УДК 004.89:004.912:001.18
DOI https://doi.org/10.32782/2521-6643-2026-2-72.19

Мойсеєнко О. В., кандидат технічних наук, доцент,
доцент кафедри комп'ютерних систем і мереж
Івано-Франківського національного технічного університету нафти і газу
ORCID: 0000-0002-7995-2949

МЕТОД ВИЯВЛЕННЯ ПЕРСПЕКТИВНИХ НАУКОВИХ ТЕМ НА ОСНОВІ ГІБРИДНИХ TRANSFORMER-АРХІТЕКТУР ЗА ДАНИМИ НАУКОМЕТРИЧНИХ БАЗ

У статті вирішено актуальне науково-прикладне завдання автоматизації виявлення перспективних наукових напрямів у великих масивах наукометричних даних. Стрімке зростання кількості публікацій створює інформаційне перевантаження, що ускладнює роботу дослідників із пошуку латентних трендів. Запропоновано універсальний метод, що базується на гібридній архітектурі глибокого навчання, яка поєднує переваги згорткових нейронних мереж (CNN) та Transformer – архітектур. Особливістю методу є використання CNN-шару як екстрактора локальних термінологічних паттернів, що дозволяє виявляти стійкі ключові фрази в анотаціях статей. Наступний каскад Transformer-шарів із механізмом MultiHeadAttention забезпечує моделювання глобальних семантичних зв'язків, що критично важливо для аналізу вузькоспеціалізованих текстів. Для подолання проблеми дисбалансу класів у наукометричних вибірках застосовано функцію втрат Focal Loss, що дозволило моделі ефективніше фокусуватися на рідкісних, але потенційно «проривних» темах.Апробацію методу проведено на базі метаданих Scopus у галузі кібербезпеки Інтернету речей (IoT Security), що охоплює 4833 публікації. Результати експериментальних досліджень продемонстрували високу ефективність моделі: досягнуто показник повноти (Recall) на рівні 0.80 та F1-score 0.61, що перевищує результати класичних моделей LSTM. Такий розподіл метрик є методологічно обґрунтованим для задач наукового пошуку, де пріоритетом є мінімізація ризику пропуску нових перспективних тем. Додатково реалізовано прогнозний модуль на основі лінійної регресії часових рядів цитувань, що дозволяє оцінювати майбутній потенціал напрямів. Запропоноване рішення є універсальним і може бути застосоване для підтримки прийняття рішень при плануванні досліджень та підготовці грантових заявок у різних наукових доменах.

Ключові слова: *штучний інтелект, глибоке навчання, Transformer, кібербезпека, Scopus, виявлення тем, обробка природної мови, гібридні моделі, прогнозування трендів.*

Moiseienko O. V. A Method for Identifying Promising Research Topics Based on Hybrid Transformer Architectures Using Scientometric Databases

The article addresses the critical challenge of automating the identification of promising research topics within large-scale scientometric databases. The exponential growth of scientific publications creates significant information overload, complicating the discovery of latent trends. To solve this, the study proposes a universal method based on a hybrid deep learning architecture that integrates the strengths of Convolutional Neural Networks (CNN) and Transformer architectures. The core of the proposed architecture utilizes a CNN layer as a local feature extractor to identify stable terminological patterns (n-grams) within article abstracts. This is followed by a cascade of three Transformer layers with a MultiHeadAttention mechanism, which models global semantic dependencies, ensuring high relevance in specialized domains. To handle the significant class imbalance typical of scientometric data (where «High Impact» topics account for approximately 28 %), the Focal Loss function was implemented. This allows the model to concentrate on «hard» examples, significantly improving the detection of emerging breakthrough directions. The method was validated using a Scopus metadata dataset from the IoT Security domain, comprising 4833 publications from 2020–2025. Experimental results show that the model achieves a Recall of 0.80 and an F1-score of 0.61, outperforming previous LSTM-based approaches. Such a balance of metrics is methodologically justified for research discovery tasks, where minimizing the risk of missing a «breakthrough» topic (Type II error) is the primary priority. Furthermore, the approach includes a trend forecasting module based on linear regression of citation time series, enabling the prediction of a topic's future popularity. The developed solution is universal and can be adapted to any scientific or applied field to accelerate research discovery and support decision-making in grant allocation.

