Звіт про валідацію автоматизованої обробки новин із застосуванням GPT на контрольній вибірці

У цьому звіті представлено результати валідації автоматизованої обробки новин із застосуванням моделей GPT та спеціально розроблених запитів. Метою валідації було оцінити точність результатів обробки новин штучним інтелектом на контрольній вибірці вручну закодованих новин.

Процедура охоплює кілька кроків обробки, що виконуються за допомогою запитів. Для валідації було обрано контрольну вибірку зі 100 новин, закодованих вручну, яка слугувала “золотим стандартом” для порівняння з отриманими результатами GPT.

Для оцінки якості обробки застосовувалась стандартна метрика – відсоток збігів із ручним кодуванням. Результати цієї перевірки дозволяють зробити висновки про придатність запитів та автоматизованого підходу загалом для аналізу усього масиву новин.

Короткий підсумок

Крок		Результат
Крок 1. Валідація відбору релевантних повідомлень		89%
Крок 2. Валідація вилучення організацій		88.9%
Крок 5. Валідація 10 індикаторів	Замовник	89%
	Виконавець	89%
	Дати проведення	93%
	Генеральна сукупність	93%
	Розмір вибірки	96%
	Метод вибірки	98%
	Похибка вибірки	100%

	Зважування	100%
	Метод опитування	98%
	Текст запитання	99%
Крок 7. Валідація тем дослідження		89%

Крок 1. Валідація відбору релевантних повідомлень

Загальний відсоток збігів становить 89%. Хибно позитивних відповідей - 13.8%; хибно негативних - 5.7%.

Виявлені проблеми:

- ШІ плутає соціологічне опитування з референдумом або національним опитуванням;
- не завжди коректно відділяє абстрактні згадки про опитування від конкретних.

Запит:

"Persona: Ти — уважний аналітик, що спеціалізується на перевірці джерел.

*Завдання: Проаналізуй наступний текст і визнач, чи згадуються в ньому результати *конкретного*, *вже проведеного* соціологічного опитування або дослідження (хоча б одного).*

Контекст (Правила): 1. Якщо згадується абстрактне опитування ('соціологи кажуть', 'опитування показують'), відповідай 'Ні'. 2. Якщо згадується майбутнє опитування ('ми плануємо опитати'), відповідай 'Ні'. 3. Якщо згадується конкретне опитування (є назва, організатор, дата або чіткі результати), відповідай 'Так'.

Формат: Відповідай лише одним словом: 'Так' або 'Ні'.

Текст повідомлення:"

Крок 2. Валідація вилучення організацій

Запит комбінований, але валідація здійснювалась лише для кроку щодо вилучення організацій. Загальний відсоток збігів - 88.9%. Хоча наведений у звіті запит є фінальною версією, було оцінено три різні варіанти запитів. Фінальний запит продемонстрував найвищу точність, тоді як два попередні варіанти показали нижчі результати - 80.1% та 66.9% відповідно.

Основною проблемою є хибні включення організацій, які не мають стосунку до проведення або замовлення соціологічного дослідження. Це зазвичай трапляється у великих за обсягом текстах, зокрема в аналітичних статтях, де згадка про власне опитування є незначною частиною матеріалу, тоді як у тексті можуть перелічуватися численні інші установи, які не пов’язані з дослідженням.

Запит:

"Ти — аналітичний асистент. Твоє завдання — вилучити з тексту дані про походження соціологічного дослідження.

КРОК 1. Знайди організації, які відіграють одну з двох ролей:

1. ВИКОНАВЕЦЬ (хто провів опитування, збирав дані, соціологічна служба/центр).

2. ЗАМОВНИК (хто ініціював, фінансував або на чиє замовлення робили опитування).

Критерії пошуку:

- Шукай фрази: 'на замовлення...', 'проведено компанією..., 'дослідження центру...', 'спільно з...'.

- Якщо організація лише ОПРИЛЮДНИЛА новину (наприклад, ЗМІ, телеканал, новинний сайт), але не є замовником чи виконавцем дослідження — НЕ включай її.

- Якщо знайдено більше двох організацій (наприклад, альянс замовників і виконавців) — обери дві найважливіші (першоджерела).

КРОК 2. Сформулюй тему опитування одним реченням.

ФОРМАТ ВИВОДУ (суворо два рядки):

Хто проводив: <Назва Виконавця>; <Назва Замовника>

Тема: <Тема опитування>

Якщо в тексті не вказано ні виконавця, ні замовника, поверни: 'Хто проводив:'

Більше ніяких пояснень."

