УДК 004.42:004.8
DOI https://doi.org/10.36994/2788-5518-2025-01-09-18

МУЛЬТИАГЕНТНА СИСТЕМА ДЛЯ ВИЯВЛЕННЯ ІНФОРМАЦІЙНИХ МАНІПУЛЯЦІЙ В ОНЛАЙН-СЕРЕДОВИЩІ

Постовий О. В., *Національний університет «Львівська політехніка», Львів, Україна. https://orcid.org/0009-0002-3935-6912, o.postovyi@gmail.com*

Шкраб Р. Р., *ст. викл., Національний університет «Львівська політехніка», Львів, Україна. https://orcid.org/0000-0001-9813-2877, roman.r.shkrab@lpnu.ua*

Анотація. *Поширення інформаційних маніпуляцій становить значну загрозу для сучасного суспільства, особливо в контексті війни та політичної нестабільності, з якою стикається Україна. У статті досліджено сучасні методи автоматизованого виявлення, аналізу й оцінювання достовірності новинного контенту. Запропоновано новий, мультиагентний підхід, де спеціалізовані програмні агенти, чий потік виконання керується великими мовними моделями, беруть на себе чітко визначені різні ролі для всебічного дослідження актуального новинного контенту. Використання декількох агентів, що конкурують між собою та намагаються проаналізувати різні погляди на новинний контент, продемонструвало ефективність мультиагентного підходу в галузі дослідження інформаційного простору й виявлення інформаційних маніпуляцій. Система інтегрує великі мовні моделі для ретельного виконання широкого спектру завдань, пов'язаних із глибоким аналізом текстового контенту, розумінням контексту, факт-чекінгом, формуванням обґрунтованих висновків і генерацією звітів, а також використовує передові бібліотеки для оркестрації та відстеження дій агентів, що керуються мовними моделями. Для забезпечення всебічного аналізу новинного контенту система звертається до різноманітних зовнішніх ресурсів: пошукових систем для знаходження релевантної інформації в інтернеті, новинних агрегаторів для доступу до статей передових ЗМІ, платформ для збору текстових даних із соціальних мереж. Отримані результати підтверджують, що впровадження великих мовних моделей та агентно-орієнтованого підходу в рамках N-рівневої архітектури дає змогу підвищити рівень аналізу інформаційного середовища й удосконалити процес виявлення інформаційних маніпуляцій.*

Метою статті є визначення необхідності використання мультиагентних систем для масштабування й поглиблення методів виявлення інформаційних маніпуляцій.

Ключові слова: *дезінформація, інформаційні маніпуляції, штучний інтелект, великі мовні моделі, мультиагентні системи, аналіз новин, Langchain, Langgraph, виявлення фейків.*

MULTI-AGENT SYSTEM FOR DETECTING INFORMATION MANIPULATION IN THE ONLINE ENVIRONMENT

Oleksii Postovyi, *Lviv Polytechnic National University, Lviv, Ukraine. https://orcid.org/0009-0002-3935-6912, o.postovyi@gmail.com*

Shkrab Roman, *Sr. Lect., Lviv Polytechnic National University, Lviv, Ukraine. https://orcid.org/0000-0001-9813-2877, roman.r.shkrab@lpnu.ua*

Abstract. *The spread of information manipulation poses a significant threat to modern society, especially in the context of war and political instability faced by Ukraine. This paper investigates modern methods of automated detection, analysis, and evaluation of news content reliability. The paper proposes a new, multi-agent approach, where specialized software agents, whose execution flow is guided by large language models, take on well-defined different roles to comprehensively investigate relevant news content. The use of several agents competing with each other and trying to analyze different views on news content has demonstrated the effectiveness of the multi-agent approach in the field of information space research and detection of information manipulation. The system integrates large language models to thoroughly perform a wide range of tasks related to in-depth analysis of textual content, contextual understanding, fact-checking, drawing reasonable conclusions and generating reports, and uses advanced libraries to orchestrate and track the actions of agents guided by language models. To provide a comprehensive analysis of news content, the system uses a variety of external resources: search engines to find relevant information on the Internet, news aggregators to access articles from leading media outlets, and platforms to collect text data from social networks. The results obtained confirm that the introduction of large language models and an agent-based approach within the framework of the N-level architecture allows to increase the level of analysis of the information environment and improve the process of detecting information manipulation.*

The purpose of the article is to determine the need to use multi-agent systems to scale and deepen the methods of detecting information manipulation.

Key words: *disinformation, information manipulation, artificial intelligence, large language models, multi-agent systems, news analysis, Langchain, Langgraph, fake detection.*

Вступ

У сучасному світі Україна стикається з проблемою інформаційної війни, яка є частиною російської агресії. Інформаційні маніпуляції використовуються для впливу на громадську думку, дестабілізації суспільства та підриву демократичних процесів. Маніпулятивний контент активно поширюються через соціальні мережі, месенджери, новинні сайти й навіть традиційні медіа, створюючи загрозу інформаційній безпеці українського суспільства.