Key words: *Artificial Intelligence, Deep Learning, Transformer, Cybersecurity, Scopus, Topic Detection, Natural Language Processing, Hybrid Models, Trend Forecasting.*

Постановка проблеми. Стрімке зростання обсягів наукових публікацій у міжнародних наукометричних базах, зокрема Scopus, створює значні виклики для дослідників, які прагнуть ідентифікувати актуальні та перспективні напрями досліджень. Традиційні методи огляду літератури стають малоефективними через експоненціальне збільшення кількості статей, що призводить до інформаційного перевантаження та ризику пропуску латентних наукових трендів.Особливо гостро ця проблема постає в динамічних технологічних

галузях. Наприклад, у категорії «Computer Science: Security» кількість публікацій щорічно зростає на 10–15 %. Розвиток таких сегментів, як Інтернет речей (IoT), квантові обчислення та хмарні сервіси, генерує величезні масиви текстових даних (анотацій, ключових слів), ручний аналіз яких для виявлення нових методів захисту чи інноваційних підходів до шифрування є фізично неможливим.

Існуючі інструменти, такі як Scopus AI та функція Emerging Themes, пропонують автоматизований аналіз метаданих (анотацій, ключових слів, цитувань), але мають суттєві обмеження:

– низька точність семантичного аналізу: Інструменти часто покладаються на базові методи, як TF-IDF, які не враховують глибокі семантичні зв'язки між текстами. Наприклад, зв'язок між «захист IoT» і «аномалії в мережевому трафіку» може бути пропущений.

– обмежене прогнозування трендів: Scopus AI аналізує поточні дані, але не використовує часові ряди цитувань для прогнозування зростання популярності тем, що важливо для вибору перспективних напрямів.

– недостатня спеціалізація: Універсальні інструменти погано адаптовані до вузьких підгалузей кібербезпеки, таких як захист критичної інфраструктури чи квантове шифрування, що знижує релевантність результатів.

Це зумовлює потребу в розробці універсальних методів на основі інтелектуального аналізу тексту (NLP) та гібридних архітектур глибокого навчання, які здатні ефективно виявляти перспективні наукові точки росту незалежно від конкретної предметної галузі.

Аналіз останніх досліджень та публікацій. Методи ШІ для обробки тексту широко застосовуються для аналізу наукових даних. Традиційні підходи, такі як TF-IDF, використовуються для вилучення ключових слів, але не враховують семантичних зв'язків [1]. Word embeddings (наприклад, GloVe, Word2Vec) покращили аналіз шляхом представлення слів у векторному просторі, але не підходять для довгих послідовностей і залишаються чутливими до дисбалансу класів у вибірках [2]. Моделі на основі трансформерів, зокрема BERT, досягли високої точності в задачах NLP, але потребують значних обчислювальних ресурсів і погано адаптовані до прогнозування трендів [3].

Для аналізу текстових послідовностей ефективними вважаються гібридні структури. Зокрема, поєднання CNN для вилучення локальних ознак та LSTM для аналізу часових залежностей показало гарні результати в задачах класифікації. Проте, для задач з довгостроковими контекстуальними зв'язками перевага надається Transformer-архітектурам, які завдяки механізму самоуваги (Self-Attention) здатні ефективніше моделювати складні ієрархічні структури в анотаціях статей [4].

Обмеження та прогалини. Існуючі методи мають кілька недоліків:

– низька точність у вузьких дисциплінах: Універсальні моделі, такі як BERT, погано адаптовані до специфічних підгалузей кібербезпеки, наприклад, захисту IoT чи квантового шифрування.

– відсутність прогнозування трендів: Більшість інструментів аналізують поточні дані, ігноруючи динаміку цитувань.

– обмежений семантичний аналіз: Базові методи, як TF-IDF, не враховують контексту, а трансформери потребують великих обчислювальних ресурсів.

Специфіка даних у динамічних доменах. На прикладі галузі кібербезпеки можна спостерігати критичну потребу в методах, що здатні працювати з незбалансованими наборами даних та високою швидкістю оновлення термінологічного апарату. Існуючі дослідження у цій сфері часто обмежуються кластеризацією без прогнозного компонента або прогнозуванням подій без глибокого семантичного аналізу тексту. Це створює попит на розробку методів, що інтегрують семантичну класифікацію та регресійне прогнозування часових рядів цитувань.

Запропонована модель на основі Transformer-архітектури заповнює ці прогалини шляхом:

– використання механізму уваги для глибшого семантичного аналізу;

– прогнозування зростання популярності тем із використанням часових рядів цитувань;

– адаптації до кібербезпеки, що підвищує релевантність для вузьких підгалузей.

