

## ІНТЕЛЕКТУАЛЬНА СИСТЕМА ВИЗНАЧЕННЯ СЕМАНТИЧНОЇ УЗГОДЖЕНОСТІ РОЗПОДІЛЕНИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ

<sup>1</sup>Національний університет «Львівська політехніка»

*Розглянуто проблему оцінювання якості розподілених інформаційних ресурсів у контексті зростання обсягів, різноманітності та динаміки даних, що циркулюють у сучасних інформаційних системах. Показано, що для розподілених середовищ традиційних підходів до контролю якості даних, які базуються переважно на точності, повноті та узгодженості, уже недостатньо, оскільки вони не враховують семантичну неоднорідність, контекст використання, реплікацію, варіативність представлення сутностей та відмінності у правилах інтерпретації даних. Особливу увагу приділено семантичній узгодженості як ключовій характеристиці, від якої залежить коректність інтеграції даних, достовірність аналітичних висновків і якість автоматизованих рішень у гетерогенних середовищах.*

*Розроблено інтелектуальну систему визначення семантичної узгодженості розподілених інформаційних ресурсів, побудовану на гібридному підході, що поєднує синтаксичні метрики швидкого відсіву та глибокий семантичний аналіз на основі векторних представлень. Така архітектура дає змогу ефективно працювати як з очевидно узгодженими парами даних, так і зі складними випадками, де однаковий зміст подано різними лексичними засобами, скороченнями або доменними синонімами. Додатково реалізовано адаптивне порогове прийняття рішень, що дозволяє налаштовувати суворість перевірки залежно від типу атрибута та предметної області застосування, а також формувати інтегральний індекс семантичної узгодженості для подальшого моніторингу якості даних.*

*Експериментальна перевірка підтвердила доцільність впровадженого підходу для підвищення точності виявлення семантично еквівалентних, але формально різних записів. Отримані результати показали, що поєднання синтаксичних і семантичних методів дає змогу суттєво зменшити кількість хибних відмов і помилкових збігів у процесі порівняння атрибутів, а також створює методологічне підґрунтя для побудови масштабованих систем контролю якості розподілених інформаційних ресурсів. Практичне значення роботи полягає в можливості інтеграції розробленої системи до платформ моніторингу, валідації та керування якістю даних у фінансових, медичних, наукових та інших гетерогенних інформаційних середовищах.*

**Ключові слова:** розподілені інформаційні ресурси, якість даних, семантична узгодженість, інтелектуальна система, гібридний підхід, векторні представлення, валідація даних, інтеграція даних.

### Вступ

Умови сучасного цифрового середовища характеризуються стрімким зростанням обсягів, різноманітності та динаміки даних, що зберігаються й обробляються в розподілених інформаційних системах. За таких обставин проблема забезпечення якості даних виходить за межі традиційної перевірки коректності записів і набуває системного характеру: від якості семантичного узгодження атрибутів залежить цілісність інтеграційних процесів, коректність аналітики, достовірність рішень і стійкість цифрових сервісів [1], [2]. Особливо актуальною ця проблема є для доменів, де дані надходять з багатьох автономних джерел і повинні узгоджуватися не лише на рівні формату, а й на рівні змісту, наприклад у фінансових, медичних, адміністративних та дослідницьких системах [3], [4]. Саме тому розроблення інтелектуальних методів визначення семантичної узгодженості розподілених інформаційних ресурсів є важливим науковим і практичним завданням, що безпосередньо пов'язане з підвищенням якості управління даними, зменшенням помилок інтеграції та підтримкою надійності інформаційної інфраструктури [5], [6].

Аналіз останніх досліджень показує, що класичні підходи до оцінювання якості даних



здебільшого ґрунтуються на універсальних вимірах — точності, повноті, узгодженості, своєчасності та доступності — і добре працюють для структурованих наборів даних, проте виявляються недостатніми для багатовузлових і семантично неоднорідних середовищ [2], [7], [8]. У 2024 році обґрунтовано систему критеріїв оцінювання якості даних у розподілених інформаційних системах, де акцентовано на специфіці розподіленого середовища та необхідності розмежування базових і контекстуальних показників [3]. Подальші дослідження 2025 року розвивають цю лінію, пропонуючи моделі оцінювання якості даних у семантичному середовищі великих даних, наголошуючи на ролі метаданих, походження даних, надійності та відповідності схемам якості [10]—[12]. Водночас у роботах з оцінювання semantic consistency показано, що традиційні overlap- і edit-distance-метрики не здатні адекватно відбити змістову узгодженість, а отже потрібні спеціалізовані метрики, семантичні векторні представлення та контекстно залежні правила прийняття рішення [7], [10]. Це особливо важливо для розподілених інформаційних ресурсів, де та сама сутність може бути подана різними словами, скороченнями, мовними варіантами або доменними синонімами [7], [8].