Одним із основних джерел маніпуляцій є пропагандистські ресурси, які поширюють вигідні наративи, спрямовані на формування викривленого уявлення про події в Україні та світі. Такі наративи часто містять емоційно забарвлені меседжі, спекуляції на соціально чутливих темах, викривлення історичних фактів і нав'язування недостовірних тверджень. Дослідження показують, що джерела маніпуляцій мають інший стиль написання, спрямований на легкість прочитання й поширення [1]. Дезінформація легша для обробки з погляду когнітивного зусилля (на 3% легша для читання, на 15% менш лексично різноманітна) і більш емоційна (у 10 разів більше покладається на негативний сентимент, на 37% більше апелює до моралі) (рис. 1), ніж надійні джерела [1].

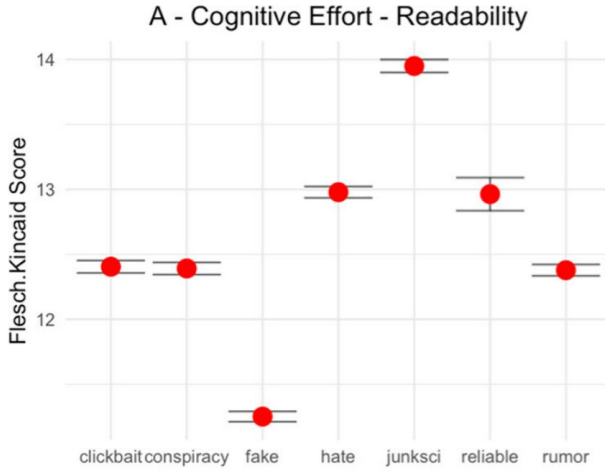

Рис. 1. Показники читабельності для різних типів інформаційних маніпуляцій

Чим легше читати новинний контент, тим імовірніше він належить до категорій клікбейту, теорій змови, фейкових новин або чуток [1]. Наприклад, фейкові новини в середньому на 8% простіші для читання й на 18% простіші в плані лексики, а також на 18% більше покладаються на негативні емоції та на 45% більше апелюють до моралі, ніж фактологічні новини [1]. Використання моральної та емоційної мови є одним із основних факторів комунікаційних стратегій маніпуляції [1]. Фактологічні новини, на відміну від них, є, по суті, нейтральними за емоційним забарвленням (-0.19). Для порівняння, розпалювання ненависті має найвище негативне значення емоційного забарвлення (-6.01), за ним ідуть фейкові новини (-3.51) і теорії змови (-3.41) (рис. 2).

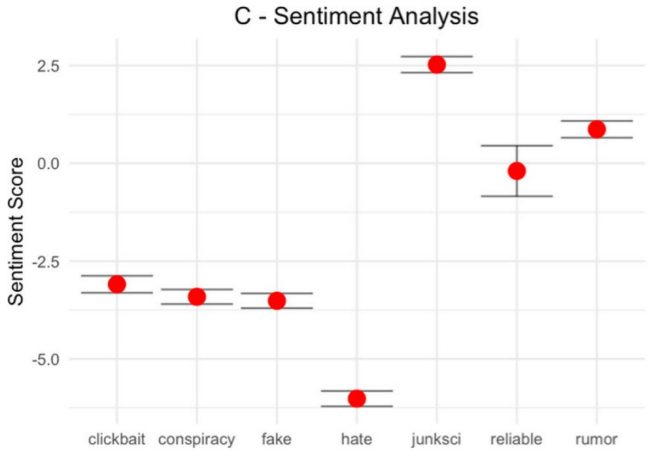

Рис. 2. Показники емоційного забарвлення для різних типів інформаційних маніпуляцій

Інформаційні маніпуляції в українському інформаційному просторі можуть стосуватися різних тем – від війни та міжнародної політики до економічних криз і внутрішньополітичних процесів. Окремо варто виділити маніпуляції, спрямовані на підрив довіри до державних інституцій, зокрема до армії, правоохоронних органів та уряду. Часто такі матеріали подаються під виглядом «аналітики» або «експертних думок» у соціальних мережах, що ускладнює їх викриття без спеціальних методів аналізу й додаткового пошуку інформації про джерело новини. Платформи на кшталт Facebook, Instagram, X, Telegram, TikTok і YouTube містять значний обсяг контенту, який важко модерується та використовується для інформаційних атак.

Виявлення інформаційних маніпуляцій в онлайн-середовищі є критично важливим завданням, що вимагає автоматизованих підходів через безпрецедентний обсяг контенту, який щодня публікується у вищезгаданих соціальних мережах, новинних ресурсах і блогах [1; 2].

Автоматизовані методи аналізу контенту дають змогу ефективно знаходити потенційно маніпулятивний контент шляхом аналізу стилістичних, лексичних і семантичних особливостей [1].

Одним із ключових етапів є аналіз тексту й класифікація інформаційних маніпуляцій. Дослідження вказують на можливість використання моделей машинного навчання для виявлення фейкових новин [12]. У публікації автори порівнюють Naïve Bayes, SVM і KNN для класифікації новин українською мовою [12]. Цей підхід передбачає наявність датасету з текстовими даними, препроцесинг тексту у вигляді токенізації та векторизації за допомогою TF-IDF, що дає змогу натренувати вищезгадані моделі машинного навчання та використати їх для визначення фейкової новини на раніше не баченому моделлю контенті [12].

