УДК 004.8 DOI: 10.18372/2073-4751.84.20898

Мельничук О.В.,
orcid.org/0009-0005-2768-4312,
e-mail: 5213201@stud.kai.edu.ua

МЕТОДОЛОГІЯ БАГАТОАГЕНТНОЇ СИСТЕМИ ДЛЯ ДЕТЕКЦІЇ МАНІПУЛЯЦІЙ ТА ДЕЗІНФОРМАЦІЇ

Державний університет "Київський авіаційний інститут"

Вступ

Стрімкий розвиток штучного інтелекту (ШІ) та великих мовних моделей (LLM) докорінно змінює інформаційний ландшафт, створюючи безпрецедентні виклики для верифікації контенту в цифровому просторі. Масштаб проблеми досягає критичних значень: швидкість поширення неправдивої та маніпулятивної інформації в соціальних мережах перевищує правдиву у 6,0 разів, створюючи системні ризики для демократичних процесів, громадського здоров'я та національної безпеки.

Проблема набуває глобального масштабу: за даними досліджень, понад 70% користувачів України споживають новини через соціальні мережі. У таких умовах соціальні платформи стають основним каналом для координованих інформаційних операцій, що використовують мережі Sybil-акаунтів для маніпуляції суспільною думкою.

Поява великих мовних моделей, таких як GPT-4 та Gemini, суттєво загострила проблему дезінформації. ШІ-згенерована пропаганда виявилася не лише складнішою для виявлення, а й ефективнішою: вона отримує на 37% більше взаємодій, ніж контент, написаний людиною. Дослідження показують, що користувачі правильно визначали ШІ-генерований контент лише у 42% випадків.

Мета

Метою статті є розроблення методології багатоагентної системи верифікації інформації для детекції дезінформації та маніпулятивного контенту в цифровому просторі, дослідження принципів декомпозиції складних завдань верифікації на спеціалізовані компоненти з використанням малих мовних моделей (SML), а також оцінка ефективності запропонованого підходу порівняно з комерційними LLM.

Обмеження традиційних ML-методів

Традиційні методи машинного навчання для детекції дезінформації стикаються з новими викликами в епоху LLM, що проявляється в кількох ключових обмеженнях. До них належить обмежене розуміння контексту, оскільки класифікатори, що базуються на поверхневих лексичних ознаках, не здатні виявляти складні маніпулятивні техніки та контекстуальні викривлення.

Також існує сильна залежність від датасету, через що моделі, треновані на історичних даних, демонструють низьку ефективність на нових типах маніпуляцій та адаптивних стратегіях ботів.

Крім того, спостерігається разливість до LLM-генерованого контенту: текстові класифікатори на основі трансформерних архітектур, такі як RoBERTa, раніше успішно розпізнавали ботів, але з появою LLM цей підхід майже втратив ефективність.

Важливим недоліком є відсутність механізмів верифікації джерел, адже традиційні підходи не інтегрують перевірку надійності джерел та фактів, що є критичним для комплексної оцінки достовірності.

Обмеження використання LLM

Незважаючи на те, що комерційні LLM на кшталт GPT-4 та Gemini забезпечують високу якість аналізу, їх впровадження обмежене суттєвими ресурсними бар'єрами. Висока вартість інференсу моделей, що налічують сотні мільярдів параметрів, робить економічно неефективною масову верифікацію контенту як для інституційних, так і для приватних користувачів. Ситуація ускладнюється значними часовими затримками: базовий час відгуку становить 2–10 секунд, проте використання методів покращення якості, таких як Chain-of-Thoughts, може збільшувати очікування до кількох хвилин. Це фактично унеможливлює використання таких систем у задачах реального часу, зокрема для оперативного моніторингу інформаційного простору.

Окрім часових та фінансових витрат, існують і структурні недоліки використання універсальних комерційних моделей. Залежність від хмарних API нівелює повний контроль над процесом верифікації та створює потенційні загрози для конфіденційності оброблюваних даних. Водночас, наявність у моделях надлишкових знань, нерелевантних для детекції маніпуляцій, призводить до нераціонального використання обчислювальних потужностей порівняно зі спеціалізованими рішеннями.