Окреме значення для формування проблематики статті мають попередні власні дослідження, у яких закладено методологічну основу для оцінювання якості та узгодженості розподілених інформаційних ресурсів. У праці «Інтелектуальна система виявлення плагіату в технічних текстах» продемонстровано практичну ефективність гібридного підходу, що поєднує синтаксичне зіставлення та семантичний аналіз на основі нейромережових представлень [11]. У статті «Система критеріїв оцінки якості даних в розподілених інформаційних системах» сформовано систему критеріїв якості, орієнтовану на особливості розподілених середовищ, та показано необхідність переходу від окремих метрик до комплексної моделі оцінювання [5]. У публікації «Оцінка семантичної та часової узгодженості розподілених інформаційних ресурсів» безпосередньо поставлено задачу кількісного вимірювання змістовної та темпоральної узгодженості як ключових характеристик якості PIP [12]. У праці «Comparison analysis of metrics and evaluation scales of distributed information resources quality parameters» розвинуто підхід до порівняння метрик і шкал оцінювання, що є важливим для побудови узгодженої моделі вимірювання якості [13]. У сукупності ці роботи створюють базу для подальшого розвитку інтелектуальних методів верифікації даних, але водночас показують, що проблема потребує узагальненішої моделі, здатної поєднати семантичну подібність, контекстну адаптацію та практичну придатність до застосування в багатьох предметних областях [7], [10], [11].

Попри наявні результати, у сучасних дослідженнях зберігаються невирішені питання щодо формалізації семантичної узгодженості саме для розподілених інформаційних ресурсів із гетерогенними джерелами, де важливо враховувати не лише текстову подібність, а й доменні онтології, варіативність термінології, різні рівні суворості перевірки та потребу в масштабованій архітектурі прийняття рішень [7]—[10]. Недостатньо опрацьованими залишаються підходи, які дозволяють поєднати швидке синтаксичне відсівання очевидно несумісних пар із глибоким семантичним аналізом складних випадків, а також побудувати індекс узгодженості, придатний для практичного використання в інтеграційних і моніторингових системах [10], [11], [14]. Саме ці аспекти зумовлюють потребу в інтелектуальній системі, що здатна виконувати двоступеневу перевірку узгодженості, адаптувати пороги рішення до типу атрибутів і забезпечувати відтворюваний результат оцінювання у реальних розподілених середовищах [1], [14].

*Метою статті* є розроблення та обґрунтування інтелектуальної системи визначення семантичної узгодженості розподілених інформаційних ресурсів на основі гібридного підходу, що поєднує синтаксичні метрики, семантичні векторні представлення та адаптивне порогове прийняття рішень [1]. Для досягнення цієї мети необхідно: формалізувати модель обчислення семантичної узгодженості; запропонувати архітектуру системи з поділом на швидкий і глибокий етапи перевірки; експериментально оцінити ефективність розробленого підходу; та показати його придатність для застосування в умовах інтеграції гетерогенних даних [1]. Очікуваним результатом є підвищення точності виявлення семантично еквівалентних, але формально різних записів, що створює підґрунтя для зменшення помилок синхронізації, підвищення якості інтеграції та надійності розподілених інформаційних ресурсів [1], [7].

## Результати дослідження

### Математична модель визначення семантичної узгодженості розподілених інформаційних ресурсів

Процес визначення семантичної узгодженості даних у розподілених інформаційних ресурсах

(PIP) потребує побудови формалізованої математичної моделі, яка здатна оперувати як структурованими символьними послідовностями, так і абстрактними семантичними одиницями. В основі запропонованої моделі лежить концепція гібридного оцінювання, де загальний показник узгодженості  $SC$  (Semantic Consistency) визначається через сукупність синтаксичних та семантичних метрик подібності. Такий підхід дозволяє нівелювати обмеження класичних алгоритмів порівняння рядків та враховувати контекстуальну близькість термінів, що є критичним для гетерогенних даних.

Початковим етапом моделювання є формалізація вхідних даних. Нехай  $S = \{s_1, s_2, \dots, s_n\}$  — множина значень атрибутів з джерела-еталона, а  $T = \{t_1, t_2, \dots, t_n\}$  — відповідна множина значень із цільового ресурсу. Завданням моделі є обчислення функції подібності  $F(s_i, t_i) \rightarrow [0, 1]$ , де значення 1 відповідає повній ідентичності, а 0 — абсолютній розбіжності. Для забезпечення оперативності системи на першому рівні («Fast Path») застосовуються синтаксичні метрики, що базуються на редакторській відстані та теоретико-множинному підході.

Однією з ключових синтаксичних метрик є нормалізована відстань Левенштейна ( $L$ ). Вона визначає мінімальну кількість операцій вставки, видалення та заміни символів, необхідних для перетворення рядка  $s$  у  $t$ . Математично подібність за Левенштейном  $Sim_L$  обчислюється як:

$$Sim_L(s, t) = 1 - \frac{lev(s, t)}{\max(|s|, |t|)},$$

де  $lev(s, t)$  — функція розрахунку відстані. Ця метрика є ефективною для виявлення друкарських помилок, розбіжностей у реєстрах або транслітерації, проте вона нездатна розпізнати синоніми з різною символьною структурою.

Для врахування лексичного складу без прив'язки до порядку символів модель використовує індекс Джаккара ( $J$ ). Ця метрика розглядає рядки як множини токенів (слів)  $A$  та  $B$ . Коефіцієнт подібності Джаккара визначається як відношення потужності перетину цих множин до потужності їх об'єднання:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

Використання максимуму з цих двох метрик  $Sim_{syn} = \max(Sim_L, J)$  дозволяє системі швидко вальдувати пари з високим ступенем візуальної схожості. Якщо  $Sim_{syn} \geq \theta_{fast}$  (де  $\theta_{fast} \approx 0,98$ ), пара вважається узгодженою без залучення нейромережових обчислень.

