УДК 004.056.5:004.8 **DOI:10.15330/itee.2025.2.02**

І. М. ЛАЗАРОВИЧ, канд. техн. наук, А. Д. КВАСНЮК

ПОРІВНЯЛЬНИЙ АНАЛІЗ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ ТА ГІБРИДНИХ АРХІТЕКТУР У ЗАДАЧАХ ВИЯВЛЕННЯ ФІШИНГОВОГО КОНТЕНТУ

Вступ

Зі стрімким розвитком інформаційно-комунікаційних технологій та широким впровадженням цифрових сервісів фішингові атаки залишаються однією з найбільш поширених і небезпечних загроз у сфері кібербезпеки. Фішинг ґрунтується на методах соціальної інженерії та передбачає створення повідомлень або веб-ресурсів, що імітують легітимні сервіси з метою отримання конфіденційних даних користувачів, зокрема облікових даних, платіжної інформації та персональних відомостей [1]. Постійна еволюція фішингових технік ускладнює їх своєчасне виявлення та знижує ефективність традиційних захисних механізмів.

Класичні підходи до виявлення фішингового контенту, які базуються на сигнатурному аналізі, списках заборонених доменів або ручному формуванні правил, демонструють обмежену ефективність в умовах динамічного розвитку атак. Такі методи не здатні адекватно реагувати на нові варіанти фішингових повідомлень, що відрізняються стилістикою, структурою та мовними особливостями. У зв’язку з цим актуальним напрямом досліджень є застосування методів машинного та глибокого навчання для автоматизованого аналізу текстового та HTML-контенту.

Сучасні методи обробки природної мови (Natural Language Processing, NLP) дозволяють ефективно аналізувати семантичні та контекстуальні особливості текстів. Особливу увагу привертають трансформерні моделі, зокрема BERT, які завдяки механізму самоуваги здатні враховувати глобальні залежності у тексті. Паралельно активно досліджуються гібридні архітектури [2], що поєднують згорткові нейронні мережі (CNN), рекурентні мережі з довготривалою пам’яттю (LSTM) та механізми уваги. Такі моделі дозволяють одночасно аналізувати локальні текстові патерни та довготривалий контекст повідомлень.

Аналіз останніх досліджень та публікацій

У ранніх дослідженнях проблеми виявлення фішингових атак застосовувалися класичні алгоритми машинного навчання, такі як наївний баєсівський класифікатор [3], метод опорних векторів [3] та логістична регресія [3]. Дані підходи вимагали попереднього ручного відбору ознак, зокрема частот слів, n-грам та структурних характеристик веб-сторінок. Хоча такі методи демонстрували прийнятні результати на обмежених наборах даних, вони не забезпечували достатнього рівня узагальнення та погано працювали з новими типами атак.

Подальший розвиток отримали нейромережеві підходи, зокрема згорткові та рекурентні нейронні мережі [4, 5]. CNN зарекомендували себе як ефективний інструмент для виявлення локальних шаблонів у текстах, таких як характерні фрази або словосполучення, притаманні фішинговим повідомленням. LSTM [6], у свою чергу, дозволили моделювати довготривалі залежності між словами та враховувати загальний зміст повідомлень.

Значний прорив у галузі NLP був досягнутий із появою трансформерних моделей. Модель BERT [7] фактично стала загальноприйнятим стандартом для багатьох задач класифікації тексту завдяки двонапрямленому моделюванню контексту та використанню механізму самоуваги. Разом із тим, висока обчислювальна складність трансформерів та

вимоги до великих обсягів навчальних даних стимулювали подальші дослідження гібридних архітектур, які поєднують переваги різних типів нейронних мереж.

Формулювання цілей статті

Основною метою дослідження є експериментальне порівняння ефективності трансформерної моделі BERT та гібридних нейромережевих архітектур CNN та LSTM і CNN, LSTM та Attention у задачі виявлення фішингового контенту. Для досягнення цієї мети було проведено серію експериментів із використанням K-Fold крос-валідації та стандартних метрик оцінювання якості класифікації.