Переваги використання SML

Ключовою перевагою малих мовних моделей (SLM) обсягом 4–12 мільярдів параметрів є їхня виняткова здатність до адаптації під конкретні прикладні задачі при збереженні повної автономності користувача. На відміну від універсальних комерційних гігантів, SLM дозволяють реалізувати стратегію глибокої спеціалізації: за допомогою методів ефективного донавчання (PEFT) базову модель можна швидко оптимізувати під специфічний домен, досягаючи точності, співмірної з GPT-4, але у вузькому сегменті. При цьому архітектурна легкість таких моделей робить їх доступними для широкого кола користувачів — від незалежних дослідників до малих організацій. Можливість локального розгортання на споживчих відеокартах (рівня RTX 3080/4080) нівелює потребу у дороговартісній інфраструктурі, забезпечуючи вартість інференсу в десятки разів нижчу за хмарні API та гарантуючи повну приватність даних безпосередньо на пристрої користувача.

Методологія багатоагентної системи

Розроблена система базується на принципі багатоагентності, яка декомпозує складний процес верифікації інформації на спеціалізовані компоненти. На відміну від монолітних підходів, що намагаються вирішити задачу в рамках єдиного запиту, багатоагентна система забезпечує модульність, спеціалізацію та можливість незалежної оптимізації кожного компонента.

У Таблиці 1 представлено концептуальну модель системи, яка ілюструє основні функціональні блоки та їх взаємозв'язки.

Таблиця 1. Концептуальна модель функціональних агентів системи верифікації

Агент	Призначення
Детекція маніпуляцій	Виявлення маніпулятивних технік у вхідному контенті
Витягування наративів	Ідентифікація ключових тверджень та наративів
Перевірка фактів	Верифікація фактичних тверджень через зовнішні джерела та веб-пошук
Фінальна верифікація	Синтеза результатів всіх агентів у структурований висновок з поясненнями

Така декомпозиція дозволяє використовувати різні моделі та підходи для кожного етапу, оптимізуючи співвідношення точності та вартості.

Система складається з чотирьох основних агентів, які працюють у комбінації паралельного та послідовного виконання для оптимізації часу обробки. Рисунок 1 демонструє загальний шлях даних під час роботи даної системи. Manipulation Classifier та Fact Checker запускаються паралельно для зменшення загального часу роботи системи. Після завершення класифікації маніпуляцій активується Narrative Extractor, який використовує результати класифікації для контекстуального витягування наративів. Фінальний Verifier агрегує результати всіх попередніх агентів для формування структурованого висновок.

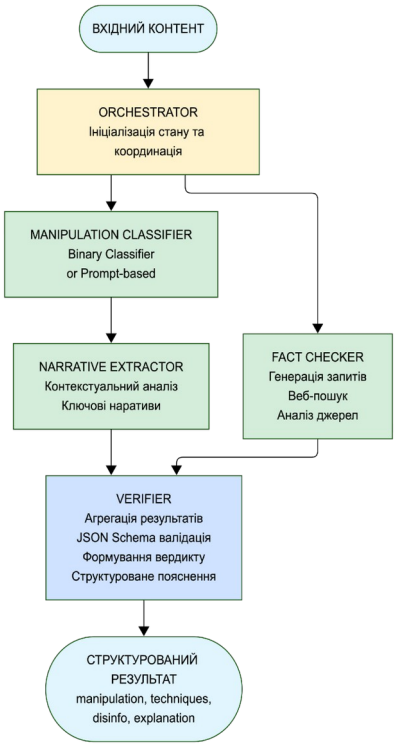

Рис. 1. Схема потоку даних у багатоагентній системі

Опис компонент системи

Процес верифікації інформації реалізується через взаємодію чотирьох ключових компонентів. Першою ланкою в ланцюзі обробки виступає Manipulation Classifier, який аналізує вхідний контент на предмет наявності маніпулятивних технік відповідно до таксономії UNLP 2025 [15]. Архітектура агента передбачає гнучку реалізацію цього модуля з підтримкою двох підходів: використання попередньо навчених трансформерів (наприклад, ModernBERT, RoBERTa) для швидкої бінарної класифікації або застосування Prompt-based класифікатора на базі основної моделі агентної системи. Останній метод використовує інженерію запитів зі структурованими шаблонами для ідентифікації підозрілих фрагментів тексту та їх подальшого зіставлення з категоріями таксономії.

Послідовно після етапу класифікації активується модуль Narrative Extractor, що спеціалізується на виокремленні ключових наративів. Для підвищення точності екстракції агент використовує контекстуальні промпти, які враховують результати попереднього етапу - зокрема, розраховану ймовірність маніпуляції та специфіку виявлених технік.