Інші ж підходи передбачають використання агента з певною послідовністю етапів аналізу [13]. В основі такого агента використовується велика мовна модель, яка керує потоком управління програми. Агент обладнаний інструментами пошуку та вебскрапінгу для пошуку інформації про новину в зовнішніх джерелах. LLM аналізує зібрану інформацію, оцінює її достовірність, синтезує докази й, зрештою, виносить вердикт щодо правдивості новини, надаючи пояснення. Цей процес може бути ітеративним, де LLM уточнює запити та збирає додаткові дані за потреби. Стаття демонструє ефективність FactAgent на трьох реальних датасетах (Snopes, PolitiFact та GossipCop) і порівнює її з іншими методами виявлення фейкових новин. Результати показують, що цей агент перевершує інші підходи, зокрема ті, що ґрунтуються на навчених моделях, стандартних запитах і ланцюгах думок [13].

Архітектура проєктного рішення

У розробленій програмній системі вирішено використовувати агентно-орієнтований підхід з декількома кроками для формування фінального висновку.

Головною особливістю розробленої системи є використання мультиагентної взаємодії та створення конкуренції агентів для всебічного аналізу поданого контенту, на відміну від запропонованого підходу, що передбачає використання лише одного агента [13]. Також у цій системі є інструменти для аналізу соціальних мереж, таких як X (Twitter) і Telegram. Крім того, система може виявляти на помічати інформаційні маніпуляції наступними позначками: «Fake News», «Hate Speech», «Clickbait», «Conspiracy», «Rumours». За основу для класифікації взято наведені в дослідженні The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions типи маніпуляцій [1].

Незважаючи на використаний агентно-орієнтований підхід, за архітектуру взято N-рівневу систему. Сама ж взаємодія агентів відбувається на рівні бізнес-логіки.

Рівень бізнес-логіки включає такі сервіси:

• Сервіс аналізу новин. Центральний сервіс, що покроково реалізує процес аналізу новини. Він ініціює та координує роботу ШІ-агентів, а також відповідає за збереження результатів дослідження в БД.

• Агенти. За допомогою мовних моделей контролюють потік виконання програми й вирішують, які інструменти та сервіси використовувати. Типи створених агентів:

– Агент-«суддя» – головний агент, який керує процесом аналізу. Він отримує новину, ініціює роботу інших агентів і на основі їхніх звітів формує фінальний висновок, базуючись на звітах інших агентів. Агент використовує LLM o4-mini від OpenAI. Для реалізації цього типу агента, що виносить фінальний вердикт щодо інформаційних маніпуляцій у новині, у системі використано патерн створення агентів Orchestrator [9].

– Агент-аналітик. Є спільним класом для «прокурора» (шукає ознаки маніпуляції) й «адвоката» (шукає підтвердження правдивості новини). Кожен із цих агентів використовує LLM від Google (Gemini Flash 2.5) і набір інструментів для збору інформації. Принциповою різницею між «адвокатом» і «прокурором» є системний промпт, який задає тон для взаємодії аналітика із середовищем. Для реалізації агентів-аналітиків використано ReAct патерн [11]. У процесі розробки граф цього типу агента модифіковано для гарантованого написання фінального звіту й витягування використаних джерел для подальшого зберігання в базі даних.

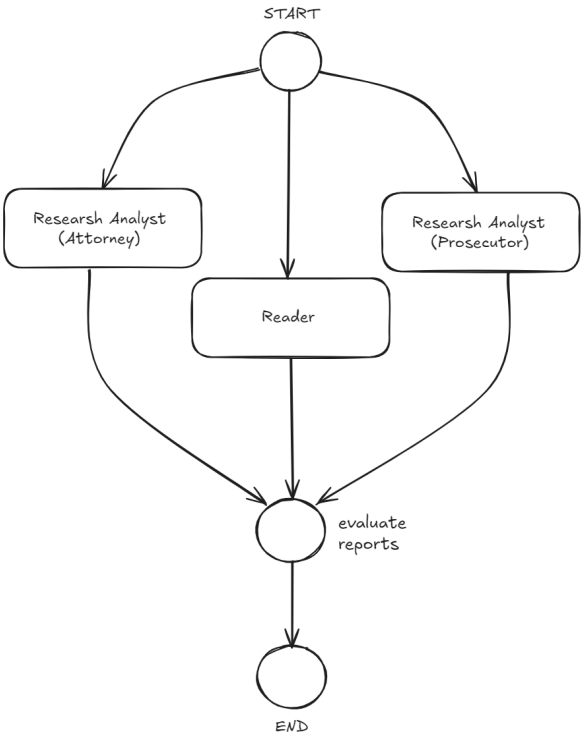

Рис. 3. Архітектура агента-«судді»