У випадках, коли синтаксична подібність є низькою, модель переходить до етапу глибокого аналізу («Deep Analysis»), що базується на векторному представленні тексту. Кожне текстове значення  $s_i$  та  $t_i$  відображається у багатовимірний семантичний простір за допомогою трансформерної моделі  $\phi: \text{Text} \rightarrow \mathbb{R}^d$ . В результаті отримуються векторні представлення  $e_s = \phi(s)$  та  $e_t = \phi(t)$ . Основною перевагою такого представлення є те, що координати векторів відображають не написання, а контекстуальне значення слів.

Центральною метрикою семантичного блоку моделі є косинусна подібність ( $Sim_{cos}$ ), яка вимірює косинус кута між двома векторами. На відміну від Евклідової відстані, ця метрика ігнорує довжину векторів (норму), фокусуючись лише на їхньому спрямуванні в семантичному просторі

$$Sim_{cos}(e_s, e_t) = \cos(\theta) = \frac{e_s \cdot e_t}{|e_s| |e_t|} = \frac{\sum_{i=1}^d e_{si} e_{ti}}{\sqrt{\sum_{i=1}^d e_{si}^2} \sqrt{\sum_{i=1}^d e_{ti}^2}}.$$

Ця формула дозволяє виявити глибокі семантичні зв'язки, такі як синонімія між термінами «Фізична особа» та «Приватна особа», які мають нульовий або низький синтаксичний збіг.

Ефективність побудови векторів  $e$  безпосередньо залежить від вагових коефіцієнтів нейронної мережі  $\phi$ . Базова модель, навчена на загальному корпусі текстів, часто демонструє недостатню точність у специфічних доменах (фінанси, юриспруденція). Для коригування векторного простору в моделі передбачено етап донавчання (Fine-tuning) на основі функції втрат Multiple Negatives Ranking Loss. Математична мета навчання полягає в мінімізації відстані між парами векторів, що позначені

як позитивні синоніми, та максимізації відстані до негативних прикладів у межах навчального пакету (batch).

Для ухвалення остаточного рішення про узгодженість вводиться бінарна функція вердикту  $V$ . Вона порівнює обчислений показник подібності  $Sim$  із заданим пороговим значенням  $\theta$ , яке визначається адміністратором системи залежно від критичності даних

$$V(s, t) = \begin{cases} 1, & \text{якщо } Sim \geq \theta, \\ 0, & \text{якщо } Sim < \theta. \end{cases}$$

Важливо зазначити, що поріг  $\theta$  може бути динамічним для різних типів атрибутів. Наприклад, для ідентифікаторів (РНОКПП) він наближається до 1,0, тоді як для описових полів (призначення платежу) може бути знижений до 0,85.

Агрегована оцінка якості набору даних (Aggregate Semantic Consistency Index) розраховується як відношення кількості узгоджених пар до загальної кількості перевірених атрибутів  $N$

$$SC_{index} = \frac{1}{N} \sum_{i=1}^N V(s_i, t_i).$$

Цей індекс дає кількісну характеристику ступеня інтегрованості та чистоти розподілених ресурсів.

Перед обчисленням будь-яких метрик вхідні дані проходять етап попередньої обробки  $P(s)$ , що включає токенізацію, приведення до нижнього регістру, видалення стоп-слів та спеціальних символів. Це дозволяє зменшити рівень «шуму» та підвищити точність векторного представлення

$$Sim_{final} = F(P(s), P(t)).$$

Така багаторівнева математична конструкція забезпечує стійкість системи до варіативності природної мови та гарантує високу достовірність результатів перевірки в умовах реальних розподілених ресурсів.

Розроблена інтелектуальна система (IC) функціонує на базі автономного модуля Semantic Consistency Module, архітектура якого спроектована для забезпечення високої відмовостійкості та точності в умовах роботи з неструктурованими даними розподілених ресурсів. Автономність модуля дозволяє використовувати його як незалежний мікросервіс, що легко інтегрується в наявні аналітичні платформи, виконуючи вузькоспеціалізоване завдання — верифікацію семантичної тотожності атрибутів.

### Проектування та реалізація інтелектуальної системи визначення семантичної узгодженості розподілених інформаційних ресурсів

Функціонування системи базується на гібридному підході до обробки даних, що реалізує концепцію дворівневої перевірки: «Швидкий шлях» (Fast Path) та «Глибокий аналіз» (Deep Analysis). На першому етапі система застосовує детерміновані синтаксичні методи — розрахунок нормалізованої відстані Левенштейна та індексу Джаккара. Це дозволяє миттєво валідувати пари даних з високим ступенем символічної схожості (наприклад, за наявності одруку чи різного регістру), мінімізуючи навантаження на обчислювальні потужності. Перехід до ресурсомісткого етапу глибокого аналізу відбувається лише за умови, що показники синтаксичного збігу є нижчими за встановлений критичний поріг, що оптимізує використання інтелектуального ядра системи. Процес потоку даних показаний на рис. 1.