Паралельно з класифікатором маніпуляцій функціонує Fact Checker - найбільш ресурсоємний компонент, відповідальний за верифікацію фактів через зовнішні джерела. Така паралелізація дозволяє суттєво знизити загальну латентність системи. Алгоритм роботи агента складається з трьох етапів: генерації оптимізованих пошукових запитів на основі аналізу контенту, виконання веб-пошуку через зовнішні API (Tavily або DuckDuckGo) з фільтрацією джерел, та агрегації доказів. На фінальному етапі знайдені дані ранжуються за авторитетністю, формуючи матрицю перехресного аналізу (cross-analysis matrix), що містить як підтверджуючі, так і спростовуючі посилання для кожного факту.

Завершує цикл обробки модуль Verifier, який синтезує результати роботи всіх попередніх агентів у єдиний структурований висновок. На вхід цього компонента подається оригінальний контент, кількісна оцінка ймовірності маніпуляції, перелік ідентифікованих

технік, витягнуті наративи та результати фактчекінгу, що стає основою для фінального рішення системи. Вихідний формат включає:

- manipulation: бульове значення присутності маніпуляцій
- techniques: список виявлених технік
- disinfo: додаткові індикатори дезінформації
- explanation: структуроване пояснення висновку з посиланнями на джерела (grounding)
- confidence: впевненість системи у висновку (0.0 1.0)

Система використовує JSON Schema для валідації структури відповіді та fallback механізми для обробки некоректних відповідей LLM.

Управління станом системи

Для ефективної координації роботи агентів застосовується граф-орієнтований підхід, який дозволяє декларативно формалізувати залежності та автоматизувати оптимізацію виконання завдань. Фундаментом інформаційного обміну виступає централізована система управління станом, що інтегрує три ключові категорії даних: вхідну інформацію (оригінальний контент та метадані), конфігураційний шар (параметри моделей та умови виконання) та вихідні дані (агреговані результати роботи кожного компонента). Така організація процесу не лише забезпечує модульність і гнучкість системи, але й гарантує прозорість та можливість детального аудиту проходження даних на кожному етапі пайплайну.

Експерементальна валідація системи

Для оцінки ефективності розробленої системи було проведено експериментальне дослідження на спеціалізованому датасеті, підготовленому організацією Texty.org.ua для UNLP Workshop [15].

Вибірка для тестування включала 400 повідомлень з Telegram - каналів зібраних після 24.02.2022 (100 маніпулятивних та 100 нейтральних, 100 укр. мовою та 100 рос. мовою). Оцінювалася виключно коректність бінарної класифікації наявності маніпуляцій на фінальному етапі роботи Verifier Agent.

У дослідженні порівнювалися два типи моделей:

1. **Gemini 2.5 Flash** - комерційна хмарна модель з доступом через API.
2. **Gemma 3 4B** - локальна мала мовна модель (SLM), розгорнута через LM Studio.

У Таблиці 2 представлено результати порівняння моделей за основними метриками класифікації.

*Таблиця 2.*Порівняння ефективності детекції маніпуляцій

Модель	**Precision**	**Recall**	**F1-score**
Gemini Flash	0.67	1.00	0.80
Gemma 3 4B	0.58	0.90	0.71

На основі даних результатів можна стверджувати, що малі мовні моделі у У Таблиці 2 представлено результати порівняння моделей за основними метриками класифікації.

Обмеження дослідження

Незважаючи на перспективні результати, отримані в ході дослідження, поточна реалізація системи має низку обмежень, що окреслюють вектори подальшої роботи. Насамперед це стосується методології оцінки якості: експериментальна валідація на даному етапі фокусується виключно на метриках точності бінарної класифікації маніпуляцій. Поза

межами аналізу залишилися такі критичні аспекти, як інтерпретованість та зрозумілість згенерованих пояснень для кінцевого користувача, а також точність екстракції наративів. Для комплексного вирішення цієї проблеми необхідна розробка розширеного валідаційного датасету з еталонними поясненнями та імплементація підходу *LLM-as-a-Judge* для автоматизованої оцінки семантичної якості генерації.

Іншим суттєвим обмеженням є текстоцентричність запропонованого рішення. Система наразі не враховує мультимодальний характер сучасної дезінформації, зокрема маніпуляції в зображеннях, інфографіці, а також аудіо- та відеоконтенті (deepfakes). Проте закладена модульність системи створює фундамент для майбутньої інтеграції спеціалізованих мультимодальних агентів без необхідності перебудови ядра системи. Окрім того, дослідження зосереджено на контентному аналізі й не охоплює аналіз соціальних сигналів [17], таких як метадані акаунтів, історія активності та патерни розповсюдження інформації. Інтеграція агента для аналізу графових структур взаємодії та мережевих метрик є необхідним кроком для виявлення складних, скоординованих інформаційних операцій.

