

АЛГОРИТМ ВИЯВЛЕННЯ ФІШИНГОВИХ ЗАГРОЗ НА ОСНОВІ ТРАНСФОРМЕРНОЇ АРХІТЕКТУРИ DISTILBERT

Вінницький національний технічний університет

Анотація

У роботі представлено алгоритм виявлення фішингових загроз у веборієнтованих системах на основі донавченої трансформерної моделі DistilBERT. Розглянуто архітектуру системи, що включає модулі попередньої обробки даних, класифікації та пояснення рішень. Описано методіку генерації синтетичних навчальних даних для адаптації до нових типів атак.

Ключові слова: DistilBERT, трансформери, виявлення фішингу, глибоке навчання, синтетичні дані.

Abstract

The paper presents an algorithm for detecting phishing threats in web-oriented systems based on a fine-tuned DistilBERT transformer model. The system architecture including data preprocessing, classification, and decision explanation modules is examined. A methodology for generating synthetic training data to adapt to new attack types is described.

Keywords: DistilBERT, transformers, phishing detection, deep learning, synthetic data.

Вступ

Сучасні фішингові атаки характеризуються високим ступенем складності та персоналізації, що робить традиційні методи виявлення малоефективними. Застосування трансформерних архітектур глибокого навчання дозволяє аналізувати семантичні та контекстуальні особливості текстових повідомлень і URL-адрес, виявляючи приховані патерни шкідливого контенту.

Модель DistilBERT, як оптимізована версія BERT, забезпечує баланс між точністю класифікації та швидкістю обробки, що є критично важливим для систем захисту реального часу. Актуальною проблемою залишається адаптація моделей до постійно змінюваних тактик зловмисників, зокрема до використання генеративного штучного інтелекту для створення правдоподібного фішингового контенту [1].

Архітектура системи виявлення

Розроблена система виявлення фішингових загроз складається з чотирьох основних компонентів, що взаємодіють послідовно для забезпечення комплексного аналізу вхідних даних. Модуль попередньої обробки здійснює токенізацію текстових даних, нормалізацію URL-адрес та перетворення їх у числові представлення, сумісні з вхідним форматом моделі DistilBERT. Основний класифікаційний модуль використовує донавчену модель DistilBERT з додатковим повнозв'язним шаром для бінарної класифікації, що дозволяє розрізняти легітимні та фішингові повідомлення на основі аналізу контекстуальних залежностей у тексті.

Модуль генерації синтетичних даних базується на великих мовних моделях і забезпечує створення навчальних зразків для адаптації системи до нових типів атак без необхідності збору реальних інцидентів. Механізм пояснення на базі Layer Integrated Gradients дозволяє визначати найбільш впливові слова та фрази, що сприяли прийняттю рішення про класифікацію, підвищуючи прозорість та довіру користувачів до системи [2].

Алгоритм класифікації

Процес виявлення фішингових загроз реалізується через послідовність етапів обробки вхідних даних та їх аналізу засобами нейронної мережі. На початковому етапі вхідний текст або URL-адреса

проходить токенизацію за допомогою WordPiece токенизатора, що розбиває послідовність на підслова та додає спеціальні токени початку й кінця послідовності. Отримані токени перетворюються на числові ідентифікатори та подаються на вхід моделі DistilBERT разом з масками уваги, що вказують на позиції реальних tokenів.

Трансформерний енкодер моделі обробляє послідовність через шість шарів самоуваги, формуючи контекстуалізовані векторні представлення кожного токена з урахуванням їхніх взаємозв'язків у реченні. Представлення спеціального токена класифікації агрегується через повнозв'язний шар з функцією активації sigmoid для отримання ймовірності належності до класу фішингових повідомлень. Додатково система генерує оцінку впливовості кожного слова на фінальне рішення, що реалізується через обчислення градієнтів виходу відносно вхідних tokenів і їх інтегрування вздовж шляху від базового до реального входу [4].

Методика адаптації до нових загроз

Для забезпечення стійкості системи до еволюції фішингових тактик розроблено методику автоматичної генерації синтетичних навчальних даних на основі великих мовних моделей. Процес генерації включає формування промптів, що містять характеристики типових фішингових сценаріїв, такі як терміновість, запити конфіденційної інформації та імітація відомих сервісів. Блок-схема алгоритму замкнутої генерації та валідації даних зображено на рис. 1.

