УДК 004.8+004.056

Л. М. Куперштейн[1]
В. О. Сороколіт[1]
С.О. Прокопенко[2]

АНАЛІЗ МОЖЛИВОСТЕЙ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗАЦІЇ ФАКТЧЕКІНГУ

[1] Вінницький національний технічний університет;
[2]Харківський національний економічний університет ім. С. Кузнеця

Анотація
Виконано аналіз можливостей великих мовних моделей для автоматизації процесу фактчекінгу. Визначено перспективи подальшого дослідження.
Ключові слова: штучний інтелект, дезінформація, фейки, виявлення фейків, фактчекінг.

Abstract
The article analyzes the capabilities of artificial intelligence to detect fake news. Prospects for further research are identified.
Keywords: AI, disinformation, fakes, fake detection, fact-checking.

Вступ

У сучасному інформаційному середовищі поширення фейкової інформації стає все більшою проблемою, що порушує довіру людей до медіа та впливає на формування громадської думки. Будь хто і сам може перевірити інформацію за офіційними джерелами або за допомогою різноманітних сервісів для фактчекінгу. Проте людині досить складно читати та аналізувати інформацію з безлічі інформаційних джерел в усьому медіа-просторі. Для автоматизації фактчекінгу ефективним помічником можуть стати великі мовні моделі (Large Language Models - LLM). Вже сьогодні LLM вміють виконувати різноманітні задачі, в тому числі й аналізувати текстову інформацію на предмет достовірності та об'єктивності [1].

Метою даного дослідження є аналіз можливостей LLM для автоматизації процесу фактчекінгу.

Результати дослідження

Кожна людина хоче володіти достовірною інформацією. З розвитком інформаційних технологій, все частіше почала поширюватись недостовірна інформація у всьому медіа-просторі. Недостовірною вважається інформація, яка не відповідає дійсності або викладена неправдиво, тобто містить відомості про події та явища, яких не існувало взагалі або які існували, але відомості про них не відповідають дійсності (неповні або перекручені) [2].

Фейкові новини мають значний вплив на суспільство, викликаючи ряд соціальних, політичних та економічних наслідків. Їхня шкода може виявлятись у різних формах та сферах життєдіяльності людей.

Великі мовні моделі представляють собою одні з найсучасніших досягнень в області штучного інтелекту, спрямовані на розуміння та генерацію людської мови [3]. Вони базуються на глибинних нейронних мережах і використовують великі обсяги текстових даних для навчання, що дозволяє їм виконувати складні мовні завдання з високою точністю. Основна концепція LLM полягає в тому, щоб навчити модель прогнозувати наступне слово в реченні на основі контексту попередніх слів. LLM можуть виконувати широкий спектр завдань, а саме [1]:

- генерувати текст, що може бути корисно для написання статей, створення контенту або навіть написання коду;
- перекладати текст з однієї мови на іншу, враховуючи контекст і особливості кожної мови;
- визначати емоційний тон тексту, що може бути корисно для аналізу відгуків, коментарів та інших текстових даних;
- перевіряти інформацію на достовірність, завдяки своїй здатності обробляти великі обсяги тексту та співставляти інформацію з різними джерелами.

LLM мають значний потенціал для використання у фактчекінгу завдяки своїй здатності обробляти та аналізувати великі обсяги текстових даних. Вони можуть виконувати низку ключових завдань, що роблять їх ефективними інструментами для перевірки достовірності інформації:

- швидко аналізувати текстові матеріали, ідентифікуючи потенційно сумнівні або неправдиві твердження;
- порівнювати твердження з наявними достовірними джерелами інформації;
- враховувати контекст, у якому було зроблено твердження, що допомагає уникнути вирваних з контексту цитат та маніпуляцій;
- автоматизувати частину процесу перевірки, що значно скорочує час і зусилля, необхідні для виявлення дезінформації.

Порівняння технічних характеристик та можливостей найсучасніших LLM наведене в таблиці 1[4-6] .

Таблиця 1 – Технічні характеристики та можливості сучасних LLM

LLM	Розмір контекстного вікна	Пошук інформації в інтернеті	Можливість підключення файлів	Підтримка RAG	Видання знайдених джерел	Подання відповіді в заданій формі
GPT 3.5	4096 токенів	Ні	Так	Обмежена	Ні	Так
GPT 4	8192 токенів (базова), до 32K (Pro)	Так	Так	Так	Так	Так
Llama 2	4096 токенів	Ні	Так	Ні	Ні	Ні
Claude 3	200K токенів (до 1 млн для вибраних)	Ні	Так	Так	Так	Так
Mistral	32K токенів	Ні	Так	Так	Так	Так
Cohere	8K токенів	Ні	Так	Так	Так	Так
PaLM 2	4096 токенів	Так	Так	Так	Так	Так
Falcon 180B	3.5 трлн токенів	Ні	Так	Так	Так	Так
Gemini 1.5	1 млн токенів	Так	Так	Так	Так	Так
Mixtral 8x22B	141 млрд параметрів	Ні	Так	Так	Так	Так
Vicuna	125K розмов	Ні	Так	Ні	Так	Так

Проаналізувавши можливості різних LLM, можна виділити GPT 4, яка надає досить зручний, багатофункціональний та добре документований API [7]. Це найсучасніша модель від OpenAI з високою точністю, мультимодальністю (текст і зображення) та можливістю пошуку в інтернеті. Підтримує RAG для отримання зовнішніх даних і генерує точні відповіді.

