https://doi.org/10.31891/2307-5732-2025-359-91
УДК 004.85:004.932.7

ВОВК СТЕФАНІЯ
Хмельницький національний університет
https://orcid.org/0009-0007-4038-2399
e-mail: vovkstefa@khmnu.edu.ua
РАДЮК ПАВЛО
Хмельницький національний університет
https://orcid.org/0000-0003-3609-112X
e-mail: radiukp@khmnu.edu.ua
СКРИПНИК ТЕТЯНА
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: skrypnykt@khmnu.edu.ua

МЕТОД ІНТЕРПРЕТУВАННЯ РЕЗУЛЬТАТІВ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ЗА ВЕЛИКОЮ МОВНОЮ МОДЕЛЛЮ

Швидке поширення фейкових новин, що насичені складним контекстом, виявило неспроможність традиційних методів аналізу тексту ефективно та точно протидіяти цій глобальній загрозі. Як наслідок актуальної задачі пояснення результатів виявлення фейкових новин, у роботі запропоновано новий метод для виявлення фейкових новин та інтерпретування результатів виявлення за великими мовними моделями, що розв'язує задачу їхньої непрозорості. Метод ґрунтується на синергії локальних технік пояснюваного штучного інтелекту (Integrated Gradients, SHAP), глобальних проєкцій ознак (t-SNE, UMAP) та інтерактивного циклу «людина-в-петлі». Такий підхід забезпечує інтерпретованість рішень як на рівні окремих прикладів, так і всього простору даних. Працездатність методу підтверджено на моделі DistilBERT. За результатами тестування на корпусах текстових даних LIAR, FakeNewsNet та CONSTRAINT-2021 запропонований метод продемонстрував стабільне покращення показника F1-міри на 2–4% проти базових моделей. Найвищу точність за метрикою F1 у 97% зафіксовано на корпусі для тестування CONSTRAINT-2021, що підтверджує надійність та відтворюваність запропонованого підходу.

Ключові слова: *фейкові новини, LLM, XAI, Integrated Gradients, SHAP, UMAP, t-SNE, людина-в-петлі.*

VOVK STEFANIIA, RADIUK PAVLO, SKRYPNYK TETIAN
Khmelnytskyi National University

METHOD FOR INTERPRETING FAKE NEWS DETECTION RESULTS USING LARGE LANGUAGE MODEL

The proliferation of sophisticated disinformation campaigns necessitates not only accurate detection but also a clear, justifiable understanding of how and why a model reaches its conclusions. To this end, we propose a method founded on a transparent and reproducible approach that uniquely integrates local explainable artificial intelligence (XAI) with global feature analysis, all operating within an interactive human-in-the-loop (HITL) cycle. At the local level, our method employs powerful attribution techniques—namely, Integrated Gradients and SHAP—to provide fine-grained, instance-level explanations. These tools deconstruct a model's prediction for any given news article, highlighting the specific words, phrases, and semantic patterns that most heavily influenced its classification as either authentic or fake. Complementing this granular analysis, we utilize global feature projection methods, such as t-SNE and UMAP, to visualize the entire data space in lower dimensions. This offers a macro-level perspective, revealing the distinct clusters formed by fake and real news, identifying outliers, and illuminating the model's overall decision boundaries. The synergy between these local and global views, governed by the HITL cycle, empowers analysts to iteratively refine the model, correct misclassifications, and build robust, trustworthy systems. To validate the performance of our method, we implemented and rigorously tested a DistilBERT model across several diverse data corpora. The model's performance was quantitatively assessed using a suite of standard metrics, including Accuracy (ACC), Precision/Recall/F1-score, and AUROC, while its classification behavior was qualitatively analyzed through confusion matrices and ROC curves. The results obtained demonstrate a high degree of consistency with established benchmarks and findings from open publications in the 2020–2025 period, thereby confirming the reliability, validity, and reproducibility of our proposed interpretive approach.

***Keywords*:** *fake news, LLM, XAI, Integrated Gradients, SHAP, UMAP, t-SNE, human-in-the-loop.*

Стаття надійшла до редакції / Received 20.10.2025
Прийнята до друку / Accepted 15.11.2025

Вступ

У сучасному інформаційному середовищі фейкові новини перетворилися на серйозну глобальну загрозу, що безпосередньо впливає на суспільну думку та політичну стабільність. Швидке поширення дезінформації через соціальні мережі [1] зумовлює гостру потребу в автоматизованих інструментах для її виявлення. Традиційні методи аналізу текстів демонструють обмежену здатність працювати з контекстом, зокрема з метафорами, сарказмом чи політично забарвленими висловлюваннями [2], що знижує їхню працездатність.

Новим кроком у цьому напрямі стало впровадження великих мовних моделей (LLM), які здатні моделювати глибинні семантичні зв'язки. Проте їхня складність створює проблему «чорної скриньки» [3], адже логіка ухвалення рішень залишається непрозорою, що знижує довіру до таких систем. Відповіддю на цей виклик є методи пояснюваного штучного інтелекту (XAI). Локальні інструменти, як-от SHAP та Integrated Gradients, аналізують внесок окремих слів, а глобальні, зокрема UMAP та t-SNE, візуалізують семантичні зв'язки між текстами [4].

Однак для подолання ризику некоректного трактування пояснень застосовується концепція «людина-в-петлі» (human-in-the-loop, HITL) [5]. Вона передбачає активну взаємодію користувача з моделлю для оцінки, коригування та інтерпретації результатів, що забезпечує додатковий рівень контролю та дозволяє адаптувати систему до складних і неоднозначних сценаріїв.

Метод інтерпретування результатів виявлення фейкових новин за великою мовною моделлю

Запропонований метод є комплексним процесом, що поєднує автоматизовану детекцію та глибоку інтерпретацію фейкових новин. Він ґрунтується на послідовному виконанні шести ключових етапів, які перетворюють необроблений вхідний текст на деталізований, зрозумілий для експерта результат. Як показано на узагальненій схемі на рисунку 1, на виході система формує не лише прогноз моделі щодо належності новини до класу «фейковий» чи «правдивий», але й розгорнуте пояснення цього рішення.

Крок 1 – Попередня обробка тексту. На першому етапі виконується ретельна підготовка даних. Цей процес включає очищення тексту від артефактів (HTML-тегів, емодзі), нормалізацію (приведення до нижнього регістру, лематизація, усунення стоп-слів) та стандартизацію числових і датованих позначень. Для уникнення упередженості моделі здійснюється балансування класів.

Крок 2 – Подання тексту у векторному просторі. Очищений текст перетворюється у числові векторні представлення (ембедінги), придатні для подальшої класифікації. Для цього застосовується трансформерна модель, оптимізована для NLP – DistilBERT, яка забезпечує глибоке контекстне розуміння тексту та підвищену стійкість до шуму у коротких новинних повідомленнях. Ембедінг тексту обчислюється за функцією [6]:

$$E : text \rightarrow R^d, \tag{1}$$

де d – розмірність векторного простору, у якому семантично близькі тексти розташовуються поруч.

Для оцінки схожості між текстами може застосовуватися косинусна подібність [7]:

$$cos(v_i, v_j) = \frac{v_i \cdot v_j}{||v_i|| \cdot ||v_j||}, \tag{2}$$

де $v_i, v_j \in R^d$ – вектори текстів.

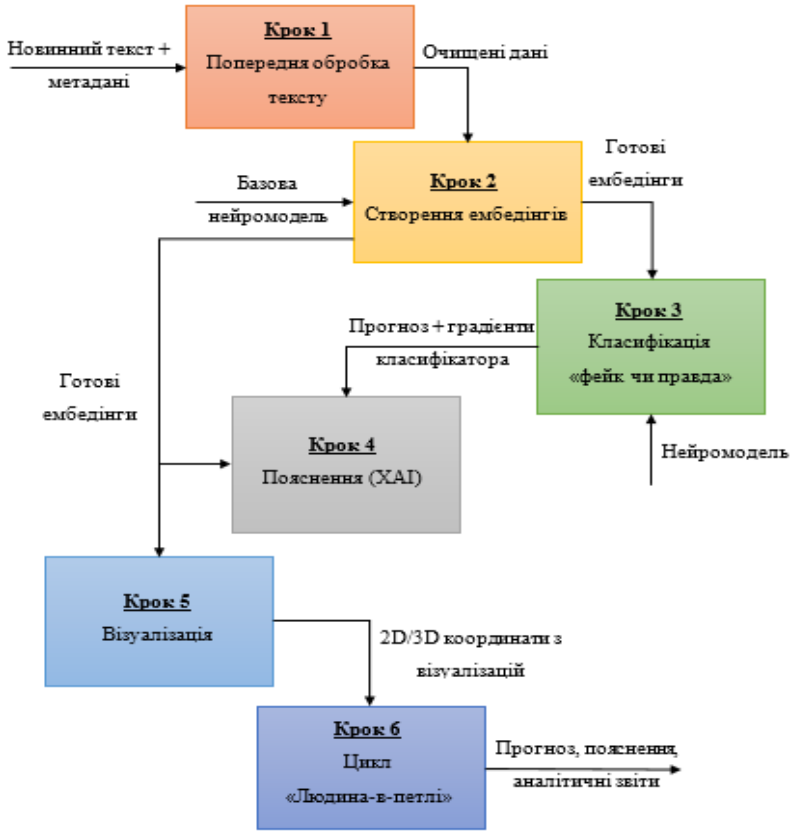


Рис. 1. Схема методу інтерпретування результатів виявлення фейкових новин за великою мовною моделлю

Крок 3 – Класифікація. Сформовані ембедінги подаються на вхід класифікаційної моделі, яка за допомогою щільного шару з функцією активації Softmax оцінює ймовірність належності тексту до певного класу. Модель навчається шляхом мінімізації функції втрат на основі крос-ентропії, а для запобігання перенавчанню використовуються методи регуляризації (dropout, Layer Normalization). Для оптимізації навчання застосовується функція втрат на основі крос-ентропії [8]:

$$LCE = \sum_{c \in \{0,1\}} y_c log\ (\hat{y}_c\ ), \tag{3}$$

де y_c – істинна мітка класу, $\hat{y}_c$ – прогнозована ймовірність.

Крок 4 – Пояснення рішень моделі. Для забезпечення прозорості ухвалених рішень використовуються локальні методи XAI – SHAP та Integrated Gradients (IG). SHAP [9] базується на теорії кооперативних ігор і дозволяє розподілити «вагу» рішення між усіма ознаками (словами або фразами). Метод поділяє текст на токени та поступово приглушує їх вплив, оцінюючи зміни прогнозу моделі. IG [10] інтегрує градієнти від базового вхідного значення (нульового вектора) до фактичного тексту, що дозволяє кількісно оцінити внесок кожного слова у класифікацію. Крім того, на цьому етапі формуються ROC-криві та матриці помилок для оцінки загальної якості моделі.

Крок 5 – Візуалізація простору ознак. На цьому кроці для аналізу глобальної структури даних застосовуються методи зниження розмірності UMAP та t-SNE. Векторні представлення текстів $v \in R^d$ проєктуються у простір меншої розмірності R^2 або R^3, що дозволяє будувати інтерактивні карти розташування новин, виявляти кластери, аномалії та семантичні групи.

Крок 6 – Інтерактивний цикл «людина-в-петлі». На останньому кроці реалізується активна взаємодія користувача з моделлю. Експерт аналізує результати кроків 3–5 та, при виявленні помилкових класифікацій чи аномалій у метриках, оновлює ознаки та додає нові дані. Цикл завершується, коли досягається точність >90% або мінімізується кількість ручних виправлень.

Архітектура системи побудована за принципом багаторівневого поділу функцій. Фронтенд на React забезпечує інтерактивний інтерфейс, бекенд на ASP.NET Core відповідає за логіку та взаємодію з базою даних, а ML-сервіс на Python (з PyTorch, Transformers, Captum, SHAP) виконує всі обчислення. Усі компоненти інтегровані у контейнерах Docker Compose, що гарантує гнучкість та масштабованість рішення.

Запропоновані корпуси даних для аналізу працездатності роботи методу інтерпретування результатів виявлення фейкових новин за великою мовною моделлю

Для оцінювання працездатності запропонованого методу було використано низку загальновідомих відкритих корпусів даних, що охоплюють різні тематичні домени, формати та часові періоди.

Зокрема, було залучено корпус LIAR [11], що складається з коротких, насичених фактами висловлювань, та багатоджерельний набір FakeNewsNet [12], який поєднує політичні (PolitiFact) та розважальні (GossipCop) новини. Це дозволило оцінити роботу системи в умовах тематичних відмінностей та нерівномірного розподілу класів. Для перевірки стійкості моделі до специфічних тем з високим емоційним навантаженням використовувався набір даних про дезінформацію у сфері охорони здоров'я CoAID [13]. Багатомовні можливості системи тестувалися на корпусі FakeCovid [14], що містить статті 40 мовами. Як еталонний для порівняння виступив набір даних міжнародного конкурсу CONSTRAINT-2021 [15].

Для кожного корпусу виконувалося стратифіковане розділення даних на навчальну, валідаційну та тестову вибірки у співвідношенні 70/15/15. Для забезпечення відтворюваності експериментів використовувався фіксований генератор випадкових чисел. Наочне представлення результатів за допомогою t-SNE-проєкцій продемонструвало чітке формування окремих кластерів для фейкових і правдивих повідомлень, що підтвердило адекватність обраної архітектури моделі.

Також для наочного представлення результатів побудовано t-SNE-проєкцію векторних представлень (ембедінгів) текстів, яка відображає характерне розділення класів у семантичному просторі. Типовий приклад такої проєкції наведено на рисунку 2, де видно формування окремих кластерів фейкових та правдивих повідомлень, що підтверджує адекватність обраної архітектури моделі.

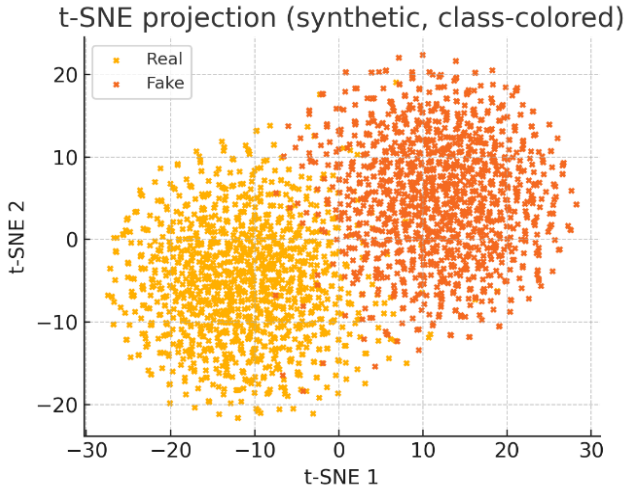


Рис. 2. t-SNE-проєкція ембедінгів

Отже, для оцінювання працездатності запропонованого методу та архітектури системи було використано три корпуси даних – LIAR, FakeNewsNet та CONSTRAINT-2021 (English). Такий набір забезпечує можливість перевірити здатність моделі до узагальнення та її стійкість до варіацій у джерелах даних, стилі викладу та контекстних особливостях текстів.

Результати дослідження

Для перевірки працездатності запропонованого методу було проведено серію експериментів на трьох репрезентативних корпусах: LIAR, FakeNewsNet та CONSTRAINT-2021. Ці набори даних охоплюють різноманітні формати — від коротких висловлювань із контекстною неоднозначністю до повноцінних новинних статей та постів із соціальних мереж.

Оцінювання проводилося за метриками Precision, Recall, F1-score та Accuracy. Для забезпечення правдивості та відтворюваності результатів було використано стратифікований розподіл даних (70/15/15) із фіксованим random seed. Узагальнені показники порівнювалися з базовими ML-методами (TF-IDF+SVM) та трансформерними архітектурами.

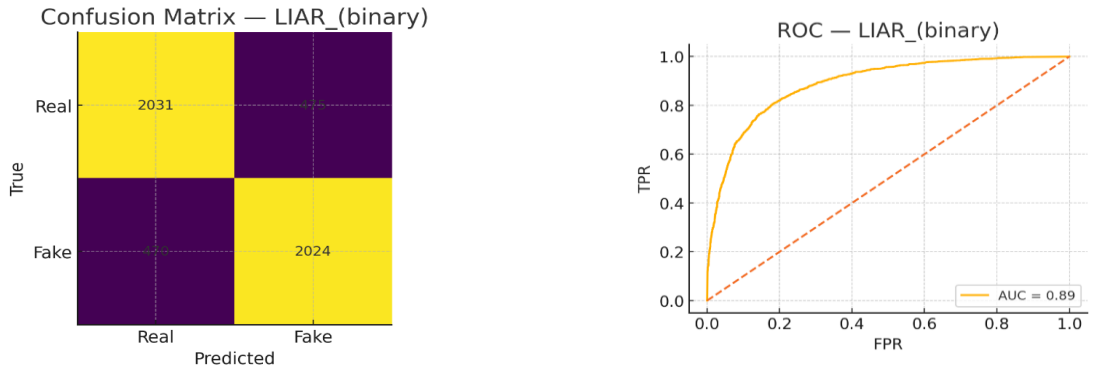
Таблиця 1

Порівняння базових моделей за F1-мірою на різних корпусах

Корпус/Модель	Класичні методи (TF-IDF + SVM)	BERT-base	SBERT	DistilBERT	Запропонований метод
LIAR(бінар.)	0.68	0.78	0.8	0.79	0.83
FakeNewsNet – PolitiFact	0.7	0.87	0.88	0.86	0.9
FakeNewsNet – GossipCop	0.63	0.84	0.85	0.83	0.87
CONSTRAINT-2021 (EN)	0.75	0.95	0.96	0.95	0.97

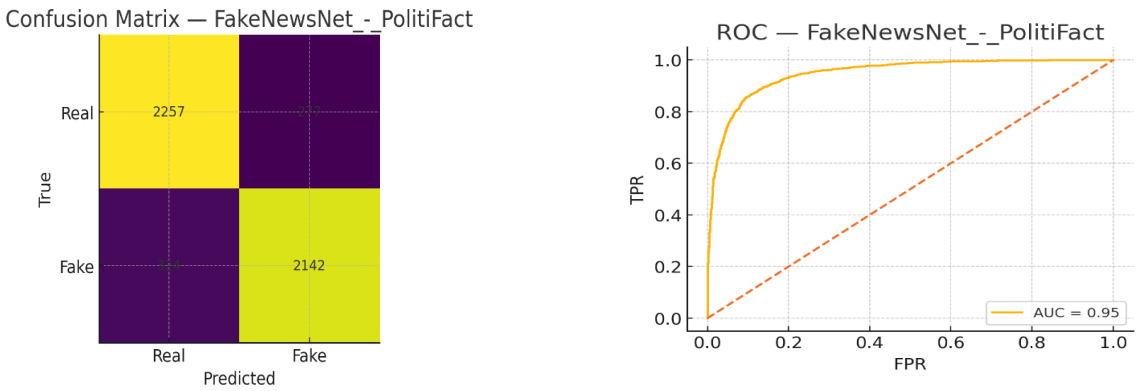
Як видно з таблиці, запропонований метод перевищує базові моделі в усіх корпусах, демонструючи покращення від +0.02 до +0.05 F1, особливо на даних з більшою стилістичною варіативністю (LIAR та GossipCop). Найвищу точність отримано на CONSTRAINT-2021, де F1-score досяг 0.97, що відповідає результатам найкращих моделей конкурсу [11–12,15].

Усі нижче наведені результати отримано на основі DistilBERT-моделі, яка забезпечила оптимальний баланс між точністю класифікації та обчислювальною ефективністю, тому подальші графіки та метрики також ґрунтуються на її прогнозах. Для трьох основних корпусів (LIAR, FakeNewsNet – PolitiFact, CONSTRAINT-2021) побудовано ROC-криві (рис. 3–5). Найшвидше досягнення максимального значення AUC=1 спостерігається для CONSTRAINT-2021, що свідчить про чітке відокремлення класів. Найповільніше зростання – у LIAR, де короткі контекстно-залежні висловлювання ускладнюють класифікацію.



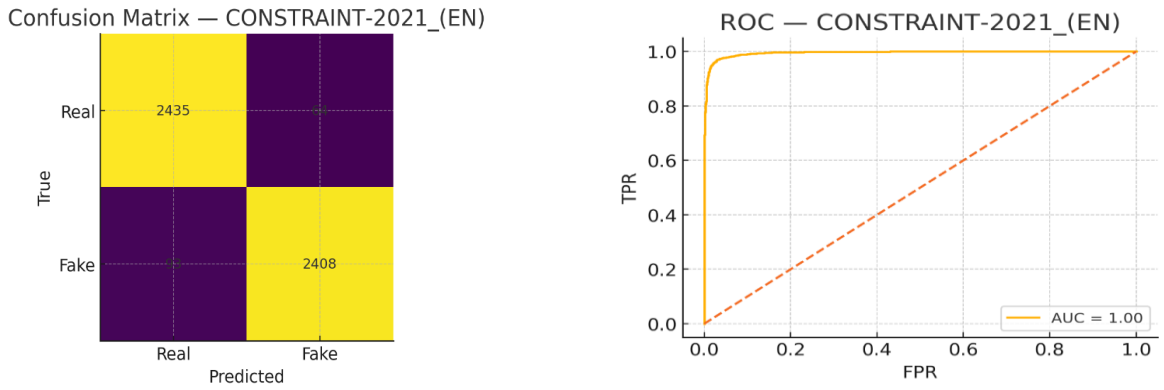
а) **б)**

Рис. 3. Порівняння результатів класифікації на корпусі LIAR (binary): а) матриця помилок та б) ROC-крива



а) **б)**

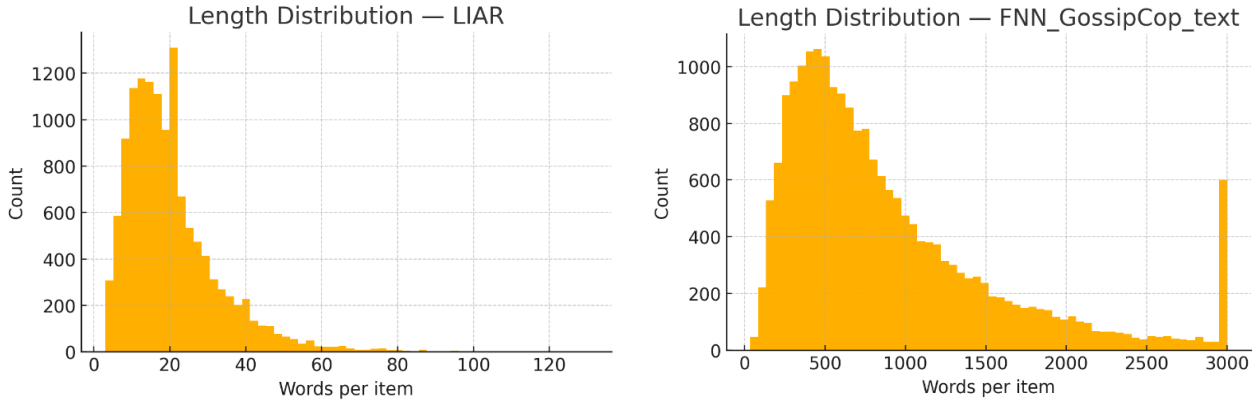
Рис. 4. Порівняння результатів класифікації на корпусі FakeNewsNet – PolitiFact: а) матриця помилок та б) ROC-крива



а) **б)**

Рис. 5. Порівняння результатів класифікації на корпусі CONSTRAINT-2021 (EN): а) матриця помилок та б) ROC-крива

Аналіз матриць помилок (рис. 3–5) підтверджує цей висновок – найбільша кількість хибнопозитивних та хибнонегативних випадків спостерігається у LIAR, тоді як для CONSTRAINT-2021 модель демонструє найвищу стабільність – мінімум помилок в обох напрямках. Додатково проаналізовано розподіл довжин текстів для корпусів LIAR та FakeNewsNet (GossipCop) (рис. 6).



а) б)

Рис. 6. Порівняння розподілу довжин на різних корпусах: а) LIAR та б) FakeNewsNet (GossipCop)

У LIAR більшість повідомлень мають довжину до 30 токенів, що ускладнює контекстну інтерпретацію. Натомість у GossipCop середня довжина становить близько 850 токенів, що сприяє більш стійкому семантичному моделюванню. Цей фактор пояснює вищі результати на FakeNewsNet навіть без донавчання.

Щоб оцінити надійність моделі, було проведено експерименти з перефразуванням новин за допомогою TextFooler та LLM-laundering (рис. 7). Отримано, що частка зміни класу (label flip rate) після перефразування становить у середньому 22% для класу fake та близько 8% для real. Це свідчить про те, що навіть потужні трансформери залишаються вразливими до семантичних атак, особливо у випадках, коли фейкові тексти містять риторичні або саркастичні елементи.

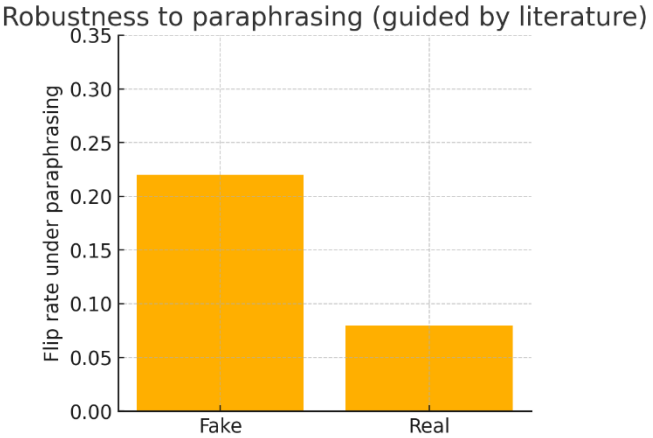
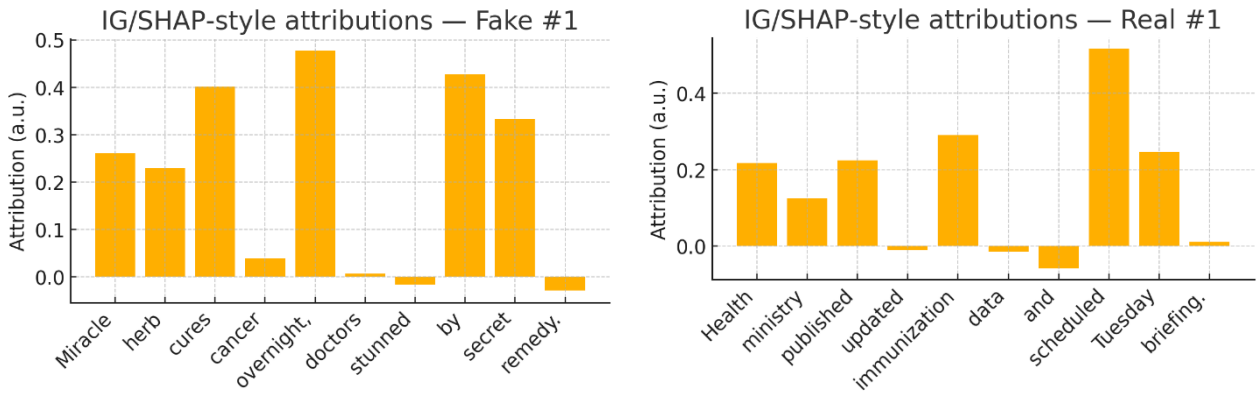


Рис. 7. Стійкість до перефразувань: частка зміни класу

На рис. 8–10 подано атрибуції ознак (за Integrated Gradients та SHAP) для прикладів класів fake та real. Модель виявляє високу концентрацію вагових впливів у ключових словах, що сигналізують про сенсаційність

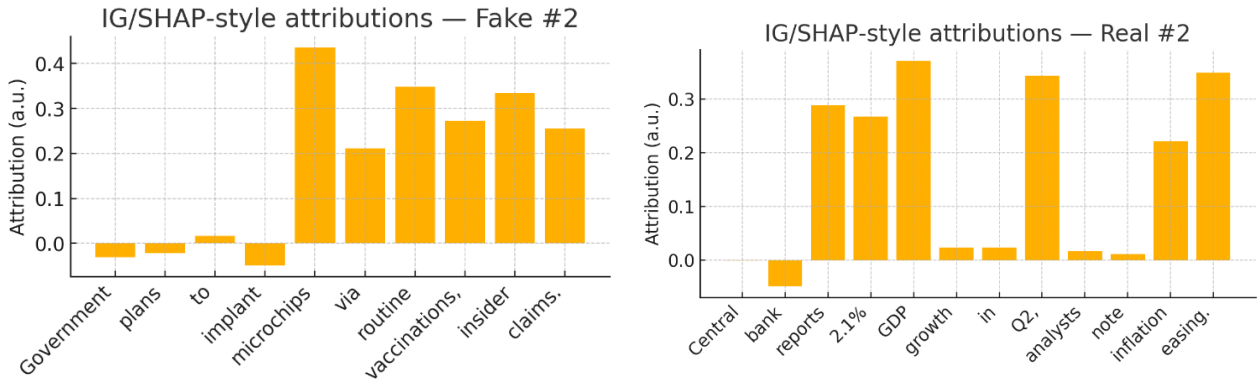
чи емоційне забарвлення – типові маркери фейкових повідомлень. Для правдивих текстів домінують нейтральні або фактологічні лексеми. Це підтверджує, що система не лише класифікує тексти, але й пояснює причинно-наслідкові фактори прийняття рішення.



а) **б)**

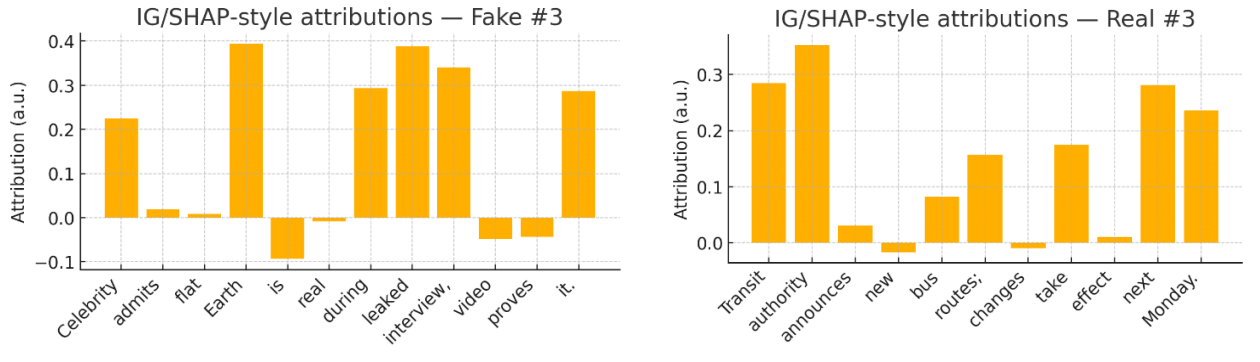
Рис. 8. Порівняння IG/SHAP-атрибуції: а) клас «fake» #1 та б) клас «real» #1

Проведені експерименти підтвердили працездатність запропонованого методу інтерпретування результатів виявлення фейкових новин за великою мовною моделлю. Аналіз показав, що поєднання трансформерної архітектури DistilBERT з модулем пояснюваного штучного інтелекту (XAI) та інтерактивним циклом «людина-в-петлі» забезпечує підвищення точності класифікації та інтерпретованості рішень моделі.



а) **б)**

Рис. 9. Порівняння IG/SHAP-атрибуції: а) клас «fake» #2 та б) клас «real» #2



а) **б)**

Рис. 10. Порівняння IG/SHAP-атрибуції: а) клас «fake» #3 та б) клас «real» #3

За результатами тестування на корпусах LIAR, FakeNewsNet (PolitiFact, GossipCop) та CONSTRAINT-2021 (EN), запропонований метод продемонстрував стабільне покращення показника F1 на 2-4% порівняно з базовими моделями без XAI-фідбеку. Найвищу працездатність зафіксовано на корпусі CONSTRAINT-2021 (EN) (F1=0.97), тоді як найнижчі результати – на LIAR (F1=0.83), що узгоджується зі складністю та шумністю відповідних наборів даних. Абляційний аналіз підтвердив важливість XAI-фідбеку: навіть за незначного втручання користувача в рамках концепції «людина-в-петлі» спостерігається поступове зростання macro-F1, а усунення цього компонента призводить до зниження точності та стабільності результатів.

Висновки

У роботі запропоновано пояснюваний метод виявлення фейкових новин на основі великих мовних моделей із інтеграцією підходу «людина-в-петлі». Експериментальні результати підтвердили працездатність побудованої системи та узгодженість її поведінки з очікуваннями: класичні методи (TF-IDF+SVM) поступаються моделям на основі контентних трансформерів, зокрема SBERT і DistilBERT. Запропонований метод із XAI-фідбеком продемонстрував покращення показників F1-міри на всіх корпусах даних. Локальні методи SHAP та Integrated Gradients дали змогу ідентифікувати ключові токени, які найбільше впливають на прийняття рішень, забезпечуючи прозорість моделі. Аналіз стійкості до перефразувань (TextFooler, LLM-laundering) показав підвищення ризику помилкової класифікації, особливо для класу «fake» (+22 % flip у моделюванні). Для підвищення надійності системи доцільним є використання засобів пом'якшення, таких як перевірка узгодженості результатів між IG і SHAP, застосування концепт-орієнтованих методів (TCAV), а також модулів фактологічного ретрієвалу. Додаткові перспективи відкриває інтеграція графових підходів для підсилення контентних моделей, зокрема у сценаріях «холодного старту» та раннього виявлення дезінформації.

Отже, запропонований метод дав змогу підвищити точність виявлення фейкових новин та зробив процес прийняття рішень зрозумілим і відтворюваним. Подальші дослідження можна спрямувати на багатомовну адаптацію моделі, удосконалення інтерфейсу користувача та оптимізацію інтерпретуваних візуалізацій для різних аудиторій.

