ВПЛИВ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ НА ЕФЕКТИВНІСТЬ МОДЕЛЕЙ РОЗПІЗНАВАННЯ ФІШИНГОВИХ ЛИСТІВ

Леонова А.О., Сидоренко З.М.
Харківський національний університет радіоелектроніки, Харків, Україна

Сучасні фішингові атаки значно змінилися через появу генеративного штучного інтелекту (Gen-AI). Завдяки Gen-AI зловмисники можуть автоматично створювати велику кількість різних повідомлень, які виглядають природно, написані без помилок і добре підлаштовані під конкретну ситуацію. Через це такі повідомлення стають більш переконливими й їх складніше розпізнати як шахрайські. Раніше фішингові листи часто можна було впізнати за помилками або дивною мовою. Тепер же, завдяки великим мовним моделям (LLM), такі повідомлення виглядають майже так само, як звичайне офіційне листування, і їх набагато складніше відрізнити від справжніх. Через це виникло своєрідне технологічне протистояння, у якому ті самі інструменти використовуються і для захисту, і для здійснення атак. У результаті традиційні методи виявлення фішингу стають менш ефективними, що потребує їх переоцінки та вдосконалення [1]. Попри це, фішинг залишається однією з найпоширеніших кіберзагроз, що підтверджує необхідність подальших досліджень у цій сфері [2].

Метою доповіді є аналіз впливу генеративного штучного інтелекту на ефективність сучасних моделей виявлення фішингових повідомлень, а також визначення перспективних методик підвищення їх стійкості з огляду на нові типи атак.

Традиційно фішингові листи виявляють за допомогою методів машинного та глибокого навчання. Сучасні моделі на основі трансформерів, такі як BERT і RoBERTa, показують дуже високу точність, приблизно 98,9–99,1% на стандартних наборах даних [3]. Це пов’язано з тим, що вони добре розуміють контекст тексту і можуть знаходити складні смислові зв’язки, наприклад ознаки терміновості, маніпуляції або спроби видати себе за надійне джерело.

Аналіз досліджень показує, що моделі глибокого навчання значно ефективніші за класичні алгоритми, оскільки вони самі визначають важливі ознаки та здатні працювати з неструктурованими даними [2]. Отже, на сьогодні саме моделі глибокого навчання є більш результативним підходом для виявлення фішингових листів.

Водночас поява Gen-AI значно ускладнила виявлення фішингових повідомлень. Дослідження показують, що навіть невелике перефразування таких повідомлень за допомогою LLM може помітно знизити точність традиційних методів їх виявлення [1]. Це пояснюється тим, що генеративні моделі прибирають типові ознаки фішингу, наприклад, граматичні помилки чи неприродні мовні конструкції, на яких раніше часто базувалося навчання моделей. У результаті системи, які показували дуже високі результати на класичних наборах даних, працюють значно гірше в реальних умовах. Це

призводить до так званої «ілюзії безпеки», коли модель здається точною, але насправді не враховує сучасні загрози, створені за допомогою штучного інтелекту [2]. Класичні моделі машинного та глибокого навчання залишаються швидкими та стабільними, проте вони обмежені у здатності розпізнавати складні контекстуальні маніпуляції та визначати справжній намір повідомлення.

Сучасні великі мовні моделі, особливо квантовані середньорозмірні, краще справляються з аналізом змісту тексту та можуть виявляти приховані стратегії соціальної інженерії [1]. Однією з їхніх важливих переваг є здатність пояснювати, чому було прийнято те чи інше рішення, що підвищує довіру до систем виявлення та робить їх більш корисними на практиці. Разом із тим, їхня точність поки що трохи нижча, ніж у спеціалізованих моделей глибокого навчання, тому для ефективного застосування потрібна додаткова інтеграція з іншими методами.

Одним із перспективних напрямів є застосування квантованих великих мовних моделей (LLM), які забезпечують оптимальний баланс між продуктивністю та обчислювальними витратами [1]. Такі моделі краще адаптуються до нових видів атак і можуть проводити більш глибокий аналіз контексту повідомлень. Також великі можливості дають гібридні підходи, що поєднують високу точність моделей глибокого навчання з пояснювальними властивостями LLM. Не менш важливим є оновлення навчальних наборів даних із додаванням прикладів, згенерованих штучним інтелектом, що дозволяє робити оцінку моделей більш реалістичною та підвищує їх стійкість до сучасних загроз [2, 3].

Генеративний штучний інтелект змінює сам підхід до виявлення фішингу. Якщо раніше це була відносно проста задача визначити, чи є повідомлення шкідливим чи ні, то тепер усе складніше. Сучасні системи повинні не просто класифікувати повідомлення, а аналізувати його зміст, розуміти контекст і намагатися визначити справжній намір. Тобто йдеться вже не тільки про перевірку ознак, а про глибше розуміння сенсу повідомлення. Незважаючи на високі показники сучасних моделей глибокого навчання, їх ефективність знижується у разі використання LLM для генерації атак. Це свідчить про необхідність розробки адаптивних, комбінованих і пояснюваних систем, здатних враховувати контекст, протистояти перефразуванню та працювати з актуальними даними.

Список літератури

1. Thakur K., Ali M. L., Obaidat M. A., Kamruzzaman A. A Systematic Review on Deep-Learning-Based Phishing Email Detection. Electronics. 2023. Vol. 12, No. 21. Art. No. 4545. DOI: https://doi.org/10.3390/electronics12214545.

2. Thapa J., Chahal G., Gabreanu Ş. V., Otoum Y. Phishing Detection in the Gen-AI Era: Quantized LLMs vs Classical Models. arXiv preprint. 2025. arXiv:2507.07406v1. URL: https://arxiv.org/abs/2507.07406v1.

3. Alhuzali A., Alloqmani A., Aljabri M., Alharbi F. In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. Applied Sciences. 2025. Vol. 15, No. 6. Art. No. 3396. DOI: https://doi.org/10.3390/app15063396.