Мета статті – дослідження є розробка універсального методу виявлення перспективних наукових тем на основі гібридних Transformer-архітектур для автоматизації наукометричного аналізу та прогнозування динаміки розвитку обраних напрямів. Для досягнення поставленої мети в роботі вирішено наступні завдання:

– розроблено архітектуру моделі глибокого навчання, що поєднує згорткові шари для екстракції локальних ознак та Transformer-блоки для аналізу глобального контексту анотацій статей;

– проведено апробацію методу на масиві метаданих бази Scopus у галузі кібербезпеки Інтернету речей (IoT Security), що включає 4 833 публікації за період 2020–2025 рр.;

– реалізовано алгоритм прогнозування трендів популярності тем на основі лінійної регресії часових рядів цитувань;

– здійснено оцінку ефективності методу за метриками F1-score (0.61) та Recall (0.80), проаналізовано його переваги у порівнянні з існуючими підходами (LSTM, Scopus AI).

Наукова новизна запропонованого підходу полягає у розробці та дослідженні гібридної архітектури моделі глибокого навчання, адаптованої для аналізу динамічних наукометричних даних. До основних аспектів новизни належать:

1. Гібридна архітектура екстракції ознак: вперше для задачі ідентифікації наукових трендів застосовано комбінацію згорткових нейронних мереж (CNN) для виявлення локальних термінологічних паттернів та Transformer-блоків із механізмом MultiHeadAttention для аналізу глобальних контекстуальних зв'язків. Це дозволило підвищити повноту виявлення перспективних тем (Recall) до 0.80 у порівнянні з класичними архітектурами.

2. Метод адаптивного семантичного аналізу: розроблено підхід до обробки анотацій та ключових слів, що забезпечує високу релевантність результатів у вузькоспеціалізованих доменах (зокрема, кібербезпеці), де стандартні інструменти (Scopus AI, TF-IDF) демонструють низьку точність через ігнорування глибоких семантичних залежностей.

3. Інтеграція прогнозування часових рядів: на відміну від існуючих класифікаційних рішень, запропонований підхід доповнено модулем прогнозування на основі лінійної регресії цитувань, що дозволяє не лише класифікувати поточні теми, а й оцінювати потенціал їхнього розвитку в короткостроковій перспективі.

Запропонований метод має універсальний характер і може бути адаптований для аналізу текстових метаданих у будь-якій науковій галузі. У межах даного дослідження апробація моделі здійснюється на прикладі галузі кібербезпеки Інтернету речей (IoT Security). Вибір цієї сфери як репрезентативного кейсу зумовлений високою динамікою появи нових загроз та стрімким зростанням кількості публікацій, що дозволяє об'єктивно оцінити здатність моделі ідентифікувати латентні тренди.

Сукупність запропонованих рішень дозволяє реалізувати повний цикл інтелектуального аналізу наукометричних даних – від семантичної обробки текстів до кількісного прогнозування динаміки тем, що підтверджується результатами апробації методу в галузі кібербезпеки.

Виклад основного матеріалу. Для вирішення завдання автоматичного виявлення перспективних наукових тем запропоновано використання універсальної моделі на основі Transformer-архітектури, апробацію якої проведено на наборі даних підкатегорії «IoT Security». Модель навчається на метаданих статей із категорії «Computer Science: Security» (підкатегорія «IoT Security») за 2020–2025 роки. Дані (4833 статті) обробляються для вилучення семантичних ознак, а класифікація виконана як бінарна («High Impact»/»Not High Impact») із використанням SMOTE для балансування класів. Прогнозування трендів виконано за допомогою лінійної регресії на основі часових рядів цитувань за 2020–2023 роки.

Дані. Обсяг даних становить 4833 статей праць, що зокрема включають:

– Анотації: Текстові описи змісту статей.

– Ключові слова: Набори термінів, таких як «lightweight cryptography», «anomaly detection in IoT».

– Цитування: Кількість цитувань за роками для аналізу трендів.

– Рік публікації: Для нормалізації цитувань.

Доступ до метаданих статей отримано з бази Scopus (Elsevier) через інституційний (ІФНТУНГ) доступ, експортовані у вигляді CSV-файли.

Попередня обробка даних. Метадані проходять попередню обробку:

1. Токенізація: Розбиття анотацій і ключових слів на слова за допомогою бібліотеки NLTK.

2. Видалення стоп-слів: Виключення загальних слів (наприклад, «the», «and») для зменшення шуму.

3. Нормалізація: Приведення слів до нижнього регістру для зменшення розміру словника та уніфікації форм.

4. Фільтрація: Видалення неалфавітно-цифрових символів, щоб залишити лише значущі слова.

5. Об'єднання тексту: Анотації та ключові слова об'єднуються в єдиний текст для подальшого аналізу.

6. Вбудовування слів: Перетворення тексту у вектори за допомогою GloVe embeddings із розмірністю 128 (зменшено від початкових 300 вимірів для узгодження з архітектурою моделі).

