

ЗАСТОСУВАННЯ LLM В ІНФОРМАЦІЙНІЙ СИСТЕМІ ДЛЯ ТЕМАТИЧНОГО АНАЛІЗУ КОРИСТУВАЦЬКИХ КОМЕНТАРІВ

Новаковський А.В., Яловега І.Г.

e-mail: anton.novakovskiy@nure.ua, iryna.ialoveha@nure.ua

Харківський національний університет радіоелектроніки,
кафедра прикладної математики
м. Харків, Україна

This work presents a prototype of information system for LLM-powered thematic analysis of user comments. The second phase of thematic analysis – codebook-based coding – is implemented as an auditable sequence of five operations with stored input, intermediate and output artifacts. A PoC was run on the set of 50 Reddit comments using two models (GPT-4o-mini, GPT-5.2). Results were compared with a researcher-created golden coded set. The paper acknowledges the limitations of a single-coder gold standard and outlines next steps, including involving independent coders and scaling to larger datasets.

Метод тематичного аналізу [1] широко використовується в дизайн-дослідженнях, зокрема для аналізу користувацьких коментарів. Його рефлексивна форма забезпечує гнучкість та чутливість до контексту проблеми. Використання книги кодів (codebook) підтримує методологічну строгість та відтворюваність результатів. Сучасні дослідження демонструють, що великі мовні моделі (LLM) значно прискорюють тематичний аналіз, автоматизуючи обробку даних (кодування, створення книги кодів, формування тем) та залишаючи людині контроль якості та інтерпретацію результатів [2,3]. Розробляються такі архітектурні рішення для тематичного аналізу за участю LLM, як ланцюжки послідовних запитів (chain-of-prompts) до моделей [4] та системи на основі декількох LLM-агентів із розподілом ролей між ними [5].

Сьогодні можна виділити декілька проблемних аспектів застосування LLM у тематичному аналізі: (1) складність підтримки гнучкої та ітеративної взаємодії дослідника з моделлю при аналізі великих за обсягом даних; (2) висока вартість масштабування підходів, що потребують значного залучення людини; (3) низька простежуваність (traceability) процесу та складність аудиту результатів (auditability) – не завжди можна встановити, з яких фрагментів даних випливає конкретне твердження у підсумковому звіті або на якому кроці виникла помилка.

Метою дослідження є розробка й апробація підходу до тематичного аналізу користувацьких коментарів за участю LLM, який підтримує обробку значних обсягів даних та ітеративну взаємодію дослідника з моделлю, а процедура кодування є формалізованою, відтворюваною та придатною до аудиту. Практичною реалізацією підходу є прототип інформаційної

системи, що розбиває кодування на послідовні кроки з фіксацією вхідних артефактів, параметрів запуску та проміжних результатів.

Розроблено та апробовано прототип інформаційної системи для тематичного аналізу користувацьких коментарів у взаємодії дослідника з LLM. У поточній версії реалізовано другу фазу тематичного аналізу – кодування за книгою кодів. Рішення підтримує різні вхідні дані (набір коментарів, книга кодів, інструкції для моделі) та різні параметри запуску (гранулярність аналізу, вибір LLM-моделі). Специфікація визначає тематичний аналіз як один з методів у межах дизайн-циклу, що закладає основу для подальшого розширення на інші етапи та методи дизайну.

Системний глосарій містить узгоджений набір базових понять, що забезпечує єдину мову опису методології, функціональних можливостей та артефактів. Він охоплює базові сутності (об'єкт, актор, операція, інструмент, запуск), а також концепти, специфічні для дизайн-процесу (проект, цикл, стадія, метод, процес, артефакт, роль). Метод тематичного аналізу представлено як нелінійний ітеративний процес, що складається з шести фаз: ознайомлення з даними, кодування, побудова кандидатів тем, перегляд тем, визначення та іменування тем, підготовка звіту.

