Огольцова Н.Д. Использование метода TF-IDF для детектирования вредоносных PDF файлов // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 5(43) с. 13–17

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.056

ИСПОЛЬЗОВАНИЕ МЕТОДА TF-IDF ДЛЯ ДЕТЕКТИРОВАНИЯ ВРЕДОНОСНЫХ PDF ФАЙЛОВ

Огольцова Н.Д.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия, (119454, г. Москва, просп. Вернадского, 78, стр. 4.), e-mail: og.nata@inbox.ru

В статье рассматривается применение метода TF-IDF (Term Frequency-Inverse Document Frequency) для обнаружения вредоносных PDF файлов. Исследуется, как этот метод может быть использован для анализа текста внутри PDF документов, чтобы определить, содержит ли файл вредоносный код или нет. Метод TF-IDF позволяет извлекать ключевые слова из текста, что делает его эффективным инструментом для анализа больших объемов данных. В статье подробно описывается процесс интеграции TF-IDF с алгоритмами машинного обучения, что позволяет значительно улучшить точность и эффективность обнаружения вредоносных файлов. Также рассматриваются преимущества и ограничения предложенного подхода, а также возможности интеграции с другими извлекаемыми признаками из PDF документов для детектирования их вредоносности.

Ключевые слова: TF-IDF, PDF, машинное обучение, классификация документов, извлечение признаков.

USING THE TF-IDF METHOD TO DETECT HARMFUL PDF FILES

Ogoltsova N.D.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue. Vernadsky, 78, b. 4), e-mail: og.nata@inbox.ru

The article discusses the application of the TF-IDF (Term Frequency-Inverse Document Frequency) method for detecting malicious PDF files. We are investigating how this method can be used to analyze text inside PDF documents to determine whether a file contains malicious code or not. The TF-IDF method allows you to extract keywords from text, which makes it an effective tool for analyzing large amounts of data. The article describes in detail the process of integrating TF-IDF with machine learning algorithms, which significantly improves the accuracy and efficiency of detecting malicious files. The advantages and limitations of the proposed approach are also considered, as well as the possibility of integration with other extracted features from PDF documents to detect their harmfulness.

Keywords: TF-IDF, PDF, machine learning, document classification, feature extraction.

PDF (Portable Document Format) — это формат документа, разработанный компанией Adobe Systems в 1990-х годах. Цель создания формата PDF заключалась в создании стандарта для представления документов и других справочных материалов в формате, который не зависит от прикладного программного обеспечения, аппаратного обеспечения и операционной системы. PDF может содержать текст, изображения, гиперссылки, поля форм, мультимедийные материалы, цифровые подписи, вложения, метаданные, геопространственные функции и 3D-объекты. Формат PDF широко используется для обмена

информационными данными между пользователями, так как поддерживается большинством операционных систем, мобильной и компьютерной техники. [1]

По данным подразделения Netskope Threat Labs Stats, американской софтверной компании NetScope, основанной в 2012 году, за ноябрь 2023, формат PDF документа является самым часто используемым форматов для использования в киберпреступности.

