КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1092
УДК 004.934

Терейковський Ігор Анатолійович
доктор технічних наук, професор,
професор кафедри системного програмування і спеціалізованих комп'ютерних систем
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID: 0000-0003-4621-9668
tereikovskyi@kpi.ua

Терейковська Людмила Олексіївна
доктор технічних наук, професор,
професор кафедри інформаційних технологій проєктування та прикладної математики
Київський національний університет будівництва і архітектури, Київ, Україна
ORCID: 0000-0002-8830-0790
tereikovska.lo@knuba.edu.ua

ВИЯВЛЕННЯ ФІШІНГУ В КАНАЛАХ ЕЛЕКТРОННОЇ КОМУНІКАЦІЇ

Анотація. Стаття присвячена підвищенню ефективності засобів захисту інформації вітчизняного кіберпростору шляхом автоматизованого виявлення фішингу у текстових повідомленнях каналів електронної комунікації. Показано, що фішинг залишається одним із домінуючих векторів кібератак, зокрема в умовах активного застосування методів соціальної інженерії та психоемоційного впливу, що суттєво ускладнює його своєчасне виявлення традиційними сигнатурними та статистичними методами. Встановлено, що більшість відомих нейромережевих рішень характеризуються високою ресурсоємністю, потребують формування значних обсягів розмічених навчальних даних та є недостатньо адаптованими до специфіки вітчизняного контенту, який характеризується обмеженістю розмічених корпусів, морфологічною варіативністю та чутливістю до контекстних соціально-інженерних впливів. Для подолання зазначених обмежень у статті запропоновано метод виявлення фішингу, який ґрунтується на використанні апробованих діалогових інтелектуальних помічників на базі великих мовних моделей у режимі формалізованої діалогової взаємодії. Метод передбачає автоматизований аналіз текстових повідомлень шляхом подачі стандартизованих запитів, сформованих з урахуванням типових ознак фішингових атак. Запропоновано класифікацію фішингових повідомлень за вектором ознак виявлення. Для кожного виду сформовано цільові предикати запитів, що забезпечують детермінований та інтерпретований аналіз результатів. Розроблений метод передбачає попередню обробку тексту, формування та подачу запитів до діалогових інтелектуальних помічників, агрегацію відповідей та оцінювання наявності фішингу з використанням механізму зваженої лінійної згортки. Проведені експериментальні дослідження засвідчили, що середній показник точності класифікації при застосуванні запропонованого методу становить 97,5%, що відповідає рівню найкращих відомих рішень аналогічного призначення. Водночас реалізація запропонованого методу не потребує ресурсоємного навчання нейромереж, формування великих обсягів розмічених даних і створення спеціалізованого апаратно-програмного забезпечення, що забезпечує оперативність створення ефективної системи виявлення фішингу у вітчизняних каналах електронної комунікації.

Ключові слова: безпека інформації; виявлення; діалоговий інтелектуальний помічник; електронна комунікація; захист інформації; нейронна мережа; фішинг.

ВСТУП

Аналіз динаміки кіберінцидентів за останній період, а також результати сучасних науково-технічних досліджень [9, 10, 16] підтверджують, що однією з пріоритетних задач у сфері кібербезпеки є підвищення ефективності захисту користувачів і корпоративних ресурсів від фішингових атак. У сучасному розумінні фішинг розглядається як різновид кібернетичного шахрайства, що реалізується у формі текстових, мультимедійних або комбінованих повідомлень та ґрунтується на застосуванні методів соціальної інженерії з метою маніпулювання поведінкою адресата і спонукання його до виконання дій, вигідних зловмиснику, зокрема, розкриття конфіденційної інформації або ініціювання несанкціонованих транзакцій у каналах електронної комунікації. Ключовим напрямом протидії фішингу є вдосконалення засобів автоматичного виявлення шахрайських повідомлень, оскільки саме своєчасна ідентифікація загрози становить основу побудови ефективних систем захисту. Незважаючи на різноманіття підходів до фільтрації контенту, на практиці найбільш поширеними залишаються методи сигнатурного аналізу та перевірки за «чорними списками», що пояснюється відносною простотою реалізації та інтеграції таких рішень у вже існуючу інфраструктуру. Разом із тим, у роботах [11, 12, 17] наголошується на суттєвих обмеженнях сучасних засобів виявлення фішингу. Насамперед, це низька ефективність проти атак без шкідливих посилань, що базуються на психологічних маніпуляціях, а також відсутність дієвих механізмів семантичного аналізу тексту. Крім того, відзначається висока ресурсоємність таких систем та їх слабка адаптивність до нових сценаріїв шахрайства. Окремо підкреслюється доцільність застосування для задач виявлення фішингових повідомлень сучасних нейромережевих технологій, зокрема моделей глибокого навчання та великих мовних моделей, здатних аналізувати зміст повідомлень з урахуванням семантики, прагматики та контексту комунікації. Водночас зазначається, що недостатня ефективність існуючих рішень значною мірою зумовлена недосконалістю методологічних підходів до їх побудови, налаштування та адаптації до специфіки вітчизняного контенту, який характеризується обмеженістю розмічених корпусів, морфологічною варіативністю та чутливістю до контекстних соціально-інженерних впливів.

Постановка проблеми. Вдосконалення методологічного базису виявлення фішингу у текстових повідомленнях засобів електронної комунікації.