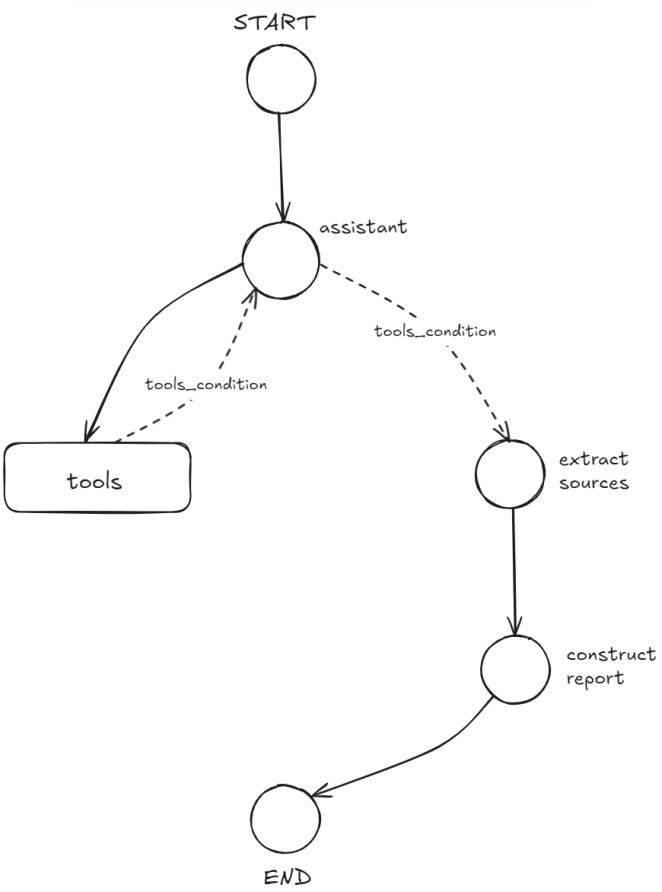

Рис. 4. Архітектура агента-«аналітика»

– Агент-«читач». Аналізує наданий текст новини безпосередньо для виявлення можливих типів маніпуляцій на основі визначених промптів для LLM. Нині реалізує лінійний потік виконання, проте в майбутньому може бути наділений спеціалізованими інструментами й керувати потоком виконання самостійно.

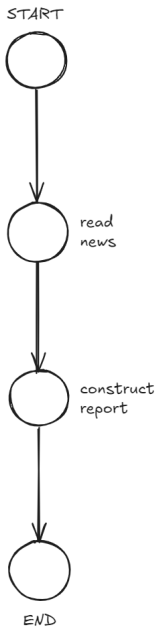

Рис. 5. Діаграма архітектури агента-«читача»

• Інструментарії для агентів. Сервіси, що містять різні типи інструментів для агентів. Для реалізації за основу взято концепцію Toolkit (інструментарія) із фреймворку Langchain [10]:

– Інструментарій для веб-пошуку використовує Tavily Search API й OpenAI Search API для пошуку релевантної інформації в Інтернеті.

– Інструментарій для скрапінгу новин взаємодіє з NewsAPI для пошуку новинних статей, пов’язаних з аналізованим контентом.

– Інструментарій для соціальних мереж призначений для збору даних із соціальних мереж, таких як Telegram та X, за допомогою платформи Apify.

Виконання досліджень

Для демонстрації роботи системи й оцінювання її ефективності у виявленні інформаційних маніпуляцій розроблено декілька сценаріїв дослідження. Кожен сценарій моделює типову ситуацію, з якою може зіткнутися користувач при аналізі новинного контенту.

Сценарій 1. Аналіз новини з ознаками «мови ворожнечі» (hate speech). Користувач надає системі текстовий фрагмент новини, яка містить ознаки мови ворожнечі й розпалювання ненависті (рис. 6).

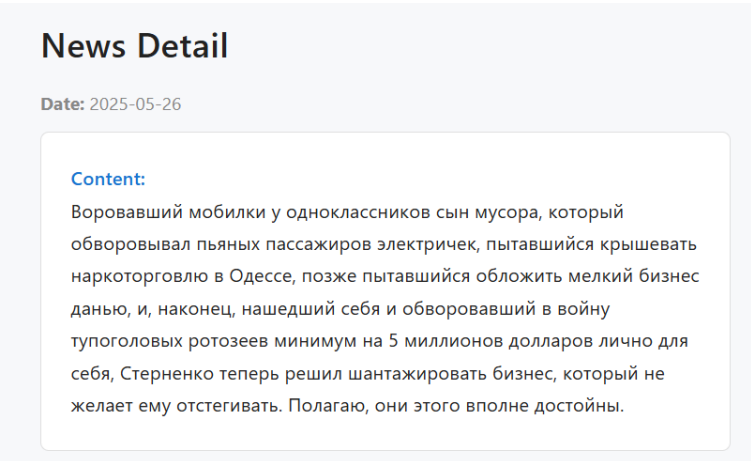

Рис. 6. Новина з мовою ворожнечі

Очікуваний результат: система класифікує новину як мову ворожнечі, надає докази її неправдивості й посилання на авторитетні джерела, що її спростовують.

Сценарій 2. Аналіз новини, що ймовірно поширює теорію змови (Conspiracy) або чутки (Rumours). Користувач надає текст новини, яка просуває необґрунтовану теорію змови або поширює неперевірені чутки. Така новина часто посилається на анонімні джерела, «експертів» без чіткої ідентифікації або будує складні, але логічно не підкріплені зв'язки між подіями (рис. 7).