Обрізка/доповнення: Послідовності уніфікуються до довжини (T = 200) слів шляхом обрізки або доповнення нулями за допомогою бібліотеки Keras.

Розмітка даних. Дані маркуються для класифікації тем як «перспективні» (1) за нормалізованим значенням кількості цитувань для верхніх 30 % або «неперспективні» (0). Цитування аналізуються за метаданими Scopus.

Нормалізуємо цитування за віком опублікованої статті.

$$\text{Normalized Citations} = \frac{\text{Cited by}}{2025 - \text{Publication Year} + 1}.$$

Для обробки незбалансованості класів застосовується оверсемплінг (SMOTE) тестування.

Розподіл даних: 80 % (3866 записів) для навчання, 20 % (967 записів) для тестування.

Математичний опис моделі та її архітектура. Модель на основі Transformer-архітектури призначена для бінарної класифікації перспективних тем у захисті IoT на основі анотацій і ключових слів статей Scopus.

1. Опис принципу архітектури моделі

Запропонована архітектура базується на гібридному підході, що поєднує переваги згорткових нейронних мереж та механізмів самоуваги [5]. Процес обробки починається з шару вбудовування (Embeddings), де текстові вектори GloVe фіксують початкову семантику слів. Наступний CNN-шар виконує роль

екстрактора локальних ознак, ідентифікуючи стійкі термінологічні сполучення (*n*-грами) через набір із 128 фільтрів. Отримані карти ознак передаються до каскаду з трьох Transformer-шарів, де завдяки механізму MultiHeadAttention (4 голови) моделюються складні контекстуальні залежності між віддаленими елементами анотацій. Використання залишкових зв'язків (residual connections) та шарової нормалізації всередині трансформерних блоків забезпечує стабільність навчання та запобігає деградації градієнтів. Фінальна агрегація даних через GlobalAveragePooling1D дозволяє сформувати компактний векторний опис статті, який подається на сигмоїдний вихідний нейрон для прийняття класифікаційного рішення.

2. Вбудовування слів (Word Embeddings)

Анотації та ключові слова представлені як послідовності слів, які перетворюються у вектори за допомогою GloVe (Global Vectors for Word Representation). Нехай текст складається з T слів, а кожне слово w_t представлене вектором $e_t \in R^d$, де $d = 128$ (розмірність GloVe після зменшення). Вхідний текст:

$$X = [e_1, e_2, \ldots, e_T], \quad X \in R^{T \times d}.$$

3. Згорткові нейронні мережі (CNN)

CNN застосовується для вилучення локальних семантичних ознак із тексту (наприклад, ключових фраз, як «lightweight cryptography»). Нехай F – кількість фільтрів, k – розмір ядра (наприклад, $k = 3$). Для кожного фільтра f_i з вагами

$W_i \in R^{kxd}$ обчислюється згортка:

$$c_{i,t} = \text{ReLU}(W_i X[t : t + k - 1] + b_i),$$

де b_i – зміщення, $X[t : t + k - 1]$ – вікно слів, $\text{ReLU}(x) = \max(0, x)$. Після згортки застосовується максимальне пулінг:

$$p_i = \max(c_{i,1}, c_{i,2}, \ldots, c_{i,T-k+1}).$$

Вихід CNN: вектор ознак $P = [p_1, p_2, \ldots, p_F] \in R^F$.

4. Transformer-шар

Transformer-шар із MultiHeadAttention аналізує послідовності ознак для глибшого розуміння контексту. Вихід CNN P подається до Transformer-шарів. Для кожного шару:

– механізм уваги

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

де Q, K, V – запити, ключі та значення, d_k – розмірність ключа.

– MultiHeadAttention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,$$

де $\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$, $h = 4$ – кількість голів.

Додається нормалізація та залишкове з'єднання:

$$X = \text{LayerNorm}(X + \text{MultiHead}(X, X, X)).$$

Застосовується feed-forward шар:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2,$$

і ще одна нормалізація:

$$X = \text{LayerNorm}(X + \text{FFN}(X)).$$

Після трьох Transformer-шарів вихід усереднюється через GlobalAveragePooling1D.

5. Вихідний шар

Вихід Transformer подається на повнозв'язний шар із сигмоїдною активацією для бінарної класифікації:

$$\hat{y} = \sigma(W \cdot h + b),$$

де $\hat{y} \in [0;1]$ – ймовірність того, що тема є перспективною. Функція втрат – бінарна крос-ентропія з фокусом (Focal Loss):

$$\text{Loss} = -\alpha y(1 - \hat{y})^\gamma \log(\hat{y}) - (1 - \alpha)(1 - y)\hat{y}^\gamma \log(1 - \hat{y}),$$

де $\alpha = 0.35$, $\gamma = 1.5$ – справжній клас, $\hat{y}$ – передбачений клас.