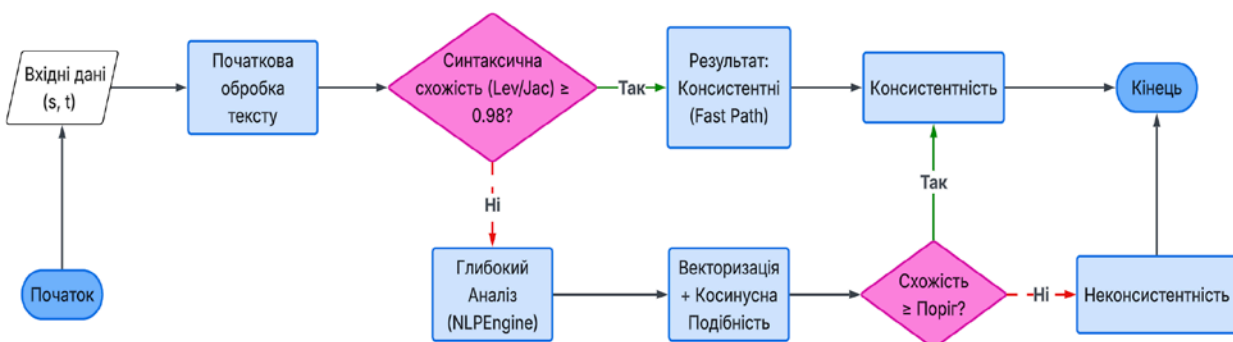


Рис. 1. Блок-схема процесу визначення семантичної узгодженості розподілених інформаційних ресурсів

Технологічний стек системи вибрано з огляду на необхідність підтримки багатомовності та високу швидкість обробки векторних представлень. Ключовим елементом інтелектуальної складової є модель *paraphrase-multilingual-MiniLM-L12-v2*. Вибір цієї архітектури сімейства Sentence-BERT обґрунтований її здатністю ефективно працювати з українською мовою та оптимальним співвідношенням між розміром моделі (що критично для розгортання в Docker-контейнерах) та якістю генерації семантичних векторних представлень. Середовище розробки базується на мові програмування Python 3.9, що забезпечує стабільну роботу бібліотек машинного навчання (PyTorch, Sentence-Transformers) та дозволяє реалізувати асинхронну обробку запитів через фреймворк FastAPI.

Основним структурним елементом системи є *Semantic Consistency Module*, внутрішня архітектура якого розділена на три функціональні рівні: рівень інтерфейсної взаємодії (API Layer), рівень координації бізнес-логіки (Service Layer) та рівень інтелектуальних обчислень (Intelligence Core). Взаємодія між цими рівнями організована таким чином, щоб мінімізувати затримки під час обробки великих масивів даних (рис. 2).

*Рівень інтерфейсної взаємодії (API Layer)* базується на використанні високопродуктивного фреймворку FastAPI. Він відповідає за прийом вхідних запитів, первинну валідацію схем даних та формування структурованих відповідей. На цьому рівні реалізовано механізми асинхронної обробки, що дозволяє системі одночасно опрацьовувати декілька потоків валідації від різних розподілених джерел.

*Рівень координації (Service Layer)* представлений компонентом Consistency Validator. Його основним завданням є реалізація стратегії «двошвидкісної перевірки» (Fast Path & Deep Analysis). Валідатор аналізує кожну пару атрибутів і ухвалює рішення про необхідність залучення нейронної мережі. Якщо синтаксичні метрики, обчислені модулем Metrics Engine, перевищують критичний поріг, система миттєво повертає результат, уникаючи ресурсомістких векторних операцій.

*Рівень інтелектуальних обчислень (Intelligence Core)* включає два критичні компоненти: NLP Engine, Metrics Engine та Model Trainer.

NLP Engine ініціалізує трансформерну модель MiniLM і забезпечує векторизацію вхідних значень. Для оптимізації споживання пам'яті цей компонент реалізований за шаблоном Singleton.

Model Trainer відповідає за доменну адаптацію системи. Він забезпечує можливість перенавчання моделі на специфічних наборах даних (наприклад, фінансових чи юридичних парах), що дозволяє системі еволюціонувати та підвищувати точність без зміни основного коду.

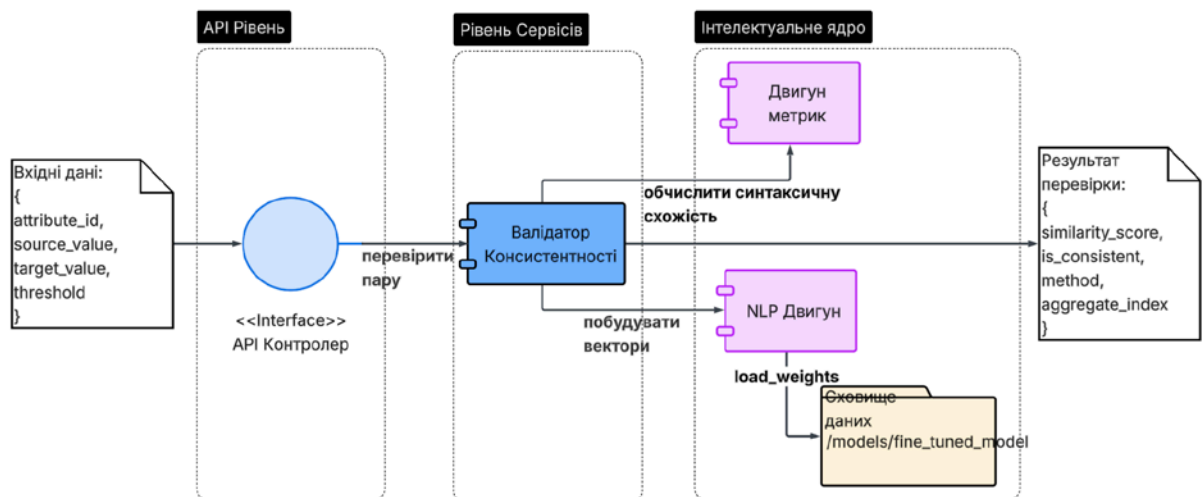


Рис. 2. Діаграма компонентів інтелектуальної системи визначення семантичної узгодженості розподілених інформаційних ресурсів

