

Методологія проєкту “Моніторинг медійного висвітлення соціологічних опитувань”

Мета дослідження: проаналізувати, які опитування потрапляють у медіа, хто їх замовляє і проводить, як подається інформація про опитування та які теми отримують найбільшу увагу.

Метод дослідження: автоматизований контент-аналіз новинних текстів.

Одиницею аналізу є одна новинна публікація. До вибірки включаються лише ті матеріали, у яких згадується конкретні результати соціологічного дослідження.

Географія збору: Україна.

Мова публікацій: українська та російська.

Часові межі: проєкт працює з масивами новин за один квартал.

Первинний масив відбирається за ключовими словами:

“опитування”, “соціологічне дослідження” та “дослідження громадської думки” в різних відмінках, з різними закінченнями та мовами написання.

До аналізу включені публікації з інтернет-ЗМІ, інформаційних агентств та друкованої преси. Матеріали цих джерел містять якісні текстові версії, тому їх можна коректно обробляти за допомогою автоматизованих методів і отримувати надійні результати.

На першому етапі аналізу з первинного масиву за допомогою ШІ відбираються лише ті публікації, які містять результати конкретних проведених соціологічних опитувань. Цей крок дозволяє відсіяти матеріали, де слова-маркери згадуються в іншому контексті (наприклад, анонси майбутніх опитувань, загальні роздуми про методологію досліджень, згадки опитувань без конкретних результатів тощо).

Ідентифікація організацій

У кожній публікації за допомогою ШІ за контекстом тексту і текстовими маркерами фіксувались замовник і виконавець опитування. Якщо організацій було більше, ніж дві, то обирались дві найважливіші.

Виконавцем опитування ми вважаємо організацію, що проводила опитування і збирала дані. Замовником опитування є організація, що ініціювала, фінансувала або на чие замовлення організація-виконавець проводила опитування.

Організації, що лише оприлюднили результати, але не замовником чи виконавцем опитування, ми відсіюємо.

Після фіксації назв ми проводили їхню нормалізацію. Спочатку застосовували базові текстові правила для очищення та уніфікації написання і використовували попередньо сформований [словник](#) із найпопулярнішими організаціями, їх найпоширенішими варіантами написання та стандартизованими відповідниками.

Організації поза словником групували і попарно порівнювали назви за допомогою ШІ, щоб уніфікувати однакові формулювання, написані по-різному. Далі формували перелік з найчастіше згадуваних назв і повторно перевіряли їх на збіги, також за допомогою ШІ. Після цього назви кластеризувалися, і для кожного кластера обиралася найкоротша та найбільш усталена форма як репрезентативна назва організації.

Тематична класифікація опитувань

Кожна публікація за коротким змістом публікації відноситься до однієї з **17 тем**, адаптованих на основі класифікації тем контенту Comparative Agendas Project. Перелік тем є фіксованим і заданим наперед. Comparative Agendas Project — це міжнародний дослідницький проєкт, який розробляє стандартизовану систему класифікації політичних та суспільних тем для порівняльного аналізу контенту в різних країнах.

Категорії охоплюють основні сфери: політика, міжнародні питання, економіка, соціальні теми, освіта, охорона здоров'я, довкілля, технології тощо.

Таблиця 1. Тематика опитувань та їх визначення

Категорія	Що охоплює
Довіра до політиків та інститутів, і врядування	Довіра до органів влади, політичних лідерів, державних інституцій, політичні настрої, врядування.
Зовнішні відносини / Міжнародні справи	Міжнародна дипломатія, зовнішня політика, відносини між країнами, геополітика, міжнародні організації.
Військо та оборона	Збройні сили, національна безпека, бойові дії, військова політика, ставлення до армії.
Економічний та бізнес-клімат	Економічні тенденції, бізнес-середовище, інвестиції, підприємництво, ринок.
Зайнятість і ринок праці	Рівень зайнятості, ринок праці, безробіття, умови праці, трудові права.
Соціальні проблеми та добробут	Бідність, нерівність, соціальний захист, демографія, вразливі групи, якість життя.
Охорона здоров'я	Медичні послуги, здоров'я населення, інфраструктура, страхування, захворювання.
Освіта	Школи, університети, якість освіти, грамотність, політика в освіті, інфраструктура.
Інфраструктура та міське планування	Транспорт, дороги, житло, розвиток міст, комунальні послуги, просторове планування.
Громадянська участь	Участь у виборах, громадська активність, волонтерство, взаємодія з владою.
Культура та ідентичність	Національна ідентичність, традиції, мова, історична пам'ять, культурні практики.
Правопорядок і судочинство	Злочинність, правоохоронна система, суди, безпека громадян, поліція, тюрми.
Довкілля та клімат	Екологія, зміна клімату, забруднення, охорона природи, стає використання ресурсів.
Медіа та інформація	Медіа, довіра до інформації, споживання новин, дезінформація, медіаграмотність.

