УДК 004.8+004.056

Л. М. Куперштейн
В. О. Клименко

АРХІТЕКТУРА МУЛЬТИАГЕНТНОЇ СИСТЕМИ КОМПЛЕКСНОГО ВИЯВЛЕННЯ ФЕЙКОВОГО КОНТЕНТУ: ОГЛЯД КОМПОНЕНТІВ ТА ПІДХОДІВ

Вінницький національний технічний університет

Анотація

У тезах представлено огляд компонентів та підходів до побудови архітектури мультиагентної системи комплексного виявлення фейкового контенту на основі великих мовних моделей із доступом до зовнішніх інструментів. Розглянуто класичний конвеєр автоматизованої перевірки фактів із п'яти етапів та його відображення на ролі окремих агентів. Проаналізовано сучасні агентні системи – з оснащенням мовної моделі інструментами, багатоагентною дискусією та пояснюваними мультимодальними детекторами. Виокремлено типові компоненти архітектури: оркестратор, агенти-аналізатори за модальностями, модуль пошуку доказів, модуль агрегації та модуль генерації пояснень. Виявлено відкриті проблеми напряму, зокрема відсутність стандартизованої архітектури, що об'єднувала б різномодальні детектори як інструменти під керуванням єдиного оркестратора.

Ключові слова: мультиагентна система; великі мовні моделі; виявлення фейкового контенту; перевірка фактів; доповнена пошуком генерація; програмні агенти; інформаційна безпека.

Abstract

The theses present a review of components and approaches to building the architecture of a multi-agent system for comprehensive fake-content detection based on large language models with access to external tools. The classical five-stage automated fact-checking pipeline and its mapping onto the roles of individual agents are considered. Modern agentic systems are analysed – equipping a language model with tools, multi-agent debate, and explainable multimodal detectors. Typical architectural components are identified: an orchestrator, modality-specific analyser agents, an evidence-retrieval module, an aggregation module, and an explanation-generation module. Open problems of the field are highlighted, in particular the absence of a standardised architecture that would unify cross-modal detectors as tools under a single orchestrator.

Keywords: multi-agent system; large language models; fake content detection; fact-checking; retrieval-augmented generation; software agents; information security.

Вступ

Сучасні засоби штучного інтелекту досягли високої точності у виявленні фейкового контенту в межах окремих модальностей, проте здебільшого працюють як вузькоспеціалізовані, ізольовані детектори – окремо для тексту, зображень, відео чи аудіо. Такі рішення погано узагальнюються на реальні дані та не здатні узгоджено аналізувати повідомлення, що поєднують кілька модальностей одночасно. Водночас поява великих мовних моделей (ВММ) загострила проблему генерування дезінформації та водночас відкрила нові можливості для протидії їй [1].

Перспективним напрямом є побудова мультиагентних систем на основі великих мовних моделей із доступом до зовнішніх інструментів – пошуку доказів, перевірки фактів та доповненої пошуком генерації (RAG). Метою цих тез є систематизація компонентів та підходів до побудови архітектури мультиагентної системи комплексного виявлення фейкового контенту, а також виявлення відкритих проблем цього напряму.

Результати дослідження

Основою більшості систем автоматизованої перевірки фактів є конвеєр із п'яти етапів, що природно розподіляються між спеціалізованими агентами [2]:

1) виявлення та виокремлення тверджень, які потребують перевірки;
2) пошук доказів із зовнішніх джерел за допомогою RAG, пошукових та fact-checking API;

3) класифікація достовірності (підтверджує / спростовує / недостатньо даних);
4) агрегація рішень окремих агентів у підсумковий вердикт;
5) генерація зрозумілого людині пояснення.

Щодо оснащення мовної моделі інструментами, то ключовою ідеєю агентного підходу є надання великій мовній моделі набору інструментів. Так, система FactAgent на основі моделі без додаткового навчання поєднує внутрішні інструменти (аналіз стилю, мовних ознак, здорового глузду) та зовнішні (пошук через API, перевірка достовірності джерела за URL) і досягає точності близько 0,88 на наборі PolitiFact, перевершуючи донавчену модель BERT [3].

