Аналіз витоків паролів на наявність патернів та можливості їх субслівної токенізації

УДК 004.056

Сергій Бабич[1], Петро Голуб[2], Богдан Слив'як[3]

Національний водного господарства та природокристування,
[1]s.v.babych@nuwm.edu.ua, [2]p.p.holub@nuwm.edu.ua, [3]slyviak_ak23@nuwm.edu.ua

Паролі використовуються повсюдно як засіб автентифікації користувачів, відповідно складність паролів є одним із основних факторів захищеності інформаційних систем. Кореляції ж у рамках однієї із останніх збірок витоків паролів, а саме RockYou2024 демонструють стабільну наявність популярних патернів, що стосуються персональної інформації чи цифрового сліду людини [1]. Урядові та відомчі мережі, включно з індустріальними системами управління (SCADA), також залишаються вразливими до використання передбачуваних паролів, що становить загрозу критичні інфраструктурі [2, 3].

Метою роботи є дослідження тенденцій та кореляцій у масштабних витоках паролів, а також використання отриманих результатів для подальшого дослідження граматичних особливостей парольних фраз та їх субслівної токенізації, задля верифікації складності паролів базуючись на цифровому сліді людини. Відповідні результати можуть бути надзвичайно важливими для аудиту парольних політик всередині організації або створення релевантних цільових словників паролів на основі розвідки з відкритих джерел (OSINT) [4].

Новизна даного дослідження полягає у розширенні класичних підходів до побудови словників паролів та інтеграції українського соціокультурного і лінгвістичного контексту. Сучасні дослідження підтверджують, що традиційні моделі генерації паролів ігнорують вплив зовнішніх семантичних факторів (соціальних трендів, популярної лексики), що суттєво знижує їхню адаптивність [5]. Запропонований підхід субслівної токенізації враховуватиме унікальні регіональні аспекти, такі як транслітерацію кириличних слів, крос-розкладне введення та використання специфічних абревіатур.

Очікується, що це дозволить підвищити точність та швидкість підбору паролів. Даний фактор є важливим, оскільки грунтуючись на аналізі витоків паролів [2], варто відзначити, що за останні 15 років середня довжина паролів зросла з 8 до 10 символів, а їхня середня інформаційна ентропія — із 40 до 50 біт. Оцінка ентропії H здійснюється за класичною формулою Шеннона [3]:

$$E = -\sum_{i=1}^{n} P_i log_2 P_i \quad (1)$$

де P_i — імовірність появи i-го символу чи токена з доступного простору.

Загалом існує сукупність факторів, що дозволяє прослідкувати ускладнення паролів, проте залишається конструктивна слабкість через наявність передбачуваної інформації.

Для розв'язання поставленої задачі пропонується використовувати алгоритмічну модель генерації та скорингу токенів. Вона опрацьовує неструктурований масив вхідних OSINT-даних (імена, дати, локації, специфічний сленг) та виконує їх багаторівневу обробку. Процес включає вилучення базових сутностей, їх розбиття на субслівні одиниці (склади,

біграми), застосування алгоритмів транслітерації та додаткових мутацій. З метою формування оптимального словника, кожен токен отримує скорингову оцінку від 0 до 100 балів (табл. 1).

Таблиця 1

Критерії скорингової оцінки субслівної токенізації

Критерій	Опис параметрів токена	Макс. бал
Пріоритет джерела	Прямі збіги (імена, дати), ініціали	40
Локалізація	Транслітерація, крос-розкладне введення (kbswap), N-грами	30
Частотність	Мультипликатор ваги за частоту появи токена у вхідних OSINT-даних	20
Довжина	Оптимальний розмір токена для формування паролю (найвища вага 3-6 символів)	10

Отже, ускладнення паролів не усуває людського фактору, користувачі продовжують спиратися на власний цифровий слід та локальний контекст. Виокремлення граматичних особливостей формування парольних фраз і поєднання їх із субслівною токенізацією та OSINT-розвідкою дозволить створювати високорелевантні цільові словники паролів. Запропонована скорингова модель генерації токенів може бути ефективним інструментом для тестування на проникнення та комплексного аудиту безпеки інформаційних систем.

1. Слатвінська В., Бевза В., Вплив збою CrowdStrike на мега-витік паролів: чи є зв’язок? Ч. 1. Вісник Хмельницького національного університету. Технічні науки. – 2024. – № 4 (339). – С. 332-338. DOI: 10.31891/2307-5732-2024-339-4-52.
2. Rodrigues G.A.P. et al., From RockYou to RockYou2024: Analyzing Password Patterns Across Generations, Their Use in Industrial Systems and Vulnerability to Password Guessing Attacks. Journal of Internet Services and Applications. – 2025. – Vol. 16, No. 1. DOI: 10.5753/jisa.2025.5041.
3. Собина В.О., Тарадуда Д.В., Демент М.О., Захист інформації відомчої інформаційно-телекомунікаційної мережі за допомогою парольної системи. Проблеми надзвичайних ситуацій. – Харків: НУЦЗ України, 2021. – № 2 (34).
4. Різак М., Котик О., Використання OSINT для захисту персональних даних. The 14th International Scientific Conference «ITSec». – Тернопіль, 2025. – С. 165-167.
5. Yang X. et al., KAPG: Adaptive Password Guessing via Knowledge-Augmented Generation. – 2025.