Аналіз останніх досліджень і публікацій. Проблема автоматизованого виявлення фішингу та соціальної інженерії залишається у фокусі уваги світової наукової спільноти та державних інституцій. Згідно зі звітом Європейського агентства з кібербезпеки, електронна пошта продовжує залишатися одним із основних векторів кібератак, причому спостерігається зміщення акцентів від масових розсилок до цільового фішингу [5]. В Україні ця негативна тенденція відображається у аналітичних звітах Державного центру кіберзахисту Держспецзв'язку за 2024 рік та перше півріччя 2025 року. При цьому офіційна статистика фіксує стабільне зростання активності кіберугруповань та визначає фішинг як одну з домінуючих категорій інцидентів, що вимагає вказує на необхідність вдосконалення відповідних систем захисту [13, 14].

Еволюція методів захисту від фішингу пройшла шлях від сигнатурних фільтрів до нейромережевих архітектур. У роботах [3, 4, 6] досліджено базові підходи до фільтрації спаму та фішингу з використанням класичних методів машинного навчання (Support Vector Machines, Naïve Bayes, Random Forest). Автори довели ефективність цих методів для ідентифікації атак, що базуються на фіксованих шаблонах та наявності відомих шкідливих посилань. Водночас суттєвим обмеженням зазначених алгоритмів є їх залежність від ручного конструювання ознак та нездатність глибоко аналізувати

семантичний контекст повідомлення. Це робить їх вразливими до нових сценаріїв атак з використанням методів соціальної інженерії, де зловмисники уникають використання типових ключових слів, покладаючись на психологічні маніпуляції.

Альтернативою рішенням на основі класичних методів машинного навчання стало використання нейромережевих технологій. Вважалось, що такі технології дозволять уникнути ручного формування шаблонів та забезпечить можливість врахування семантики повідомлення. Наприклад, у роботі [2] за результатами експериментальних досліджень нейромережевих систем на базі CNN та LSTM емпірично підтверджено, що моделі глибокого навчання здатні автоматично виділяти складні залежності в текстових даних, забезпечуючи високу точність класифікації без використання попередньо визначених шаблонів. В роботах [1, 10] показано можливість виявлення фішингових повідомлень, що містять шкідливі URL-адреси, за рахунок використання нейромережевих моделей, оптимізованих для аналізу технічних і структурних ознак. Разом з тим, аналіз цих рішень вказує на їхню обмеженість у протидії фішинговим атакам типу BEC (Business Email Compromise) через фокусування на формальних атрибутах та недостатню ефективність виявлення маніпуляцій, прихованих у семантиці текстового повідомлення. Для подолання цих обмежень світовий тренд досліджень в області нейромережевих технологій виявлення фішингу 2024-2025 років змістився у бік використання великих мовних моделей та трансформерів, пристосованих до визначення семантики текстових повідомлень. У дослідженні [2] показано, що застосування засобів на базі великої мовної моделі BERT дозволяє досягти точності виявлення фішингу близько 98%, що перевищує точність засобів розпізнавання на базі Random Forest та Naive Bayes. Розвиваючи цей напрям, автори [4] запропонували систему виявлення на основі гібридної моделі BERT-LSTM, що забезпечила досягнення точності 99,55%. При цьому вказується на високу ресурсоємність навчання такої моделі.

Вагомим фактором для вітчизняного кіберпростору є мовна специфіка. У дослідженні [8] показано, що стандартні моделі, ефективні для англійської мови, демонструють високий рівень помилкових спрацювань на мовах зі складною морфологією, до яких належить і українська. У вітчизняній науковій літературі ця проблематика активно досліджується. Зокрема у праці [7] проаналізовано методи протидії атакам, згенерованим засобами штучного інтелекту, а роботу [17] присвячено оцінці ефективності нейромережевих засобів для виявлення фейкового україномовного контенту. При цьому особливу увагу приділено використанню великих мовних моделей для ідентифікації шахрайських схем та автоматизації кібербезпеки. Результати дослідження свідчать про необхідність комплексного підходу до захисту від атак соціальної інженерії, який передбачає синергію між технологічними інструментами та людським фактором. Разом з тим результати аналізу, проведеного у працях [9, 17], свідчать, що інструментальні засоби виявлення текстових фішингових повідомлень, характерних для вітчизняних каналів електронної комунікації, не повною мірою враховують специфіку семантичної побудови сучасних загроз та динаміку змін у сценаріях соціотехнічних атак.

Мета статті. Розробка ефективного методу виявлення фішингу у текстових повідомленнях вітчизняних каналів електронної комунікації.

РОЗРОБКА ТА ДОСЛІДЖЕННЯ

У рамках цього дослідження під фішинговим текстовим повідомленням будемо розуміти текстовий набір даних, зміст яких містить лінгвістичні та стилістичні ознаки маніпулятивного впливу, сформовані з використанням методів соціальної інженерії та спрямовані на введення адресата в оману з метою спонукання його до виконання дій, вигідних зловмиснику, зокрема розкриття конфіденційної інформації або ініціювання несанкціонованих операцій у каналах електронної комунікації.