У межах прототипу реалізовано другу фазу тематичного аналізу – кодування за книгою кодів. Вхідними артефактами для цієї фази є: набір користувацьких коментарів, книга кодів, інструкції для LLM, конфігурація запуску (визначає обсяг і гранулярність вхідних даних, а також вибір LLM-моделі). Процедuru кодування реалізовано як послідовність п'яти операцій: підготовка вхідних даних і контексту, формування запиту до LLM, отримання результатів у нормованому форматі, перевірка структури й обмежень, збереження вихідних артефактів і метаданих запуску. Основним вихідним артефактом фази є «розмітка коментарів кодами», що містить результати конкретного запуску.

Пілотну апробацію прототипу виконано на наборі коментарів з платформи Reddit. Вибірка містить 2 дискусійні гілки із загальною кількістю 50 коментарів на тему використання штучного інтелекту в UX/UI та дизайн-практиках. Для пілотної вибірки підготовлено книгу кодів. На її основі дослідник виконав розмітку коментарів – отримані результати надалі використано як еталонне кодування (golden coded set) для зіставлення з результатами запусків прототипу. Автоматизоване кодування виконано з використанням двох моделей: GPT-4o-mini та GPT-5.2. Отримані результати зіставлено з еталонним кодуванням на рівні «коментар → множина кодів» із розрахунком метрик точності (precision) та повноти (recall).

Отримані результати демонструють значну різницю між моделями: GPT-5.2 забезпечує високу точність (precision = 0.894) за помірної повноти (recall = 0.544), тоді як GPT-4o-mini зберігає прийнятну точність (precision = 0.779), але має суттєво нижчу повноту (recall = 0.218). Це означає, що компактна модель у режимі кодування за книгою кодів схильна

пропускати значну частину релевантних кодів, тоді як більш потужна модель відтворює їх істотно краще, зберігаючи водночас низький рівень «зайвих» призначень. Водночас ці результати слід інтерпретувати обережно – метрики відображають, наскільки кодування, виконане LLM, збігається з еталонним кодуванням, яке виконав дослідник. Такий підхід є зручним для порівняння результатів між собою, але не відображає міру «правильності» кодування. У практиці надійність кодування зазвичай підсилюють залученням кількох незалежних кодерів і перевіркою узгодженості їхніх рішень між собою. Крім того, тематичний аналіз, особливо рефлексивний, допускає кілька обґрунтованих інтерпретацій. Тому навіть за однакових даних можливі різні, але змістовно прийнятні варіанти кодування.

Пілотна апробація прототипу підтвердила життєздатність запропонованого підходу до проєктування інформаційної системи для тематичного аналізу за участю LLM. Запропоноване рішення підвищує простежуваність процесу та придатність результатів до аудиту: з'являється можливість зіставляти запуски між собою, відстежувати, на якому кроці виникають розбіжності, і пов'язувати підсумкові висновки з конкретними фрагментами даних. Подальші кроки включають: (1) розширення вибірки, повторення експериментів на різних наборах даних; (2) дослідження впливу параметрів запуску (гранулярність аналізу, обсяг контексту, інструкції) на баланс точності та повноти; (3) уточнення книги кодів, правил застосування та інструкцій для LLM на основі аналізу типових розбіжностей; (4) розширення прототипу на наступні фази тематичного аналізу; (5) залучення додаткових незалежних кодерів для підвищення надійності еталонного кодування.

Список використаних джерел:

1. Alhojailan M. Thematic Analysis: A Critical Review of Its Process and Evaluation. *West East Journal of Social Sciences*. 2012. 1. P. 39–47.
2. Dai S. et al. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. *EMNLP*. 2023. DOI: 10.18653/v1/2023.findings-emnlp.669.
3. Wang Q. et al, M. L. LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences. *CSCW*. 2025. 9(2). DOI: 10.1145/3711022.
4. Raza M. et al. LLM-TA: An LLM-Enhanced Thematic Analysis Pipeline for Transcripts from Parents of Children with Congenital Heart Disease. *arXiv preprint*. 2025. arXiv:2502.01620.
5. Qiao T. et al. Thematic-LM: A LLM-based Multi-agent System for Large-scale Thematic Analysis. *Web Conference 2025 (WWW '25)*. 2025. P. 649–658. DOI: 10.1145/3696410.3714595.