УДК 004.7

Башманівський М. О., магістрант,
Чижмотря О. В., ст. викладач,
Дмитренко І. А., ст. викладач
Державний університет «Житомирська політехніка»

ГІБРИДНА МЕТОДОЛОГІЯ АНАЛІЗУ КОНТЕНТУ REDDIT: ПОЄДНАННЯ КЛАСИЧНОГО NLP ТА ГЕНЕРАТИВНИХ МОДЕЛЕЙ ДЛЯ УЗАГАЛЬНЕННЯ ТА ВИЯВЛЕННЯ ТОНАЛЬНОСТІ

Аналіз Reddit, як масштабного джерела суспільних настроїв, ускладнений обсягами, динамічністю та неструктурованістю даних. Традиційні методи нездатні впоратися зі складністю мови, сарказмом та контекстуальними нюансами при спробах виявити тональність чи узагальнити дискусії, що вимагає нових, гнучких методологій аналізу.

Існуючі підходи мають суттєві обмеження. Класичні методи NLP, такі як статистичний аналіз (TF-IDF) чи лексикон-орієнтовані підходи до тональності , ефективні для виявлення частотних закономірностей, але демонструють низьку точність при роботі зі складними семантичними явищами та нездатні до абстрактного узагальнення дискусій. З іншого боку, сучасні великі мовні моделі (LLM) чудово розуміють контекст та можуть генерувати змістовні резюме, однак їх пряме застосування до всього масиву сирих даних є обчислювально дорогим та може призводити до втрати фактичної точності без належної "доказової канви".

Метою даної роботи є представлення гібридної методології аналізу, яка вирішує ці проблеми шляхом поєднання сильних сторін обох підходів. Запропонований аналітичний конвеєр використовує класичні NLP-алгоритми для швидкої попередньої обробки, фільтрації та виділення ключових статистичних ознак (n-грам, тем). Далі, ці структуровані дані разом із репрезентативними фрагментами тексту подаються до генеративної моделі, яка виконує високорівневі завдання: абстрактивне узагальнення (TL;DR) та глибоке семантичне визначення тональності. Такий гібридний підхід дозволяє досягти балансу між обчислювальною ефективністю, точністю та глибиною аналізу.

Запропонована методологія ґрунтується на послідовному аналітичному конвеєрі, що поєднує швидкість класичної обробки природної мови (NLP) з семантичною глибиною генеративних моделей. На першому етапі закладається NLP-фундамент. Після збору даних через офіційний Reddit API, весь текстовий контент проходить жорсткий процес попередньої обробки. Цей процес включає

нормалізацію (усунення HTML-артефактів, емодзі, вирівнювання регістру), визначення мови, токенізацію та лематизацію для зменшення варіативності словоформ. Важливою частиною цього етапу є фільтрація шумів: застосовуються евристики та швидкі класифікатори для виявлення та видалення спаму, дублікатів та низькоінформативних повідомлень.

Після очищення база нормалізованих текстів проходить етап статистичного аналізу для формування так званого "профілю" дискусії. На цьому кроці розраховується частотність n-грам (наприклад, зважених за TF-IDF), виділяються ключові фрази та сутності, а також оцінюються базові показники, такі як лексична різноманітність . Результатом цього фундаменту є структурований набір даних: ключові теми, статистичні розподіли та "доказова канва", що складається з найбільш репрезентативних фрагментів тексту. Цей структурований профіль є основою для наступного, інтерпретаційного етапу.

Отримані структуровані дані використовуються для формування "розумного" запиту (промпту) до локальної великої мовної моделі. Замість того, щоб подавати на вхід весь масив сирого тексту, модель отримує комбінований запит, що містить ключові NLP-показники, такі як теми чи ключові фрази, та репрезентативні приклади оригінальних коментарів . При вирішенні задачі узагальнення, LLM, спираючись на надану "доказову канву", генерує змістовне резюме (TL;DR), що зменшує ризик "галюцинацій" моделі . Для аналізу тональності, LLM виходить за межі простих класифікаторів, надаючи більш тонку оцінку емоційного фону та ідентифікуючи сарказм, що важко для класичних методів [1].

Класичний NLP надає LLM фактичну основу, що "заземлює" її генеративні здібності та підвищує точність узагальнень. Водночас LLM збагачує сухі статистичні дані якісною, людсько-читабельною інтерпретацією, надаючи користувачу готові аналітичні висновки та інсайти, а не лише графіки. Крім того, досягається значна обчислювальна ефективність, оскільки основна маса даних обробляється швидкими NLP-алгоритмами, а ресурсоємна LLM залучається лише на фінальному етапі для інтерпретації вже підготовлених даних.

Список використаних джерел:

1. Katta K. Analyzing User Perceptions of Large Language Models (LLMs) on Reddit: Sentiment and Topic Modeling of ChatGPT and DeepSeek Discussions. *arXiv preprint*, arXiv:2502.18513, 2025. URL: https://arxiv.org/abs/2502.18513