Використавши результати [12, 15] та авторські напрацювання в області розробки нейромережевих засобів виявлення кібератак, в основу методу виявлення фішингу у вітчизняних каналах електронної комунікації покладено положення щодо використання в якості інструментальних засобів виявлення апробованих діалогових інтелектуальних помічників (ДІП) на базі великих мовних нейромережевих моделей. Доцільність такого підходу обумовлена тим, що апробовані ДІП типу ChatGPT забезпечують готовий до використання інструментарій для глибокого семантичного аналізу текстових повідомлень. Це дозволяє виявляти приховані ознаки фішингу без потреби складної розробки спеціалізованих нейромережевих моделей та відлагодження програмного забезпечення. Крім того, застосування таких ДІП у значній мірі нівелює труднощі, пов'язані з необхідністю формування репрезентативних наборів даних, у яких відображено різні типи маніпулятивних впливів у текстових повідомленнях українською мовою. Також зазначимо, що застосування в якості ядра виявлення фішингу ДІП типу ChatGPT в певній мірі являється реалізацією наведеної в [9] концепції використання великих мовних моделей для автоматизованого виявлення маніпулятивної складової у текстових повідомленнях засобів масової інформації у контексті захисту вітчизняного кіберпростору. У відповідності до означеної концепції, з урахуванням обмеженості спрямування фішингових повідомлень, функціонал засобів виявлення запропоновано описувати виразом виду:

$$T_{NN}(X, R) = \langle Y \rangle, \tag{1}$$

$$\langle Y \rangle = \langle A, B, C \rangle, \tag{2}$$

де T_{NN} – функція, що описує отримання відповіді від ДІП; X – текст, що підлягає аналізу; R – множина запитів до ДІП; $\langle Y \rangle$ – кортеж, що містить результат виявлення фішингу; A – результат оцінювання наявності фішингу; B – вид (спрямування) фішингу; C – пояснення результату оцінювання.

Зазначимо, що, відповідно до наведеної концепції [9], реалізація функціоналу, означеного виразами (1, 2), передбачає використання наперед сформованих запитів до ДІП для виявлення фішингу із наперед визначеної множини видів фішингу.

Враховуючи можливості великих мовних моделей, визначено доцільність класифікувати фішингові атаки не за тематикою, а за вектором аналізу, необхідним для їх виявлення. Відповідно до цього підходу в базовому випадку, виділено сім класів фішингових повідомлень, для ідентифікації яких формуються специфічні запити до ДІП. Опис означених видів фішингу наведено в таблиці 1, а характеристики цільових предикатів запитів до ДІП для виявлення фішингу наведено в таблиці 2.

Таблиця 1

Характеристика видів фішингу

Позначення	Спрямування фішингу	Характерні ознаки
b_1	Компрометація автентифікаційних даних	Вимоги повторної авторизації, імітація сторінок входу, запити на введення облікових даних у зовнішніх формах
b_2	Ексфільтрація платіжних реквізитів	Обіцянки соціальних виплат чи повернення коштів, форми для «верифікації» картки, імітація банківських запитів
b_3	Пряме ініціювання транзакцій	Надання номерів карток/криптогаманців у тексті, QR-коди для оплати, сценарії «родич у біді» або «терміновий збір»
b_4	Збір персональних даних	Анкети для «актуалізації даних», запити скан-копій документів, імітація перевірок від держорганів
b_5	Перехоплення сесії або керування акаунтом	Пропозиції перейти за посиланням для «скасування видалення акаунту», запити на прив'язку нового пристрою, маніпуляції з OAuth-токенами
b_6	Доставка шкідливого програмного забезпечення	Наявність вкладень (архіви, APK, exe), посилання на файлообмінники, легенди про «оновлення безпеки» або «перегляд документів»
b_7	Психоемоційний вплив	Штучне створення дефіциту часу («діє тільки 10 хвилин»), залякування блокуванням або штрафами, спекуляції на темах війни чи здоров'я

Таблиця 2

Характеристика цільових предикатів запитів для виявлення фішингу

Позначення предикату	Позначення виду фішингу	Формулювання цільового предикату
r_1	b_1	Чи містить текст спроби спонукання адресата до передачі логінів, паролів або кодів доступу?
r_2	b_2	Чи наявні в повідомленні запити на отримання повних даних банківської картки (PAN, CVV, термін дії) або доступу до банкінгу?
r_3	b_3	Чи містить повідомлення заклики до негайного переказу коштів на вказані реквізити?
r_4	b_4	Які саме персональні ідентифікатори (ПІБ, ІПН, адреса, паспортні дані) намагається отримати відправник?
r_5	b_5	Чи спонукає інструкція до дій, що можуть призвести до втрати контролю над обліковим записом без передачі пароля?
r_6	b_6	Чи містить повідомлення заклики до завантаження файлів або встановлення сторонніх додатків?
r_7	b_7	Чи використовуються в тексті агресивні емоційні тригери (страх, терміновість, жалість) для пригнічення критичного мислення?