Інший підхід полягає в організації обміну аргументами між агентами: у системі MAD-Sherlock кілька мультимодальних агентів незалежно оцінюють пару «зображення–підпис», а потім дискутують до досягнення консенсусу, використовуючи зовнішній модуль пошуку. Це підвищує точність виявлення до 90,8 % на наборі NewsCLIPpings, причому сам механізм дискусії покращує базову модель із 70,7 % до 90,2 % [4].

Окремий напрям становлять детектори на основі мультимодальних великих мовних моделей. Система SNIFFER поєднує внутрішню перевірку узгодженості зображення й тексту із зовнішньою перевіркою за знайденими доказами та формує не лише вердикт, а й його обґрунтування, досягаючи точності 88,4 % на наборі NewsCLIPpings [5].

Узагальнення розглянутих систем дозволяє виокремити характерні складники: оркестратор на основі ВММ, агенти-аналізатори за модальностями, модуль пошуку доказів (RAG), модуль агрегації рішень та модуль генерації пояснень. Перспективним є залучення спеціалізованих детекторів – виявлення дипфейків, антиспуфінгу мовлення, виявлення згенерованого тексту – як зовнішніх інструментів, що викликаються оркестратором за потреби.

Відкриті проблеми. Попри обнадійливі результати, напрям має низку невирішених задач: відсутність стандартизованої архітектури, яка об'єднувала б різномодальні детектори як інструменти під керуванням єдиного оркестратора; обмежені можливості динамічного добору інструментів; вразливість до змагальних маніпуляцій із твердженнями та доказами; ризик галюцинацій ВММ під час верифікації; а також значні обчислювальні витрати й затримки у відповіді.

Висновок

Проведений огляд показує, що мультиагентні системи на основі великих мовних моделей із доступом до зовнішніх інструментів є перспективною архітектурою для комплексного виявлення фейкового контенту, оскільки поєднують міркування мовної моделі, пошук доказів та пояснюваність рішень. Такий підхід безпосередньо відповідає на виявлену раніше прогалину – брак комплексних мультимодальних засобів виявлення. Подальші дослідження доцільно спрямувати на розроблення уніфікованої архітектури, у якій оркестратор на основі ВММ узгоджено залучає спеціалізовані детектори за модальностями як інструменти, забезпечуючи водночас стійкість, пояснюваність та прийнятну обчислювальну вартість.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Chen C. Combating Misinformation in the Age of LLMs: Opportunities and Challenges / C. Chen, K. Shu // AI Magazine. – 2024. – Vol. 45, № 3. – P. 354–368. – DOI: 10.1002/aaai.12188.
2. Akhtar M. Multimodal Automated Fact-Checking: A Survey / M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, A. Vlachos // Findings of the Association for Computational Linguistics: EMNLP 2023. – 2023. – P. 5430–5448. – DOI: 10.18653/v1/2023.findings-emnlp.361.
3. Li X. Large Language Model Agent for Fake News Detection / X. Li, Y. Zhang, E. C. Malthouse // arXiv preprint arXiv:2405.01593. – 2024. – DOI: 10.48550/arXiv.2405.01593.
4. Lakara K. MAD-Sherlock: Multi-Agent Debate for Visual Misinformation Detection / K. Lakara, G. Channing, C. Rupprecht, J. Sock, P. Torr, J. Collomosse, C. Schroeder de Witt // arXiv preprint arXiv:2410.20140. – 2025. – DOI: 10.48550/arXiv.2410.20140.
5. Qi P. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection / P. Qi, Z. Yan, W. Hsu, M. L. Lee // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024. – P. 13052–13062. – DOI: 10.48550/arXiv.2403.03170.

Куперштейн Леонід Михайлович — к. т. н., доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця, email: kupershtein@vntu.edu.ua

Клименко Володимир Олександрович — аспірант групи F5-25аз, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: vovaklim2000@gmail.com

Leonid Kupershtein — PhD (eng), associated professor of information protection department, Vinnytsia National Technical University, Vinnytsia, email: kupershtein@vntu.edu.ua

Volodymyr Klymenko — Faculty of information technologies and computer engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: vovaklim2000@gmail.com