DOI: 10.34185/1991-7848.itmm.2024.01.067

АВТОМАТИЗОВАНЕ ВИЯВЛЕННЯ ПОТЕНЦІЙНО НЕБЕЗПЕЧНИХ URL-АДРЕС З ВИКОРИСТАННЯМ БІБЛІОТЕКИ SCIKIT-LEARN

Каштан В.Ю., Панферова Я.В., Бешта Л.В., Ващук Д.О.

Національний технічний університет «Дніпровська політехніка», Дніпро, Україна

Анотація. *Розглянуто методологію автоматизованого виявлення потенційно небезпечних URL-адрес за допомогою бібліотеки scikit-learn. Запропонована методологія включає підготовку даних, генерацію ознак та оцінку моделі на основі алгоритму випадкового лісу для класифікації URL-адрес на фішингові та бещпечні. Реалізація методології здійснюється з використанням мови програмування Python та бібліотеки scikit-learn. Результати експериментів показують ефективність моделі у виявленні потенційно небезпечних URL-адрес, що може відігравати важливу роль у захисті користувачів від шахрайства та інших онлайн-загроз.*

Ключові слова: *небезпечні URL-адреси, фішинг, scikit-learn, машинне навчання.*

Вступ. Онлайн-послуги та веб-сайти стали необхідною складовою сучасного життя в різних сферах, включаючи бізнес, освіту, банківську справу та особисте життя. Зі зростанням популярності і використання онлайн-ресурсів зростає і кількість шкідливих веб-сайтів. Шкідливий веб-сайт може містити небажаний вміст з метою збору конфіденційних даних або встановлення шкідливого програмного забезпечення на комп'ютер користувача. Часто це відбувається без уведення користувача, зокрема під час драйвового завантаження, коли шкідливе програмне забезпечення автоматично встановлюється без його дозволу.

Захист від таких атак складний, оскільки іноді навіть обережне користування Інтернетом не є достатнім. Зловмисники можуть використовувати вразливості у веб-додатках для впровадження шкідливого коду без відома власника. Згідно з дослідженням Webroot Threat Report за 2019 рік, 40% шкідливих URL-адрес було виявлено на добропорядних доменах [1]. Це означає, що навіть легітимні веб-сайти можуть стати потенційною загрозою для користувачів.

У зв'язку з цим виникає необхідність розробки методів та інструментів, які допоможуть відрізнити шкідливі URL-адреси від безпечних. Одним зі широко використовуваних методів захисту є створення чорних списків [2], але

цей підхід має свої обмеження. Іншим методом є застосування методів машинного навчання для виявлення шаблонів у шкідливих URL-адресах [3, 4, 5].

Мета роботи полягає в розробці програмного інструменту для виявлення потенційно шкідливих URL-адрес з використанням керованого навчання. Для цього використовується алгоритм випадковий ліс та бібліотека scikit-learn.

Методологія. На рис.1 наведено структурну схему інструменту виявлення потенційно небезпечних URL-адрес. Ефективна підготовка даних вважається ключовою складовою процесу машинного навчання і приводить до досягнення кращих результатів. Метою цього дослідження є навчання та оцінка різних моделей машинного навчання. Для цього створено інструмент на мові програмування Python та бібліотеки scikit-learn для обробки необроблених даних URL-адрес.

Представлений інструмент включає чотири етапи обробки даних: попередню обробку, генерацію ознак, зберігання даних та пост обробку. Кожен з цих етапів може конфігуруватися шляхом додавання або видалення компонентів відповідно до потреб дослідження. Компоненти інструменту динамічно завантажуються згідно з конфігурацією програми. Крім того, деякі компоненти можуть підтримувати багато потоковість для оптимізації швидкодії обробки даних.