З метою уніфікації процесу обробки даних структуру запитів було стандартизовано. Зокрема, ініціалізуюча частина, що задає фокусну рамку, формалізована виразом (3), а уточнюючий компонент – виразом (4).

$$q_1 = \text{"Проаналізуй текст і в числовому вигляді } (0-1) \text{ без коментарів сформулюй чітку відповідь на запитання щодо наявності фішингу. Запитання."}, \quad (3)$$

$$q_2 = \text{" Лаконічно поясни свою відповідь."} \quad (4)$$

$$q_3 = \text{" Надай відповідь у форматі: Розглянутий вид фішингу } - \text{ "} \quad (5)$$

$$q_4 = \text{" , Результат оцінювання } - \text{ , Пояснення відповіді } - \text{ ."} \quad (6)$$

Зазначимо, що ініціалізуюча частина запиту (3) призначена для мінімізації стохастичності генерації та запобігання виникненню «галюцинацій» ДІП. Зміст цієї

частини формулює жорсткі граничні умови роботи ДІП, вимагаючи від нього функціонування в режимі детермінованого експертного аналізу. Інструкція «без коментарів» слугує фільтром для відсікання надлишкової інформації, що забезпечує отримання лаконічної та однозначної відповіді, придатної для подальшої автоматизованої обробки.

Передбачено, що, відповідно до семи означених в табл. 1 видів фішингу, повний аналіз тексту передбачає виконання семи запитів, цільові предикати яких наведено в таблиці 2. Повний текст запиту для виявлення одного із видів фішингу визначається так:

$$Prompt_i = q_1 + r_i + q_2 + q_3 + b_i + q_4 + X, \quad (7)$$

$$X = \text{Текст:[текст для аналізу]}, \quad (8)$$

де $Prompt_i$ – запит для виявлення *i*-го виду фішингу; r_i – позначення цільового предикату для виявлення *i*-го виду фішингу.

Для підвищення достовірності результатів виявлення фішингу передбачено паралельну обробку результатів запиту (7) декількома ДІП. Для прикладу в якості таких ДІП можливо використовувати ChatGPT, Gemini, Grok, Claude. У цьому випадку узагальнений результат оцінювання наявності фішингу певного виду визначається з використанням моделі зваженої лінійної згортки:

$$a_i = \sum_{k=1}^{K} w_{i,k} s_{i,k}, \quad (9)$$

$$\sum_{k=1}^{K} w_{i,k} = 1, \qquad w_{i,k} > 0, \quad (10)$$

де a_i – результат оцінювання наявності *i*-го виду фішингу ($a_i \in \boldsymbol{A}$); K – кількість використаних ДІП; $w_{i,k}$ – ваговий коефіцієнт компетентності *k*-го ДІП щодо виявлення *i*-го виду фішингу; $s_{i,k}$ - результат оцінювання наявності *i*-го виду фішингу *k*-им ДІП.

В базовому випадку прийнято, що вагові коефіцієнти компетентності різних ДІП є однаковими. Тобто

$$w_{i,k} = 1/K. \quad (11)$$

Результати проведених досліджень дозволили запропонувати наступну реалізацію методу, що відповідно до виразів (1, 2) забезпечує виявлення фішингу у текстових повідомленнях вітчизняних каналів електронної комунікації.

Етап 1. Попередня обробка текстового повідомлення. На вхід етапу подається текст, що підлягає аналізу на предмет виявлення фішингу. Обробка реалізується для зменшення обсягу текстового повідомлення у відповідності до вимог апробованих ДІП. Відповідно до виразу (1), виходом етапу являється X – текст, адаптований до аналізу за допомогою ДІП.

Етап 2. Визначення переліку ДІП, що будуть використані для виявлення фішингу. На вхід етапу подається D_a – множина доступних апробованих ДІП, а виходом етапу являється D_e – множина ефективних ДІП. Визначення множини D_e реалізується експертним шляхом з урахуванням технічних та економічних ресурсів, що виділені для

виявлення фішингу, а також з урахуванням вимог щодо безпечного використання доступних ДІП.

Етап 3. Формування запитів до ДІП. Етап виконується відповідно до виразів (3-8) з використанням даних табл. 1, 2 та X, отриманого в результаті виконання першого етапу даного методу. Входом етапу являється X – текстове повідомлення, що підлягає аналізу, а виходом – множина запитів до ДІП, $Prompt$.

Етап 4. Реалізація запитів до ДІП. На вхід етапу подається множина $Prompt$ та множина D_e. Виходом етапу являється S, B, C – множини відповідей ДІП у форматі визначеному виразами (5, 6). Формування відповідей здійснюється за рахунок подання запитів до ДІП.

Етап 5. Оцінювання наявності фішингу. На вхід етапу подаються S – результати оцінювання кожного ДІП та W – множина значень вагових коефіцієнтів компетентності використаних ДІП. Виходом етапу являється A – множина, що містить результати оцінювання наявності кожного виду фішингу, представленого у табл. 1. Оцінювання реалізується відповідно до виразів (9-11).

Результатом виконання запропонованого методу являється дещо модифікований кортеж $\langle Y \rangle$:

$$\langle Y \rangle = \langle A, S, B, C \rangle \tag{12}$$

Відносно заданого виразом (2) базового варіанту, модифікація результату виявлення фішингу полягає у додаванні до складу кортежу $\langle Y \rangle$ множини S, що містить результати оцінювання наявності кожного із видів фішингу кожним із ДІП із складу D_e.

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДУ ВИЯВЛЕННЯ ФІШИНГУ

Для емпіричного підтвердження ефективності запропонованого методу було проведено експериментальні дослідження, спрямовані на перевірку його функціональних можливостей щодо аналізу легітимних та фішингових повідомлень з використанням різних ДІП. При цьому до множини ефективних ДІП було віднесено ChatGPT, Gemini, Grok та Claude.