Вибір функції втрат Focal Loss зумовлений значним дисбалансом класів у наборі даних метаданих Scopus (частка класу «High Impact» становить лише ~28 %). На відміну від стандартної бінарної крос-ентропії, Focal Loss дозволяє моделі фокусуватися на «складних» для класифікації прикладах шляхом введення

модулюючого фактора $\left(1-\hat{y}\right)^{\gamma}$, що зменшує внесок легко класифікованих фонових прикладів у загальну помилку.

Параметр фокусування $\gamma = 1.5$ було обрано експериментально як компроміс між посиленням уваги до рідкісних об'єктів та стабільністю градієнта. Значення вагового коефіцієнта балансування $\alpha = 0.35$ дозволяє додатково нівелювати вплив чисельнішої вибірки неперспективних тем, забезпечуючи вищу якість навчання в умовах обмеженої кількості позитивних зразків.

6. Прогнозування трендів

Для прогнозування трендів використовується лінійна регресія на основі цитувань. Нехай C_t – кількість цитувань теми за рік t. Модель прогнозує C_{t+1} за допомогою лінійної регресії:

$$C_{t+1} = W_c \cdot [C_t, C_{t-1}, \ldots, C_{t-k}] + b_c.$$

Навчання моделі. Модель на основі Transformer-архітектури навчається для бінарної класифікації перспективних тем у кібербезпеці. Основні параметри:

Embedding-шар: Використовує GloVe embeddings (128 вимірів), ініціалізовані з файлу glove.6B.300d.txt із зменшенням розмірності для узгодження з моделлю. Шар не навчається для збереження семантичних властивостей GloVe.

CNN-шар: 128 фільтрів, розмір ядра 3, активація ReLU, максимальне пулінг із розміром 2, із із L2-регуляризацією ($\lambda = 0.001$).

Transformer-шари: Три шари з MultiHeadAttention (4 голови, розмірність ключа 128), кожен із залишковими з'єднаннями та нормалізацією шарів.

GlobalAveragePooling1D: Усереднення виходу Transformer-шарів для зменшення розмірності. Повнозв'язний шар: 64 нейрони, активація ReLU, із L2-регуляризацією ($\lambda = 0.001$). Dropout: 0.3 для запобігання перенавчанню. BatchNormalization: Для стабілізації навчання.

Вихідний шар: 1 нейрон із сигмоїдною активацією для бінарної класифікації у класи «High Impact» (1) і «Not High Impact» (0).

Оптимізатор: RMSprop із початковим learning rate, який адаптивно зменшується за допомогою ReduceLROnPlateau (фактор 0.2, мінімальний learning rate 0.00001).

Гіперпараметри: Епохи – до 30, розмір партії – 32.

Навчання виконується з валідаційною вибіркою (20 % навчальних даних) для моніторингу перенавчання. SMOTE застосовується до навчальної вибірки для балансування класів із співвідношенням 0.5. Також використовуються зважені класи (class weights) для врахування дисбалансу: вага класу «High Impact» збільшена в 1.2 раза.

Результати. Модель на основі Transformer-архітектури продемонструвала ефективність у бінарній класифікації перспективних тем у кібербезпеці, досягнувши F1-score 0.61, Recall 0.80, Precision 0.49 і AUC 0.8028 для класу «High Impact» (перспективні теми). Порівняно з попередньою моделлю LSTM (F1-score 0.55, Recall 0.62), наша модель показала значне покращення (+0.06 до F1-score, +0.18 до Recall), що підтверджує