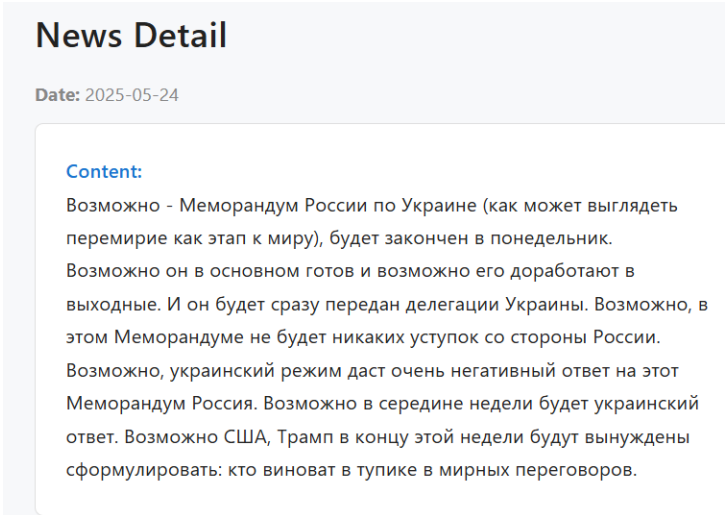

Рис. 7. Новина, що поширює слухи та/або теорії змови

Очікуваний результат: система описує новину, яка поширює необґрунтовану теорію змови або чутки, пояснюючи відсутність доказів і можливі мотиви поширення такої інформації, або в разі недостатніх доказів на користь присутності теорії змови пояснює, чому новина є правдивою.

Сценарій 3. Аналіз фактологічної новини з надійного джерела. Користувач надає системі текст новини, опублікованої відомим та авторитетним ЗМІ. Новина висвітлює реальну подію, подає факти збалансовано, посилається на конкретні джерела або офіційних осіб, уникає надмірної емоційності й сенсаційності (рис. 8).

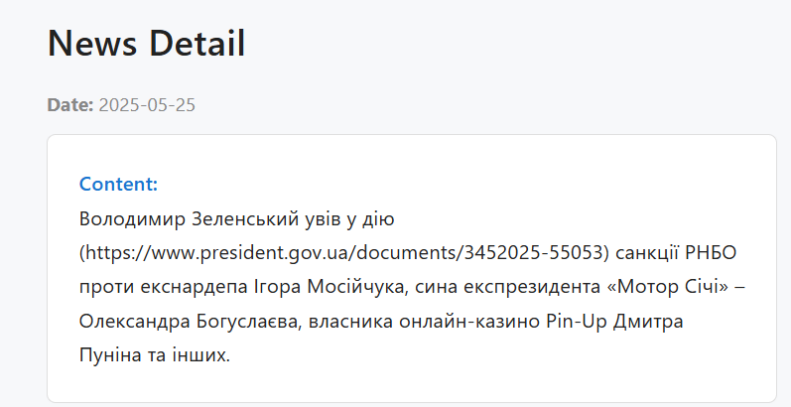

Рис. 8. Приклад новини без інформаційних маніпуляцій

Очікуваний результат: система описує новину як достовірну й таку, що не містить інформаційних маніпуляцій. У звіті наводяться посилання на інші авторитетні джерела, що підтверджують викладені факти, і зазначається відсутність виявлених маніпулятивних технік.

Метою дослідження було оцінити потенційну ефективність системи у виявленні різних типів інформаційних маніпуляцій, а також її здатність коректно ідентифікувати фактологічно достовірні новини. У ході дослідження також перевірено роботу системи у різних сценаріях.

На основі детального розбору кожного сценарію (фейкові новини, клікбейт, теорії змови/чутки, фактологічна новина) система продемонструвала високу ефективність у досягненні поставлених цілей.

У випадку сценарію 1 (розпалювання ненависті) система успішно ідентифікувала новину як таку, що містить мову ворожнечі. Ключову роль відіграв агент «Читач», який виявив ознаки Hate Speech

у самому тексті. Фінальний висновок агента «Судді» чітко аргументував розпалювання ненависті в тексті новини (рис. 9).

Рис. 9. Висновок про новину, що поширює мову ворожнечі

У процесу аналізу також успішно ідентифіковано маніпуляції Hate Speech і Fake News (рис. 10).

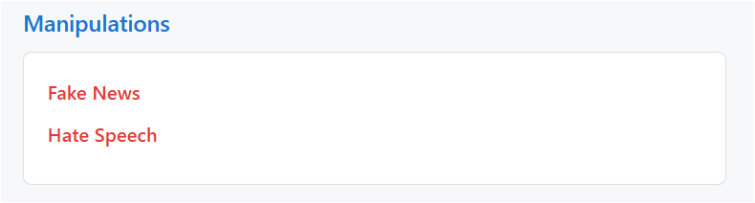

Рис. 10. Виявлені маніпуляції

Також помітно, що агент-«адвокат» для захисту тверджень у новині більше схилявся до проросійських джерел (рис. 11).

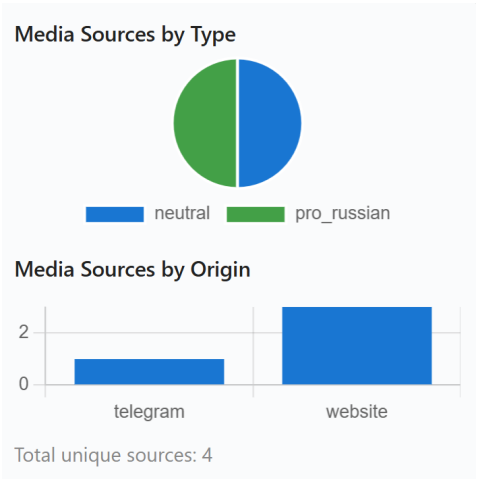

Рис. 11. Джерела, до яких звертався агент-«адвокат» для аналізу новини зі сценарію 1

Агент-«прокурор», у свою чергу, проаналізував більше джерел, серед яких були проукраїнські.