У першій серії експериментів було проаналізовано 100 легітимних повідомлень, отриманих авторами у вигляді листів електронної пошти. Ілюстрацією застосування ChatGPT, Gemini для аналізу таких повідомлень на предмет виявлення різних видів фішингу являються рис. 1, 2.

ISSN 2663 – 4023

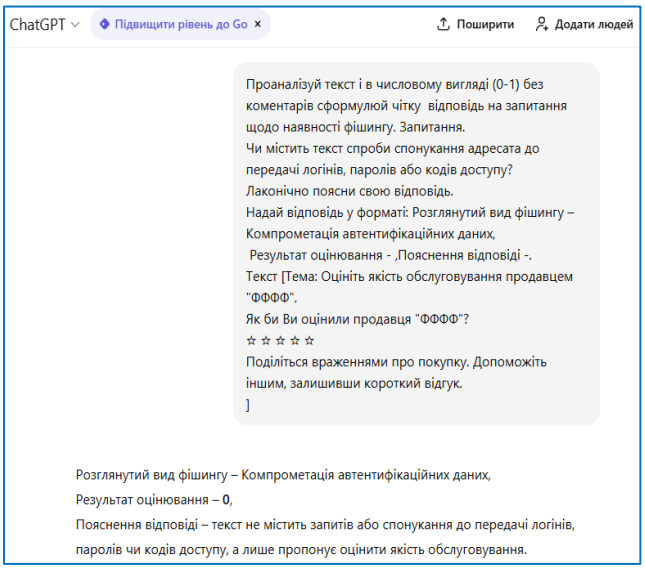

Рис. 1. Ілюстрація застосування ChatGPT для аналізу легітимного листа на предмет наявності фішингу виду b_1

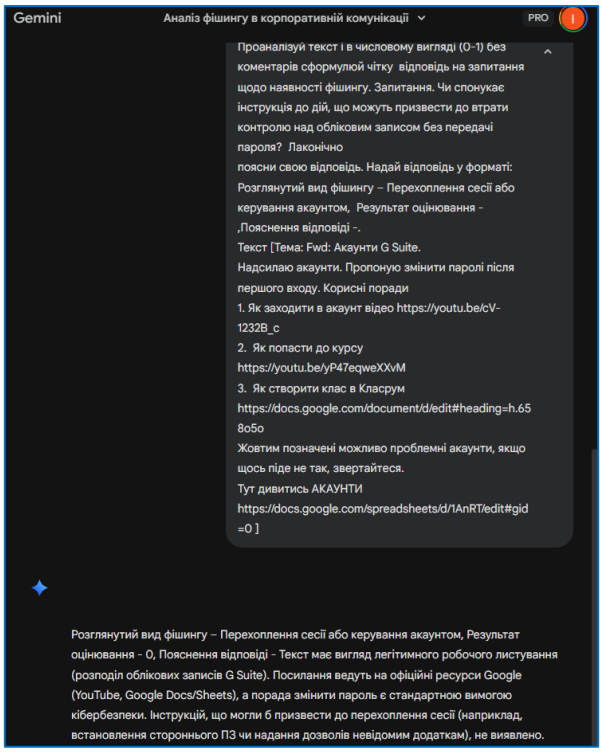

Рис. 2. Ілюстрація застосування Gemini для аналізу легітимного листа на предмет наявності фішингу виду b_5

У другій серії експериментів було проаналізовано 100 фішингових повідомлень, зібраних авторами шляхом моніторингу відкритих джерел інформації, аналізу звітів CERT-UA та агрегації знеособлених листів електронної пошти. Ілюстрацією застосування Grok та Claude для аналізу таких повідомлень являються рис. 3, 4.

ISSN 2663 – 4023

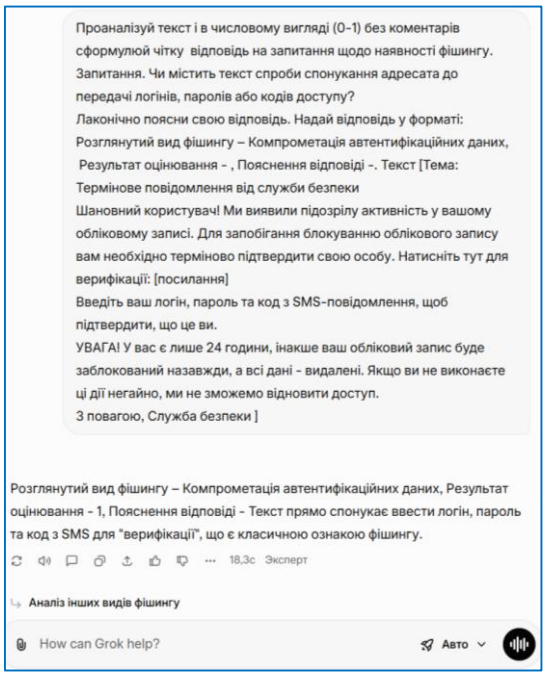

Рис. 3. Ілюстрація застосування Grok для аналізу повідомлення на предмет наявності фішингу виду b_1

