УДК 004.056.53:004.8

ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ВИЯВЛЕННЯ ФІШИНГОВИХ ПОВІДОМЛЕНЬ

[1]Тетяна СУВОРОВА, [2]Жанна ДЕЙНЕКО
[1]магістрантка, кафедра МСТ, ХНУРЕ,
[2]доцент, к.т.н., зав. каф. МСТ, ХНУРЕ

Abstract. *The paper examines intelligent phishing detection systems that combine the analysis of message content, sender information and links. A three-component verification scheme with explainable risk assessment is proposed and illustrated with hypothetical examples. The limitations of individual indicators are shown, and requirements for data preparation and evaluation criteria for Ukrainian-language email datasets are outlined.*

Keywords: *phishing detection, email security, artificial intelligence, machine learning, text analysis, explainability.*

Фішингове повідомлення спонукає людину виконати небезпечну дію: передати пароль, відкрити шкідливе вкладення або перейти на підроблений сайт. Для цього зловмисник видає себе за банк, службу доставки чи працівника установи. Завдання інтелектуальної системи – виявити такі повідомлення до того, як користувач виконає запропоновану дію, і водночас не блокувати справжні листи лише через наявність слів про оплату, пароль або терміновість.

Мета роботи – обґрунтувати комплексний підхід до виявлення фішингових повідомлень, який поєднує аналіз змісту листа з його технічними ознаками. Основну увагу приділено електронній пошті, для якої, крім тексту, можна перевіряти адресу відправника, службові заголовки та посилання.

Штучний інтелект у кібербезпеці застосовують як для вдосконалення атак, так і для їх виявлення, а ефективна протидія загрозам потребує поєднання автоматизованих рішень з експертним аналізом [1]. Генеративні моделі дають змогу створювати грамотні й переконливі тексти, тому відсутність помилок сама по собі не свідчить про безпечність повідомлення. Перевірка має враховувати, яку дію пропонує відправник і чи узгоджується вона з іншими відомостями про лист.

Одним із підходів до виявлення фішингу є класифікація на основі методу опорних векторів, у якій ключову роль відіграє виділення ознак, що дають змогу розрізняти фішингові та легітимні об’єкти [2]. Для виявлення фішингових сайтів використовують також алгоритми Random Forest і XGBoost [3]. Інший підхід ґрунтується на вейвлет-когерентності, яка дає змогу оцінити взаємну динаміку характеристик фішингових атак і виявити приховані тенденції в їх розвитку [4]. Водночас результати, отримані для вебресурсів, не можна безпосередньо переносити на класифікацію листів, проте аналіз посилань доцільно використовувати як окремий складник перевірки повідомлення.

На основі зіставлення цих підходів пропонується розділити перевірку на три складники: зміст повідомлення, походження листа та характеристики посилань. Текстовий модуль виявляє ознаки психологічного тиску, запити конфіденційних даних і невідповідність між заявленою метою листа та запропонованою дією. Технічний модуль перевіряє відправника й результати поштової автентифікації (SPF, DKIM, DMARC). Модуль аналізу посилань зіставляє показану користувачеві

адресу з фактичною та перевіряє домен. Приклади ознак і межі їх застосування наведено в табл. 1.

Табл. 1.

Ознаки для комплексної перевірки фішингових повідомлень

Складник аналізу	Приклад ознаки	Обмеження
Зміст	Тиск на одержувача, запит пароля, незвична дія	Терміновість трапляється і в справжніх листах
Відправник	Невідповідність домену, помилки автентифікації	Справжня скринька може бути зламана
Посилання	Розбіжність видимої та фактичної адреси	Відомий сайт може бути скомпрометований

Поєднання ознак потребує обережної інтерпретації. Невдала перевірка поштової автентифікації може бути наслідком помилки налаштування, а успішна не виключає надсилання шахрайського листа зі скомпрометованої скриньки. Так само наявність HTTPS свідчить лише про захищене з'єднання, але не про добросовісність власника сайту. Тому жодну з цих ознак не слід розглядати як самодостатній доказ безпечності листа або його фішингового характеру.

Розглянемо два умовні повідомлення з проханням підтвердити платіж. У першому випадку користувач очікує рахунок від постачальника, адреса відправника відповідає відомій організації, а реквізити збігаються з раніше погодженими. У другому відправник погрожує блокуванням облікового запису, вимагає негайно ввести пароль і пропонує посилання на сторонній домен. Обидва листи можуть містити слово «терміново», однак сукупність ознак у них суттєво відрізняється. Цей приклад показує обмеженість фільтра, що спирається лише на окремі слова.