Методологія та основний матеріал досліджень

У дослідженні використано загальнодоступний набір даних [8], що містить HTML-веб-сторінки, класифіковані як фішингові або легітимні. Початково дані були представлені у форматі SQL-архіву та згодом експортовані у формат CSV для подальшої обробки у середовищі Python.

Для аналізу використовувалися дві основні ознаки: htmlContent, що містить HTML-код сторінок, та isPhish, яка визначає клас веб-контенту. Після попередньої обробки набір даних складав 10 373 записи, серед яких переважали фішингові сторінки, що вказує на наявність дисбалансу класів.

Попередня обробка тексту є критично важливою для задач обробки природної мови [9], особливо у контексті виявлення фішингового контенту. У межах роботи застосовувалися очищення HTML-коду, токенізація, лематизація, видалення стоп-слів та формування біграм. Такий підхід дозволив зменшити шум у даних і зберегти значущі лексичні та контекстні ознаки.

Підготовлені текстові дані були використані для формування векторних представлень [10], які подавалися на вхід досліджуваних моделей, забезпечуючи коректні умови для подальшого порівняльного аналізу трансформерних та гібридних архітектур у задачі виявлення фішингового контенту.

У межах дослідження проведено порівняльний аналіз трансформерної моделі BERT та двох гібридних нейромережевих архітектур на основі згорткових і рекурентних шарів. Обрані моделі відрізняються способом вилучення та узагальнення текстових ознак, що дозволяє оцінити їх ефективність у задачі виявлення фішингового контенту.

Трансформерна модель BERT реалізована на основі полегшеної архітектури DistilBERT, яка забезпечує баланс між якістю класифікації та обчислювальною складністю. Модель приймає на вхід токенізований текст та формує контекстуальні векторні представлення за допомогою механізму self-attention. Для задачі класифікації використовується проміжний повнозв’язний шар із нелінійною активацією ReLU та шар Dropout з коефіцієнтом 0.3, після чого вихідний шар із функцією Softmax формує ймовірнісну оцінку належності контенту до відповідного класу.

Базова гібридна архітектура CNN та LSTM поєднує локальний та послідовний аналіз тексту. На першому етапі токени перетворюються у векторні представлення за допомогою попередньо навчених ембеддінгів GloVe. Згортковий шар Conv1D з 128 фільтрами забезпечує вилучення локальних патернів, характерних для фішингових повідомлень, після чого застосовується шар MaxPooling для зменшення розмірності ознак. Отримані послідовності передаються до рекурентного шару LSTM з 128 юнітами та L2-регуляризацією, який моделює довготривалі контекстуальні залежності. Для зменшення ризику перенавчання використовується Dropout, а фінальна класифікація здійснюється за допомогою шару Softmax.

Розширена гібридна архітектура CNN, LSTM та Attention є подальшим розвитком базової моделі та доповнюється механізмами уваги. Після етапу згорткової обробки застосовується self-attention, який дозволяє моделі зосереджуватися на найбільш інформативних фрагментах тексту. Додатковий контекстуальний шар уваги після LSTM

підсилює релевантні ознаки перед фінальною класифікацією. Такий підхід забезпечує більш гнучке узагальнення інформації та підвищує здатність моделі виявляти приховані фішингові патерни.

Таким чином, у дослідженні охоплено три різні підходи до обробки текстового веб-контенту: трансформерний, гібридний згортково-рекурентний та гібридний із механізмом уваги, що створює основу для об'єктивного порівняльного аналізу їх ефективності.

Для забезпечення об'єктивності результатів застосовано K-Fold крос-валідацію [11], що дозволяє оцінити здатність моделей до узагальнення та зменшити вплив випадкового поділу даних.