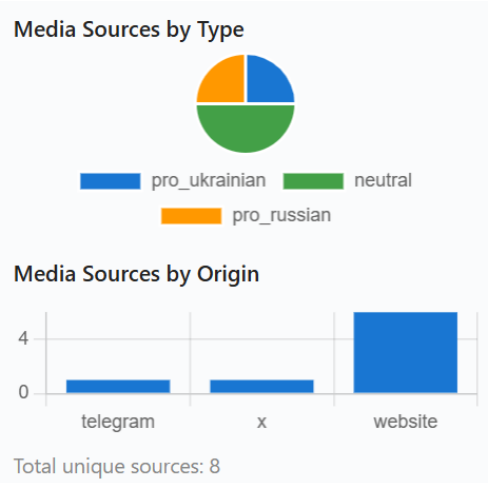

Рис. 12. Джерела, до яких звертався агент-«прокурор» для аналізу новини зі сценарію 1

У випадку сценарію 2 (теорія змови/чутки) система ефективно виявила ознаки чуток. Агент-«читач» указав на характерні риси Rumours. Його висновок став основою для висновку агента-«судді», оскільки агент-«читач» зміг виявити доволі необґрунтовані твердження, які не підтверджені ні агентом-«прокурором», ні агентом-«адвокатом» (рис. 13).

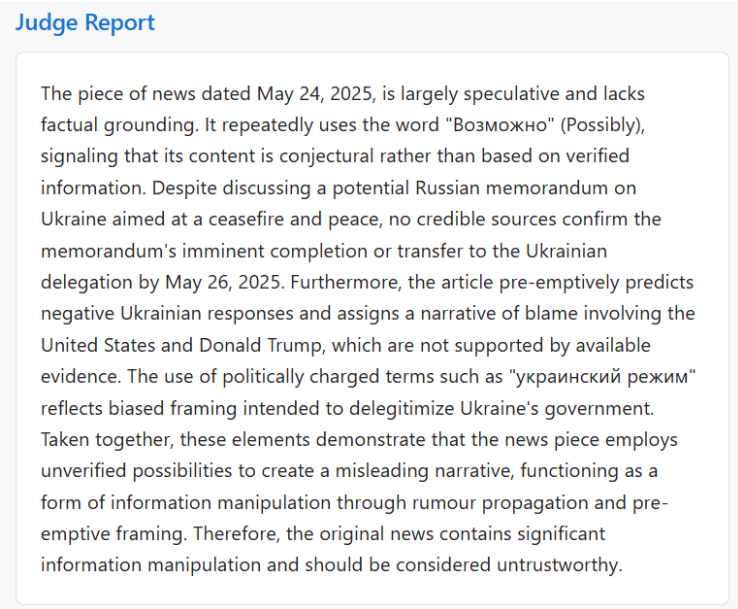

Рис. 13. Висновок системи про новину, що поширює теорії змови й/або чутки

У процесу аналізу також успішно ідентифіковано маніпуляцію Rumours (рис. 14).

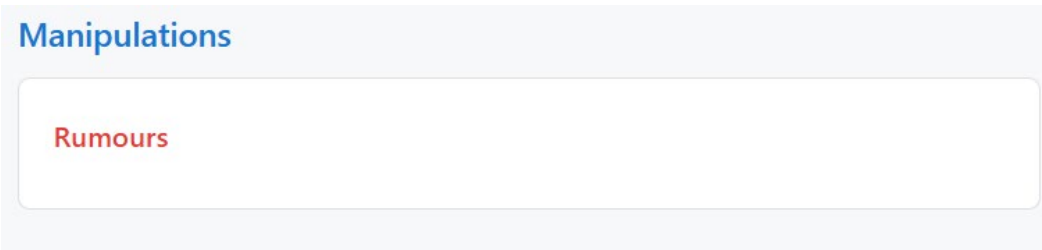

Рис. 14. Виявлені маніпуляції

Агент-«адвокат» намагався знайти підтвердження новини в нейтральних джерелах (рис. 15).

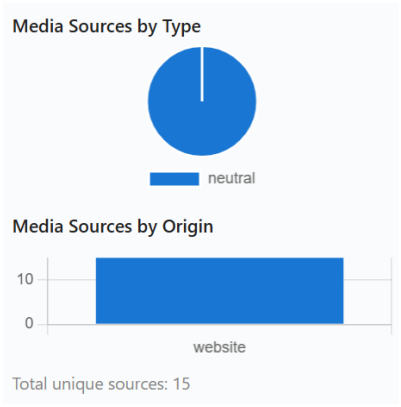

Рис. 15. Джерела, до яких звертався агент-«адвокат» для аналізу новини зі сценарію 2

Агент-«прокурор» теж використав лише нейтральні джерела (рис. 16).

