УДК 004.8

Чумак М.В.[1], Мазурець О.В.[2]

[1] Ліцей №4 ім. Павла Жука Хмельницької міської ради
[2] Хмельницький національний університет

ПОЯСНЕННЕ ВИЯВЛЕННЯ ФІШИНГОВИХ ПОВІДОМЛЕНЬ НЕЙРОМЕРЕЖЕВИМИ ЗАСОБАМИ В СИСТЕМАХ КІБЕРЗАХИСТУ

Робота присвячена поясненому виявленню фішингових повідомлень, у якому нейромережева модель не лише класифікує текст як «фішинг / нефішинг», а й ідентифікує домінантний тип маніпуляції (терміновість дії, апеляція до авторитету, емоційний або раціональний тиск). Запропонована інформаційна технологія базується на трансформерній архітектурі глибокого навчання, що аналізує текстові представлення повідомлень і виконує бінарну та багатокласову класифікацію. Поясненність результатів досягається завдяки використанню механізмів уваги та градієнтно орієнтованих атрибуцій, які виділяють фрагменти тексту, що найбільше вплинули на прийняте рішення. Інтеграція такого модуля в системи кіберзахисту підвищує якість виявлення соціоінженерних атак і рівень довіри до нейромережевих засобів захисту.

The paper addresses explainable detection of phishing messages, where a neural model not only classifies text as «phishing / non-phishing» but also identifies the dominant type of manipulation (urgency, authority appeal, emotional or rational pressure). The proposed information technology relies on a transformer-based deep learning architecture that operates on textual representations and performs both binary and multiclass classification. Explainability is achieved through attention mechanisms and gradient-based attribution methods that highlight text fragments with the strongest impact on the decision. Integration of the module into cybersecurity systems improves the detection of social engineering attacks and increases trust in neural network–based defence tools.

Проблематика виявлення фішингових повідомлень традиційно розглядалася крізь призму технічних характеристик атаки: підозрілих URL, доменів, вкладень, сигнатур і структурних аномалій електронних листів [1, 2]. Проте сучасні фішингові кампанії дедалі частіше ґрунтуються не на шкідливому коді, а на стратегіях соціальної інженерії, де ключову роль відіграють маніпулятивні мовні патерни [3]. Повідомлення будуються таким чином, щоб викликати у користувача відчуття терміновості, штучно сконструйованої довіри або неминучої втрати, а також спонукати до негайної дії без критичного аналізу змісту [4]. За таких умов рішення, що орієнтуються лише на формальні параметри, виявляються недостатніми, оскільки не охоплюють приховану комунікативну природу загрози.

Зростання кількості кібератак та поширення фішингових повідомлень становлять серйозну загрозу безпеці інформаційних систем і користувачів. Традиційні методи виявлення фішингу, що базуються на жорстко визначених правилах або сигнатурах, часто не встигають за швидко змінюваними техніками шахраїв і не здатні ефективно працювати з великими потоками електронної кореспонденції та повідомлень. У цьому контексті застосування нейромережевих методів обробки природної мови [4] та машинного навчання [5] відкриває нові можливості для автоматизованого і масштабованого виявлення фішингових атак.

Нейромережеві моделі дозволяють аналізувати лінгвістичні [6], синтаксичні [7] та семантичні [8] патерни повідомлень, виявляючи ознаки маніпулятивного або шахрайського контенту [9], що не завжди очевидні при поверхневому аналізі. Використання трансформерних архітектур [10] і моделей глибокого навчання [11] забезпечує контекстне розуміння тексту [12], дозволяючи системам не лише класифікувати повідомлення як фішингові або безпечні, а й ідентифікувати ключові елементи ризику, такі як підроблені домени, соціальна інженерія або загрози фінансового характеру [13].

Особливо важливим є інтегрування пояснюваності у нейромережеві моделі, що дозволяє фахівцям з кібербезпеки розуміти, на яких ознаках базується класифікація, підвищуючи довіру до автоматизованих систем і полегшуючи процес реагування на інциденти [14, 15].

Метою роботи є розроблення поясненної нейромережевої технології виявлення фішингових повідомлень за маніпулятивними патернами для інтеграції в системи кіберзахисту. У межах такої технології автоматизоване рішення повинно не лише визначати, чи є повідомлення фішинговим, а й ідентифікувати домінантний вид маніпуляції, за допомогою якого здійснюється психологічний вплив. Йдеться насамперед про стратегії, що апелюють до терміновості дії, авторитету відправника або переконання через емоційні та раціональні аргументи; у реальних сценаріях вони можуть поєднуватися у межах одного тексту, формуючи складну структуру мовного тиску.

Для досягнення поставленої мети пропонується інформаційна технологія, у межах якої текстові повідомлення проходять поетапну обробку: нормалізацію і токенізацію, перетворення в векторні представлення, подальший аналіз трансформерною архітектурою глибокого навчання, що виконує одночасно бінарну класифікацію «фішинг / нефішинг» і багатокласову або багатоміткову класифікацію типів маніпулятивного впливу.

