https://doi.org/10.31891/2307-5732-2025-355-102
УДК 004.85:004.912:025.4.03

ШУПТА АНДРІЙ
Хмельницький національний університет
https://orcid.org/0009-0000-9771-5579
e-mail: shuptaao@khmnu.edu.ua

МЕТОД ФОРМАЛІЗОВАНОЇ ПРОЦЕДУРИ СИНТЕЗУ ТА ОБЧИСЛЕННЯ ОЗНАК ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН

У роботі запропоновано новий метод, що формалізує процедуру виявлення фейкових новин, яка ґрунтується на можливостях великих мовних моделей (LLM) для синтезу підозрілих текстових атрибутів та їхнього перетворення на числові вектори, що придатні для класифікації. Завдання дослідження полягає в уточненні процесу перетворення текстових сигналів на числові ознаки, що покращує інтеграцію лінгвістичних сигналів з глибокими контекстуальними векторами ознак. Експерименти проводилися за англомовним (FakeNewsNet) та україномовним (Ukrainian news) наборами даних, де запропонований метод перевершив базові підходи, досягнувши точності до 89.6% для англійської та 88.3% для української мови. Ключові результати показують, що поєднання числових індикаторів (наприклад, коефіцієнтів перефразування та тональності) з генерацією за LLM забезпечує вищу повноту виявлення оманливих новинних статей. Запропонована процедура обчислення ознак успішно підвищує точність виявлення, зберігаючи прозорість прийняття рішень моделлю. Дослідження підкреслює важливість систематично розроблених числових ознак, які доповнюють генерації за LLM, пропонуючи шлях до більш надійних, адаптивних та пояснюваних систем виявлення фейкових новин.

Ключові слова: виявлення фейкових новин, великі мовні моделі, процедура обчислення ознак, обробка природної мови, класифікація текстів.

SHUPTA ANDRII
Khmelnytskyi National University

METHOD OF FORMALIZED PROCEDURE FOR SYNTHESIS AND COMPUTATION OF FEATURES FOR FAKE NEWS DETECTION

The pervasive and evolving nature of digital disinformation necessitates the development of sophisticated detection systems that are accurate, transparent, and adaptable to novel deceptive strategies. While Large Language Models (LLMs) have demonstrated considerable prowess in discerning nuanced textual patterns, their application in fake news detection often results in "black-box" systems, limiting trust and hindering the ability to respond to emergent manipulative techniques. This paper introduces a novel method designed to bridge this gap. We present a structured procedure for systematically synthesizing suspicious textual attributes, guided by LLM-driven insights, and their subsequent transformation into a robust set of quantifiable, interpretable numerical features. These features, encompassing aspects such as paraphrase intensity, sentiment polarity, stylistic anomalies, and fact-checking congruity, are then synergistically integrated with the deep contextual embeddings generated by LLMs. Rigorous experimental validation was conducted on diverse English (FakeNewsNet) and Ukrainian (Ukrainian news) datasets. The proposed method outperformed established baseline approaches, achieving substantial accuracy improvements, with figures reaching up to 89.6% for English and 88.3% for Ukrainian language texts. Key findings reveal that explicitly incorporating these engineered numeric indicators significantly enhances recall rates for deceptive articles, a critical factor in mitigating the societal impact of misinformation. Furthermore, the method's modularity fosters adaptability, enabling the incorporation of newly identified deceptive patterns as additional numeric features without necessitating the complete retraining of the foundational LLM. This study unequivocally underscores the significant value of systematically engineered, interpretable numeric features as a vital complement to the powerful, yet often opaque, embeddings of LLMs.

Keywords: fake news detection, large language models, feature computation procedure, natural language processing, text classification.

Стаття надійшла до редакції / Received 22.05.2025
Прийнята до друку / Accepted 26.06.2025

Вступ

Стрімке поширення фейкових новин через соціальні медіа та онлайн-платформи стало глобальною проблемою. Дезінформація може суттєво впливати на суспільну думку, підривати довіру до інституцій та навіть провокувати конфлікти. Традиційні методи виявлення фейкових новин часто не встигають за еволюцією тактик створення оманливого контенту. Сучасні підходи, що використовують великі мовні моделі (LLM), демонструють значний потенціал, проте часто страждають від браку прозорості та інтерпретованості результатів. Виникає потреба у розробці методів, які мали б змогу виявляти фейкові новини з високою точністю, надавали би зрозуміле пояснення прийнятих рішень, а також були б адаптивними до нових видів маніпуляцій.