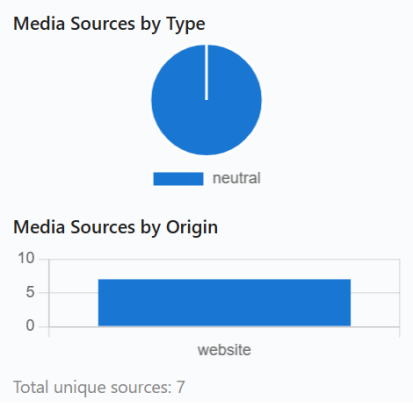

Рис. 16. Джерела, до яких звертався агент-«прокурор» для аналізу новини зі сценарію 2

У випадку сценарію 3 (фактологічна новина) система продемонструвала здатність правильно ідентифікувати достовірну інформацію. «Читач» не виявив маніпуляцій, «прокурор» зміг довести, що новина правдива, надавши численні підтвердження з авторитетних джерел, а «адвокат» не зміг знайти переконливих доказів неправдивості новини. Висновок «судді» підтвердив фактологічність і відсутність маніпуляцій (рис. 17).

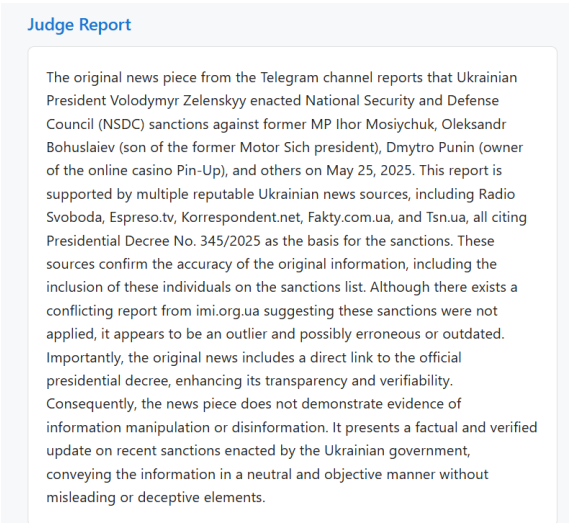

Рис. 17. Фінальний висновок системи для новини без інформаційних маніпуляцій

Агент-«адвокат» використав більше нейтральних джерел, хоча шукав інформацію також і в проукраїнських джерелах (рис. 18).

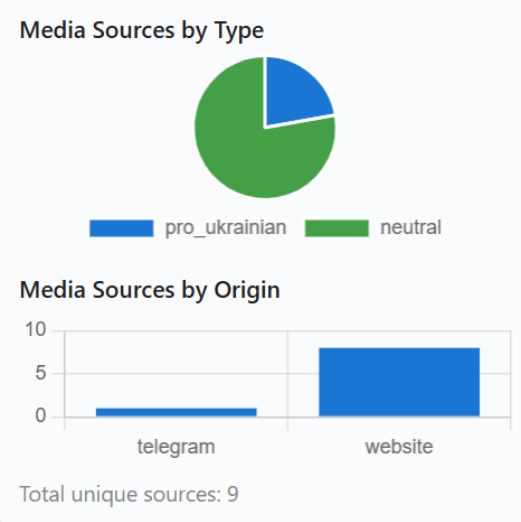

Рис. 18. Джерела, до яких звертався агент-«адвокат» для аналізу новини зі сценарію 3

Агент-«прокурор» використав більше проукраїнських джерел, ніж агент-«адвокат», проте теж переважно переглядав нейтральні джерела (рис. 19).

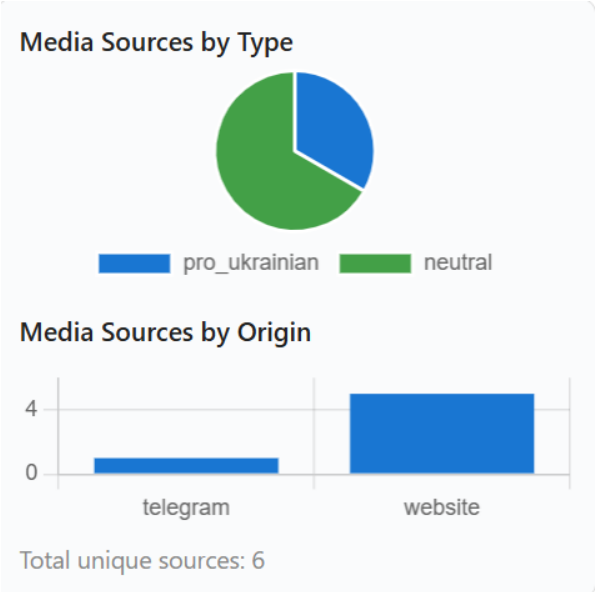

Рис. 19. Джерела, до яких звертався агент-«прокурор» для аналізу новини зі сценарію 3

Висновок

Дослідження дало змогу встановити, що мультиагентний підхід для виявлення інформаційних маніпуляцій є доволі ефективним. У всіх трьох сценаріях, які змодельовано в ході дослідження, система видала правильний результат, надавши детальний розбір новини з підкріпленням із різноманітних джерел.

Розроблена мультиагентна архітектура також дає змогу гнучко інтегрувати різноманітні нові джерела даних та інструменти, даючи змогу забезпечити ще глибше дослідження кожної новини.

Література

1. Carrasco-Farré C. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities & Social Sciences Communications*. 2022. Vol. 9. Art. 162. DOI: 10.1057/s41599-022-01174-9.

2. Golovchenko Y. Measuring the scope of pro-Kremlin disinformation on Twitter. *Humanities & Social Sciences Communications*. 2020. Vol. 7. Art. 176. DOI: 10.1057/s41599-020-00659-9.

