UDC 004.738 DOI: https://doi.org/10.64076/iedc250614.14

Automated system for analyzing publicly available information using artificial intelligence

Vladyslav Shulha
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv
https://orcid.org/0009-0008-8981-3214
Serhii Shylo
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv
https://orcid.org/0000-0001-5782-552X
Serhii Khmelevskiy
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv
https://orcid.org/0000-0001-6216-3006
Oleksandr Pershyn
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv
https://orcid.org/0000-0003-0022-8972

Abstract. *A concept for an automated system designed for collecting, processing, and analyzing data from open-source intelligence sources is presented. The system automatically aggregates information from various sources, such as news websites, social media platforms, and forums, providing analysts with a user-friendly interface for searching and reviewing current materials. Key capabilities of the platform include classifying the collected data, generating concise analytical reports, and adding topical tags utilizing large language models.*
Keywords: *information classification, data collection, large language models, prompt engineering.*

Автоматизована система аналізу публічно доступної інформації з використанням штучного інтелекту

Владислав Шульга
Харківський національний університет
Повітряних Сил імені Івана Кожедуба, м. Харків
https://orcid.org/0009-0008-8981-3214
Сергій Шило
Харківський національний університет
Повітряних Сил імені Івана Кожедуба, м. Харків
https://orcid.org/0000-0001-5782-552X
Сергій Хмелевський
Харківський національний університет
Повітряних Сил імені Івана Кожедуба, м. Харків
https://orcid.org/0000-0001-6216-3006
Олександр Першин
Харківський національний університет
Повітряних Сил імені Івана Кожедуба, м. Харків
https://orcid.org/0000-0003-0022-8972

Анотація. *Представлено концепцію автоматизованої системи, призначеної для збору, обробки та аналізу даних з відкритих джерел (OSINT). Система автоматично агрегує інформацію з різноманітних джерел, таких як новинні сайти, соціальні мережі та форуми, надаючи аналітикам зручний інтерфейс для пошуку та перегляду актуальних матеріалів. Ключовими можливостями платформи є класифікація отриманих даних, створення стислих аналітичних звітів і додавання тематичних тегів із застосуванням великих мовних моделей (LLM).*
Ключові слова: *класифікація інформації, збір даних, великі мовні моделі, інженерія запитів.*

Зазвичай аналітики стикаються з величезним потоком інформації з відкритих джерел (OSINT), включаючи новини, соціальні мережі, супутникові знімки та звіти очевидців [1]. У сучасних умовах гібридних конфліктів та інформаційних війн стає дедалі складніше оперативно відстежувати події, ідентифікувати загрози та аналізувати їхній вплив.

У 2025 року обсяг даних, які генеруються, зберігаються та споживаються, вже перевищує 180 зеттабайт, що значно ускладнює їхню обробку та використання. Для порівняння, у 2020 році цей показник становив 64,2 зеттабайта. При цьому частка інформації, придатної для аналітичної обробки, зросла з 22% у 2012 році до 37% у 2020 році. Джерела цих даних охоплюють різні сфери, зокрема соціальні платформи, системи моніторингу, розвідувальні звіти та інші джерела [2-3].

Зі зростанням обсягів OSINT-даних виникає потреба у розробці платформи, здатної автоматизувати процеси збору, фільтрації та аналізу інформації у військовій сфері. Така система повинна:

– агрегувати дані з відкритих джерел у режимі реального часу;

– класифікувати інформацію за тематичними категоріями (наприклад, військова техніка, переміщення військ, кібератаки, політичні події);

– автоматично створювати короткі аналітичні звіти на основі зібраних матеріалів;

– генерувати релевантні теги для спрощеного пошуку та структуризації даних;

– здійснювати автоматизований переклад контенту з іноземних мов для розширення аналітичного охоплення.

Розробка подібної платформи значно підвищить швидкість та ефективність аналізу інформації, дозволяючи експертам оперативно реагувати на потенційні загрози та приймати обґрунтовані рішення.

Рішення щодо вибору конкретної архітектури для платформи, що розробляється може залежати від деяких факторів.

У першу чергу, платформа має відповідати вимогам високої швидкості обробки великої кількості джерел. Це особливо важливо, оскільки в сучасних умовах обсяги даних зростають експоненційно [2-3]. Не менш важливою є масштабованість – архітектура повинна дозволяти системі легко масштабуватися разом із зростанням кількості користувачів та джерел інформації. Без належної адаптивності система може швидко застаріти. Надійність також відіграє важливу роль – архітектура повинна забезпечувати стабільну роботу платформи за високих навантажень, при необхідності розподіляти завдання між іншими вузлами та елементами системи.

Однією з основних вимог для платформи, що розробляється є можливість її легкого розгортання та забезпечення найкращої сумісності з різними операційними системами. У цьому контексті оптимальним рішенням є

використання технології контейнеризації із застосуванням Docker та інструмента управління контейнерами Docker Compose.Серед найбільш поширених типів архітектури програмного забезпечення (ПЗ) виділяють такі: монолітна, мікросервісна, багаторівнева, сервіс-орієнтована та клієнт-серверна. Зважаючи на вимоги до OSINT-платформи, мікросервісна архітектура є найбільш вдалим рішенням. Вона дозволяє розробляти, оновлювати та масштабувати окремі модулі без ризику для всієї системи. Контейнеризація (Docker) та оркестрація контейнерів (Docker Compose або Kubernetes) забезпечать ефективне управління ресурсами та високу стійкість до навантажень.

Автоматизована система для збору, обробки та аналізу даних з відкритих джерел поєднує технології аналізу тексту із використанням великих мовних моделей (large language models, LLM) та інформаційні технології з метою автоматизації збору, аналізу та представлення актуальної інформації.