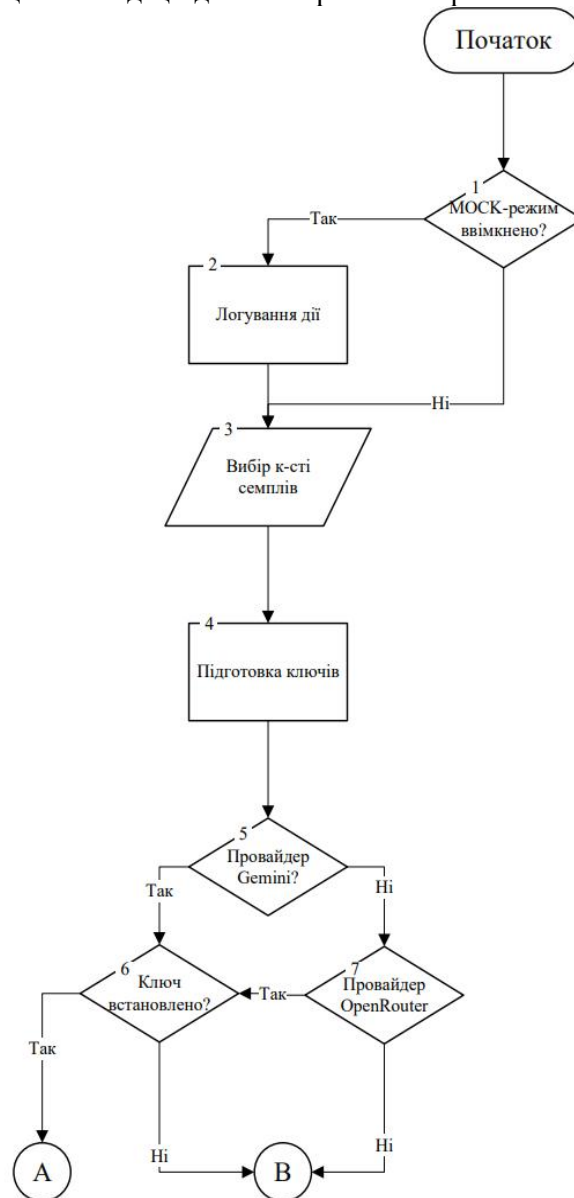
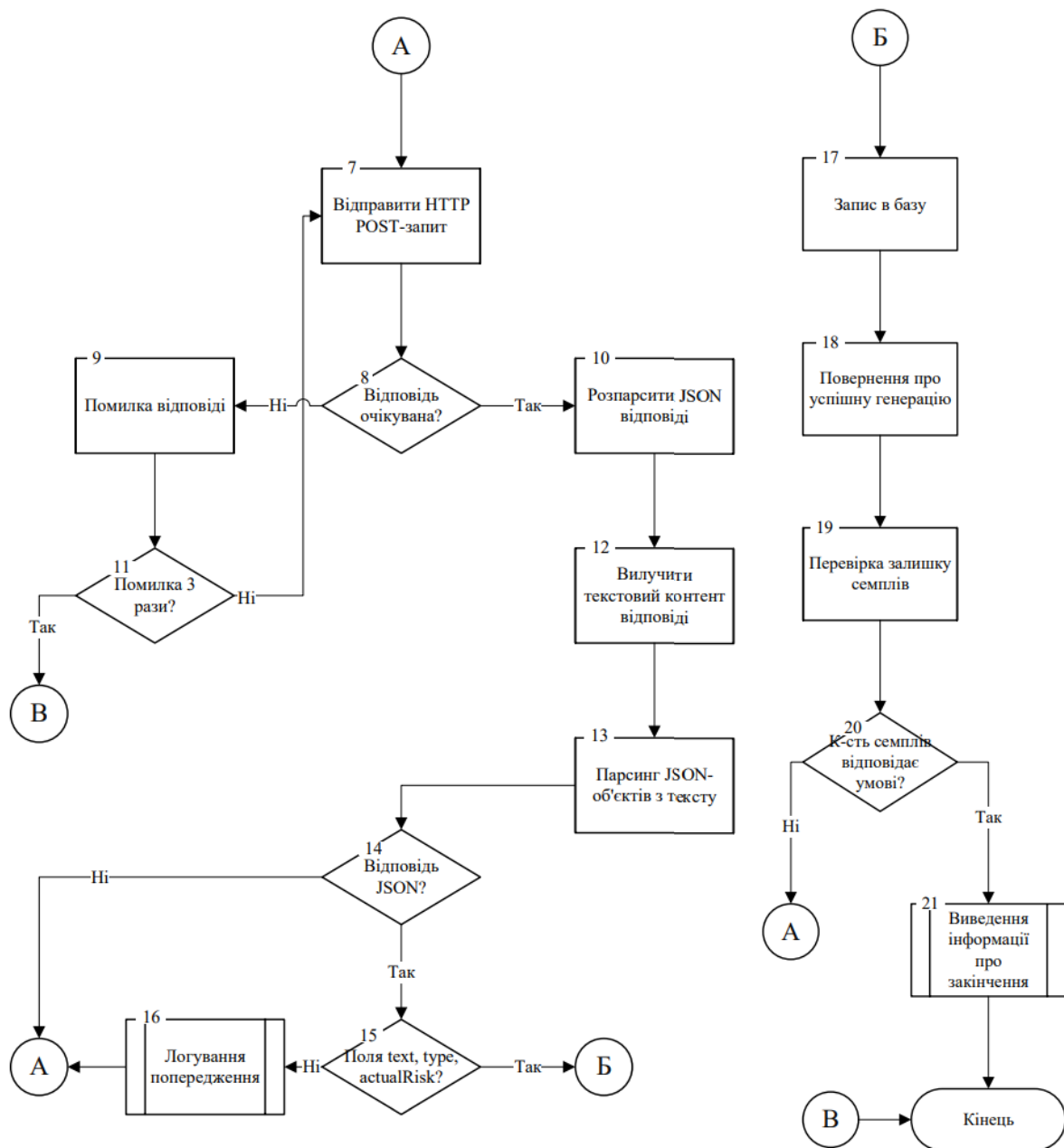