3. Vicari R., Komendatova N. Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanities & Social Sciences Communications*. 2023. Vol. 10. Art. 332. DOI: 10.1057/s41599-023-01838-0.

4. Mantis Analytics. AI-Powered Early Warning System for Information Environment Threats : веб-сайт. URL: https://mantisanalytics.com/ (дата звернення: 07.05.2025).

5. Reality Defender. The Only Comprehensive Deepfake & AI-Generated Content Detection Platform : веб-сайт. URL: https://www.realitydefender.com/ (дата звернення: 07.05.2025).

6. LangGraph. Introduction. *LangGraph Documentation* : веб-сайт. URL: https://langchain-ai.github.io/langgraph/ (дата звернення: 08.05.2025).

7. Apify. Welcome to Apify Docs. *Apify Documentation* : веб-сайт. URL: https://docs.apify.com/ (дата звернення: 08.05.2025).

8. SQLAlchemy. SQLAlchemy 2.0 Documentation : веб-сайт. URL: https://docs.sqlalchemy.org/en/20/ (дата звернення: 08.05.2025).

9. LangGraph. Workflows and Agents. *LangGraph.js Documentation* : веб-сайт. URL: https://langchain-ai.github.io/langgraphjs/tutorials/workflows/#orchestrator-worker (дата звернення: 08.05.2025).

10. Langchain. Toolkits. *Langchain Python Documentation* : веб-сайт. URL: https://python.langchain.com/docs/concepts/tools/#toolkits (дата звернення: 12.05.2025).

11. Langchain. create_react_agent. *Langchain Python API Reference* : веб-сайт. URL: https://python.langchain.com/api_reference/langchain/agents/langchain.agents.react.agent.create_react_agent.html (дата звернення: 08.05.2025).

12. Джоші А., Павленко В. І. Порівняння методів машинного навчання для визначення фейкових новин. *Інфокомунікаційні та комп'ютерні технології*. 2024. Vol. 1, № 07. С. 46–55. DOI: 10.36994/2788-5518-2024-01-07-06.

13. Li X., Zhang Y., Malthouse E.C. Large Language Model Agent for Fake News Detection. arXiv. 2024. URL: https://arxiv.org/html/2405.01593v1 (дата звернення: 01.05.2025).

References

1. Carrasco-Farré C. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities & Social Sciences Communications*. 2022. Vol. 9. Art. 162. DOI: 10.1057/s41599-022-01174-9 [in English].

2. Golovchenko Y. Measuring the scope of pro-Kremlin disinformation on Twitter. *Humanities & Social Sciences Communications*. 2020. Vol. 7. Art. 176. DOI: 10.1057/s41599-020-00659-9 [in English].

3. Vicari R., Komendatova N. Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanities & Social Sciences Communications*. 2023. Vol. 10. Art. 332. DOI: 10.1057/s41599-023-01838-0 [in English].

4. Mantis Analytics. AI-Powered Early Warning System for Information Environment Threats : website. URL: https://mantisanalytics.com/ (accessed: 07.05.2025) [in English].

5. Reality Defender. The Only Comprehensive Deepfake & AI-Generated Content Detection Platform : website. URL: https://www.realitydefender.com/ (accessed: 07.05.2025) [in English].

6. LangGraph. Introduction. *LangGraph Documentation* : website. URL: https://langchain-ai.github.io/langgraph/ (accessed: 08.05.2025) [in English].

7. Apify. Welcome to Apify Docs. *Apify Documentation* : website. URL: https://docs.apify.com/ (accessed: 08.05.2025) [in English].

8. SQLAlchemy. SQLAlchemy 2.0 Documentation : website. URL: https://docs.sqlalchemy.org/en/20/ (accessed: 08.05.2025) [in English].

9. LangGraph. Workflows and Agents. *LangGraph.js Documentation* : website. URL: https://langchain-ai.github.io/langgraphjs/tutorials/workflows/#orchestrator-worker (accessed: 08.05.2025) [in English].

10. Langchain. Toolkits. *Langchain Python Documentation* : website. URL: https://python.langchain.com/docs/concepts/tools/#toolkits (accessed: 12.05.2025) [in English].

11. Langchain. create_react_agent. *Langchain Python API Reference* : website. URL: https://python.langchain.com/api_reference/langchain/agents/langchain.agents.react.agent.create_react_agent.html (accessed: 08.05.2025) [in English].

12. Dzhoshi A., Pavlenko V.I. Porivniannia metodiv mashynnoho navchannia dlia vyznachennia feikovykh novyn [Comparison of machine learning methods for detecting fake news]. *Infokomunikatsiini ta kompiuterni tekhnolohii* [Infocommunication and computer technologies]. 2024. Vol. 1. № 07. P. 46–55. DOI: 10.36994/2788-5518-2024-01-07-06 [in Ukrainian].

13. Li X., Zhang Y., Malthouse E.C. Large Language Model Agent for Fake News Detection. arXiv. 2024. URL: https://arxiv.org/html/2405.01593v1 (accessed: 01.05.2025) [in English].

Дата першого надходження рукопису до видання: 27.05.2025
Дата прийнятого до друку рукопису після рецензування: 23.06.2025
Дата публікації: 25.07.2025