ФЕЙКОВІ НОВИНИ В ЕПОХУ ШІ: РИЗИКИ ТА ТЕХНОЛОГІЇ ВИКОРИСТАННЯ

Лубенець Є.О., Хайрова Н.Ф.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Наше дослідження спрямоване на створення синтетичного датасету фейкових/правдивих новин за допомогою генеративного штучного інтелекту та їх класифікації з використанням методів машинного навчання. Останнім часом ШІ-моделі такі як GPT, Bard, Lama здатні генерувати текст високої якості, що сприяє автоматизації, але й збільшує ризики поширення фейкових новин та дезінформації для. Для генерації шкідливого контенту використовують спеціальні запити. Щоб обійти правила регуляції мовної моделі, які не дозволяють створювати потенційно шкідливий контент використовують адвесаріальний запит (adversarial prompting). Також застосовують few-shot learning - наведення кількох прикладів того типу дезінформації, яку потрібно створити, або ж серію запитів (

1. Запит теми: задає тему для подальших кроків;
2. "Глибокий" запит (deep prompt): генерує фейкову новину;
3. Запит доповнення (News Augmentation Prompt): доповнює текст деталями (час, локація, джерело), щоб підвищити його реалістичність.) для створення новин різних категорій. [2] Наприклад фабрикація, зміна числових показників у статті, щоб навмисно перебільшити ситуацію (наприклад, "25 людей померли від холери в Нью-Делі" може бути переписано як "Десятки людей померли від холери в Нью-Делі"), misrepresentation (внесення упередженості до висновку, при цьому технічно зберігаючи оригінальну історію).[1] Численні експерименти показують, що мовні моделі здатні генерувати високоякісні фейкові новини та дезінформацію, які важко відрізнити від людських (GPT-3 є мечем з двома лезами: у порівнянні з людьми він може генерувати точну інформацію, яку легше сприймати, але водночас може створювати більш переконливу дезінформацію [3]), що створює серйозні ризики і вимагає розробки ефективних механізмів виявлення згенерованого штучним інтелектом контенту та встановлення етичних норм його використання.

Література:

1. Shrey Satapara та ін. Fighting Fire with Fire: Adversarial Prompting to Generate a Misinformation Detection Dataset. 2024р. URL: https://arxiv.org/pdf/2401.04481 (дата звернення 16.04.2025)
2. Yue Huang та ін. FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models. 2024р. URL: https://arxiv.org/pdf/2310.05046v2 (дата звернення 16.04.2025)
3. Giovanni Spitale та ін. AI model GPT-3 (dis)informs us better than humans. 2023р. Т. 11., № 15, С.1-19 URL: https://www.science.org/doi/10.1126/sciadv.adh1850 (дата звернення 16.04.2025)