Ключовою відмінністю запропонованого підходу є акцент на поясненності результатів. Нейромережева модель доповнюється модулем інтерпретації, який на основі механізмів уваги, градієнтно-орієнтованих атрибуцій або інтегрованих градієнтів будує візуалізацію важливості окремих фрагментів повідомлення для

прийнятого рішення. У результаті користувач або аналітик системи кіберзахисту отримує не лише мітку «фішингове повідомлення з домінуванням стратегії терміновості» чи «апеляції до авторитету», а й текст із виділенням тих словосполучень і конструкцій, які зумовили таку класифікацію. Це забезпечує можливість верифікації та критичного аналізу роботи моделі, а також полегшує інтеграцію технології у регламентовані процеси реагування на інциденти.

Запропонована технологія спирається на попереднє маркування корпусу фішингових і нефішингових повідомлень з урахуванням типів маніпулятивного впливу, що створює підґрунтя для навчання і об'єктивної оцінки якості моделі. Порівняння з підходами, орієнтованими лише на технічні індикатори або плоскі текстові ознаки, дає змогу продемонструвати приріст ефективності для сценаріїв, де зловмисники свідомо мінімізують формальні маркери загрози, натомість ускладнюючи мовні стратегії соціальної інженерії.

Практична значущість роботи полягає у можливості інтеграції поясненної нейромережевої технології в існуючі системи кіберзахисту як окремого інтелектуального модуля для аналізу вхідних повідомлень. Такий модуль дозволяє знизити ймовірність успішного фішингу за рахунок раннього виявлення текстових загроз, підвищити обґрунтованість рішень щодо блокування або маркування повідомлень, а також сформувати базу для навчальних і просвітницьких заходів, які демонструють користувачам типові маніпулятивні патерни. Поясненність результатів у поєднанні з високими показниками якості класифікації є ключовою передумовою для зростання довіри до нейромережевих засобів кіберзахисту та їхнього системного впровадження в організаціях різного масштабу.

Перелік посилань

1. Elberri, M. A., Tokeşer, Ü., Rahebi, J., & Lopez-Guede, J. M. (2024). A cyber defense system against phishing attacks with deep learning game theory and LSTM-CNN with African vulture optimization algorithm (AVOA). International Journal of Information Security, 23(4), 2583-2606.
2. Abed, A. K. (2025). Utilizing artificial intelligence in cybersecurity: A study of neural networks and support vector machines. Babylonian Journal of Networking, 2025, 14-24.
3. Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. (2024). The application of a novel neural network in the detection of phishing websites. Journal of Ambient Intelligence and Humanized Computing, 15(3), 1865-1879.
4. Murava V., Zalutska O., Didur V., Mazurets O. Software architecture of information system for exchanging LLM thematic prompts. Global Trends in the Development of Information Technology and Science. Proceedings IV International Scientific and Practical Conference. June 25-27, 2025. Stockholm, Sweden. Pp. 121-127.
5. Юрченко Д.Ю., Овчарук О.М., Мазурець О.В., Шевчук П.О. Метод використання нейромережі гібридної архітектури для визначення емоційної тональності текстових повідомлень. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 2, 2025. с. 136-141.

6. Віт Р.В., Мазурець О.В. Метод виявлення психологічного цифрового перевантаження за аналізом текстових даних нейромережевими моделями глибокого навчання. Науковий журнал «Вісник Херсонського національного технічного університету». 2025. №2 (93). Т. 2. С. 107-114.
7. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. June 5-7, 2024. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.
8. Овчарук О.М., Мазурець О.В. Нейромережевий метод діагностування психологічних розладів за аналізом повідомлень на основі роздільного підходу до класифікації. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 1, 2025. с. 210-216.
9. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». May 31, 2024. Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.
10. O. Mazurets, R. Vit, M. Molchanova, I. Tymofiiev, O. Sobko, Context-enriched approach to students depression monitoring in education using BERT-GPT hybrid model, CEUR Workshop Proceedings 4096 (2025) 167-176.
11. Molchanova M., Didur V., Sobko O., Mazurets O. Detection of Web Propaganda Patterns by Transformer Neural Networks: Improving Efficiency via Dataset Balancing, CEUR Workshop Proceedings, 2025, vol. 3988, pp. 112-126.
12. E. A. Manziuk, O. V. Sobko, I. O. Podhorniuk, M. O. Molchanova, O. V. Mazurets, Multifactorial analysis of mobbing behavioral signs in educational environments posts by NLP means, Journal of Physics Conference Series 3105(1) (2025) 012025.
13. Мазурець О.В., Козенко О.В., Собко О.В. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю. Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання». Київ – 2023. с. 106-109.
14. Кок І.А., Мазурець О.В., Молчанова М.О. Поясненний підхід на основі трансферного навчання Vision Transformers до виявлення куріння у публічних просторах. Матеріали XIII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2025». 24-26.09.2025. Одеса. 2025. С.147-149.
15. Юрченко Д.Ю., Мазурець О.В., Залуцька О.О, Безпрозвана Ю.Г. Підхід до візуального пояснення результатів нейромережевого аналізу емоційної тональності повідомлень у соціальних мережах. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 565-571.