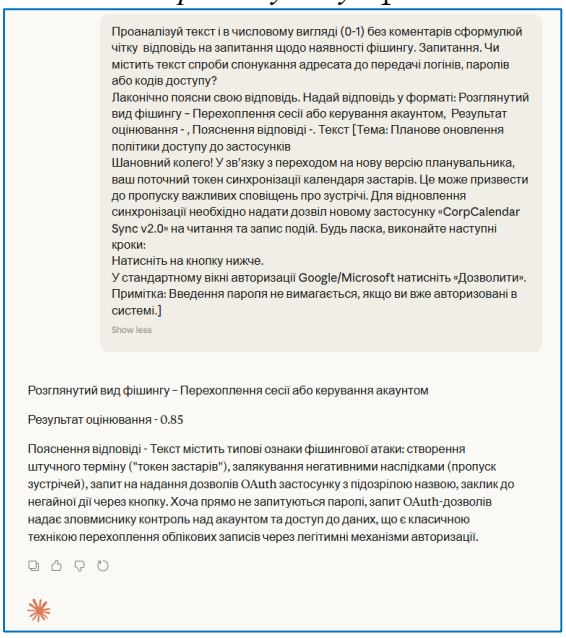

Рис. 4. Ілюстрація застосування Claude для аналізу повідомлення на предмет наявності фішингу виду b_5

Узагальнюючи результати експериментальних досліджень на вибірці із 100 легітимних та 100 фішингових повідомлень, встановлено, що середній показник точності класифікації при використанні запропонованого методу склав 97,5%. При цьому у групі легітимних повідомлень коректно класифіковано 98 із 100 зразків. У групі фішингових повідомлень успішно виявлено 97 із 100 зразків. Порівняння отриманих результатів із задекларованими у літературі показниками точності класифікації повідомлень, отриманими з використанням відомих нейромережевих засобів, зокрема моделей класу BERT (заявлена точність близько 98 %), свідчить про те, що запропонований метод

забезпечує досягнення точності на рівні найкращих відомих рішень. Водночас, на відміну від традиційних рішень до виявлення фішингових повідомлень із застосуванням нейромережевих засобів, які потребують складної адаптації нейромережевої моделі до специфічних умов застосування, формування значних обсягів розмічених навчальних даних та реалізації ресурсоємної процедури навчання, запропонований метод дозволяє забезпечити оперативність розробки системи виявлення фішингу за рахунок використання апробованого нейромережевого інструментального забезпечення, що істотно знижує складність означених процесів.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В результаті проведених досліджень розроблено метод виявлення фішингу у текстових повідомленнях вітчизняних каналів електронної комунікації, який передбачає реалізацію автоматичного аналізу повідомлень за рахунок подачі формалізованих запитів до діалогових інтелектуальних помічників на базі великих мовних моделей.