Дослідження в галузі виявлення фейкових новин пройшли шлях від використання лексичних ознак та TF-IDF [1] до застосування статичних векторних представлень слів (Word2Vec, FastText) та глибоких нейронних мереж [2]. Архітектура нейронних мереж трансформер, зокрема BERT [3], забезпечили значний прогрес завдяки здатності фіксувати контекстуальні зв'язки у тексті. З появою LLM, як от LLaMA [4] та GPT [5], відкрилися нові можливості для аналізу тексту на ще більш глибокому рівні. Однак, попри високу точність, LLM часто функціонують як "чорні скриньки". Наукові роботи [**Помилка! Джерело посилання не знайдено.**, **Помилка! Джерело посилання не знайдено.**] вказують на важливість синтезу інтерпретованих ознак та процедур їхнього обчислення [6]. Запропонований у цій статті метод призначений для подолання

розриву між високим рівнем результативності LLM та потребою в пояснюваних і адаптивних системах виявлення.

Метою даної роботи є формалізування та покращення процедури синтезу та обчислення текстових ознак для виявлення фейкових новин з використанням LLM. Ця процедура використовує можливості LLM для ідентифікації підозрілих характеристик тексту та їхнього перетворення на числові значення, придатні для використання в моделях машинного навчання. Ключовими аспектами є підвищення точності виявлення, забезпечення інтерпретованості результатів та адаптивності методу до нових форм дезінформації.

Метод визначення механізму обчислення ознак для виявлення фейковості новини

Запропонований метод визначення механізму обчислення ознак для виявлення фейковості новини ґрунтується на структурованому підході, що складається з чотирьох взаємопов'язаних завдань: а) синтез характеристик фейкових новин, б) формулювання процедури обчислення ознак, в) побудова моделі машинного навчання та г) генерування експертних висновків. Опис робочого процесу наведено нижче покроково.

Блок 1. Синтез характеристик фейкових новин. На цьому кроці ідентифікуються потенційно підозрілі текстові елементи. Це може здійснюватися через запити до LLM з використанням методів інженерії промтів та ланцюжкового мислення. Дослідники або експерти в предметній області виявляють нові атрибути, що розвиваються, та можуть вказувати на оманливий контент. Ці характеристики документуються з попередніми метаданими, що вказують на можливі способи їхнього кількісного вимірювання. Приклади таких характеристик (текстових ознак $F = \{f_1, \ldots, f_5\}$) наведено в таблиці 1.

Блок 2. Формулювання процедури обчислення ознак. Це завдання перетворює набір підозрілих або "характерних" елементів на чисельно обчислений вектор ознак, зберігаючи пряме відображення на текстові свідчення. Процедура формально визначає відображення вхідного тексту T на набір об'єктивних ознак O:

$$M: T \to O, \tag{1}$$

де M – це механізм (модель) обчислення ознаки.

Модель, що формалізована у формулі (1), складається з таких кроків:

Крок 2.1. Збір метаданих з характерних ознак. Кожна ідентифікована характеристика $f_i \in F$ має пов'язану назву, опис та метадані.

Крок 2.2. Визначення формули для кожної ознаки. Кожну ознаку f_i представляємо через алгоритм або формулу ALG_i. Наприклад, для коефіцієнта перефразування (f_1):

$$f_1(T) = \frac{\sum_{j=1}^{k-1} \sum_{l=j+1}^{k} sim(s_j, s_l)}{k(k-1)/2}, \tag{2}$$

де $T = \{s_1, s_2, \ldots, s_k\}$ – набір речень у тексті, k – кількість речень, $sim(s_j, s_l)$ – функція схожості між реченнями s_j та s_l (косинусна схожість їхніх векторів ознак).