Рис. 1. Блок-схема алгоритму замкнутого циклу генерації даних



Продвоження рис. 1.

Як видно з блок-схеми, згенеровані зразки проходять автоматичну фільтрацію на основі попередньо визначених критеріїв якості та різноманітності, після чого використовуються для періодичного донавчення моделі DistilBERT методом fine-tuning зі зниженою швидкістю навчання для запобігання катастрофічному забуванню раніше набутих знань. Такий підхід дозволяє системі адаптуватися до нових типів атак із мінімальною затримкою та без необхідності ручного збору та розмічання великих обсягів реальних фішингових повідомлень [4].

Висновки

Представлений алгоритм виявлення фішингових загроз на основі трансформерної архітектури DistilBERT демонструє ефективний баланс між точністю класифікації та обчислювальною ефективністю. Інтеграція модуля генерації синтетичних даних забезпечує можливість адаптації системи до постійно змінюваних тактик зловмисників без значних витрат на збір та розмічання навчальних даних.

Механізм пояснення рішень на основі Layer Integrated Gradients підвищує довіру користувачів та полегшує аналіз помилок класифікації для подальшого вдосконалення моделі. Подальший розвиток алгоритму може включати інтеграцію мультимодального аналізу для обробки візуальних компонентів фішингових атак та оптимізацію архітектури для розгортання на мобільних пристроях з обмеженими обчислювальними ресурсами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. *Knowledge and Information Systems*, 64(6), 1457-1500.
 2. Yang, W., Zuo, W., & Cui, B. (2019). Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network. *Ieee Access*, 7, 29891-29900.
 3. Tarapiah, S., Abbas, L., Mardawi, O., Atalla, S., Himeur, Y., & Mansoor, W. (2025). Evaluating the effectiveness of large language models (LLMs) versus machine learning (ML) in identifying and detecting phishing email attempts. *Algorithms*, 18(10), 599.
- Syed, S. A. (2025). Adversarial AI and cybersecurity: defending against AI-powered cyber threats. *Iconic Research And Engineering Journals*, 8(9), 1030-1041.

Гуменчук Едуард Сергійович – студент групи 1KITC-24м, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця, e-mail: hasesed1@gmail.com

Науковий керівник: **Грицак Анатолій Васильович** – кандидат технічних наук, провідний фахівець Центру ІТ та захисту інформації, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця.

Eduard S. Humenchuk student of group 1KITS-24m, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia.

Supervisor: **Anatolii V. Hrytsak** Candidate of Technical Sciences (Ph.D.), Leading Specialist of the Center for IT and Information Protection, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia.