УДК 004.85:004.932.7

Вовк С.В., Радюк П.М., Скрипник Т.К.

Хмельницький національний університет

МЕТОД ІНТЕРПРЕТУВАННЯ РЕЗУЛЬТАТІВ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ЗА ВЕЛИКОЮ МОВНОЮ МОДЕЛЛЮ

Запропоновано пояснюваний метод виявлення фейкових новин, що інтегрує великі мовні моделі з модулями пояснюваного штучного інтелекту та концепцією «людина-в-петлі». Метод ґрунтується на послідовній обробці тексту за допомогою трансформерної архітектури DistilBERT для глибокого семантичного аналізу, подальшої класифікації та інтерпретації рішень моделі з використанням SHAP та Integrated Gradients. Це забезпечує високу точність класифікації та прозорість й довіру до системи, дозволяючи експертам розуміти фактори, що впливають на виявлення дезінформації.

An explainable method for detecting fake news is proposed, integrating large language models with explainable artificial intelligence modules and the "human-in-the-loop" concept. The method is based on sequential text processing using the DistilBERT transformer architecture for deep semantic analysis, subsequent classification, and interpretation of model decisions with SHAP and Integrated Gradients. This ensures not only high classification accuracy but also transparency and trust in the system, allowing experts to understand the factors influencing disinformation detection and actively intervene to enhance its effectiveness.

Сучасний інформаційний простір зіткнувся з безпрецедентним викликом – експоненціальним поширенням дезінформації, зокрема фейкових новин. Це явище перетворилося на серйозну загрозу, що здатна дестабілізувати суспільну думку, політичні процеси та навіть національну безпеку. Швидкість, з якою неправдива інформація розповсюджується через соціальні медіа, значно випереджає можливості традиційних методів фактчекінгу, що породжує потребу в створенні ефективних автоматизованих інструментів для протидії [1]. Особливої актуальності ця проблема набуває в умовах гібридних конфліктів і криз, де оперативність та достовірність інформації стають критично важливими.

Традиційні підходи до обробки природної мови (NLP) демонструють обмежену працездатність, оскільки їм бракує здатності глибоко аналізувати контекст, іронію та емоційне забарвлення тексту [2]. Поява великих мовних моделей (LLM) відкрила нові горизонти в автоматизованому аналізі, однак водночас створила проблему «чорної скриньки» – непрозорості процесу ухвалення рішень [3]. Ця особливість підриває довіру до таких систем та ускладнює їх застосування у таких відповідальних сферах, як політика чи охорона здоров'я.

Для подолання цих викликів активно розвиваються методи пояснюваного штучного інтелекту (XAI), зокрема інструменти SHAP та Integrated Gradients, що

дозволяють інтерпретувати рішення моделей [4]. Разом із концепцією «людина-в-петлі» (HITL), яка поєднує автоматизацію з експертним контролем, це створює синергію для підвищення надійності системи [5]. Попри значний прогрес, залишаються невирішеними питання адаптації моделей до багатомовних та емоційно насичених контекстів.

Метою дослідження є підвищення рівня інтерпретованості систем виявлення фейкових новин через проєктування нового методу, який дає змогу експертам інтерактивно аналізувати, валідувати та ітеративно вдосконалювати простір текстових ембедінгів, що лежить в основі класифікаційних рішень.

Запропонований метод поєднує автоматизоване виявлення фейкових новин з інтерпретацією результатів та втручанням людини в процес навчання. Його робота включає шість основних етапів:

1. Попередня обробка тексту – очищення від шумів, нормалізація, лематизація та балансування класів для уникнення упередженості.
2. Перетворення тексту у векторний простір за допомогою DistilBERT, що забезпечує глибоке контекстне розуміння.
3. Класифікація зі застосування щільного шару з функцією активації Softmax та мінімізацією функції втрат на основі крос-ентропії.
4. Пояснення рішень моделі за допомогою локальних методів SHAP та Integrated Gradients для виявлення ключових слів/фраз, що впливають на рішення моделі. Також формуються ROC-криві та матриці помилок для загальної оцінки якості моделі.
5. Візуалізація простору ознак за допомогою UMAP та t-SNE, для проєкції векторних представлень текстів у 2D/3D простір, що дозволяє виявляти кластери фейкових/правдивих новин та аномалії.
6. Інтерактивний цикл «людина-в-петлі», в межах якого експерт аналізує пояснення, коригує помилки й оновлює дані для підвищення точності (>90%).

Для експериментальної перевірки працездатності методу використано репрезентативні корпуси даних: LIAR [6], FakeNewsNet (PolitiFact, GossipCop) [7] та CONSTRAINT-2021 (EN) [8]. Ці набори охоплюють різні типи новин (короткі висловлювання, повноцінні статті), тематичні домени та часові періоди, дозволяючи оцінити здатність моделі до узагальнення.