Таблиця 1

Приклади узагальнених текстових ознак для виявлення фейкових новин

Ознака	Опис	Тип значення
f_1: перефразування	Ступінь перефразування тексту, вимірює схожість між реченнями	Числовий
f_2: суб'єктивність	Частка суб'єктивних висловлювань у тексті	Числовий
f_3: тональність	Загальна емоційна забарвленість тексту (позитивна, негативна, нейтральна)	Числовий
f_4: незвична мови	Наявність незвичної, недоречної або маніпулятивної лексики	Числовий
f_5: підтвердження фактів	Ступінь відповідності наведених фактів відомим джерелам	Числовий

Крок 2.3. Реалізація прямої та оберненої процедур: а) пряме обчислення (числове) $ALG_i(T)$ – перетворює текст T на числову оцінку $o_i \in [0,1]$; б) обернене пояснення (текстове) $ALG_i^{-1}(o_i, T)$ – реєструє, які конкретні слова, речення або фрази $T_{\text{sub}} \subset T$ найбільше вплинули на обчислену оцінку o_i.

Крок 2.4. Включення арифметичних або логічних операцій. Деякі ознаки f_c можуть бути похідними (композитними) від множини базових ознак $\{f_{b_1}, \ldots, f_{b_m}\}$. Наприклад, "Оцінка маніпулятивної мови" (Manipulative Score), $f_c(T)$:

$$f_c(T) = \sum_{j=1}^{m} w_j \cdot f_{b_j}(T), \tag{3}$$

де w_j – вагові коефіцієнти, $\sum w_j = 1$.

Крок 2.5. Формування кінцевого набору ознак. Процедура генерує вектор обчислених ознак $O = (o_1, o_2, \ldots, o_N)$, де $o_i = ALG_i(T)$. Кожен елемент o_i пов'язаний з ідентифікатором ознаки, описовою назвою, алгоритмом ALG_i та метаданими ALG_i^{-1} для інтерпретації.

Формалізація механізму обчислення ознаки f_i для виявлення фейковості новини та включає такі пункти.

1. Визначення вхідних даних: Текст новини T.

2. Вибір лінгвістичного аспекту: Характеристика C_i (наприклад, “наявність емоційної лексики”).

3. Розробка алгоритму ALG_i: Послідовність операцій для кількісної оцінки C_i в T: а) токенізація $T \to \{t_1, \ldots, t_p\}$; б) використання зовнішніх лексиконів L_i (наприклад, словник емоційної лексики); в) застосування правил або моделей машинного навчання до ідентифікації релевантних текстових фрагментів $T_{\text{rel}} \subset T$; г) агрегація та нормалізація для отримання числового значення $o_i \in [0,1]$.

Наприклад, для ознаки “Коефіцієнт тональності” (f_3):

$$ALG_3(T) = \text{normalize}\left(\sum_{s \in T} sentiment\_core(s)\right), \quad (4)$$

де $\text{sentiment\_core}(s)$ – оцінка тональності речення s, отримана за допомогою LLM або спеціалізованої бібліотеки, а normalize – функція нормалізації (масштабування до інтервалу [0,1]).

Блок 3. Побудова моделі машинного навчання. Після формування вектора ознак O наступним кроком є навчання класифікатора Clf для маркування новини як фейкової ($y = 1$) або істинної ($y = 0$).

$$Clf: O \to y, \text{ де } y \in \{0,1\}.$$

У цьому дослідженні розглядаються як класичні алгоритми, так і нейронні архітектури. Паралельно використовуються вектори ознак за LLM, $E_{LLM}(T)$, як додаткові ознаки. Комбінований вектор ознак X_{comb} для класифікатора формується як конкатенація:

$$X_{\text{comb}} = O \oplus E_{LLM}(T). \quad (5)$$

Отже, кожна новинна стаття подається як комбінацією обчислених оцінок, так і глибоким латентним вбудуванням, що отриманий за LLM.