У цьому дослідженні проведено порівняльний аналіз основних моделй з використанням кількох метрик оцінки для визначення їх ефективності в задачі класифікації веб-контенту. Зокрема, моделі порівнювалися за такими критеріями:

Precision та Recall. Ці метрики використовувалися для оцінки ефективності кожного методу векторизації у правильному виявленні шкідливого контенту. Precision та Recall визначаються як:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}. \tag{1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}. \tag{2}$$

F1-міра. F1-міра обчислюється за формулою:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3}$$

Точність (Accuracy). Загальна точність класифікації визначається як:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Samples}. \tag{4}$$

Окрім зазначених метрик, для більш повної оцінки та порівняння загальної ефективності класифікації моделей використовувалася метрика AUC. Вона відображає здатність моделі розрізняти релевантний (цільовий) та нерелевантний (нецільовий) контент при різних порогах класифікації.

AUC обчислюється як площа під ROC-кривою (Receiver Operating Characteristic), що відображає залежність рівня істинних позитивних від рівня хибнопозитивних спрацьовувань:

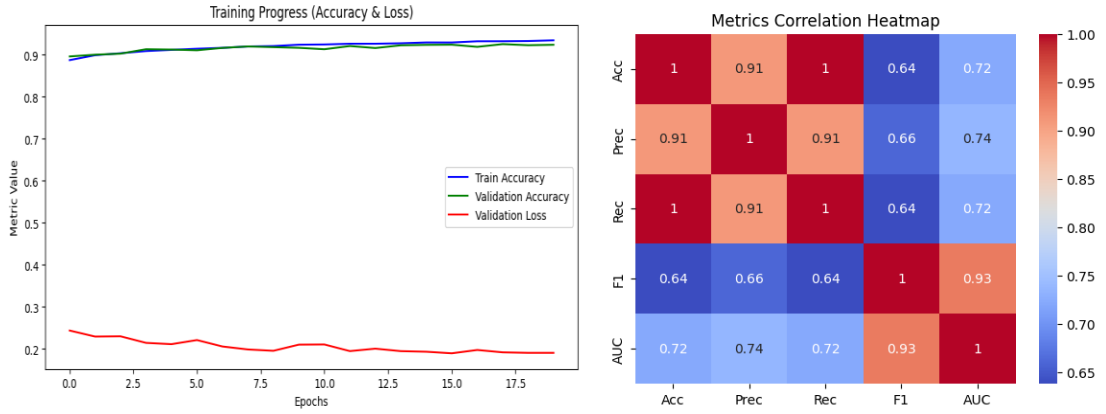
$$AUC = \int_0^1 TRP(t)d(FPR(t)). \tag{5}$$

Результати експериментів для трансформерної моделі BERT наведені на рис. 1 а) та рис. 1 б). Аналіз кривих навчання див. рис. 1 а) демонструє швидку конвергенцію моделі та досягнення високих значень як тренувальної, так і валідаційної точності на рівні близько 0.9. Це свідчить про здатність BERT ефективно відокремлювати фішингові повідомлення від легітимних. Водночас надто висока тренувальна точність може вказувати на потенційний ризик перенавчання, що підтверджується стабілізацією валідаційної втрати без подальшого покращення.

Теплокарта кореляції метрик див. рис. 1 б) показує сильний взаємозв'язок між Accuracy, Precision та Recall, що свідчить про узгоджену поведінку моделі за основними показниками якості. Водночас слабша кореляція між Accuracy та F1-score може вказувати на певну нерівномірність у роботі з різними класами.

Результати для гібридної архітектури CNN та LSTM наведені на рис. 2 а) та рис. 2 б). Криві навчання див. рис. 2 а) демонструють майже максимальну тренувальну точність, що