Водночас перший лист не можна автоматично визнати безпечним лише через знайоме ім'я відправника. Якщо рахунок надіслано зі справжньої, але зламаної скриньки, ознаки відправника не викличуть підозри. У такому разі система має виявити зміну реквізитів або незвичний запит і запропонувати підтвердити його через незалежний відомий канал зв'язку. Отже, корисним результатом роботи системи є конкретне пояснення ризику (наприклад, «реквізити відрізняються від попереднього рахунку»), а не лише загальна позначка «підозрілий лист».

Текстовий модуль можна реалізувати як класифікатор, навчений на розмічених прикладах листів, а велику мовну модель (LLM) використовувати для доповнення аналізу змісту та формування пояснень. Доцільність її застосування слід перевіряти на тій самій тестовій вибірці: складніша модель не обов'язково пропускатиме менше атак або даватиме менше хибних спрацювань.

Пропонується передбачити три варіанти рішення: пропуск повідомлення, попередження з поясненням ризику та переміщення до карантину з можливістю перегляду. Порогові значення для прийняття рішень доцільно визначати на окремій розміченій вибірці. У разі нестачі відомостей або суперечливих оцінок модулів лист передається на додатковий розгляд фахівцеві.

Окремою вимогою є підготовка даних. Навчальна вибірка має містити як фішингові, так і легітимні повідомлення, зокрема українською мовою. Необхідно

вилучати дублікати та розділяти навчальні й тестові дані так, щоб близькі варіанти одного листа не потрапляли до обох частин, інакше оцінка відображатиме лише впізнавання знайомих шаблонів. Для перевірки стійкості до нових атак варто використовувати повідомлення, отримані пізніше за навчальні дані, а результати для різних мов аналізувати окремо.

Якість системи доцільно оцінювати за сукупністю показників. Повнота (recall) показує, яку частку фішингових повідомлень виявлено. Точність (precision) характеризує частку справді фішингових листів серед позначених як такі. Частка хибних спрацювань (FPR) показує, скільки легітимних листів помилково віднесено до небезпечних. Додатково слід враховувати час перевірки та кількість листів, переданих фахівцеві. Загальна частка правильних відповідей (accuracy) на незбалансованих даних може приховувати пропуск рідкісних атак, тому сама по собі вона не є достатньою.

Передавання листів зовнішній мовній моделі може призвести до розкриття персональних або службових даних, тому слід мінімізувати обсяг переданої інформації та розглядати можливість локального оброблення. Крім того, система має бути стійкою до впровадження інструкцій (prompt injection): вказівки, що містяться в тексті листа, не повинні впливати на правила її роботи. Пояснення ризику має спиратися на виявлені ознаки, а не на довільні висновки моделі.

Таким чином, поєднання аналізу змісту, походження листа та характеристик посилань дає змогу компенсувати обмеження окремих ознак і зменшити кількість як пропущених атак, так і хибних спрацювань. Запропонована схема з трьома варіантами рішення та поясненням ризику робить роботу системи прозорішою для користувача. Подальші дослідження передбачають формування україномовного набору даних та експериментальне порівняння класифікаторів і мовних моделей за наведеними показниками.

Список літератури

1. Яковлев, М., & Любчак, В. (2025). Можливості штучного інтелекту у виявленні та запобіганні фішингу й кібератакам. *Кібербезпека: освіта, наука, техніка, 1*(29), 298-309. https://doi.org/10.28925/2663-4023.2025.29.840.

2. Гайдур, Г. І., Гахов, С. О., Марченко, В. В., & Гайдур, К. В. (2024). Концептуальна модель виявлення фішингових атак на основі використання методів опорних векторів. *Сучасний захист інформації, 2*(58), 24-33. https://doi.org/10.31673/2409-7292.2024.020003.

3. Трохименко, Ю. О. (2026). Виявлення фішингових сайтів за допомогою машинного навчання. *Радіоелектроніка та молодь у XXI столітті*. Т. 5. (с. 214-215). https://openarchive.nure.ua/handle/document/35476.

4. Abu Aladas, F. E., Abu-Jassar, A., Ismail, S. M., Deineko, Z., & Lyashenko, V. (2026). Wavelet coherence data for preventing phishing attacks in cyber security. *In Proceedings of the 5th International Computer Sciences and Informatics Conference (ICSIC 2024) (AIP Conference Proceedings), 3406(1), 020047*. AIP Publishing. https://doi.org/10.1063/5.0318447.