Вхідні дані системи подаються у форматі JSON-об'єкта, який містить не лише значення для порівняння, а й контекстуальні метадані. Важливим параметром вхідних даних є *threshold* (поріг узгодженості) — це динамічний коефіцієнт від 0 до 1, який визначає рівень суворості перевірки. Використання індивідуальних порогів для різних типів атрибутів дозволяє налаштовувати систему під конкретні вимоги бізнес-логіки (наприклад, встановлення  $\theta = 0,99$  для паспортних даних та  $\theta = 0,80$  для описів товарів). Також вхідний пакет включає *attribute\_id*, що забезпечує простежуваність (traceability) даних у разі їх подальшої інтеграції в звітні якості.

Вихідні дані системи (Validation Response) є деталізованим звітом, який, окрім бінарного

вердикту `is_consistent`, містить низку інтелектуальних показників:

1) `similarity_score` — нормалізоване значення подібності, заокруглене до сотих знаків, що дає кількісну оцінку близькості значень;

2) `validation_method` — індикатор того, який саме алгоритм спрацював (синтаксичний чи семантичний). Це дозволяє проводити аудит роботи системи та оцінювати ефективність «швидкого шляху»;

3) `aggregate_sc_index` — підсумковий показник для всього пакета даних, що відображає загальний рівень якості інтеграції між ресурсами.

Окрему увагу в архітектурі приділено контейнеризації за допомогою технології Docker. Кожен компонент системи разом з навченими вагами моделі ізолюється в окремому контейнері. Це гарантує відтворюваність середовища обчислень незалежно від хост-системи та дозволяє масштабувати ПС у хмарних інфраструктурах. Взаємодія з моделлю відбувається через абстрагований інтерфейс, що дає змогу в майбутньому замінювати модель MiniLM на потужніші архітектури (наприклад, RoBERTa або GPT-подібні моделі) без перебудови всього архітектурного каркаса.

Таким чином, архітектура ПС забезпечує не лише високу точність семантичного аналізу, а й має необхідну гнучкість для подальшої інтеграції в повномасштабну екосистему контролю якості даних. Вона виступає інтелектуальним фільтром, який перетворює сирі дані розподілених ресурсів у верифіковану інформацію, готову до ухвалення управлінських рішень.

Розроблений інтелектуальний модуль в подальшому буде інтегрований у ширшу екосистему контролю якості як функціональний компонент платформи Verification Hub. У межах цієї архітектури Verification Hub виступає центральним оркестратором, який автоматично вилучає нормалізовані дані з реляційних баз даних, забезпечує високий рівень паралелізації обчислювальних процесів шляхом одночасного запуску спеціалізованих модулів валідації та здійснює моніторинг стану кожної перевірки в реальному часі. Після завершення всіх детермінованих та інтелектуальних процедур аналізу, оркестратор ініціює фінальний етап обробки — генерацію консолідованого звіту про якість ресурсу, де результати семантичної узгодженості стають фундаментом для прийняття рішення про достовірність інформації в розподіленому середовищі.

### Експериментальне дослідження та аналіз результатів

*Метою експериментального дослідження є верифікація запропонованого методу валідації узгодженості даних та обґрунтування доцільності доменної адаптації нейромережевої моделі для підвищення точності прийняття рішень у межах інтелектуальної системи. У межах роботи проведено комплексний експеримент, спрямований на порівняльний аналіз базової та донавченої моделей (A/B Test). Дослідження орієнтоване на оцінку точності семантичного зіставлення специфічних фінансових термінів після етапу Fine-tuning, а також на визначення здатності моделі долати високу лексичну дивергенцію атрибутів, що є критичним для забезпечення функціонування сучасних систем обробки даних.*

Програмна реалізація експериментального стенда базується на архітектурі мікросервісів з використанням фреймворку FastAPI. Як базову модель (Base Model) вибрано багатомовний трансформер `paraphrase-multilingual-MiniLM-L12-v2`, який пройшов процедуру донавчання на сформованому корпусі фінансових пар (5000+ записів) з використанням функції втрат `CosineSimilarityLoss`. Обчислення проводилися на вузлі з підтримкою архітектури CUDA для мінімізації латентності нейромережевого виводу.

Перший етап дослідження полягав у порівняльній оцінці трьох стратегій обробки запитів: синтаксичного аналізу (Fast Path), базової моделі (Deep Base) та адаптованої моделі (Deep Fine-tuned).

Для забезпечення репрезентативності результатів сформовано контрольний набір даних розміром 100 пар, який включає специфічні фінансові атрибути, аббревіатури та професійні синоніми. Структура вхідних даних та приклади тестових пар подані у табл. 1.

Таблиця 1

Приклад вхідних даних експерименту 1

| ID      | Назва атрибута         | Значення джерела (Source) | Значення цілі (Target)          |
|---------|------------------------|---------------------------|---------------------------------|
| SEM_001 | <code>fin_id</code>    | РНОКПП                    | ідентифікаційний код            |
| SEM_002 | <code>ownership</code> | кінцевий бенефіціар       | UBO (Ultimate Beneficial Owner) |
| SEM_003 | <code>account</code>   | поточний рахунок          | IBAN клієнта                    |
| SEM_004 | <code>credit</code>    | кредитна лінія            | ліміт овердрафту                |
| SEM_005 | <code>deposit</code>   | депозитний сертифікат     | вкладне свідоцтво               |

Для кожного методу розраховано показник косинусної подібності (*Sim*), результати якого для ключових атрибутів подано у табл. 2. Для автоматичного прийняття рішення про консистентність даних у межах ІС встановлено критичний поріг  $\theta = 0,85$ .