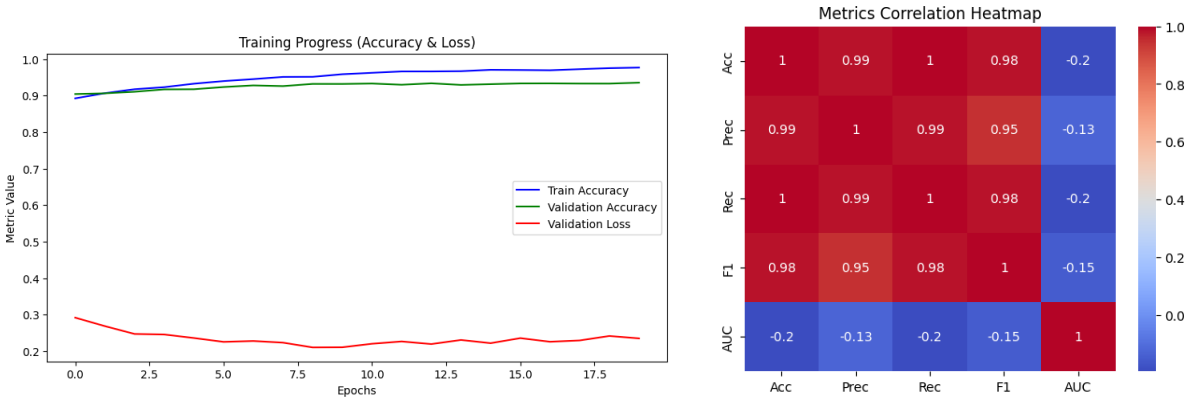
поєднується зі стабілізацією валідаційної точності на рівні близько 93%. Така різниця між тренувальними та валідаційними показниками свідчить про наявність ознак перенавчання. Валідаційна втрата залишається відносно стабільною, проте її значення вказує на можливість подальшої оптимізації архітектури.



а) б)

Рис. 1. а) Процес навчання: зміна точності та втрат для BERT. б) Аналіз кореляції між Accuracy, Precision, Recall, F1 та AUC після проведення проведення K-Fold крос-валідації для BERT

Аналіз кореляції метрик див. рис. 2 б) виявляє слабку або навіть негативну кореляцію AUC з іншими показниками, що означає обмежену здатність моделі ефективно розрізняти класи при зміні порогу класифікації. Це є важливим обмеженням у контексті незбалансованих наборів даних, характерних для задач виявлення фішингового контенту.



а) б)

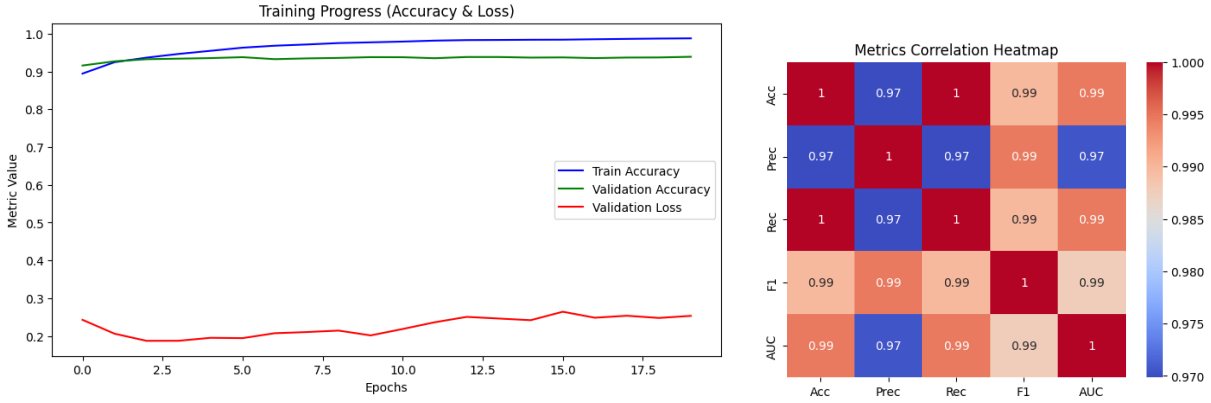
Рис. 2. а) Процес навчання після проведення K-Fold крос-валідації для CNN та LSTM. б) Аналіз кореляції між Accuracy, Precision, Recall, F1 та AUC після проведення K-Fold крос-валідації для CNN та LSTM

Результати для моделі CNN, LSTM та Attention наведені на рис. 3 а) та рис. 3 б). Аналіз процесу навчання див. рис. 3 а) показує зростання тренувальної точності без суттєвого покращення валідаційної, що знову вказує на ризик перенавчання. Водночас використання attention-механізму дозволяє моделі зосереджуватися на найбільш інформативних фрагментах тексту, що позитивно впливає на основні метрики якості.