Також варто зазначити, що OpenAI представила технологію агентів OpenAI – так звані GPTs [8]. Цей інструмент дозволяє користувачам створювати кастомізовані версії ChatGPT для різних застосувань. Користувачі можуть налаштовувати асистентів без необхідності програмування, вказуючи конкретні інструкції та правила поведінки. Це включає створення спеціалізованих відповідей, інтеграцію з інструментами та додатками, а також використання зовнішніх API для розширення функціональності. Такий підхід забезпечує гнучкість та масштабованість, дозволяючи адаптувати асистентів до специфічних потреб, на приклад – для автоматизації фактчекінгу.

Крім того, OpenAI API надає можливості створення власних віртуальних асистентів на основі мов програмування високого рівня [9]. Вони можуть використовувати не тільки власні знання, але і агрегувати результати глобального пошуку. Це є досить корисним інструментом для перевірки фактів на різних веб-медіапорталах. Також поєднання технології RAG із LLM дозволяє аналізувати власні дані, а саме ті, на яких LLM не навчалася. Це може бути дуже корисним для автоматизації фактчекінгу даних отриманих із соціальних мереж та месенджерів, наприклад, із Telegram-каналів. Вцілому на основі асистентів OpenAI можна побудувати свій власний сервіс для виконання задач та необхідним фукціоналом.

В подальшому дослідженні планується розробка автоматизованої системи перевірки інформації на достовірність із використанням саме OpenAI Assistant.

Висновки

У цій роботі було проаналізовано можливості використання LLM для автоматизації процесу фактчекінгу в контексті сучасного інформаційного середовища. Досліджено предметну область дезінформації. Виявлено, що великі мовні моделі мають значний потенціал для автоматизації процесу фактчекінгу. Проведено порівняльний аналіз найбільш поширених LLM. Вони здатні ефективно аналізувати великі обсяги тексту, ідентифікувати фейкову інформацію та порівнювати дані з різними джерелами. Це робить LLM важливими інструментами в боротьбі з дезінформацією. На основі проведеного аналізу розроблено перспективний напрямок подальших досліджень, що передбачає використання OpenAI Assistant.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. LLM: що це таке і які відкриває можливості? URL: https://aw.club/global/uk/blog/what-are-llms-and-what-opportunities-do-they-offer (date of access: 09.05.2024).
2. Відповідальність за розповсюдження недостовірної інформації. Міністерство юстиції України. URL: https://minjust.gov.ua/m/vidpovidalnist-za-rozpovsyudjennya-nedostovirnoi-informatsii (дата звернення: 09.05.2024).
3. Що таке велика мовна модель? URL: https://thetransmitted.com/adlucem/shho-take-velyka-movna-model-large-language-model-llm/ (дата звернення: 11.05.2024).
4. Найкращі LLM з відкритим кодом: що це таке, огляд топових великих мовних моделей. Apix-Drive. URL: https://apix-drive.com/ua/blog/reviews/najkrashi-llm-z-vidkritim-kodom (дата звернення: 12.05.2024).
5. Google представив Gemini 1.5 Pro та 1.5 Flash – нові моделі, які можуть аналізувати різні форми медіа. Mezha.Media. URL: https://mezha.media/2024/05/15/google-gemini-1-5-pro-flash/ (дата звернення: 13.05.2024).
6. Mistral 7B LLM – Nextra. URL: https://www.promptingguide.ai/models/mistral-7b (date of access: 14.05.2024).
7. OpenAI API. URL: https://openai.com/index/openai-api/ (date of access: 16.05.2024).
8. OpenAI GPTs. URL: https://openai.com/index/introducing-gpts/ (date of access: 16.05.2024).
9. OpenAI Assistants. URL: https://platform.openai.com/docs/assistants/how-it-works/objects/ (date of access: 17.05.2024).

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

Сороколіт Володимир Олександрович — студент групи 1БС-20б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: sorokolitvovan@gmail.com@gmail.com

Прокопенко Сергій Олександрович – аспірант, Харківський національний економічний університет ім. С. Кузнеця, email: prokopenko.serhii@gmail.com

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

Sorokolit Volodymyr — student of group 1БС-20б, Faculty of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: sorokolitvovan@gmail.com

Prokopenko Serhii – PhD student, Semen Kuznets Kharkiv National University of Economics, email: prokopenko.serhii@gmail.com