Технології	Технології, цифровізація, інтернет, кібербезпека, інновації, діджитал-інфраструктура.
Енергетика	Політика в енергетиці, джерела енергії, ринки палива, відновлювана енергетика.
Міграція	Міграція, біженці, інтеграція, політика щодо іммігрантів, громадська думка про мігрантів.

Оцінювання дотримання стандартів розкриття інформації про опитування

Для кожної новинної публікації ми перевіряли, чи містить вона ключову інформацію, яка має бути присутні в матеріалах про результати соціологічних опитувань згідно зі стандартами розкриття AAPOR.

Перевірка здійснювалася автоматично: ШІ отримував перелік із 10 пунктів із їхніми визначеннями та аналізував повний текст новини, щоб визначити, чи міститься в ньому кожний з цих елементів.

До аналізованих елементів належать: замовник і виконавець дослідження, дати проведення опитування, опис генеральної сукупності, розмір вибірки, метод відбору респондентів, статистична похибка, застосоване зважування, метод опитування та точне формулювання запитання.

На основі цих 10 елементів був сформований *Індекс дотримання стандартів розкриття інформації про опитування*. Концепт включає три виміри:

- “Джерельна доброчесність” — наявність згадок про замовника та виконавця.
- “Статистична прозорість” — наявність згадок про генеральну сукупність, розмір та метод вибірки, похибку та зважування.
- “Інтерпретаційна точність” — наявність згадок про дати та метод опитування, а також текст запитань.

Кожен індикатор у своєму вимірі оцінюється як 1 або 0. Для кожного виміру обчислюється частка заповнених індикаторів: сума значень індикаторів ділиться на їхню кількість у цьому вимірі. У результаті формуються три окремі показники — по одному для кожного виміру.

Загальний індекс визначається як середнє арифметичне цих трьох показників, оскільки усі виміри мають однакову вагу. Після цього індекс класифікується так: низький [0–0.3), середній [0.3–0.6), високий [0.6–1].

Ідентифікація публікацій про одне й те саме опитування

Щоб визначити, чи кілька новин стосуються одного й того самого опитування, ми порівнювали їх за трьома ознаками:

1. вони мають однакову тему;
2. опубліковані в межах 5 днів;
3. містять хоча б одну спільну організацію (виконавця або замовника).

Новини об'єднувалися в одну групу, якщо збігалися принаймні два з цих трьох критеріїв. Таким чином формувалися набори публікацій про одне опитування, які можна було аналізувати як єдиний інформаційний сюжет.

Опісля ми визначаємо 10 найбільш поширюваних у медіа опитувань за аналізований період.

Перевірка якості

Для оцінки якості автоматичного кодування було відібрано 100 публікацій, які паралельно кодувалися ШІ та людиною-кодувальником. Такий контроль проводився на всіх етапах, де використовувалося автоматичне кодування: визначення релевантних новин, ідентифікація організацій, тематична класифікація та перевірка елементів стандартів розкриття.

Кодування людиною розглядалося як “золотий стандарт”. Результати ШІ порівнювалися з результатами ручного кодування, а якість оцінювали за показником точності — часткою випадків, у яких ШІ визначив інформацію правильно. Звіт про валідацію можна переглянути за [посиланням](#).

Валідація показала, що автоматизоване кодування загалом забезпечує високий рівень точності для більшості етапів. Відбір релевантних новин і вилучення організацій дали 89% збігів із ручним кодуванням. Оцінка 10 елементів стандартів розкриття варіювалася від 89% для замовника і

виконавця до 98–100% для більшості технічних параметрів (метод вибірки, похибка, зважування, метод опитування, текст запитань). Визначення тем показало 89% збігів і продемонструвало високу стабільність за повторних запусків.

Крок		Результат
Крок 1. Валідація відбору релевантних повідомлень		89%
Крок 2. Валідація вилучення організацій		89%
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%

Обмеження

- Дослідження охоплює інтернет-ЗМІ, інформагентства і пресу, але не включає телевізійні сюжети, Telegram/YouTube/соцмережі.
- Ми не можемо гарантувати, що у базу Looqme потрапляють всі новини з цих джерел.
- Відбір за ключовими словами може пропустити публікації, де результати опитування подаються без прямого маркування як “опитування”, “дослідження” тощо.
- Кожне опитування отримує тільки одну основну тему.
- Дослідження аналізує лише ті опитування, що з’явилися у медіа.

- Оцінка дотримання стандартів розкриття інформації фіксує лише наявність обов'язкових елементів, але не виявляє помилок чи маніпуляції в інтерпретації результатів.
- Теми не виводилися з масиву новин, а задавалися наперед за готовим списком.

Технічний додаток

Повний опис алгоритмів, логіки кодування, прикладів проблемних випадків і процедур перевірки якості подано у технічному додатку: [“Технічні деталі аналізу”](#).