У межах запропонованого методу розроблено класифікацію фішингових повідомлень за вектором ознак виявлення, що дозволило виділити сім базових видів фішингу, зокрема атаки, спрямовані на компрометацію автентифікаційних даних, ексфільтрацію платіжної інформації, ініціювання несанкціонованих транзакцій, збір персональних даних, перехоплення контролю над обліковими записами, доставку шкідливого програмного забезпечення та психоемоційний маніпулятивний вплив. Для кожного з означених видів сформовано цільові предикати запитів до діалогових інтелектуальних помічників, що забезпечують детермінований та інтерпретований аналіз текстових повідомлень. Запропоновано формалізовану структуру запитів, яка включає ініціалізуючий та уточнюючий компоненти і дозволяє зменшити стохастичність генерації відповідей, мінімізувати ризик виникнення галюцинацій та забезпечити отримання результатів, придатних для подальшої автоматизованої обробки. Додатково реалізовано механізм підвищення достовірності результатів за рахунок паралельного використання декількох діалогових інтелектуальних помічників та зваженої лінійної згортки їх відповідей. Проведені експериментальні дослідження, засвідчили, що середній показник точності класифікації при використанні запропонованого методу становить 97,5%, що відповідає рівню найкращих відомих нейромережевих рішень. Водночас, реалізація запропонованого методу не потребує ресурсоємного навчання нейромереж, формування великих обсягів розмічених даних та розробки спеціалізованого апаратно-програмного забезпечення, що забезпечує оперативність створення ефективної системи виявлення фішингу у вітчизняних каналах електронної комунікації. Перспективи подальших досліджень доцільно пов’язати з розробкою процедури автоматизованого формування та адаптації запитів до нових видів фішингу, а також з розробкою підходів до визначення значень коефіцієнтів компетентності різних діалогових інтелектуальних помічників, що пройшли донавчання відповідно до підходу few-shot learning. Ще одним перспективним напрямком дослідження є інтеграція запропонованого методу у систему моніторингу вітчизняного інформаційного простору.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Aldakheel, E. A., Zakariah, M., Gashgari, G. A., Almarshad, F. A., & Alzahrani, A. I. A. (2023). A deep learning-based innovative technique for phishing detection in modern security with uniform resource locators. *Sensors, 23*(9), Article 4403. https://doi.org/10.3390/s23094403
2. Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing phishing email detection: A comparative study of deep learning models. *Sensors, 24*(7), Article 2077. https://doi.org/10.3390/s24072077
3. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., & Maka, S. R. (2025). Building an intelligent phishing email detection system using machine learning and feature engineering. *European Journal of Applied Science, Engineering and Technology, 3*(2), 41–54. https://doi.org/10.59324/ejaset.2025.3(2).04
4. Dychka, I. A., Tereikovskyi, I. A., Korovii, O. S., Tereikovska, L. O., & Romankevich, V. O. (2023). Evaluation of the effectiveness of tools for recognizing emotional sentiment of text fragments. *Scientific Notes of V. I. Vernadsky Taurida National University. Series: Technical Sciences, 34*(73), 3(1), 130–135. https://doi.org/10.32782/2663-5941/2023.3.1/20
5. European Union Agency for Cybersecurity. (2020). *ENISA threat landscape: Phishing*. Publications Office of the European Union. https://doi.org/10.2824/552242
6. Firman, Tukiyat, & Wiharjo, S. (2025). Phishing email classification approach using machine learning algorithms: A literature review. *Data: Journal of Information Systems and Management, 3*(3), 135–145. https://doi.org/10.61978/data.v3i3
7. Harasymchuk, O., Oliarnyk, Y., Nestor, A., & Nakonechyy, T. (2025). Psychological methods of fraud in cyberspace and ways to counter them. *Cybersecurity: Education, Science, Technique, 2*(30), 511–529. https://doi.org/10.28925/2663-4023.2025.30.990
8. Komosny, D. (2025). Phishing detection on webpages in European non-English languages based on machine learning. *Scientific Reports, 15*, Article 37472. https://doi.org/10.1038/s41598-025-21384-w
9. Korchenko, O., Tereikovskyi, I., Dychka, I., Romankevich, V., & Tereikovska, L. (2025). Detection of manipulative component in text messages of mass media in the context of protection of domestic cyberspace. *Cybersecurity: Education, Science, Technique, 1*(29), 27–40. https://doi.org/10.28925/2663-4023.2025.29.839
10. Korchenko, O., Tereikovskyi, I., Ziubina, R., Tereikovska, L., Korystin, O., Tereikovskyi, O., & Karpinskyi, V. (2025). Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images. *Applied Sciences, 15*, Article 2553. https://doi.org/10.3390/app15052553
11. Petliak, N., Bezkorovalnyi, Y., & Kupchyk, N. (2024). Analysis of modern methods of detection of phishing e-mails. *Herald of Khmelnytskyi National University. Technical Sciences, 341*(5), 510–515. https://doi.org/10.31891/2307-5732-2024-341-5-73
12. Rangapur, A., Kanakam, T., & Dhanvanthini, P. (2022). Phish-defence: Phishing detection using deep recurrent neural networks. *arXiv*. https://doi.org/10.48550/arXiv.2110.13424
13. State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine. (2024). *Annual report of the vulnerability detection and cyber incident and cyberattack response system: 2024 (Analytical report)*.
14. State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine. (2025). *System of vulnerability detection and response to cyber incidents and cyberattacks: First half of 2025 (Analytical report)*.
15. Tereikovskyi, I. A., Chernyshev, D. O., Korchenko, O. H., Tereikovska, L. O., & Tereikovskyi, O. I. (2022). Procedure for applying neural networks for raster image segmentation. *Cybersecurity: Education, Science, Technique, 2*(18), 25–38. https://doi.org/10.28925/2663-4023.2022.18.2438
16. Tereikovskyi, I., AlShboul, R., Mussiraliyeva, S., Tereikovska, L., Bagitova, K., Tereikovskyi, O., & Hu, Z. (2024). Method for constructing neural network means for recognizing scenes of political extremism in graphic materials of online social networks. *International Journal of Computer Network and Information Security, 16*(3), 52–69. https://doi.org/10.5815/ijcnis.2024.03.05
17. Tyshchenko, V. S. (2023). Analysis of training methods and neural network tools for fake news detection. *Cybersecurity: Education, Science, Technique, 4*(20), 21–34. https://doi.org/10.28925/2663-4023.2023.20.2034

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Ihor Tereikovskyi
Doctor of Technical Sciences, Professor, Professor of the Department of Systems Programming and Specialized Computer Systems
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID: 0000-0003-4621-9668
tereikovskyi@kpi.ua

Liudmyla Tereikovska
Doctor of Technical Sciences, Professor, Professor of the Department of Information Technology Design and Applied Mathematics
Kyiv National University of Construction and Architecture, Kyiv, Ukraine
ORCID: 0000-0002-8830-0790
tereikovska.lo@knuba.edu.ua

PHISHING DETECTION IN ELECTRONIC COMMUNICATION CHANNELS

Abstract. The article is devoted to increasing the effectiveness of information protection tools in domestic cyberspace by automated detection of phishing in text messages of electronic communication channels. It is shown that phishing remains one of the dominant vectors of cyberattacks, in particular in conditions of active use of social engineering methods and psycho-emotional influence, which significantly complicates its timely detection by traditional signature and statistical methods. It is established that most of the known neural network solutions are characterized by high resource intensity, require the formation of significant volumes of marked-up training data and are insufficiently adapted to the specifics of domestic content, which is characterized by limited marked-up corpora, morphological variability and sensitivity to contextual social engineering influences. To overcome these limitations, the article proposes a method for detecting phishing, which is based on the use of proven dialogical intelligent assistants based on large language models in the mode of formalized dialogical interaction. The method involves automated analysis of text messages by submitting standardized queries, formed taking into account typical features of phishing attacks. A classification of phishing messages by a vector of detection features is proposed. For each type, target query predicates are formed, which provide deterministic and interpretable analysis of the results. The developed method involves preprocessing of the text, formation and submission of queries to dialogic intelligent assistants, aggregation of responses, and assessment of the presence of phishing using the weighted linear convolution mechanism. Experimental studies have shown that the average classification accuracy when using the proposed method is 97.5%, which corresponds to the level of the best known solutions of a similar purpose. At the same time, the implementation of the proposed method does not require resource-intensive training of neural networks, the formation of large volumes of labeled data, and the creation of specialized hardware and software, which ensures the efficiency of creating an effective phishing detection system in domestic electronic communication channels.

