РОЗВІДКА КІБЕРЗАГРОЗ З ВИКОРИСТАННЯМ NOTEBOOKLM ЯК МЕТОД ОТРИМАННЯ СТАТИСТИЧНИХ ДАНИХ ДЛЯ РОЗРАХУНКУ РИЗИКІВ КІБЕРБЕЗПЕКИ ОРГАНІЗАЦІЇ

Слончинська Н.О., Чакрян В.Х., Андрушко Д.В.

Кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Харківський національний університет радіоелектроніки,
Україна

E-mail: nadiia.slonchynska@nure.ua
vadym.chakrian@nure.ua
dmytro.andrushko@nure.ua

Abstract

The paper presents an approach to cyber threat intelligence (CTI) aggregation using Google NotebookLM as a unified analytical environment for processing heterogeneous open-source cybersecurity reports. The proposed method enables automated extraction and synthesis of statistical indicators of cyberattacks, adversary tactics, techniques, and procedures (TTPs), as well as frequency characteristics of successful intrusions. These data can be applied in quantitative and semi-quantitative cybersecurity risk assessment models to increase the accuracy of probability estimations and support data-driven risk management decisions in modern organizations.

Сучасні методи розвідки кіберзагроз (Cyber Threat Intelligence, CTI) ґрунтуються на залученні великої кількості різнорідних джерел, серед яких щорічні галузеві звіти Verizon Data Breach Investigations Report [1], IBM Cost of a Data Breach [2], регулярні аналітичні огляди SANS Institute [3], статистичні дані антивірусних лабораторій та інші відкриті звіти, приклади яких можна знайти на GitHub сторінці [4]. Сукупність таких звітів формує фундамент для формування статистики атак, інцидентів, поширених тактик і технік кіберзловмисників.

Фрагментованість джерел та відсутність єдиного інструменту для інтеграції статистики ускладнюють використання цих даних у моделях оцінювання ризиків. У зв'язку з цим у роботі запропоновано застосування Google NotebookLM як інтегрованої платформи для семантичного аналізу, класифікації та агрегації змісту відкритих CTI-зведень. NotebookLM дозволяє завантажувати значні обсяги текстових матеріалів, автоматично синтезувати зведену статистику, виявляти закономірності в описах атак, зводити інформацію щодо застосованих технік, тактик, та процедур відповідно до MITRE ATT&CK [5], а також оцінювати частоти реалізації окремих типів кіберзагроз на основі.

Серед найбільш інформативних статистичних показників, що доцільно аналізувати у межах запропонованого підходу, варто виділити частоту реалізації окремих векторів атак, а також динаміку їхньої популярності в часовій перспективі. Важливими є розподіл інцидентів за галузями та ймовірність компрометації певних типів активів, що дозволяє визначити найбільш вразливі сегменти організації. Значну аналітичну цінність мають метрики часу виявлення та реагування, тривалість прихованої присутності зловмисника, частота обходу механізмів автентифікації, а також частка інцидентів, спричинених людським фактором. Для моделювання ризиків особливо цінною є статистика щодо частоти експлуатації критичних вразливостей, частки вразливостей із переліку CISA KEV, швидкості появи експлойтів після публікації даних про вразливість, а також співвідношення типових технік бокового руху та ескалації привілеїв відповідно до MITRE ATT&CK. Не менш важливими є показники середніх збитків для різних класів атак, розміру викрадених даних, векторів пост-експлуатації, а також розподіл цілей атак у хмарних, гібридних та локальних середовищах. Сукупність цих характеристик дозволяє формувати достовірну емпіричну базу для побудови кількісних моделей ризиків та пріоритезувати заходи кіберзахисту відповідно до реальних трендів кіберзагроз.

Отримані за допомогою NotebookLM агреговані статистичні показники можуть бути інтегровані у кількісні й напівкількісні моделі оцінювання ризиків кібербезпеки. Зокрема, вони можуть підсилювати моделі на основі частотного підходу, байєсівські мережі, а також методи типу FAIR, де вхідні параметри імовірності подій та середніх збитків значною мірою залежать від достовірності вхідної статистики. Крім того, емпіричні показники частоти експлуатації можуть використовуватися у поєднанні з EPSS та CVSS-метриками, підвищуючи точність оцінки актуальності окремих вразливостей. Все це дозволить підвищити точність, якість, та об'єктивність розрахунку кіберризиків та визначення пріоритетів їх усунення.

Разом з тим, при використанні NotebookLM необхідно враховувати його технічні обмеження. Згідно з офіційною документацією, кількість джерел у межах одного блокнота може становити до 50 джерел у стандартній конфігурації інструменту [6]. Максимальний розмір кожного з текстових джерел досягає приблизно 500 000 слів або еквівалентного об'єму інших типів документів [7]; окремі типи файлів також мають ліміт ~200 МБ. Ці обмеження необхідно враховувати при побудові масштабованої CTI-моделі, зокрема при роботі з великими наборами звітів понад кілька сотень документів.