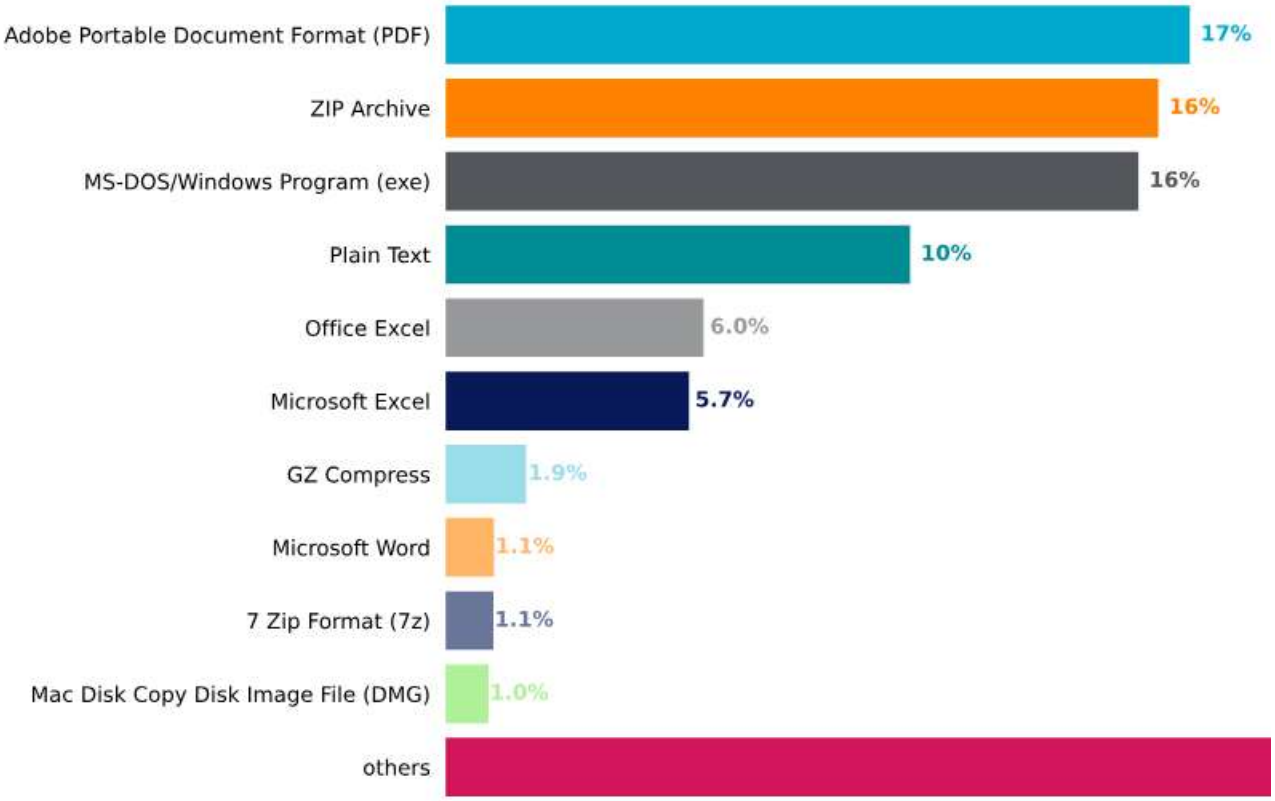

Рисунок 1 – Топ форматов документов по использованию в киберпреступлениях

Согласно статистике, почти половина всех атак вредоносного ПО направлена на малые предприятия, делая их основной мишенью для атак вредоносного ПО. Большинство малых предприятий плохо подготовлены к отражению таких атак, поскольку они обычно не имеют специализированных ИТ-специалистов или надежных систем безопасности. [2]

Вредоносные PDF-файлы могут быть распространены через различные каналы, включая электронную почту, веб-сайты и социальные сети. Киберпреступники могут использовать различные тактики, включая поддельные профили, вредоносные ссылки и обманчивую рекламу, чтобы заставить пользователей скачать или открыть эти файлы.

Методы детектирования вредоносных PDF файлов включают в себя два основных метода – это динамический и статический анализ.

Динамический анализ подразумевает использование поведенческого анализатора для обнаружения аномалий в поведении вредоносных PDF файлов, таких как попытки выполнения вредоносного кода или попытки эксплуатации уязвимостей в программном обеспечении. Этот метод использует инструменты для мониторинга поведения программы в реальном времени, чтобы выявить ошибки, уязвимости и проблемы, которые могут возникнуть только при выполнении программы.

Статический анализ нацелен на поиск известных сигнатур вредоносного кода внутри PDF файлов. Это может включать в себя поиск уникальных хешей вредоносных файлов, которые были обнаружены и анализированы в прошлом. К этому методу также относится и анализ вредоносного JavaScript в PDF файлах, спрятанного в объектах и дешифруемого другим JavaScript кодом, для обнаружения и анализа вредоносного кода.

Преимуществом динамического анализа является возможность обнаружить ошибки и уязвимости, которые возникают только при выполнении программы или открытия файла, однако он может быть более трудоемким и затратным по времени, чем статический анализ, особенно для больших и сложных программ. Статический анализ не требует большого затрата ресурсов, но он подразумевает наличие большой базы исходных данных для точного детектирования. [3]

Исходя из приведённой информации, самым лёгким для последующего внедрения и использования, является статический метод. Используя при этом методы машинного обучения, можно получить модель, способную детектировать вредоносные PDF файлы с высокой точностью.

В исследовании AlMahadeen Awss и Alkasassbeh Mouhammd «PDF Malware Detection using Machine learning» от 2023 года приведён эксперимент, в котором извлекаются сигнатурные признаки из PDF документов, общим количеством – 32, и в последующем обучаются на алгоритме случайного леса в соотношении 80:20. В своём исследовании авторы смогли получить значения точности модели равное 99.5%. [4]

Метод TF-IDF (Term Frequency-Inverse Document Frequency) может улучшить показатель точности — это статистическая мера, которая оценивает, насколько слово релевантно для документа в коллекции документов. Это достигается путем умножения двух метрик: частоты появления слова в документе и обратной частоты документа (IDF) слова в наборе документов.