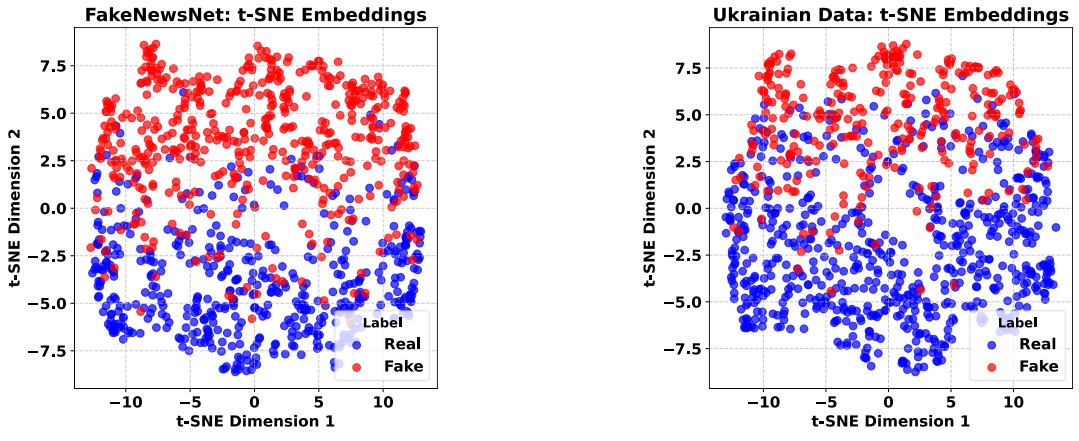
Блок 4. Побудова шаблону експертного висновку. Важливим аспектом є генерування зрозумілого “експертного висновку”. Після того, як класифікатор Clf видає мітку y, система посилається на метадані ALG_i^{-1} з блоку 2, щоб визначити, які текстові сегменти T_{sub} сприяли числовим значенням o_i, що призвели до цього рішення. Це забезпечує прозорість та інтерпретованість моделі.

Дослідження проводилося на двох наборах даних: FakeNewsNet (англійська мова) [7] та Ukrainian Fake & True News (українська мова) [8]. Дані були розділені на навчальну (80%), валідаційну (10%) та тестову (10%) вибірки. Були виділені наступні базові ознаки: TF-IDF, середні вектори ознак Word2Vec, вектори ознак BERT. Ознаки за LLM генерувалися за допомогою моделі “Llama-3.2-3B-Instruct”. Було реалізовано п’ять прикладів числових ознак, згідно з таблицею 1. Для класифікації використовувалася проста двошарова нейронна мережа прямого поширення.

Результати експериментів

У таблиці 2 наведено результати класифікації для англомовного та україномовного наборів даних. Використано метрики включають Accuracy, Precision, Recall, F1-міра та AUC-ROC. Результати класифікації у таблиці 2 демонструють, що додавання запропонованої процедури обчислення числових ознак послідовно покращує показники для всіх базових методів. Найкращі результати досягаються у випадку поєднання ознак LLM з обчисленими ознаками, що свідчить про комплементарність цих підходів.

Візуалізації t-SNE (рис. 1) показують, що доповнення стандартних векторів ознак за LLM числовими атрибутами призводить до більш чіткого розділення між кластерами фейкових та істинних новин у проєктованому 2D-просторі.



а) б)

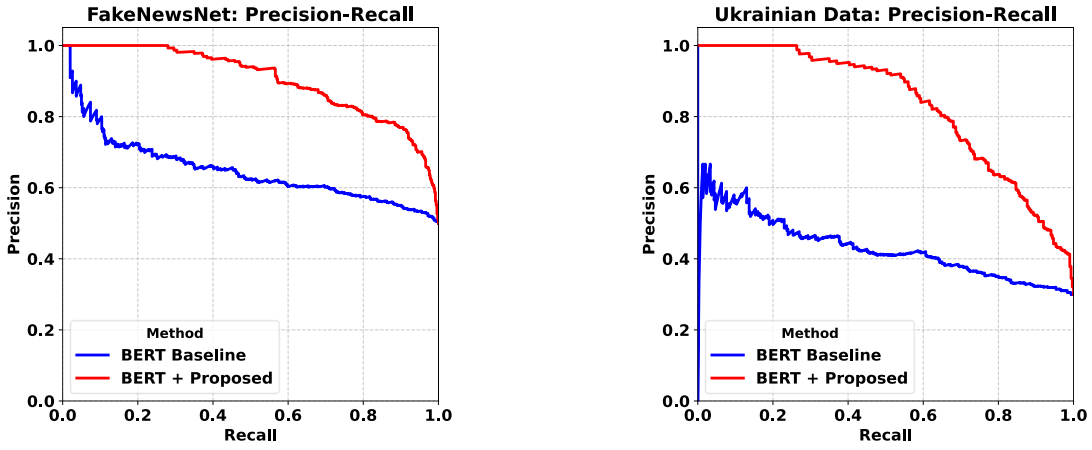
Рис. 1. t-SNE візуалізація для LLM + Запропоновані ознаки: а) FakeNewsNet (англ.);б) Ukrainian new (укр.)

Криві Precision-Recall (рис. 2) для BERT та BERT + Запропоновані ознаки демонструють покращену влучність та повноту на різних порогах класифікації у випадку використання запропонованого методу.