Кореляційний аналіз див. рис. 3 б) демонструє більш узгоджену поведінку метрик порівняно з базовою моделлю CNN та LSTM.

Результати, наведені в таблиці 1, відображають ефективність трансформерної та гібридних нейромережевих моделей у задачі виявлення фішингового контенту. Трансформерна модель BERT демонструє стабільні результати з показником Accuracy = 0.95, а також близькими значеннями Precision (0.91), Recall (0.87) та F1-score (0.88). Це свідчить

про здатність моделі ефективно враховувати контекст текстових повідомлень, однак її продуктивність поступається гібридним підходам у межах проведеного експерименту.



а) б)

Рис. 3. а) Процес навчання після проведення K-Fold крос-валідації для CNN, LSTM та Attention-механізм. б) Аналіз кореляції між Accuracy, Precision, Recall, F1 та AUC після проведення K-Fold крос-валідації для CNN, LSTM та Attention-механізм

Гібридна архітектура CNN та LSTM забезпечує покращення всіх основних метрик якості класифікації, досягаючи Accuracy = 0.97 та F1-score = 0.91. Отримані результати підтверджують ефективність поєднання локального аналізу тексту за допомогою згорткових шарів із моделюванням послідовного контексту.

Найвищі показники продемонструвала модель CNN, LSTM та Attention, яка досягла максимального значення Accuracy (0.98), а також найвищих значень Precision (0.94) і Recall (0.92). Це вказує на позитивний вплив механізму уваги, який дозволяє моделі зосереджуватися на найбільш інформативних фрагментах тексту та підвищує ефективність виявлення фішингових повідомлень. Незначна різниця у значенні F1-score (0.90) порівняно з моделлю CNN + LSTM може бути пов’язана з балансом між точністю та повнотою класифікації.

Таблиця 1

Метрики навчання моделей

Метод	Accuracy	Precision	Recall	F1-score
BERT	0.951	0.912	0.873	0.882
CNN та LSTM	0.974	0.931	0.901	0.915
CNN, LSTM та ATTENTION	0.980	0.942	0.920	0.931

Загалом результати таблиці 1 свідчать, що гібридні моделі перевершують трансформерний підхід BERT за більшістю показників, а використання Attention-механізму забезпечує додаткове підвищення якості класифікації у задачі виявлення фішингового контенту.

Висновки

Отримані результати дослідження свідчать про високу ефективність гібридних нейромережевих архітектур у задачі виявлення фішингового контенту порівняно з трансформерною моделлю BERT. Проведений аналіз показав, що BERT демонструє стабільні значення Accuracy, Precision, Recall та F1-score, що свідчить про здатність моделі враховувати контекст тексту, проте її продуктивність поступається гібридним підходам, особливо у врахуванні локальних патернів і довготривалих залежностей.

Модель CNN та LSTM забезпечує покращення всіх основних метрик, підтверджуючи ефективність поєднання згорткових шарів для виділення локальних ознак тексту та

рекурентних шарів для моделювання послідовного контексту. Найвищі показники продемонструвала модель CNN, LSTM та Attention, де додатковий механізм уваги дозволяє моделі зосереджуватися на найбільш інформативних фрагментах тексту, підвищуючи точність та повноту виявлення фішингових повідомлень. Водночас експериментальні дані показують ознаки перенавчання та коливання валідаційної втрати, що вказує на потребу додаткової регуляризації та оптимізації гіперпараметрів.