Keywords: conversational intelligent assistant; detection; electronic communication; information protection; information security; neural network; phishing.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Aldakheel, E. A., Zakariah, M., Gashgari, G. A., Almarshad, F. A., & Alzahrani, A. I. A. (2023). A deep learning-based innovative technique for phishing detection in modern security with uniform resource locators. *Sensors, 23*(9), Article 4403. https://doi.org/10.3390/s23094403
2. Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing phishing email detection: A comparative study of deep learning models. *Sensors, 24*(7), Article 2077. https://doi.org/10.3390/s24072077
3. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., & Maka, S. R. (2025). Building an intelligent phishing email detection system using machine learning and feature engineering. *European Journal of Applied Science, Engineering and Technology, 3*(2), 41–54. https://doi.org/10.59324/ejaset.2025.3(2).04

4. Dychka, I. A., Tereikovskyi, I. A., Korovii, O. S., Tereikovska, L. O., & Romankevich, V. O. (2023). Evaluation of the effectiveness of tools for recognizing emotional sentiment of text fragments. *Scientific Notes of V. I. Vernadsky Taurida National University. Series: Technical Sciences, 34*(73), 3(1), 130–135. https://doi.org/10.32782/2663-5941/2023.3.1/20
5. European Union Agency for Cybersecurity. (2020). *ENISA threat landscape: Phishing*. Publications Office of the European Union. https://doi.org/10.2824/552242
6. Firman, Tukiyat, & Wiharjo, S. (2025). Phishing email classification approach using machine learning algorithms: A literature review. *Data: Journal of Information Systems and Management, 3*(3), 135–145. https://doi.org/10.61978/data.v3i3
7. Harasymchuk, O., Oliarnyk, Y., Nestor, A., & Nakonechyy, T. (2025). Psychological methods of fraud in cyberspace and ways to counter them. *Cybersecurity: Education, Science, Technique, 2*(30), 511–529. https://doi.org/10.28925/2663-4023.2025.30.990
8. Komosny, D. (2025). Phishing detection on webpages in European non-English languages based on machine learning. *Scientific Reports, 15*, Article 37472. https://doi.org/10.1038/s41598-025-21384-w
9. Korchenko, O., Tereikovskyi, I., Dychka, I., Romankevich, V., & Tereikovska, L. (2025). Detection of manipulative component in text messages of mass media in the context of protection of domestic cyberspace. *Cybersecurity: Education, Science, Technique, 1*(29), 27–40. https://doi.org/10.28925/2663-4023.2025.29.839
10. Korchenko, O., Tereikovskyi, I., Ziubina, R., Tereikovska, L., Korystin, O., Tereikovskyi, O., & Karpinskyi, V. (2025). Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images. *Applied Sciences, 15*, Article 2553. https://doi.org/10.3390/app15052553
11. Petliak, N., Bezkorovalnyi, Y., & Kupchyk, N. (2024). Analysis of modern methods of detection of phishing e-mails. *Herald of Khmelnytskyi National University. Technical Sciences, 341*(5), 510–515. https://doi.org/10.31891/2307-5732-2024-341-5-73
12. Rangapur, A., Kanakam, T., & Dhanvanthini, P. (2022). Phish-defence: Phishing detection using deep recurrent neural networks. *arXiv*. https://doi.org/10.48550/arXiv.2110.13424
13. State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine. (2024). *Annual report of the vulnerability detection and cyber incident and cyberattack response system: 2024 (Analytical report)*.
14. State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine. (2025). *System of vulnerability detection and response to cyber incidents and cyberattacks: First half of 2025 (Analytical report)*.
15. Tereikovskyi, I. A., Chernyshev, D. O., Korchenko, O. H., Tereikovska, L. O., & Tereikovskyi, O. I. (2022). Procedure for applying neural networks for raster image segmentation. *Cybersecurity: Education, Science, Technique, 2*(18), 25–38. https://doi.org/10.28925/2663-4023.2022.18.2438
16. Tereikovskyi, I., AlShboul, R., Mussiraliyeva, S., Tereikovska, L., Bagitova, K., Tereikovskyi, O., & Hu, Z. (2024). Method for constructing neural network means for recognizing scenes of political extremism in graphic materials of online social networks. *International Journal of Computer Network and Information Security, 16*(3), 52–69. https://doi.org/10.5815/ijcnis.2024.03.05
17. Tyshchenko, V. S. (2023). Analysis of training methods and neural network tools for fake news detection. *Cybersecurity: Education, Science, Technique, 4*(20), 21–34. https://doi.org/10.28925/2663-4023.2023.20.2034

Отримано редакцією журналу / Received: 05.01.26
Прорецензовано / Revised: 21.02.26
Схвалено до друку / Accepted: 26.03.26