Таблиця 2

Порівняння результатів класифікації базових класифікаторів та класифікаторів на основі запропонованої процедури синтезу та обчислення ознак (у %) за наборами даних FakeNewsNet (англійська мова) та Ukrainian Fake & True News (українська мова). Найкращі показники виділено жирним шрифтом

Набір даних	Метод	Accuracy	Precision	Recall	F1-міра	AUC
FakeNewsNet	TF-IDF (базовий)	80.2	82.1	78.3	80.1	85.0
	TF-IDF + Запроп.	**82.5**	**84.0**	**81.1**	**82.5**	**86.3**
Ukrainian news	TF-IDF (базовий)	78.5	80.0	75.8	77.8	84.0
	TF-IDF + Запроп.	**80.8**	**82.1**	**78.5**	**80.2**	**85.5**
FakeNewsNet	Word2Vec (базовий)	78.1	79.4	76.6	78.0	83.0
	Word2Vec + Запроп.	**81.0**	**82.5**	**79.3**	**80.8**	**85.2**
Ukrainian news	Word2Vec (базовий)	75.6	77.1	74.0	75.5	81.0
	Word2Vec + Запроп.	**78.2**	**79.3**	**77.5**	**78.4**	**83.2**
FakeNewsNet	BERT (базовий)	85.0	86.0	83.1	84.5	90.0
	BERT + Запроп.	**86.9**	**87.5**	**85.7**	**86.6**	**91.2**
Ukrainian news	BERT (базовий)	82.9	83.4	82.1	82.7	88.0
	BERT + Запроп.	**84.7**	**85.2**	**84.1**	**84.7**	**89.3**
FakeNewsNet	LLM (базовий)	88.5	88.2	89.7	88.9	93.0
	LLM + Запроп.	**89.6**	**89.5**	**90.2**	**89.8**	**93.5**
Ukrainian news	LLM (базовий)	86.7	85.2	88.3	86.7	92.0
	LLM + Запроп.	**88.3**	**87.7**	**89.4**	**88.5**	**92.6**



а) б)

Рис. 2. Криві Precision-Recall для BERT та BERT + Запропоновані ознаки: а) FakeNewsNet (англ.);б) Ukrainian news (укр.)

Результати дослідження (таблиця 1) демонструють, що запропонований метод формалізованої процедури синтезу та обчислення ознак суттєво покращує виявлення фейкових новин, особливо у випадку інтеграції з LLM. Поєднання числових індикаторів, як от коефіцієнти перефразування та тональності, з глибокими контекстуальними векторами ознак LLM дало змогу досягти точності близько 90% як за англомовним (FakeNewsNet), так і за україномовним (Ukrainian news) наборами даних. Це свідчить про результативність доповнення латентних векторів ознак за LLM явними, інтерпретованими ознаками, що підвищує не тільки точність, але й повноту виявлення оманливого контенту, що особливо важливо для мінімізації пропуску фейкових новин. Візуалізації t-SNE (рис. 1) та криві Precision-Recall (рис. 2) також підтверджують перевагу комбінованого підходу, показуючи кращу роздільність класів та вищі показники на різних порогах класифікації.

Дискусія навколо запропонованої процедури обертається навколо її здатності забезпечити прозорість та адаптивність системи. Формалізований покроковий механізм обчислення ознак дає можливість зрозуміти, які саме текстові характеристики вплинули на рішення моделі. Крім того, вона забезпечує відносно легку інтеграцію нових ознак для протидії тактикам створення дезінформації, що постійно розвиваються та вдосконалюються, без необхідності повного перенавчання LLM. Однак, основними обмеженнями методу є обчислювальна складність генерації векторів ознак LLM для великих обсягів даних у реальному часі та потенційна упередженість як самих LLM, так і зовнішніх джерел, що використовуються для перевірки фактів чи визначення специфічних ознак (наприклад, емоційної лексики). Також існує залежність від якості та повноти метаданих, що документують характеристики фейкових новин.

Заразом подальші дослідження матимуть перспективи щодо аналізу поширення новин у соціальних мережах, впровадження нових методів оптимізації обчислювальних витрат та розширення крос-мовних валідацій для створення більш надійних, прозорих і глобально адаптивних систем виявлення фейкових новин.