Частота термина (Term Frequency, TF) определяет, сколько раз слово появляется в документе. Чем чаще слово встречается в документе, тем выше его значение TF.

TF-IDF используется в автоматизированном текстовом анализе и очень полезен для оценки слов в алгоритмах машинного обучения для обработки естественного языка (NLP). Он был изобретен для поиска документов и извлечения информации и работает, увеличиваясь пропорционально количеству раз, когда слово появляется в документе, но компенсируется количеством документов, содержащих это слово.

Пример использования TF-IDF в Python может включать использование метода TfidfVectorizer() из модуля sklearn.feature_extraction.text, который позволяет вычислять значения TF-IDF для слов в документах. [5]

Однако, стоит отметить, что TF-IDF имеет свои ограничения, такие как проблемы с очень редкими терминами, отсутствие понимания смысла или контекста слов, игнорирование порядка слов и трудности с интерпретацией синонимов и похожих слов. [6]

Создание модели, обученной на признаках, извлеченных с помощью парсера PDF файлов, и словах, обработанных с использованием TF-IDF, представляет собой инновационный подход к анализу и классификации PDF документов, особенно в контексте обнаружения вредоносного содержимого. Данная модель может быть внедрена в уже существующие ресурсы, например, почтовые сервисы, для анализа и обнаружения вредоносных файлов.

Этот подход может можно считать наиболее эффективным для обнаружения вредоносных PDF файлов, так как он сочетает в себе анализ структуры файла и анализ текстового содержимого, что позволяет модели лучше понимать характеристики вредоносных документов.

В заключении, объединение этих двух методов в процессе обучения модели машинного обучения может значительно улучшить ее способность к точному обнаружению вредоносных PDF файлов. Этот подход может быть особенно полезен для исследователей в области безопасности информации, специалистов по кибербезопасности и разработчиков ПО, работающих над обнаружением и предотвращением вредоносного ПО. Этот подход можно считать эффективным и многоаспектным к обнаружению вредоносных PDF файлов. Он может служить основой для разработки более продвинутых систем обнаружения вредоносного ПО, способных адаптироваться к новым угрозам и методам атаки.

Список литературы

1. «Обзор формата PDF» [Электронный ресурс] URL: https://helpx.adobe.com/ru/incopy/using/pdf.html (Дата обращения: 27.03.2024);
2. «Актуальные киберугрозы: III квартал 2023 года» [Электронной ресурс] URL: https://www.netskope.com/blog/netskope-threat-labs-stats-for-september-2023 (Дата обращения: 27.12.2023);
3. Li, Min & Zhou, Ying & Yu, Min & Liu, Chao. (2016). Combining static and dynamic analysis for the detection of malicious JavaScript-bearing PDF documents. 475-482. 10.1142/9789813200449_0059. (Дата обращения: 27.03.2024);
4. AlMahadeen, Awss & Alkasassbeh, Mouhammd. (2023). PDF Malware Detection using Machine learning. 10.20944/preprints202301.0557.v1 (Дата обращения: 27.03.2024);
5. «Understanding TF-IDF (Term Frequency-Inverse Document Frequency)» [Электронный ресурс] URL: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/ (Дата обращения: 27.03.2024);
6. Jayady, Siti & Antong, Hasmawati. (2021). Theme Identification using Machine Learning Techniques. Journal of Integrated and Advanced Engineering (JIAE). 1. 123-134. 10.51662/jiae.v1i2.24. (Дата обращения: 27.03.2024).

References

1. «Overview of the PDF format.» [Web resource] URL: https://helpx.adobe.com/ru/incopy/using/pdf.html (Date of address: 27.03.2024);
2. «Current cyber threats: third quarter 2023» [Web resource] URL: https://www.netskope.com/blog/netskope-threat-labs-stats-for-september-2023 (Date of address: 27.12.2023);
3. Li, Min & Zhou, Ying & Yu, Min & Liu, Chao. (2016). Combining static and dynamic analysis for the detection of malicious JavaScript-bearing PDF documents. 475-482. 10.1142/9789813200449_0059. (Date of address: 27.03.2024);
4. AlMahadeen, Awss & Alkasassbeh, Mouhammd. (2023). PDF Malware Detection using Machine learning. 10.20944/preprints202301.0557.v1 (Date of address: 27.03.2024);

5. «Understanding TF-IDF (Term Frequency-Inverse Document Frequency)» [Web resource] URL: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/ (Date of address: 27.03.2024);
6. Jayady, Siti & Antong, Hasmawati. (2021). Theme Identification using Machine Learning Techniques. Journal of Integrated and Advanced Engineering (JIAE). 1. 123-134. 10.51662/jiae.v1i2.24. (Date of address: 27.03.2024).