УДК 004.056.53:004.8
Д. Вихованець; М. Стадник, к. т. н.
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ВИЯВЛЕННЯ ФІШИНГОВИХ ПОВІДОМЛЕНЬ ЗА ДОПОМОГОЮ LLM

UDC 004.056.53:004.8
D. Vykhovanets; M. Stadnyk, PhD.

DETECTION OF PHISHING MESSAGES USING LLMS

Виявлення фішингових повідомлень за допомогою великих мовних моделей (LLM) є одним із найперспективніших напрямів сучасної кібербезпеки. Стрімке зростання кількості електронних комунікацій та масштабування соціоінженерних технік призвели до того, що фішинг став домінуючим вектором атак: за даними Verizon DBIR [1], 42% успішних кіберінцидентів у 2025 році починалися з фішингової взаємодії, а аналітика IBM свідчить, що середня вартість наслідків однієї фішингової атаки сягає 4,76 млн доларів США. Компанія Proofpoint [2] відзначає, що понад 90% світових організацій щороку піддаються фішинговим атакам, а частота таргетованого spear-phishing зросла більш ніж на 22% у 2023–2025 рр. Одночасно зі збільшенням кількості загроз спостерігається й підвищення їх складності, оскільки зловмисники активно використовують генеративні моделі для створення високоякісних та персоналізованих фішингових повідомлень, які складно виявити традиційними методами.

Традиційні підходи до виявлення фішингу демонструють зниження ефективності через динамічний характер атак та здатність зловмисників обходити відомі механізми захисту. Це створює потребу у впровадженні інтелектуальних систем, здатних аналізувати не лише поверхневі ознаки, а й семантичний зміст, контекст, наміри та приховані патерни в тексті повідомлень.

Поява великих мовних моделей (GPT, LLaMA, Claude, Mistral та ін.) докорінно змінила можливості аналізу текстової інформації. Архітектура Transformer забезпечує глибоке розуміння контексту, що дозволяє LLM виявляти фішингові повідомлення у zero-shot та few-shot режимах, часто без додаткового навчання на спеціалізованих вибірках. Це значно підвищує гнучкість та швидкість адаптації системи до нових типів атак, включаючи spear-phishing, clone-phishing та шахрайські повідомлення у соціальних мережах і месенджерах. Використання технологій fine-tuning, prompt engineering та RAG-архітектури відкриває можливість створення високоточних моделей класифікації фішингових текстів, які не лише визначають наявність загрози, але й пояснюють логіку прийнятого рішення.

Проте застосування LLM у сфері кібербезпеки супроводжується низкою викликів: ризиком генеративних галюцинацій, потенційними вразливостями до prompt-ін’єкцій, необхідністю захисту конфіденційних даних та оптимізації обчислювальних ресурсів під час розгортання моделі в хмарі або локальній інфраструктурі.

Експериментальні дослідження та практичні впровадження демонструють, що LLM-орієнтовані рішення можуть значно підвищити точність виявлення фішингу, зменшити кількість хибних спрацювань і забезпечити здатність системи адаптуватися до загроз.

Література

1. Verizon. (2025). 2025 Data Breach Investigations Report. https://www.verizon.com/business/resources/T555/reports/2025-dbir-data-breach-investigations-report.pdf (date of access: 03.12.2025).
2. Proofpoint. (2024). State of the Phish 2024. Proofpoint, Inc. https://www.proofpoint.com/us/resources/threat-reports/state-of-phish (date of access: 03.12.2025).