Окремою перевагою NotebookLM є його ефективність у порівнянні з іншими загальнодоступними сервісами штучного інтелекту. На відміну від універсальних мовних моделей, таких як ChatGPT, Claude чи Perplexity, NotebookLM реалізує повністю джерело-орієнтований аналіз, при якому всі відповіді моделі ґрунтуються виключно на завантажених документах. Це значно зменшує ризик галюцинацій і підвищує точність витягнутої статистики, що особливо важливо для CTI та ризик-менеджменту [6]. Крім того, NotebookLM автоматично створює міждокументні зв'язки, здійснює глибоку семантичну індексацію та формує узагальнення з багатьох джерел, що в інших сервісах великих мовних моделей вимагає розгортання спеціалізованих RAG-конвеєрів, зовнішніх баз знань і значних технічних ресурсів [8].

Попри значні переваги, запропонований підхід має низку обмежень, які необхідно враховувати під час практичного застосування. Насамперед, якість аналітики прямо залежить від репрезентативності завантажених джерел, а щорічні звіти різних компаній часто мають різні методології збору даних, різну повноту охоплення та потенційні комерційні чи регіональні упередження. Це може призводити до структурного зміщення статистики та потребує виваженого підбору джерел. Другим обмеженням є технічні ліміти NotebookLM щодо кількості джерел, розміру файлів та обсягу тексту, що ускладнює роботу з великими історичними корпусами даних і змушує виконувати попереднє нормування або сегментацію документів. Крім того, NotebookLM виконує семантичну інтерпретацію тексту, але не гарантує глибокої галузевої експертності, що може впливати на коректність складних висновків без участі аналітика. Важливою є й неможливість автоматичного оновлення даних у реальному часі, оскільки сервіс працює виключно з вручну завантаженими матеріалами, що обмежує оперативність при аналізі динамічних загроз. Нарешті, глибока залежність від структури текстових звітів означає, що дані, представлені лише у вигляді графіків, зображень або складних PDF-таблиць, можуть бути інтерпретовані неповністю або некоректно, що потребує додаткової перевірки результатів людиною. У сукупності ці фактори вказують на те, що використання NotebookLM є ефективним інструментом підтримки прийняття рішень, але не може повністю замінити експертну оцінку та класичні методи ризик-менеджменту.

Також, потрібно зазначити, що якість результатів у NotebookLM значною мірою залежить від точності та конкретності сформульованих запитів. Оскільки система працює виключно з інформацією з завантажених джерел, запити повинні бути джерело-орієнтованими, чіткими та такими, що визначають тип запитуваних даних (наприклад: частота, розподіл, динаміка чи порівняння між звітами). Рекомендовано зазначати часові межі та назви джерел, оскільки це мінімізує ризики неповних або узагальнених відповідей [9], [10]. Загальні або надто широкі запити можуть призвести до втрати деталізації, тоді як надмірно складні запити можуть бути інтерпретовані некоректно через обмеження в опрацюванні графіків, таблиць або малоструктурованих даних [11]. Додатковим ризиком є семантичне зміщення, коли довші чи більш структуровані документи мають більший вплив на результати, тож важливо забезпечувати збалансованість корпусу джерел. Дотримання цих рекомендацій дозволяє підвищити точність, повноту та релевантність отриманої аналітики.

У комплексі описані властивості дозволяють розглядати NotebookLM як високоефективний інструмент для автоматизації збору, узагальнення та аналітики статистичних даних про кіберзагрози. При правильному виокристанні, враховуючи всі обмеження і дотримуючись рекомендацій і найкра-

щих практик використання цього інструменту, такий підхід сприяє трансформації класичного якісного підходу до оцінювання ризиків у даноорієнтовану модель, що характеризується підвищеною точністю прогнозів та обґрунтованістю прийняття рішень.

Література

1. Verizon. Data Breach Investigations Report. Verizon Enterprise, 2024.
2. IBM Security. Cost of a Data Breach Report. IBM Corp., 2024.
3. SANS Institute. Threat Landscape Survey and Reports. SANS, 2024.
4. Wilson J. D. Awesome-annual-security-reports: A curated list of annual cyber security reports [Електронний ресурс]. GitHub, 2024. Режим доступу: https://github.com/jacobdjwilson/awesome-annual-security-reports.
5. MITRE Corporation. MITRE ATT&CK Framework [Електронний ресурс]. 2024. Режим доступу: https://attack.mitre.org.
6. Google Support. Google NotebookLM: Source and notebook limits [Електронний ресурс]. 2024. Режим доступу: https://support.google.com/notebooklm/answer/16213268.
7. Google Support. Google NotebookLM file size and content limits [Електронний ресурс]. 2024. Режим доступту: https://support.google.com/notebooklm/answer/16269187.
8. Google Research Blog. Google NotebookLM Technical Overview [Електронний ресурс]. 2024. Режим доступу: https://blog.google/technology/ai/notebooklm/.
9. Google Help Center. How to use NotebookLM: Get started, use chat, and create notebooks [Електронний ресурс]. 2024. Режим доступу: https://support.google.com/notebooklm.
10. Google Blog. 8 expert tips for getting started with NotebookLM [Електронний ресурс]. 2024. Режим доступу: https://blog.google/technology/ai/notebooklm-beginner-tips.
11. LearnPrompting.org. A Complete How-To Guide to NotebookLM [Електронний ресурс]. 2024. Режим доступу: https://learnprompting.org/blog/notebooklm-guide.