Запропонований метод продемонстрував стабільне покращення показника F1-міри на 2–4% порівняно з базовими моделями (TF-IDF+SVM, BERT-base, SBERT) на всіх корпусах. Найвищу працездатність було зафіксовано на корпусі CONSTRAINT-2021 (F1 = 0.97), що узгоджується з результатами найкращих моделей міжнародних конкурсів [6–8]. Це підтверджує працездатність інтеграції DistilBERT та XAI-фідбеку (таблиця 1).

Експерименти з перефразуванням новин за допомогою TextFooler та LLM-laundering виявили вразливість трансформерних моделей до семантичних атак: частка зміни класу становила 22% для фейкових та 8% для правдивих новин (рисунок 1). Це вказує на необхідність посилення стійкості моделі.

Таблиця 1 – Порівняння базових моделей за F1-мірою на різних корпусах

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

Застосування SHAP та Integrated Gradients дозволило чітко ідентифікувати ключові слова та фрази (рисунки 2–3), які модель використовує для класифікації новин як фейкових (наприклад, сенсаційні, емоційно забарвлені лексеми) або правдивих (фактологічні, нейтральні), що підвищує прозорість та довіру до системи.

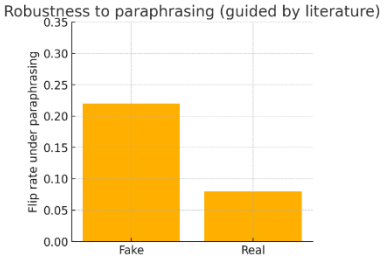
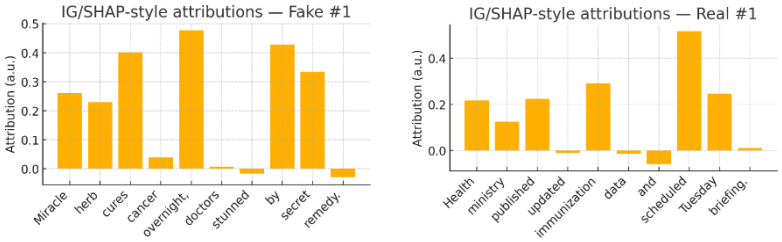
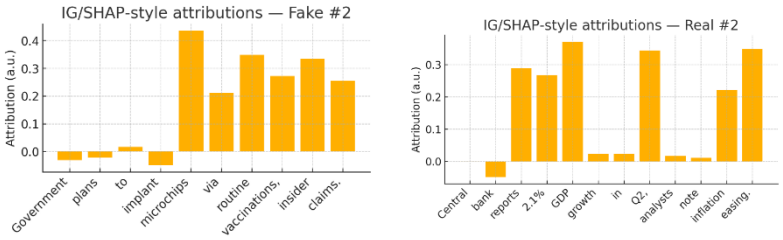


Рисунок 1 – Стійкість до перефразування контенту



а) б)

Рисунок 2 – Порівняння IG/SHAP-атрибуції: а) клас «fake» #1 та б) клас «real» #1



а) б)

Рисунок 3 – Порівняння IG/SHAP-атрибуції: а) клас «fake» #2 та б) клас «real» #2

У підсумку, у даній роботі було запропоновано та експериментально підтверджено працездатність пояснюваного методу виявлення фейкових новин на основі великих мовних моделей, інтегрованого з підходом «людина-в-петлі». Побудована система забезпечує не лише високу точність класифікації, але й прозорість прийнятих рішень завдяки модулям пояснюваного штучного інтелекту. Це дозволяє експертам розуміти логіку моделі та втручатися для покращення її роботи.

Перелік посилань

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 (date of access: 19.10.2025).
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 (date of access: 19.10.2025).
3. Survey on explainable AI: from approaches, limitations and applications aspects / W. Yang et al. Human-Centric Intelligent Systems. 2023. Vol. 3. P. 161–188. URL: https://doi.org/10.1007/s44230-023-00038-y (date of access: 19.10.2025).
4. Lyu Q., Apidianaki M., Callison-Burch C. Towards faithful model explanation in NLP: a survey. Computational Linguistics. 2024. Vol. 50, no. 2. P. 1–70. URL: https://doi.org/10.1162/coli_a_00511 (date of access: 19.10.2025).
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 (date of access: 19.10.2025).
6. 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, 30 July – 4 August 2017. Stroudsburg, PA, USA, 2017. P. 422–426. URL: https://doi.org/10.18653/v1/p17-2067 (date of access: 19.10.2025).
7. 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 (date of access: 19.10.2025).
8. 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 (date of access: 19.10.2025).