Висновки

У даній роботі представлено метод, що формалізує процедуру синтезу та обчислення ознак для виявлення фейкових новин, який інтегрує можливості LLM з чітко визначеними числовими індикаторами. Експерименти за англомовним та україномовним наборами даних підтвердили результативність запропонованого методу: поєднання векторів ознак за LLM з обчисленими ознаками дало змогу досягти точності близько 90% (89.6% для англійської та 88.3% для української мови). Ключовими перевагами методу є покращена точність класифікації, підвищена інтерпретованість результатів завдяки відстеженню внеску конкретних текстових характеристик, та адаптивність системи до нових видів дезінформації через додавання нових ознак без необхідності повного перенавчання базових LLM. Формалізований покроковий опис механізму обчислення ознак дає можливість систематично підходити до розроблення та вдосконалення систем виявлення. Попри перспективні результати, є певні обмеження, такі як обчислювальна складність генерації векторів ознак за LLM для великих обсягів даних у реальному часі та потенційна упередженість моделей або зовнішніх прикладних програмних інтерфейсів для перевірки фактів.

Перспективи подальших досліджень включають аналіз поширення новин у соціальних мережах, дослідження методів часткового донавчання LLM або дистиляції знань для зменшення обчислювальних витрат, а також розширення крос-мовних валідацій для врахування глобального характеру проблеми дезінформації.

Література

1. Wierzbicki A., Shupta A., Barmak O. Synthesis of model features for fake news detection using large language models. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume IV: Computational Linguistics Workshop* : CEUR-Workshop Proceedings, Lviv, Ukraine, 12–13 April 2024. Aachen, 2024. P. 50–65. URL: https://ceur-ws.org/Vol-3722/paper5.pdf

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

3. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* : Proceedings, Minneapolis, MN, USA, 2–7 June 2019. Stroudsburg, PA, USA, 2019. P. 4171–4186. URL: https://doi.org/10.18653/v1/N19-1423

4. Llama 2: Open foundation and fine-tuned chat models / H. Touvron et al. Ithaca, NY, USA : Cornell University, 2023. 77 p. (Preprint. Cornell University ; 2307.09288). URL: https://doi.org/10.48550/arXiv.2307.09288

5. GPT-4 technical report / OpenAI et al. Ithaca, NY, USA : Cornell University, 2024. 100 p. (Preprint. Cornell University ; 2303.08774). URL: https://doi.org/10.48550/arXiv.2303.08774

6. Радюк П. М. Підхід до прискорення навчання згорткової нейронної мережі за рахунок налаштування гіперпараметрів навчання. *Комп'ютерні системи та інформаційні технології*. 2020. Т. 2, № 2. С. 31–37. URL: https://doi.org/10.31891/CSIT-2020-2-5

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

8. Zepopo. Ukrainian news. *Kaggle.com*. URL: https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news/data

References

1. Wierzbicki A., Shupta A., Barmak O. Synthesis of model features for fake news detection using large language models. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume IV: Computational Linguistics Workshop* : CEUR-Workshop Proceedings, Lviv, Ukraine, 12–13 April 2024. Aachen, 2024. P. 50–65. URL: https://ceur-ws.org/Vol-3722/paper5.pdf

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

3. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* : Proceedings, Minneapolis, MN, USA, 2–7 June 2019. Stroudsburg, PA, USA, 2019. P. 4171–4186. URL: https://doi.org/10.18653/v1/N19-1423

4. Llama 2: Open foundation and fine-tuned chat models / H. Touvron et al. Ithaca, NY, USA : Cornell University, 2023. 77 p. (Preprint. Cornell University ; 2307.09288). URL: https://doi.org/10.48550/arXiv.2307.09288

5. GPT-4 technical report / OpenAI et al. Ithaca, NY, USA : Cornell University, 2024. 100 p. (Preprint. Cornell University ; 2303.08774). URL: https://doi.org/10.48550/arXiv.2303.08774

6. Radiuk P. M. Pidkhid do pryskorennia navchannia zghortkovoi neironnoi merezhi za rakhunok nalashtuvannia hiperparametriv navchannia. Kompiuterni systemy ta informatsiini tekhnolohii. 2020. T. 2, № 2. S. 31–37. URL: https://doi.org/10.31891/CSIT-2020-2-5

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

8. Zepopo. Ukrainian news. *Kaggle.com*. URL: https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news/data