Таблиця 2

Результати експерименту 1

| ID      | Fast Path (Lev/Jac) | Deep Base (MiniLM) | Deep Fine-tuned (Proposed) |
|---------|---------------------|--------------------|----------------------------|
| SEM_001 | 0,00                | 0,53               | 0,94                       |
| SEM_002 | 0,04                | 0,27               | 0,38                       |
| SEM_003 | 0,14                | 0,40               | 0,62                       |
| SEM_004 | 0,20                | 0,28               | 0,92                       |
| SEM_005 | 0,26                | 0,80               | 0,90                       |

Візуалізація розподілу значень подібності для часткової вибірки (перші 20 атрибутів) подана на рис. 3.

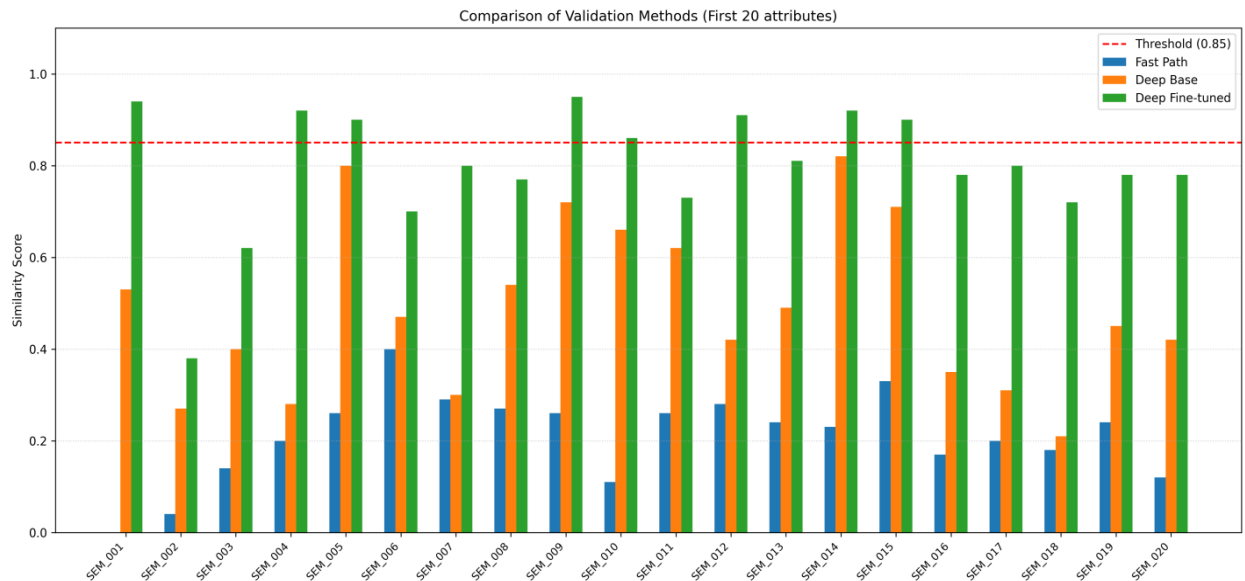


Рис. 3. Результати порівняння методів визначення семантичної узгодженості розподілених інформаційних ресурсів

Аналіз рис. 3 та табл. 2 дозволяє стверджувати, що традиційні синтаксичні методи (Fast Path) є непридатними для валідації фінансових термінів через високу лексичну дивергенцію (відхилення *Sim* від порогу становить понад 80 %). Базова нейромережева модель (Deep Base) виявляє семантичну спорідненість, проте її точності недостатньо для подолання встановленого бар'єра  $\theta$ , що призводить до помилкових відмов системи.

Використання запропонованої моделі Deep Fine-tuned дозволило суттєво підвищити показник узгодженості для складних пар. Зокрема, для термінів SEM\_001, SEM\_004 та SEM\_005 спостерігається зростання *Sim* до рівня 0,90...0,94, що свідчить про успішну інтерналізацію моделлю доменної специфіки. Низькі показники для SEM\_002 та SEM\_003 (відповідно 0,38 та 0,62) вказують на наявну межу семантичного узагальнення, де система потребує або додаткових метаданих, або подальшого розширення навчального корпусу для розпізнавання непрямих зв'язків між текстовим описом та цифровими ідентифікаторами (як-от IBAN).

Результати проведеного експерименту підтверджують працездатність розробленої інтелектуальної системи та її здатність до коректної ідентифікації семантичних зв'язків у розподілених інформаційних ресурсах. Система продемонструвала високу адаптивність до професійної термінології, забезпечивши автоматичне прийняття рішень у випадках, де традиційні методи виявляються неспроможними через високу лексичну варіативність даних. Це доводить ефективність вибраної архітектури ІС як надійного інструменту для синхронізації та валідації значущих атрибутів у складних інформаційних середовищах.

## Наукова новизна розробленого підходу

Вперше запропоновано гібридну дворівневу модель оцінювання семантичної узгодженості, яка поєднує детерміновані синтаксичні метрики швидкого відсіювання («Fast Path») з глибоким нейромережним аналізом контекстуальних векторних представлень («Deep Analysis»). Це дозволяє системі миттєво валідувати очевидні збіги та залучати ресурсомісткі обчислення лише для складних випадків лексичної дивергенції, забезпечуючи високу пропускну здатність.