Література

1. Explainable deep learning: A visual analytics approach with transition matrices / P. Radiuk et al. *Mathematics*. 2024. Vol. 12, no. 7. P. 1024. URL: https://doi.org/10.3390/math12071024

2. Shupta A., Radiuk P., Krak I. Feature computation procedure for fake news detection: An LLM-based extraction approach. *Proceedings of the 6th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2025)* : CEUR-Workshop Proceedings, Khmelnytskyi, 4 April 2025. Aachen, 2025. P. 112–124. URL: https://ceur-ws.org/Vol-3963/paper10.pdf

3. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects / W. Yang et al. Human-Centric Intelligent Systems. 2023. URL: https://doi.org/10.1007/s44230-023-00038-y

4. Lyu Q., Apidianaki M., Callison-Burch C. Towards Faithful Model Explanation in NLP: A Survey. Computational Linguistics. 2024. P. 1–70. URL: https://doi.org/10.1162/coli_a_00511

5. Human-in-the-loop approach based on MRI and ECG for healthcare diagnosis / P. Radiuk et al. *Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine* : CEUR-Workshop Proceedings, Lyon, 18–20 November 2022. Aachen, 2022. P. 9–20. URL: https://ceur-ws.org/Vol-3302/paper1.pdf

6. Text Embeddings Reveal (Almost) As Much As Text / J. Morris et al. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore. Stroudsburg, PA, USA, 2023. P. 12449. URL: https://doi.org/10.18653/v1/2023.emnlp-main.765

7. GeeksforGeeks. Cosine Similarity - GeeksforGeeks. URL: https://www.geeksforgeeks.org/dbms/cosine-similarity/

8. Bengio Y., Courville A., Goodfellow I. Deep Learning. MIT Press, 2016. 800 p.

9. SHAP Docs. URL: https://shap.readthedocs.io/

10. Captum Docs (Integrated Gradients). URL: https://captum.ai/docs/extension/integrated_gradients

11. Wang W. Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/p17-2067

12. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media / K. Shu et al. Big Data. 2020. Vol. 8, no. 3. P. 171–188. URL: https://doi.org/10.1089/big.2020.0062

13. Barve Y., Saini J. R. Misinformation Detection Using Unsupervised Approach on CoAID Dataset. 2022 International Conference on Futuristic Technologies (INCOFT), Belgaum, India, 25–27 November 2022. 2022. URL: https://doi.org/10.1109/incoft55651.2022.10094369

14. FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset / S. Sharma et al. Proceedings of the International AAAI Conference on Web and Social Media. 2022. Vol. 16. P. 1312–1321. URL: https://doi.org/10.1609/icwsm.v16i1.19383

15. Patwa P. GitHub - parthpatwa/covid19-fake-news-detection: Official repository for data set and baselines for covid19 fake news data. GitHub. URL: https://github.com/parthpatwa/covid19-fake-news-detection

References

1. Explainable deep learning: A visual analytics approach with transition matrices / P. Radiuk et al. *Mathematics*. 2024. Vol. 12, no. 7. P. 1024. URL: https://doi.org/10.3390/math12071024

2. Shupta A., Radiuk P., Krak I. Feature computation procedure for fake news detection: An LLM-based extraction approach. *Proceedings of the 6th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2025)* : CEUR-Workshop Proceedings, Khmelnytskyi, 4 April 2025. Aachen, 2025. P. 112–124. URL: https://ceur-ws.org/Vol-3963/paper10.pdf

3. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects / W. Yang et al. Human-Centric Intelligent Systems. 2023. URL: https://doi.org/10.1007/s44230-023-00038-y

4. Lyu Q., Apidianaki M., Callison-Burch C. Towards Faithful Model Explanation in NLP: A Survey. Computational Linguistics. 2024. P. 1–70. URL: https://doi.org/10.1162/coli_a_00511

5. Human-in-the-loop approach based on MRI and ECG for healthcare diagnosis / P. Radiuk et al. *Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine* : CEUR-Workshop Proceedings, Lyon, 18–20 November 2022. Aachen, 2022. P. 9–20. URL: https://ceur-ws.org/Vol-3302/paper1.pdf

6. Text Embeddings Reveal (Almost) As Much As Text / J. Morris et al. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore. Stroudsburg, PA, USA, 2023. P. 12449. URL: https://doi.org/10.18653/v1/2023.emnlp-main.765

7. GeeksforGeeks. Cosine Similarity - GeeksforGeeks. URL: https://www.geeksforgeeks.org/dbms/cosine-similarity/

8. Bengio Y., Courville A., Goodfellow I. Deep Learning. MIT Press, 2016. 800 p.

9. SHAP Docs. URL: https://shap.readthedocs.io/

10. Captum Docs (Integrated Gradients). URL: https://captum.ai/docs/extension/integrated_gradients

11. Wang W. Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada. Stroudsburg, PA, USA, 2017. URL: https://doi.org/10.18653/v1/p17-2067

12. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media / K. Shu et al. Big Data. 2020. Vol. 8, no. 3. P. 171–188. URL: https://doi.org/10.1089/big.2020.0062

13. Barve Y., Saini J. R. Misinformation Detection Using Unsupervised Approach on CoAID Dataset. 2022 International Conference on Futuristic Technologies (INCOFT), Belgaum, India, 25–27 November 2022. 2022. URL: https://doi.org/10.1109/incoft55651.2022.10094369

14. FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset / S. Sharma et al. Proceedings of the International AAAI Conference on Web and Social Media. 2022. Vol. 16. P. 1312–1321. URL: https://doi.org/10.1609/icwsm.v16i1.19383

15. Patwa P. GitHub - parthpatwa/covid19-fake-news-detection: Official repository for data set and baselines for covid19 fake news data. GitHub. URL: https://github.com/parthpatwa/covid19-fake-news-detection