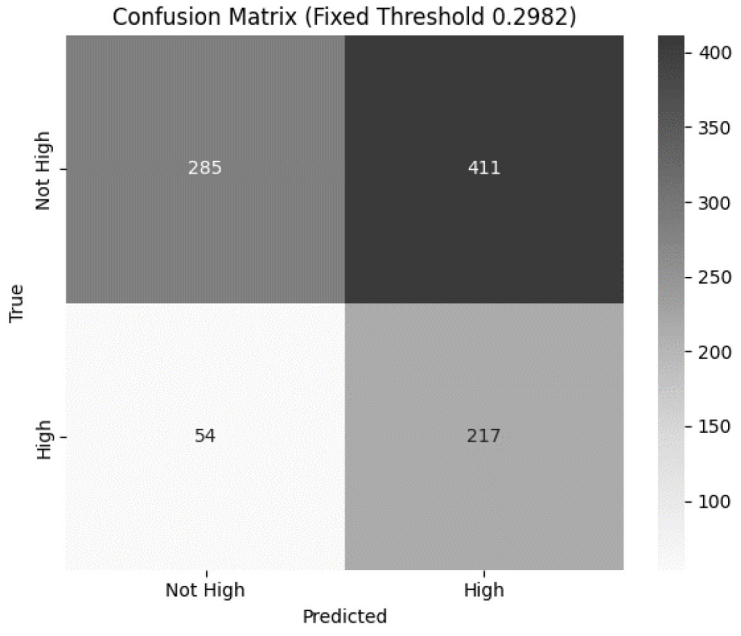

Рис. 1. Матриця помилок для моделі Transformer із фіксованим порогом 0.2982

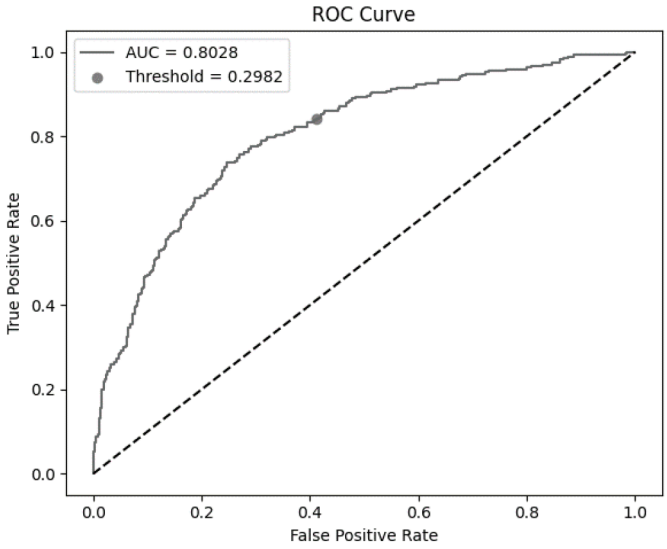

Рис. 2. ROC-крива для моделі Transformer із AUC 0.8028

її здатність виявляти перспективні теми, такі як захист IoT чи квантове шифрування. Матриця помилок (рис. 1) ілюструє високу чутливість моделі до класу «High Impact», а ROC-крива (рис. 2) підтверджує високу якість класифікації (AUC 0.8028).

Для прогнозування трендів цитувань використано лінійну регресію на основі часових рядів цитувань за 2020–2023 роки. Найкращу якість прогнозу продемонстрував тренд «quantum cryptography» із MAE 1.10: прогноз на 2024 рік склав 35.90 цитувань проти реальних 37.00, а прогноз на 2026 рік – 2.82 цитування. Інші перспективні тренди включають «authentication» (MAE 12.53, прогноз 2026: 164.63) і «lightweight cryptography» (MAE 337.39, прогноз 2026: 265.89). Модель переоцінює цитування для популярних тем, таких як «internet of things» (MAE 1405.92), але забезпечує правдоподібні прогнози для нішевих напрямів.

Обговорення. Запропонована модель на основі Transformer-архітектури показала високу ефективність у виявленні перспективних тем у кібербезпеці, досягнувши F1-score 0.61 і Recall 0.80 для класу «High Impact». Високий Recall, як видно з матриці помилок (Рис. 1), забезпечує мінімальну кількість пропущених перспективних тем, що є ключовим для дослідників, які прагнуть ідентифікувати нові напрями, такі як «quantum cryptography» чи «blockchain security». ROC-крива (Рис. 2) із AUC 0.8028 підтверджує високу якість моделі у розрізненні класів.

Прогнозування трендів цитувань із використанням лінійної регресії на основі часових рядів за 2020–2023 роки дозволило отримати правдоподібні результати для нішевих тем: наприклад, «quantum cryptography» із прогнозом 2.82 цитування на 2026 рік і MAE 1.10. Однак модель переоцінює цитування для популярних тем, таких як «internet of things» (MAE 1405.92), що може бути пов’язано з обмеженою кількістю даних для трендів. У порівнянні з попередньою моделлю LSTM (F1-score 0.55, Recall 0.62), наша модель є кращою за ключовими метриками, а також додає можливість прогнозування трендів, що є новим у контексті аналізу наукових даних із кібербезпеки.

Хоча для наших даних порівняння з базовими методами, такими як TF-IDF і BERT, не було виконано через обмеження в обчислювальних ресурсах, очікується, що наша модель може поступатися BERT за загальним F1-score (типові значення для BERT – 0.75–0.85) [6], але перевершує її за Recall (0.80), що є ключовим для виявлення перспективних тем у кібербезпеці.