Удосконалено метод ідентифікації концептуальної тотожності об'єктів в гетерогенних середовищах шляхом застосування доменно-адаптованих трансформерних моделей (типу Sentence-BERT) для побудови векторних представлень значень атрибутів. Метод забезпечує розпізнавання еквівалентних сутностей за повної відсутності лексичного збігу, враховуючи специфічну професійну термінологію, скорочення та доменні синоніми.

Розроблено механізм адаптивного прийняття рішень на основі динамічних порогів узгодженості ( $\theta$ ) та інтегрального індексу. На відміну від традиційних систем, метод дозволяє налаштовувати суворість перевірки залежно від типу атрибута та формувати кількісну характеристику ступеня чистоти всього набору даних у режимі реального часу.

## Висновки

У статті показано, що проблема семантичної узгодженості розподілених інформаційних ресурсів є однією з ключових для підвищення якості даних у гетерогенних середовищах, а її вирішення потребує поєднання класичних метрик якості з інтелектуальними методами аналізу змісту. Дослідження обґрунтовує необхідність гібридного підходу, який дозволяє поєднати швидку синтаксичну перевірку з глибоким семантичним зіставленням, а також показує, що саме така схема забезпечує практичну придатність моделі до реальних розподілених систем. Отримані результати створюють основу для подальшого розвитку засобів моніторингу, валідації та інтеграції даних у фінансових, медичних, наукових та інших інформаційних середовищах, де критично важлива достовірність і узгодженість даних.

Розроблена інтелектуальна система визначення семантичної узгодженості забезпечує якісно новий рівень автоматизації процесів верифікації розподілених інформаційних ресурсів. Завдяки впровадженню гібридної архітектури, система успішно поєднує високу швидкість попередньої обробки даних з глибоким нейромережним аналізом складних семантичних структур. Це дозволяє ефективно розв'язувати проблему неузгодженості ідентичних за змістом, але різних за формою запису даних, що є критичним для інтеграції гетерогенних джерел інформації.

Експериментальна перевірка системи підтвердила її високу результативність у реальних умовах функціонування. Розроблена ПС продемонструвала здатність до подолання значних синтаксичних розбіжностей, підвищивши точність розпізнавання фінансових термінів та специфічних аббревіатур до рівня, що відповідає вимогам промислових стандартів узгодження даних. Отримані показники подібності свідчать про те, що система мінімізує необхідність ручного втручання та знижує ймовірність виникнення помилок у разі синхронізації об'єктів у розподіленому середовищі.

Наукова новизна розробленого методу полягає у переході від суто формального порівняння структур даних до інтелектуального аналізу їхнього змістовного наповнення в умовах високої асинхронності та гетерогенності розподілених ресурсів. Авторський підхід вперше інтегрує етап швидкої синтаксичної фільтрації з механізмами глибокого машинного навчання та адаптивною логікою прийняття рішень, що дозволяє автоматизувати виявлення семантичних колізій з точністю, недоступною для класичних ETL-систем. Це створює методологічне підґрунтя для побудови самонавчальних систем контролю якості, здатних еволюціонувати разом з професійним доменом без необхідності ручного оновлення валідаційних правил.