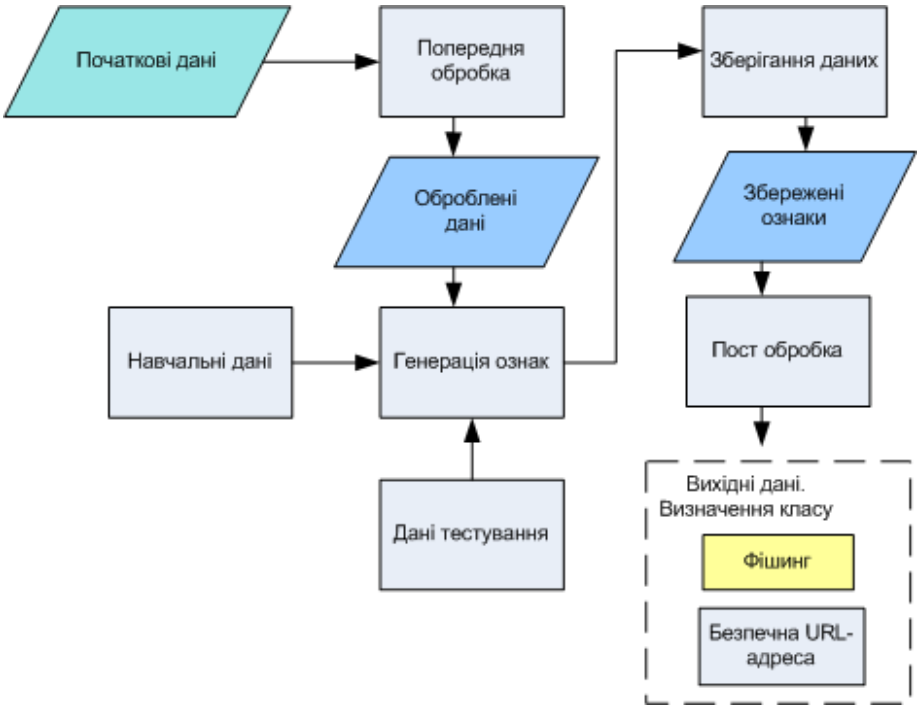


Рисунок 1 – Структурна схема методології

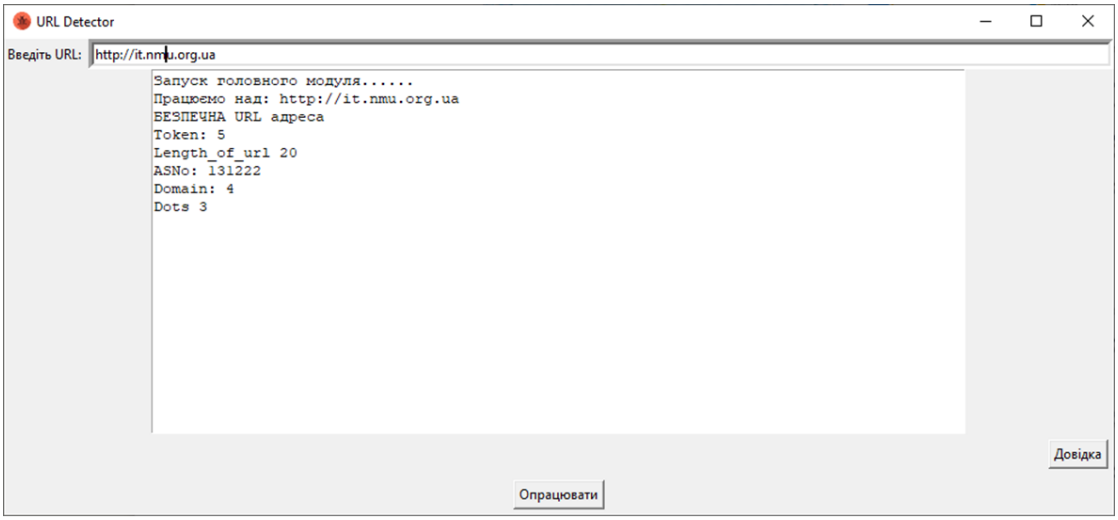
Попередня обробка є першим етапом, на якому здійснюється збір та збереження ключової інформації про вхідні дані. Цей етап включає в себе отримання всієї необхідної інформації для подальшої обробки ознак, такої як словник або значення для нормалізації даних.

Генерація ознак представляє собою основний етап обробки даних, під час якого відбувається виділення ознак з вхідних даних і їх збереження для майбутнього використання. Результатом цього етапу є список вилучених ознак разом з відповідними значеннями.