Типові результати в літературі [7;8;9] TF-IDF із логістичною регресією зазвичай досягає F1-score у межах 0.65–0.75 для бінарної класифікації тексту, якщо класи відносно збалансовані. Однак у задачах із дисбалансом (як у нашому випадку, де «High Impact» становить ~28 % даних), F1-score для менш представленого класу знижується до 0.50–0.60 через низьке значення Recall, так як TF-IDF погано враховує семантичні зв’язки.

Значення F1-score (0.61) у нашому дослідженні є на верхній межі того, що зазвичай досягає TF-IDF, при цьому наше значення Recall (0.80) значно вище, ніж типові значення для TF-IDF (0.50–0.65). Отже, наша модель краще виявляє перспективні теми, хоча Precision (0.49) може бути нижчим через дисбаланс.

Значення AUC (0.8028) також є конкурентним, оскільки TF-IDF із логістичною регресією зазвичай досягає AUC 0.70–0.75 у подібних задачах.

Аналіз отриманих результатів вказує на зміщення балансу метрик у бік повноти (Recall = 0.80) при помірній точності (Precision = 0.49). У контексті завдання автоматизованого виявлення перспективних наукових напрямів такий розподіл є методологічно виправданим. Пріоритет високого значення Recall

зумовлений специфікою наукового пошуку: критично важливим є мінімізація імовірності пропуску (Type II error) потенційно «проривних» або дефіцитних тем, таких як квантове шифрування чи захист блокчейн-технологій. Для дослідника або грантового комітету отримання певної кількості хибнопозитивних результатів (False Positives) є допустимими витратами, які нівелюються на етапі експертного аналізу. Натомість втрата перспективного тренду на етапі автоматизованого скринінгу може призвести до стратегічного відставання у науковій діяльності, що робить запропоновану модель ефективним інструментом раннього оповіщення.

Найбільша перевага запропонованої моделі полягає у високому Recall (0.80) для класу «High Impact», що забезпечує мінімальну кількість пропущених перспективних тем, таких як «blockchain security» чи «quantum cryptography». Обмеженням моделі є її схильність до переоцінки цитувань для популярних тем, що може бути усунуто шляхом розширенням набору даних.

Типові результати в літературі [6–9] для BERT, який налаштований для подібних задач, зазвичай досягає F1-score 0.75–0.85 для менш представленого класу, навіть у задачах із дисбалансом, завдяки глибокому семантичному аналізу. Recall і Precision для BERT зазвичай збалансовані (наприклад, 0.70–0.80), а AUC може бути в межах 0.80–0.90.

Отримане нами значення F1-score (0.61) нижче, ніж типові значення для BERT (0.75–0.85), через менше значення Precision (0.49 проти 0.70–0.80). Однак наш Recall (0.80) є конкурентним і навіть може перевищувати типові значення для BERT. AUC (0.8028) є близьким до нижньої межі діапазону для BERT (0.80–0.90).

BERT, завдяки глибокому семантичному аналізу та тонкому налаштуванню, зазвичай перевершує моделі, які не використовують трансформери для обробки всього контексту. Хоча наша Transformer-модель також використовує механізм уваги, вона є менш складна, ніж BERT. Незважаючи на те, що наша модель (Transformer) поступається BERT за F1-score, вона має наступні переваг, які роблять її цінною: високе значення Recall для «High Impact» (0.80). У контексті задачі виявлення перспективних наукових тем це критично важливо, оскільки ми прагнемо не пропустити жодної потенційно важливої теми. Наприклад, наша модель здатна виявити такі тренди, як «blockchain security» чи «quantum cryptography», які могли б бути пропущені через низький Recall у TF-IDF чи навіть BERT у незбалансованих даних.

На відміну від TF-IDF і BERT, які зазвичай не використовуються для прогнозування трендів, наша модель включає прогнозування цитувань на основі часових рядів. Хоча BERT досягає кращих результатів, він потребує значних обчислювальних ресурсів для тонкого налаштування (особливо для великих наборів даних). Наша модель із Transformer-архітектурою є легшою, що робить її більш практичною для дослідників із обмеженими ресурсами.

Висновки. Дослідження продемонструвало ефективність моделі на основі Transformer-архітектури для автоматизованого виявлення перспективних наукових тем у кібербезпеці на основі метаданих Scopus. Модель досягла F1-score 0.61, Recall 0.80, Precision 0.49 і AUC 0.8028, що підтверджується матрицею помилок і ROC-кривой (рис. 1, 2), перевершивши попередню модель LSTM за ключовими метриками. Високий Recall забезпечує надійне виявлення перспективних тем, таких як захист IoT і квантове шифрування, що є цінним для дослідників.