Нижче представлена концептуальна модель, що демонструє ключові компоненти та взаємозв'язки між ними. На діаграмі (рис. 1) зображено модель системи. Вона включає кілька ключових компонентів і модулів, які взаємодіють між собою для забезпечення роботи системи.

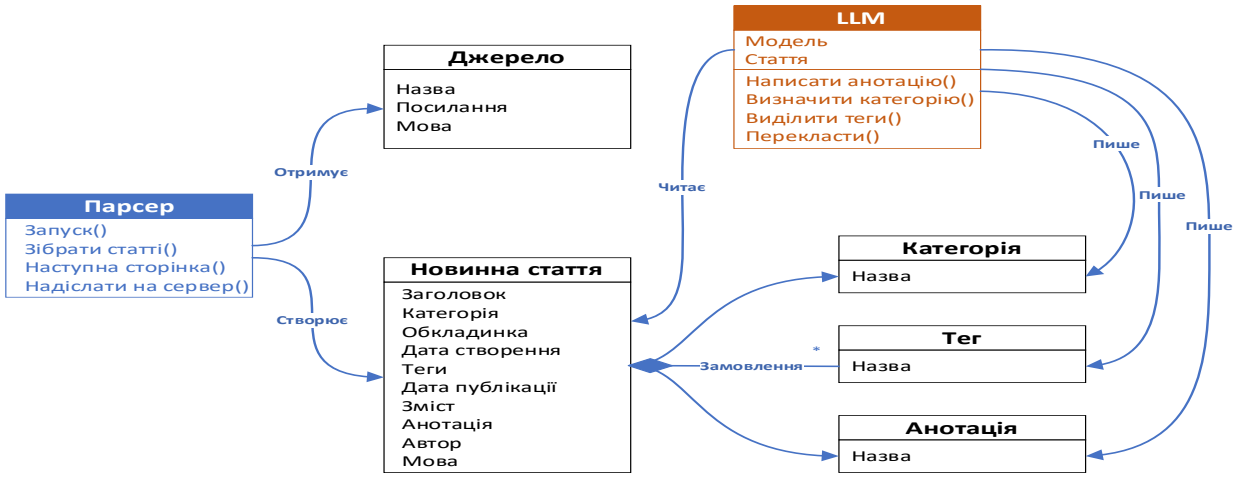

Рис. 1. Концептуальна діаграма веб-платформи

Джерело: розроблено авторами.

На діаграмі представлені такі ключові компоненти:

- Парсер: *Атрибути:* Ресурс; *Методи:* Запуск(), Зібрати статті(), Наступна сторінка(), Надіслати на сервер(); *Функції:* парсер відповідає за вилучення інформації з джерел. Він збирає посилання на статті, переходячи по сторінках інформаційного ресурсу, після чого передає їх сервісу. Сервіс створює чергу посилань на статті для подальшого отримання їхнього контексту та надсилання на сервер бази даних.

- Джерело: *Атрибути*: Назва, Посилання, Мова; *Функції*: джерело надає інформацію, з якої парсер буде вилучати статті.

- LLM: *Атрибути*: Модель, Стаття; *Методи*: Написати анотацію(), Визначити категорію(), Виділити теги(), Перекласти(); *Функції*: LLM обробляє статті, зібрані парсером, аналізує їхній вміст, створює анотації, визначає категорії, присвоює теги та виконує переклад.

- Новинна стаття: *Атрибути*: Заголовок, Категорія, Обкладинка, Дата створення, Теги, Дата публікації, Зміст, Анотація, Автор, Мова; *Функції*: цей об'єкт містить інформацію про новинну статтю, включаючи заголовок, дату створення, посилання на джерело, обкладинку, зміст, категорію, теги, анотацію та мову.

- Категорія: *Атрибути*: Назва; *Функції*: містить категорію статті.

- Тег: *Атрибути*: Назва; *Функції*: містить теги, пов’язані зі змістом статті.

- Анотація: *Атрибути*: Контекст; *Функції*: зберігає короткий опис змісту статті.

Запропонований підхід дозволяє забезпечити гнучкість, масштабованість та надійність платформи, що є критично важливим для аналізу даних у режимі реального часу. Використання автоматизованих методів збору, класифікації та обробки інформації сприяє оперативному виявленню загроз і прийняттю стратегічно важливих рішень у військовій сфері та інших галузях.

Подальший розвиток автоматизованої системи може включати оптимізацію продуктивності та її масштабованості, паралельне виконання процесів парсингу й анотування даних, а також використання кешування для прискорення доступу до часто запитуваної інформації. Існує можливість розширення кількості джерел новинних статей шляхом додавання додаткових інформаційних ресурсів, що спеціалізуються в галузі інформаційних технологій.

Також можна зосередити дослідження на впровадженні штучного інтелекту для підвищення точності аналізу, розширенні можливостей інтеграції з іншими системами та вдосконаленні алгоритмів обробки природної мови. Таким чином, створення такої автоматизованої системи сприятиме посиленню інформаційної безпеки та забезпеченню ефективного моніторингу відкритих джерел.

Список використаних джерел

1. Литвиненко М.І., Ленець В.Г., Гармаш Н.В., Шульга В.В. Аспекти впровадження штучного інтелекту у військовій справі. Збірник наукових праць Харківського національного університету Повітряних Сил. 2024. № 2(80). С. 13–18. https://doi.org/10.30748/zhups.2024.80.02.

2. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028. Statista. URL: https://www.statista.com/statistics/871513/worldwide-data-created (дата звернення: 27.05.2025).

3. Big Data Statistics 2023: How Much Data is in The World? URL: https://firstsiteguide.com/big-data-stats (дата звернення: 27.05.2025).