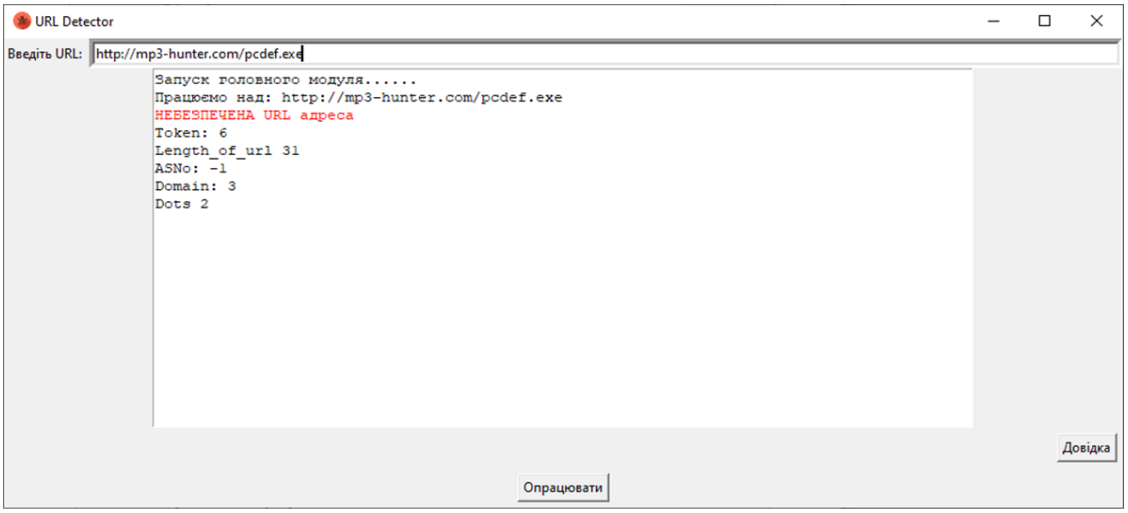
Зберігання даних - це етап, на якому здійснюється перетворення та збереження ознак у відповідному форматі, що є необхідним для подальшого використання алгоритму випадковий ліс.

Останнім етапом є пост обробка даних. Після навчання класифікатора на навчальному наборі даних для класифікації URL-адрес як фішингових або безпечних, використовується отриманий шаблон для класифікації нових вхідних даних. З URL-адреси виділяються характеристики, такі як IP-адреса, довжина URL, домен, тощо, і формується список їхніх значень. Цей список передається на вхід класифікатору випадкових лісів. Після цього оцінюється продуктивність моделей і генерується оцінка їхньої точності. Навчений

класифікатор, використовуючи згенерований список, визначає, чи є URL-адреса безпечною чи фішинговою. У списку використовуються значення 1, 0 та -1 для позначення наявності, відсутності або незастосовності ознак відповідно. Результати запропонованої методології автоматизованого виявлення потенційно небезпечних URL-адрес були реалізовані за допомогою мови програмування Python та бібліотеки scikit-learn (рис.2).



а)



б)

Рисунок 2 – Результати запропонованої методології: а) безпечна URL-адреса кафедри ІТКІ НТУ «ДП»; б) фітингова адреса

Висновки. Запропоновано методологію автоматизованого виявлення потенційно небезпечних URL-адрес за допомогою бібліотеки scikit-learn. В рамках цієї методології проведено підготовку даних, виконано генерацію ознак та оцінку моделі на основі алгоритму випадкового лісу для класифікації URL-адрес на фішингові та легальні. Реалізація запропонованої методології

здійснена за допомогою мови програмування Python та бібліотеки scikit-learn. Результати показують ефективність моделі у виявленні потенційно небезпечних URL-адрес, що може бути корисним для захисту користувачів від шахрайства та інших онлайн-загроз.

ЛІТЕРАТУРА / REFERENCE

1. INC., Webroot Threat Report. ¶[Електронний ресурс] – Режим доступу до ресурсу: https : / / www - cdn . webroot . com.
2. Sheng, Steve, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason I. Hong and Chengshan Zhang. “An Empirical Analysis of Phishing Blacklists.” International Conference on Email and Anti-Spam, 2009.
3. J. H. Ateeq and M. Moreb, “Detecting malicious URL using neural network”, in Proc. Int. Congr. Adv. Technol. Eng. (ICOTEN), Jul. 2021, pp. 1–8.
4. M. E. H. V. S. Aalla and N. R. Dumpala, “Malicious URL prediction using machine learning techniques”, Ann. Romanian Soc. Cell Biol., vol. 25, no. 5, pp. 2170–2176, 2021.
5. Y. Pingle, S. N. Bhatkar, and S. Patil, “Detection of malicioius content using AI”, in Proc. 7th Int. Conf. Computing Sustain. Global Develop., 2020, pp. 1–6.

AUTOMATED DETECTION OF POTENTIALLY DANGEROUS URL ADDRESSES USING THE SCIKIT-LEARN LIBRARY

Kashtan Vita, Panferova Yana, Beshta Liliia, Vashchuk Dmytro

Abstract. *The methodology of automated detection of potentially dangerous URLs using the sci-kit-learn library is considered. The proposed methodology includes data preparation, feature generation, and model evaluation based on the random forest algorithm for classifying URLs into phishing and safe ones. The methodology is implemented using the Python programming language and the scikit-learn library. Experimental results show the effectiveness of the model in identifying potentially dangerous URLs, which plays an essential role in protecting users from fraud and other online threats.*

Keywords: *dangerous URLs, phishing, scikit-learn, machine learning.*