Напрямки подальшого розвитку

Ключовий напрямок розвитку системи передбачає застосування Parameter Efficient Fine-Tuning для сторення LoRa адаптерів [13] для компонентів: Manipulation Classifier та Narrative Extractor.

LoRA адаптер для Manipulation Classifier спеціалізується на детекції маніпулятивних технік у новинному контенті. Провести тренування на розширеному датасеті з понад 6000 анотованими прикладами, що забезпечить покращення точності класифікації з базовими моделями.

LoRA адаптер для Narrative Extractor адаптація до журналістських текстів та контенту соціальних мереж, що забезпечить консистентне витягування наративів у форматі 2-3 речень.

Економічноефективна оркестрація. Динамічна оркестрація агентів на основі аналізу результатів попередніх етапів як от аналізу джерела та ймовірності маніпуляції. Запити з низьким рівнем ризику можуть пропускати ресурсоємні етапи перевірки зменшуючи витрати та час роботи системи.

Висновки

Представлена методологія багатоагентної системи верифікації інформації, яка пропонує новий підхід до вирішення критичної проблеми дезінформації в цифровому просторі. Дослідження пропонує багатоагентний підхід для декомпозиції складних завдань верифікації інформації. Принципи спеці алізації агентів, паралелізму та економічної ефективності можуть бути застосовані в інших доменах штучного інтелекту.

Запропоновано підхід, що поєднує спеціалізовані агенти в єдину систему з використанням оркестрації паралельної та послідовної обробки інформації. Такий підхід потенційно забезпечує економію вартості порівняно з використанням API комерційних LLM.

Система забезпечує незалежну оптимізацію кожного компонента та легку інтеграцію нових функціональних можливостей через систему управління станом та гнучку конфігурацію.

Практична застосовність системи орієнтована на практичне використання медіаорганізаціями, фактчекінговими агентствами, державними установами завдяки модульній ар хітектурі та можливості локального розгортання.

Подальша розширена експериментальна валідація дозволить підтвердити ефективність запропонованих рішень.

Література

1. *Vosoughi S., Roy D., Aral S.* The spread of true and false news online. Science. 2018. Vol. 359, no. 6380. P. 1146–1151. DOI: 10.1126/science.aap9559.

2. Ukrainians Increasingly Rely on Telegram Channels for News and Information During Wartime. USAID-Internews. 2024. URL: https://internews.org/ukrainians-increasingly-rely-on-telegram-channels-for-news-and-information-during-wartime/ (дата звернення: 28.12.2025).

3. *Melnychuk O.* Social Media Sybil Detection in the Age of AI-Generated Content: A Literature Review. Кібербезпека: освіта, наука, техніка. 2025. Вип. 1, № 29. DOI: 10.28925/2663-4023.2025.29.885.

4. *Yang K.-C., Menczer F.* Anatomy of an AI-powered malicious social botnet. Journal of Quantitative Description: Digital Media. 2024. Vol. 4. DOI: 10.51685/jqd.2024.icwsm.7.

5. *Radivojevic A., Clark J. H., Brenner P. C.* LLMs Among Us: Generative AI Participating in Digital Discourse. Proceedings of the AAAI Symposium Series. 2024. DOI: 10.1609/aaaiss.v3i1.31202.

6. *Sharma K., Qian F., Jiang H., Ruchansky N., Zhang M., Liu Y. Combating* Fake News: A Survey on Identification and Mitigation Techniques. ACM Transactions on Intelligent Systems and Technology. 2019. Vol. 10, no. 3. P. 1–42. DOI: 10.1145/3305260.

7. *Liu Y. et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. 2019. DOI: 10.48550/arXiv.1907.11692.

8. *Anderson K., Wilson R.* Evading Bot Detection in the Age of Large Language Models. IEEE Security & Privacy. 2024. URL: https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=8013 (дата звернення: 28.12.2025).

9. *OpenAI.* GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. 2023. DOI: 10.48550/arXiv.2303.08774.

10. *Gemini Team.* Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805. 2024. DOI: 10.48550/arXiv.2312.11805.

11. *Zhang W., Kumar R., Thompson A.* Cost-Effective AI: Comparing Local and Cloud-Based Language Models for Production Workloads. Proceedings of the International Conference on Machine Learning. 2024. URL: https://proceedings.mlr.press/ (дата звернення: 28.12.2025).