Практична цінність роботи полягає у створенні готового програмного інструментарію, який може бути інтегрований у наявні системи управління даними для забезпечення їхньої цілісності та узгодженості. Гнучкість налаштувань розробленої інтелектуальної системи дозволяє масштабувати її для різних предметних областей, що відкриває широкі перспективи для її використання в інтелектуальних агентах моніторингу та системах підтримки прийняття рішень. Подальший розвиток розробленої ПС вбачається у впровадженні механізмів самонавчання на основі зворотного зв'язку, а також у її глибокій інтеграції до складу комплексних систем контролю якості розподілених інформаційних ресурсів. Це дозволить забезпечити наскрізний моніторинг життєвого циклу даних та гарантувати високий рівень достовірності інформації в умовах постійного розширення гетерогенних джерел.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, p. 2, 2015. <https://doi.org/10.5334/dsj-2015-002>.
- [2] J. Declerck, et al., "Assessing data quality in heterogeneous healthcare integration: The AIDAVA framework (preprint)," *JMIR Medical Informatics*, 2025. <https://doi.org/10.2196/75275>.
- [3] Ю. М. Геряк, і А. Ю. Берко, «Система критеріїв оцінки якості даних в розподілених інформаційних системах», *Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі*, № 16, с. 191-202, 2024. <https://doi.org/10.23939/sisn2024.16.191>.
- [4] S. Vujević, M. Cerjan, and K. Rabuzin, "Data quality in distributed information systems: Conceptual challenges and empirical findings," in *Proc. Central European Conf. Information and Intelligent Systems*, 2025, pp. 105-112. [Online]. Available: <https://archive.cecis.foi.hr/public/conferences/2025/Proceedings/S3/3.pdf>.
- [5] O. Novytskyi, "Concept and evaluation of the big data quality in semantic environment," *Problemy Prohramuvannia*, no. 3-4, pp. 260-270, 2022. <https://doi.org/10.15407/pp2022.03-04.260>.
- [6] B. Heinrich, M. Klier, A. Schiller, and G. Wagner, "Assessing data quality – A probability-based metric for semantic consistency," *Decision Support Systems*, vol. 110, pp. 95-106, 2018. <https://doi.org/10.1016/j.dss.2018.03.011>.
- [7] Z. Tan, et al., "Semantic similarity distance: Towards better text-image consistency metric in text-to-image generation," *Pattern Recognition*, vol. 144, p. 109883, 2023. <https://doi.org/10.1016/j.patcog.2023.109883>.
- [8] N. Wang, and S.-G. Leem, "Modernizing data quality: Evolving dimensions for unstructured text data in big data, AI and ethical contexts," *IntechOpen*, 2026. <https://doi.org/10.5772/intechopen.1013232>.
- [9] M. Klier, A. Obermeier, C. Sparn, and T. Widmann, "Anomaly-based assessment of semantic consistency: Design and evaluation of a novel probability-based metric in cooperation with a German car manufacturer," *Journal of Data and Information Quality*, vol. 17, no. 2, art. no. 24, 2025. <https://doi.org/10.1145/3732783>.
- [10] R. Huidrom, M. Lorandi, S. Mille, C. Thomson, and A. Belz, "Assessing semantic consistency in data-to-text generation: A meta-evaluation of textual, semantic and model-based metrics," in *Proc. 18th Int. Natural Language Generation Conf.*, Hanoi, Vietnam, 2025, pp. 98-107. [Online]. Available: <https://aclanthology.org/2025.inlg-main.6/>.
- [11] Ю. М. Геряк та А. Ю. Берко, «Інтелектуальна система виявлення плагіату в технічних текстах», *Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі*, № 14, с. 235-247, 2023. <https://doi.org/10.23939/sisn2023.14.235>.
- [12] Ю. М. Геряк, і А. Ю. Берко, «Оцінка семантичної та часової узгодженості розподілених інформаційних ресурсів», *Model. Control Inf. Technol.*, № 8, с. 125-128, 2025. <https://doi.org/10.31713/MCIT.2025.036>.
- [13] Y. Heriak, and A. Berko, "Comparison analysis of metrics and evaluation scales of distributed information resources quality parameters," *Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі*, № 17, с. 130-137, 2025. <https://doi.org/10.23939/sisn2025.17.130>.
- [14] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, 2014. <https://doi.org/10.1007/s11036-013-0489-0>.

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Дата надходження 24.04.2026

Дата прийняття після доопрацювання 6.05.2026

Дата прийняття до друку після рецензування 12.05.2026

Дата публікації 2.10.2026

**Геряк Юрій Миколайович** — аспірант кафедри інформаційних систем та мереж, e-mail: [yurii.m.heriak@lpnu.ua](mailto:yurii.m.heriak@lpnu.ua) . <https://orcid.org/0009-0008-3251-2007>;

**Берко Андрій Юліанович** — д-р техн. наук, професор кафедри інформаційних систем та мереж, e-mail: [andrii.y.berko@lpnu.ua](mailto:andrii.y.berko@lpnu.ua) . <https://orcid.org/0000-0003-2892-9519>.

Національний університет «Львівська політехніка», Львів

**Yu. M. Heriak<sup>1</sup>**  
**A. Yu. Berko<sup>1</sup>**

## Intelligent System of Semantical Consistency Detection in Distributed Information Resources

<sup>1</sup>Lviv Polytechnic National University

*This article addresses the problem of data quality assessment in distributed information resources under conditions of rapidly growing data volume, its heterogeneity, and dynamism in modern information systems. It demonstrates that conventional data quality approaches, which focus mainly on accuracy, completeness, and consistency, are no longer sufficient for distributed environments because they do not fully capture semantic heterogeneity, usage context, replication effects, entity*

representation variability, and differences in data interpretation rules. Particular attention is paid to semantic consistency as a key property that determines the correctness of data integration, the reliability of analytical conclusions, and the quality of automated decisions in heterogeneous environments.

The paper implements an intelligent system for detecting semantic consistency in distributed information resources based on a hybrid approach that combines fast syntactic filtering with deep semantic analysis using vector representations. This architecture makes it possible to process both obviously consistent data pairs and more complex cases where the same meaning is expressed through different lexical forms, abbreviations, or domain-specific synonyms. In addition, the proposed solution implements adaptive threshold-based decision making, allowing the strictness of verification to be adjusted according to the attribute type and application domain, while also producing an integrated semantic consistency index for further monitoring of data quality.

Experimental evaluation confirmed the feasibility of the proposed approach for improving the detection of semantically equivalent but formally different records. The results showed that combining syntactic and semantic methods significantly reduces both false rejections and false matches during attribute comparison, while also providing a methodological basis for building scalable quality control systems for distributed information resources. The practical value of the study lies in the possibility of integrating the developed system into data quality monitoring, validation, and management platforms for financial, medical, scientific, and other heterogeneous information environments.

**Keywords:** distributed information resources, data quality, semantic consistency, intelligent system, hybrid approach, vector representations, data validation, data integration.

**Heriak Yuriy M.** — Post-Graduate Student of the Chair of Information Systems and Networks, e-mail: [yurii.m.heriak@lpnu.ua](mailto:yurii.m.heriak@lpnu.ua) . <https://orcid.org/0009-0008-3251-2007>;

**Berko Andrii Yu.** — Dr. Sc. (Eng.), Professor of the Chair of Information Systems and Networks, e-mail: [andrii.y.berko@lpnu.ua](mailto:andrii.y.berko@lpnu.ua) . <https://orcid.org/0000-0003-2892-9519>