

ПРИКЛАДНИЙ ІНСТРУМЕНТАРІЙ ШІ НА РІЗНИХ ЕТАПАХ СТВОРЕННЯ ДАТА-ІСТОРІЇ

Трансформація журналістики даних під впливом штучного інтелекту сьогодні досягла етапу, коли технології перестали бути просто допоміжними інструментами й перетворилися на цілісну екосистему, що охоплює весь життєвий цикл створення дата-історії. Сучасний журналіст більше не обмежений технічними бар'єрами написання складного коду для пошуку інформації або виснажливим ручним очищенням неструктурованих масивів. Натомість фокус змістився в бік управління інтелектуальними агентами, які здатні обробляти терабайти інформації – від текстових документів до супутникових знімків та відеоматеріалів – для виявлення закономірностей, раніше недоступних для людського ока.

Еволюція інструментарію ШІ дозволяє інтегрувати контекст і сенс безпосередньо в процес збору даних. Це означає, що дані більше не сприймаються як сухі цифри чи факти; вони стають частиною динамічного наративу, де кожен етап – від першого запиту до фінального інтерактивного графіка, мапи, візуалізації – підтримується алгоритмами, що розуміють структуру та взаємозв'язки всередині інформаційного поля.

Традиційні методи вебпошуку, що базувалися на жорстких правилах програмування, виявилися занадто вразливими до будь-яких змін у верстці вебсайтів. Тому на зміну їм прийшов семантичний скрейпінг, де ШІ-асистенти використовують логіку природної мови для ідентифікації та витягування потрібної інформації.

Для дата-журналістів багатьох розвинених країн існує такий інструмент як ScrapeGraphAI, що став флагманським і реалізує підхід, орієнтований на схему даних. Замість того, щоб вказувати шлях до конкретного елемента, журналіст використовує промпт на кшталт «Витягни всі назви департаментів та суми бюджетних асигнувань». Система автоматично аналізує структуру і формує валідовану відповідь на наданих чи знайдених файлах [1, с. 15].

Журналісти використовують такі інструменти для вилучення даних з урядових звітів, де таблиці часто приховані за розгорнутими рядками або складними навігаційними елементами. Використання інструментарію ШІ дозволяє системі самостійно шукати інформацію на декількох сторінках одночасно, що критично важливо для моніторингу наприклад державних закупівель або активності в соціальних мережах.

Особливе місце займає сервіс Apify, який пропонує понад 6000 готових «варіантів» для пошуку і моніторингу складних платформ, таких як Instagram, TikTok або Twitter/X [1, с. 21]. Це дозволяє розслідувачам аналізувати поширення дезінформації або відстежувати активність політичних фігур у реальному часі. Автоматизація через такі платформи дозволяє налаштувати регулярний збір даних, наприклад, щогодинне новлення бази новин.

Після етапу збору дані часто вимагають глибокої обробки. Журналісти витрачають до 80% часу на виправлення помилок, видалення дублікатів та стандартизацію форматів. Інструменти на кшталт Numerous.ai інтегрують можливості ШІ безпосередньо в Google Sheets або Excel. Це дозволяє журналістам використовувати формули природною мовою для виконання складних операцій очищення.

Функція Data Analyst дозволяє завантажувати великі CSV-файли й проводити комплексне очищення за допомогою Python-скриптів, які ШІ пише та виконує самостійно. Журналіст може попросити «Знайди всі аномальні значення в колонці» або «Об'єднай категорії», і система миттєво обробить тисячі рядків [2].

Важливим аспектом є перетворення хаотичних заміток або результатів інтерв'ю в структуровані бази даних. Інструменти на кшталт TextSniper та NoteThisDown використовують ШІ для витягування тексту з відеодзвінків, рукописних записів або слайдів презентацій. Це дозволяє створювати бази даних на основі візуальних джерел, що раніше вважалося надто трудомістким процесом.

У великих розслідуваннях, де джерелами є сотні різних типів файлів, використовуються спеціалізовані рішення, такі як Talend Data Preparation або Trifacta. Вони застосовують машинне навчання для розпізнавання типів даних та пропонують журналістам автоматичні рекомендації щодо структурування, наприклад, виявлення прихованих зв'язків між організаціями та особами через автоматичне тегування сутностей.

Водночас однією з проблем журналістики даних залишається те, що важлива інформація часто публікується у форматах, непризначених для аналізу, зокрема в захищених PDF. Особливо коли наявні складні таблиці або багатоколонкова верстка. Натомість ШІ-рішення, такі як Talonic, використовують візуальне розуміння документа, подібне до людського сприйняття [2].

Такі платформи, як Google Document AI та Adobe PDF Extract API, дозволяють не просто витягувати текст, а й зберігати «порядок читання» та ієрархію заголовків. Це критично важливо для аналізу судових звітів або банківських виписок, де розташування числа в таблиці визначає його сенс.

Сучасні редакції впроваджують автоматизовані конвеєри, де кожен новий PDF-документ, отриманий через запит на публічну інформацію, автоматично проходить через систему. Вона класифікує документ, витягує ключові показники і надсилає повідомлення журналісту, якщо виявлено збіг із «чорним списком» компаній або осіб. Це радикально скорочує час від отримання даних до виходу публікації.

Проте провідні медіа-організації все частіше відмовляються від загальних інструментів на користь власних розробок, адаптованих під специфічні потреби журналістів-розслідувачів. Зокрема, «Washington Post» розробила власний ШІ-інструмент під назвою Haystack, призначений для аналізу величезних масивів неструктурованих даних [3]. Його головна перевага полягає у здатності працювати одночасно з відео, фото та текстом, виявляючи патерни, які важко помітити при ручному перегляді. Вайн Кхосла, технічний директор видання, зазначає, що врахування журналістської етики та специфіки перевірки фактів є найважливішим для будь-якої моделі використання ШІ.

Отже, використання штучного інтелекту в журналістиці даних дозволяє автоматизувати більшу частину рутини та аналізувати будь-які формати – від захищених PDF до відео й аудіо. Це забезпечує перехід до автоматизованих конвеєрів аналізу в реальному часі для виявлення прихованих зв'язків. При цьому людина залишається ключовою фігурою, яка гарантує журналістську етику та адаптацію алгоритмів під конкретні завдання дата-журналіста.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. de-Lima-Santos M.-F., Salaverría R. From Data Journalism to Artificial Intelligence: Challenges Faced by La Nación in Implementing Computer Vision in News Reporting. *Palabra Clave*. 2021. Vol. 24, no. 3. P. 1–40. URL: <https://doi.org/10.5294/pacla.2021.24.3.7>
2. Kuek Ser K. K. Best Practices for Data Journalism. *Knowledge Bridge. Global Intelligence for the Digital Transition*. URL: <https://surl.li/rxqfpn>
3. Shapiro A. The Washington Post Debuts AI-Powered “Ask The Post” for User Insights. *AiNews.com*. URL: <https://www.ainews.com/p/the-washington-post-debuts-ai-powered-ask-the-post-for-user-insights>