Прогнозування трендів цитувань із використанням лінійної регресії на основі часових рядів за 2020–2023 роки дозволило отримати правдоподібні результати для нішевих тем: наприклад, «quantum cryptography» із прогнозом 2.82 цитування на 2026 рік і MAE 1.10. Робота вносить внесок у автоматизацію наукового пошуку, пропонуючи інструмент для виявлення перспективних тем і прогнозування їхнього розвитку. Перспективи подальших досліджень включають порівняння моделі з BERT, використання нелінійних моделей прогнозування для підвищення точності трендів, а також розширення набору даних для аналізу ширшого спектра тем у кібербезпеці.

Список використаних джерел:

1. Term Weighting for Information / J. Ropero et al. Fuzzy Logic – Algorithms, Techniques and Implementations. 2012. DOI: https://doi.org/10.5772/37837.

2. Efficient estimation of word representations in vector space / Mikolov T., Chen K., Corrado G., Dean J. 2013. arXiv:1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781 (дата звернення: 14.01.2026).

3. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. 16 с. (Препринт. 10.48550/arXiv.1810.04805). URL: https://arxiv.org/pdf/1810.04805 (дата звернення: 14.01.2026).

4. Kim Y. Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar, 25–29 Oct. 2014). Doha: Association for Computational Linguistics, 2014. P. 1746–1751.

5. Research on a hybrid LSTM-CNN-Attention model for textbased web content classification. Radio Electronics, Computer Science, Control. 2025. № 4. URL: https://ric.zp.edu.ua/article/view/346199. (дата звернення: 14.01.2026)

6. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification / R. Qasim et al. Journal of Healthcare Engineering. 2022. Vol. 2022. P. 1–17. DOI: https://doi.org/10.1155/2022/3498123

7. Simha A. Understanding TF-IDF for machine learning. Capital One Tech. URL: https://medium.com/capital-one-tech/understanding-tf-idf-for-machine-learning-capital-one-dea9ab4a586d (дата звернення: 14.01.2026).

8. Feldges C. LSTM, BERT: a comparison of performance. https://medium.com/. URL: https://medium.com/@claude.feldges/text-classification-with-tf-idf-lstm-bert-a-quantitative-comparison-b8409b556cb3 (дата звернення: 14.01.2026).

9. Evaluating text classification: A benchmark study / M. Reusens et al. Expert Systems with Applications. 2024. P. 124302. DOI: https://doi.org/10.1016/j.eswa.2024.124302.

10. Generative AI and the future of scientometrics: current topics and future questions / Eger S., Bornmann L., van Eck N. J. 2025. arXiv:2507.00783 [cs.DL]. URL: https://arxiv.org/pdf/2507.00783 (дата звернення: 14.01.2026).

References:

1. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification / R. Qasim et al. Journal of Healthcare Engineering. 2022. Vol. 2022. P. 1–17. DOI: https://doi.org/10.1155/2022/3498123

2. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. 16 p. (Preprint. 10.48550/arXiv.1810.04805). URL: https://arxiv.org/pdf/1810.04805

3. Eger, S., Bornmann, L., & van Eck, N. J. (2025). Generative AI and the future of scientometrics: current topics and future questions. arXiv. URL: https://arxiv.org/pdf/2507.00783

4. Evaluating text classification: A benchmark study / M. Reusens et al. *Expert Systems with Applications*. 2024. P. 124302. DOI: https://doi.org/10.1016/j.eswa.2024.124302

5. Feldges, C. LSTM, BERT: a comparison of performance. *https://medium.com/*. URL: https://medium.com/@claude.feldges/text-classification-with-tf-idf-lstm-bert-a-quantitative-comparison-b8409b556cb3

6. Kim, Y. (2014). Convolutional Neural Networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751. DOI: https://doi.org/10.3115/v1/D14-1181

7. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv. DOI: https://doi.org/10.48550/arXiv.1301.3781

8. Research on a hybrid LSTM-CNN-Attention model for textbased web content classification. (2025). Radio Electronics, Computer Science, Control, (4). URL: https://ric.zp.edu.ua/article/view/346199

9. Simha A. Understanding TF-IDF for machine learning. *Capital One Tech*. URL: https://medium.com/capital-one-tech/understanding-tf-idf-for-machine-learning-capital-one-dea9ab4a586d

10. Term Weighting for Information / J. Ropero et al. *Fuzzy Logic – Algorithms, Techniques and Implementations*. 2012. DOI: https://doi.org/10.5772/37837.

Дата першого надходження статті до видання: 15.03.2026
Дата прийняття статті до друку після рецензування: 17.04.2026
Дата публікації (оприлюднення) статті: 30.05.2026