning methods for detection of malicious docx-files, develop thecorresponding software application.

Methods: machine learning methods.

Result: the application of machine learning methods for detection of malicious docx-files has been studied. The corresponding software application has been developed.

Scope: information security field.

ВВЕДЕНИЕ

Вредоносныймакрос (такженазываемый «макровирус») - этокод, которыйиспользуетфункциональность документовOffice (обычно этоExceliиWord) длявыполнениязлонамеренныхдействийпротивсистемыпри открытиифайла. ЭтоттипвредоносногоПОбылоченьпопуляренвконце 90-х - начале 2000-хгодов. С 2014-го года макровирусы начали использоваться снова, на этот раз какметодраспространениядругихвредоносныхпрограмм.Так или иначе макровирусыпродолжаютпредставляютугрозудляпользователя. Вэтойработеизучаетсявозможностьулучшенияобнаружениямакровирусов спомощьюметодовмашинногообучения, применяемыхксвойствамкода.

Изначально макросы добавляли дополнительную функциональность к документам, предоставляя им динамические свойства, которые позволяют, например, выполнять действия над набором ячеек в документе Excel или встраивать объекты мультимедиа в файлах Word. Но к концу 90-х они стали использоваться злоумышленниками для выполнения кода в операционной системе. Злоумышленники программировали макросы, наделяя их дополнительными свойствами для выполнения вредоносных действий (например, загрузка и запуск исполняемых файлов).

В конце 1990-х одним из самых распространенных и вредных вирусов была «Мелисса» [1]. Традиционно, наиболее популярным способом распространения этого типа вредоносных программ была (и остается) электронная почта. Жертва получает электронное письмо с вложением, и при его открытии макрос, содержащийся в документе, исполняется и заражает операционную систему. Однако в последующие годы Microsoft предоставили механизм, встроенный в пакет Office, предотвращающий автоматическое выполнение макросов. Из-за этого данный вирус потерял свою актуальность. Этому также поспособствовало существование других, более прямых методов, не зависящих от конфигурации системы Office (например, использование уязвимостей). Однако в 2014 году Microsoft заявили, что макровирусы вновь стали использоваться злоумышленниками,на этот раз как метод распространения другого вредоносного ПО.

В 2014 и 2015 годах макровирусы использовались вполне успешно, несмотряна всяческиемерыпротиводействия. Такимобразом, даже спустяболее чем 15 летпослесвоегопоявлениямакровирусы все ещеостаютсяугрозойимеханизмы ихобнаруженияявсеещенеобходимы.

Многиеисследованиябылисосредоточенынаобнаружениивредоносных PDF-документов [2].Однако, исследование, фокусирующеесянафайле PDF,

неявляетсядостаточнообщим, чтобыегоможнобылоприменятьикдругимфайловымструктурам, такимкакофисныедокументы MS. Файлы PDF значительноотличаютсяотфайловdocxв двухважныхпунктах. Во-первых, файлы PDF состоятизнаборасвязанныхобъектов, афайлыdocx - этоархивныефайлы, состоящиеизпапокифайлов XML. Учитываяэто, применениеструктурногоподходадляобнаруженияредоносных PDF-файлов [3] навредоносныхфайлахdocxбылобынеудачным. Во-вторых, для файлов PDF idocxиспользуютсяразныеметодыатаки, идажекогдаиспользуетсяобщаяметодика (например, встроенныефайлы), атаказапускаетсясовершеннопо-разному, такчтооднаитажеатакавлияетнаструктуруфайлапо-разномувкаждомслучае. Следовательно, дляобнаруженияредоносныхфайловdocxтребуетсядругой подход, основанныйнаспециальнойфайловойструктуреэтихфайлов.

Цель данной работы -продемонстрировать, действительно ли обнаружение ианализнаиболеераспространенныхметодовпрограммированияредоносныхпрограммможетспособствоватьправильнойиавтоматическойклассификацииивозможности отличитьдокументы, содержащихдопустимыемакросы, оттех, что содержат вредоносные.

Обнаружение макровирусов все же является специфической задачей антивирусных систем, поэтому вовсе не предполагается, что методы машинного обучения могут полностью заменить их. Это лишь иной взгляд на проблему. В этой работе планируется провести обзор наиболее известных и удачных способов и подходов к применению техник машинного обучения для обнаружения вредоносных docx-файлов.