12. *Lu Z. et al.* Small Language Models: Survey, Measurements, and Insights. arXiv preprint arXiv:2409.15790. 2024. DOI: 10.48550/arXiv.2409.15790.

13. *Hu E. J. et al.* LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685. 2021. DOI: 10.48550/arXiv.2106.09685.

14. *Gautam A.* Multi-agent Systems for Misinformation Lifecycle: Detection, Correction And Source Identification. Proceedings of the ICWSM Workshop. 2025. URL: https://arxiv.org/abs/2505.17511 (дата звернення: 28.12.2025).

15. *Romaniuk M., Smetanin S., Shypilo M. et al.* Shared Task on Media Manipulation Detection in Ukrainian. Proceedings of the 4th Workshop on Ukrainian Natural Language Processing (UNLP). 2025. URL: https://unlp.org.ua/ (дата звернення: 28.12.2025).

16. Detector Media Manipulation Techniques with Examples. UNLP Workshop. URL: https://github.com/unlp-workshop/unlp-2025-shared-task/blob/main/data/techniques-en.md (дата звернення: 28.12.2025).

17. *Da San Martino G., Cresci S., Barrón-Cedeño A., Yu S., Di Pietro R., Nakov P. A* Survey on Computational Propaganda Detection. Proceedings of the Twenty-Ninth International Joint

Conference on Artificial Intelligence, IJCAI-20 / ed. by C. Bessiere. International Joint Conferences on Artificial Intelligence Organization, 2020. P. 4826–4832. DOI: 10.24963/ijcai.2020/672

Мельничук О.В.

МЕТОДОЛОГІЯ БАГАТОАГЕНТНОЇ СИСТЕМИ ДЛЯ ДЕТЕКЦІЇ МАНІПУЛЯЦІЙ ТА ДЕЗІНФОРМАЦІЇ

Стрімкий розвиток генеративного штучного інтелекту призвів до безпрецедентного зростання обсягів дезінформації, що значно ускладнює традиційні методи моніторингу медіапростору та вимагає впровадження нових автоматизованих інструментів верифікації. Проте ефективність існуючих систем виявлення маніпуляцій та дезінформації часто залежить від використання потужних комерційних великих мовних моделей, що пов'язане з високими експлуатаційними витратами, обмеженою доступністю та ризиками передачі конфіденційних даних на зовнішні сервери. У даній статті оцінюється потенціал використання малих малих мовних моделей як основи для побудови автономних систем протидії інформаційним маніпуляціям. Було запропоновано новий підхід до побудови системи верифікації, який базується на багатоагентній взаємодії, де складне завдання розбивається на підзадачі для спеціалізованих агентів: детектора маніпуляцій, аналітика наративів та фактчекера. Отримані результати експериментального порівняння локальних та хмарних моделей дають змогу зробити висновок, що запропонований підхід є більш безпечним та економічно ефективним, забезпечуючи високу точність аналізу при збереженні повного контролю над даними користувача.

***Ключові слова**: штучний інтелект; інформаційні технології; багатоагентні системи; верифі кація інформації; детекція маніпуляцій; великі мовні моделі.*

Melnychuk O.V.

METHODOLOGY OF MULTI-AGENT SYSTEM FOR MISINFORMATION AND DISINFORMATION DETECTION

The rapid growth of generative artificial intelligence has led to an unprecedented surge in disinformation, significantly complicating traditional media monitoring methods and necessitating the implementation of new automated verification tools. However, the effectiveness of existing manipulation and disinformation detection systems often relies on powerful commercial Large Language Models (LLMs), which entail high operational costs, limited accessibility, and risks associated with transferring confidential data to external servers. This paper assesses the potential of utilizing Small Language Models (SLMs) as a foundation for building autonomous systems to counter information manipulation. A proposed approach to constructing a verification system based on multi-agent interaction is presented, in which a complex task is decomposed into subtasks for specialized agents: a manipulation detector, a narrative analyst, and a fact-checker. The results of an experimental comparison of local and cloud-based models indicate that the proposed approach is more secure and cost-effective, achieving high analysis accuracy while maintaining full control over user data.

***Keywords**: artificial intelligence; information technology; multiagent systems; information verification; manipulation detection; large language models*

Стаття подана до редакції: 07/12/2025
Стаття прийнята до опублікування:19/12/2025
Стаття опублікована: 30/12/2025