Крок 3. Валідація фільтрації організацій

Фільтрація організацій застосовувалася на етапі пілотного дослідження як додатковий спосіб відкидання нерелевантних організацій. Однак після змін у коді її ефективність для різних запитів становила близько 50%. Тому цей крок було вилучено.

Крок 5. Валідація 10 індикаторів

Запит було застосовано двічі. Відтворюваність відповідей між двома спробами відрізняється за окремими пунктами, проте для жодного з індикаторів вона не опускається нижче 93%.

Для восьми пунктів (окрім “Замовника” та “Виконавця”) відсоток збігів становить не менше 93%, що свідчить про досить стабільну інтерпретацію цих елементів моделлю.

Найбільші проблеми трапляються під час визначення наявності/відсутності замовника та виконавця. Рівень збігів становить 89% для обох критеріїв. У більшості випадків розбіжності спричинені нечіткими, неповними або надмірно узагальненими формулюваннями у текстах новин, через що GPT складно однозначно встановити факт згадки цих критеріїв.

Запит:

"Ти — експерт з методології соціологічних досліджень. Твоє завдання — проаналізувати текст прес-релізу або звіту та визначити, чи містить він інформацію, що відповідає стандартам AAPOR (American Association for Public Opinion Research).

Для кожного з 10 пунктів нижче напиши 'так', якщо інформація наявна в тексті, або 'ні', якщо вона відсутня. Використовуй наведені визначення AAPOR для прийняття рішення:

...

ФОРМАТ ВІДПОВІДІ (суворо дотримуйся порядку):

Замовник дослідження: [так/ні]

Виконавець дослідження: [так/ні]

Дати проведення опитування: [так/ні]

Генеральна сукупність: [так/ні]

Розмір вибірки: [так/ні]

Метод вибірки: [так/ні]

Похибка вибірки: [так/ні]

Застосування вагових коефіцієнтів: [так/ні]

Метод проведення опитування: [так/ні]

Текст запитання: [так/ні]"

Крок 7. Валідація тем дослідження

Запит для визначення основної теми опитування застосовувався двічі до кожного повідомлення, що дозволило оцінити внутрішню надійність моделі. Надійність становить 99%, що свідчить про стабільність результатів за повторних ітерацій.

Загальний відсоток збігів із ручним кодуванням – 89%. Частка повідомлень, для яких модель повернула “None”, становить лише 3%.

Загалом результат можна вважати задовільним: набір тем є достатньо повним і добре охоплює зміст досліджень. Основна проблема пов’язана з опитуваннями, що лежать на перетині кількох тем. У таких випадках модель може вагатися між двома близькими темами або обирати тему, яка покриває лише частину змісту.

Запит:

"Classify the survey's primary topic from these categories:

paste(topics_definitions$Topic, collapse = "; ")

Definitions:

paste(paste0(topics_definitions$Topic, ": ", topics_definitions$Definition), collapse = "; ")

Survey description:

Return only the best-matching topic or 'None'.

Крок 8. Перевірка ідентифікації публікацій про однакові опитування (об’єднання у групи)

Для визначення оптимального способу автоматичного групування публікацій, що стосуються одного й того самого соціологічного опитування, було протестовано три підходи:

- трьохетапний (ідентифікація за посиланням; ідентифікація за ознаками; ручне об’єднання);
- одноетапний за ознаками зі “строгою” умовою до організацій;
- одноетапний за ознаками із “м’якою” умовою до організацій.

У всіх підходах ознаками були: однакова тема; дата публікації в медіа в межах 5 днів; одна спільна організація (або дві – для другого способу).

Трьохетапний метод вважався умовно еталонним, однак його застосування до всього масиву вимагало б значних часових ресурсів через необхідність ручної перевірки на фінальному етапі. Це зумовило пошук більш автоматизованого варіанту.

Під час аналізу була виявлена проблема: публікації від одного автора можуть з’являтися в один і той самий період, що створює ризик помилкового об’єднання. Це виникає у двох випадках:

- коли опитування згадується із суттєвою затримкою (наприклад, в аналітичних статтях);
- коли одна організація справді публікує два різні опитування в один часовий період.

Жоден із протестованих способів не повністю усуває вплив цих ситуацій. “Строгий” спосіб виявився менш чутливим до випадків публікації двох опитувань одночасно, оскільки є ймовірність, що друга організація зазвичай буде відрізнятися. Водночас цей спосіб сформував найменшу кількість груп через неповноту інформації в окремих новинах, де можуть бути зазначені не всі дотичні до дослідження організації. Перший і третій способи показали практично однакові результати, однак третій є повністю автоматизованим і не потребує ручної доробки.

З огляду на співвідношення якості та трудовитрат, було обрано третій спосіб ідентифікації.