Загалом результати дослідження підтверджують, що гібридні архітектури перевершують трансформерний підхід BERT за більшістю показників якості, а поєднання CNN, LSTM та Attention є перспективним напрямом для побудови практичних систем кібербезпеки. Подальші дослідження доцільно спрямувати на інтеграцію трансформерних моделей із гібридними підходами, застосування методів балансування даних та адаптацію моделей до динамічних фішингових атак.

Список літератури:

1. M. Vijayalakshmi, S. Mercy Shalinie, M. H. Yang, and R. M. U., “Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions,” *IET Netw.*, vol. 9, no. 5, pp. 235–246, Sep. 2020. doi: https://doi.org/10.1049/iet-net.2020.0078.
2. M. A. Adebowale, K. T. Lwin, and M. A. Hossain, “Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection,” in *2019 13th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA)*, Island of Ulkulhas, Maldives, Aug. 26–28, 2019. IEEE, 2019. doi: https://doi.org/10.1109/skima47702.2019.8982427.
3. S. Priya, D. Gutema, and S. Singh, “A Comprehensive Survey of Recent Phishing Attacks Detection Techniques,” in *2024 5th Int. Conf. Innovative Trends Inf. Technol. (ICITIIT)*, Kottayam, India, Mar. 15–16, 2024. IEEE, 2024. doi: https://doi.org/10.1109/icitiit61487.2024.10580446.
4. S. Atawneh and H. Aljehani, “Phishing Email Detection Model Using Deep Learning,” *Electronics*, vol. 12, no. 20, p. 4261, Oct. 2023. doi: https://doi.org/10.3390/electronics12204261.
5. U. Daniel, E. Bartholomew, and F. Egbono, “Phishing URL Attack Detection using Logistic Regression and Convolutional Neural Network,” *Int. J. Comput. Appl.*, vol. 187, no. 1, pp. 8–14, May 2025. doi: https://doi.org/10.5120/ijca2025924611.
6. B. Vishnupriya, B. Vikas, and M. D. Choudhry, “Hybrid Deep Learning Framework for ECG-Based Arrhythmia Detection Using CNN, Bi-LSTM, and Transformer,” in *2025 IEEE 4th World Conf. Appl. Intell. Comput. (AIC)*, GB Nagar, Gwalior, India, Jul. 26–27, 2025. IEEE, 2025, pp. 628–632. doi: https://doi.org/10.1109/aic66080.2025.11212074.
7. A. A. Tawil, L. Almazaydeh, D. Qawasmeh, B. Qawasmeh, M. Alshinwan, and K. Elleithy, “Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT,” *Comput., Mater. & Continua*, pp. 1–10, 2024. doi: https://doi.org/10.32604/cmc.2024.057279.
8. M. E. Maurer, "Phishload," URL: https://www.medien.ifi.lmu.de/team/max.maurer/files/phishload .
9. S. Luo, Y. Gu, X. Yao, and W. Fan, “Research on Text Sentiment Analysis Based on Neural Network and Ensemble Learning,” *Revue d'Intell. Artificielle*, vol. 35, no. 1, pp. 63–70, Feb. 2021. doi: https://doi.org/10.18280/ria.350107.
10. K. Zhou, Y. Zhou, W. X. Zhao, and J.-R. Wen, “Learning to Perturb for Contrastive Learning of Unsupervised Sentence Representations,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1–10, 2023. doi: https://doi.org/10.1109/taslp.2023.3304485.
11. M. Abdar *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021. doi: https://doi.org/10.1016/j.inffus.2021.05.008.

Надійшла до редколегії 30.05.2025

Відомості про авторів:

Лазарович Ігор Миколайович – кандидат технічних наук, доцент, доцент кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника / Vasyl Stefanyk Precarpathian National University, Україна; email: ihor.lazarovych@pnu.edu.ua; ORCID: https://orcid.org/0000-0001-5219-4714

Кваснюк Андрій Дмитрович – магістрант кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника / Vasyl Stefanyk Precarpathian National University, Україна; email: andrii.kvasniuk.20@pnu.edu.ua; ORCID: https://orcid.org/0009-0001-0653-3217