https://doi.org/10.31891/2307-5732-2026-363-14
УДК 004.056.5

КАБАК РУСЛАН
Український державний університет науки і технологій
https://orcid.org/0009-0000-8342-691X
e-mail: ruslankabak941@gmail.com
ЛЯШЕНКО ОКСАНА
Український державний університет науки і технологій
https://orcid.org/0000-0002-9983-5504
e-mail: o.a.liashenko@ust.edu.ua
РАДУЛЬ ОЛЕКСАНДР
Український державний університет науки і технологій
https://orcid.org/0009-0001-3625-0859
e-mail: o.a.radul@ust.edu.ua

ПРОЄКТУВАННЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ АНАЛІЗУ УКРАЇНОМОВНИХ КОРПОРАТИВНИХ ДОКУМЕНТІВ У ЗАДАЧАХ ЗАХИСТУ ІНФОРМАЦІЇ

У статті обґрунтовано та описано проєктування модульної архітектури інформаційної системи для дослідження підходів до виявлення чутливої інформації в україномовних корпоративних документах у контексті вимог DLP-платформ і задач захисту текстових інформаційних ресурсів.

Ключові слова: *інформаційна безпека, чутлива інформація, DLP-системи, NER, проєктування програмних систем, україномовні тексти.*

KABAK RUSLAN, LIASHENKO OKSANA, RADUL OLEKSANDR
Ukrainian State University of Science and Technologies

INFORMATION SYSTEM DESIGN FOR THE ANALYSIS OF UKRAINIAN-LANGUAGE CORPORATE DOCUMENTS IN INFORMATION SECURITY TASKS

This paper addresses the problem of designing the architecture of an information system intended for detecting sensitive information in Ukrainian-language corporate documents within the context of approaches used in data loss prevention (DLP) platforms. The relevance of this study is determined by the rapid growth of unstructured textual data in corporate information systems, as well as by the increasing need to combine information security requirements with sound principles of information system design and software architecture. An additional motivating factor is the limited number of scientific studies focused on automated processing of Ukrainian-language textual content, particularly with regard to its morphological complexity, syntactic variability, and stylistic features, which significantly influence the effectiveness of natural language processing methods. The paper provides an analytical overview of the main approaches to sensitive information detection in text documents, including rule-based and pattern-matching methods, as well as contextual methods based on named entity recognition using machine learning and deep learning models. Special attention is paid to the architectural implications of integrating heterogeneous detection methods within a single information system, which is essential for ensuring comparability, reproducibility, and extensibility of further experimental studies. Based on the conducted analysis, a modular architecture of an experimental software system is proposed. The architecture is designed to provide unified conditions for text document ingestion, preprocessing, and analysis, support for multiple detection methods with different computational characteristics, and centralized aggregation and evaluation of detection results using standard quality metrics. The proposed architectural solutions rely on the principles of streaming processing of large text documents, clear separation of responsibilities between functional components, scalability, and reproducibility of analytical procedures. The architecture also defines interfaces for the potential integration of rule-based scanners and contextual NER modules based on transformer models, forming a consistent environment for future comparative analysis. The proposed architectural approach can be used as a methodological and design foundation for the subsequent implementation and validation of an experimental information system aimed at studying methods for protecting textual information resources in corporate environments, as well as for supporting further research at the intersection of information system design and information security.

Keywords: *information security, sensitive data, DLP systems, NER, software system design, Ukrainian-language texts.*

Стаття надійшла до редакції / Received 18.01.2026
Прийнята до друку / Accepted 11.02.2026
Опубліковано / Published 26.03.2026

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Цифровізація бізнес-процесів та активне впровадження електронного документообігу зумовили стрімке зростання обсягів текстових корпоративних даних, що обробляються в інформаційних системах організацій. Значна частина таких даних містить персональні або конфіденційні відомості, які відповідно до міжнародних та національних нормативних актів підлягають особливому захисту [1]. Порушення вимог щодо обробки цих даних може призводити до фінансових санкцій, репутаційних втрат і правових наслідків для організацій.

В цих умовах захист інформаційних активів трансформується з суто технічної задачі у критично важливий елемент стратегічного управління підприємством. Значна частина даних, які сучасні організації щоденно генерують та обробляють, представлена у вигляді неструктурованих текстових документів: юридичних договорів, фінансових звітів, ділового листування, технічної документації та внутрішніх розпоряджень. За оцінками галузевих експертів, саме неструктуровані дані становлять найбільший ризик з точки зору витоку інформації, оскільки, на відміну від структурованих записів у реляційних базах даних, вони

не мають чіткої схеми зберігання, часто дублюються та важко піддаються автоматизованому контролю засобами класичного периметрального захисту. У цьому контексті системи запобігання витоку даних (Data Loss Prevention, DLP) є одним із ключових інструментів забезпечення інформаційної безпеки в корпоративних середовищах.

Одним з основних функціональних завдань DLP-систем є автоматизоване виявлення чутливої інформації в неструктурованих і напівструктурованих текстових документах. Виявлення чутливих сутностей у корпоративних документах пов'язане з низкою специфічних викликів, серед яких різноманітність форматів файлів, багатомовність контенту та необхідність врахування контексту для уникнення помилкових спрацювань [2]. Складність завдання полягає у необхідності виявляти чутливу інформацію (Personally Identifiable Information – PII, комерційні таємниці) у неструктурованих текстових документах в умовах високої варіативності контекстів, де одне й те саме слово може мати різне значення залежно від оточення.

Аналіз досліджень та публікацій

Проблема автоматичного виявлення чутливих сутностей у тексті історично вирішувалася за допомогою шаблонно-правилових (rule-based) методів [3], [4]. Ранні DLP-системи покладалися на пошук за ключовими словами та регулярними виразами. До переваг такого підходу дослідники відносять прозорість правил, чіткість результату та надвисоку швидкість обробки, що є критичним для сканування трафіку в реальному часі. Однак, як підкреслюють у своїй фундаментальній роботі I. Neamatullah та M. Douglass, варіативність запису й залежність від контексту створюють систематичні прогалини в таких системах [3]. Жорсткі правила не здатні ефективно обробляти помилки друку, нестандартне форматування або семантичну неоднозначність, що призводить до неприйнятного рівня пропусків загроз (False Negatives) у складних корпоративних документах.

Розвиток методів машинного навчання дозволив перейти до більш гнучких підходів. Зокрема, використання моделей умовних випадкових полів (Conditional Random Fields – CRF) дозволило враховувати локальний контекст слів та ймовірності переходів між тегами, що суттєво підвищило точність класифікації порівняно з простими словниковими методами [5]. Як показали Neamatullah I. та інші [3], ефективність CRF зростає за рахунок шаблонів ознак. До них можна віднести: базові, позиційні індикатори, словникові підказки. Усе це зменшує омонімічні помилки та можливі зсуви. Подальший прогрес у цій сфері пов’язаний із впровадженням архітектур глибокого навчання (Deep Learning). Комбінація двонаправлених рекурентних мереж та CRF (BiLSTM-CRF), описана Z. Liu та співавторами, стала де-факто стандартом для задач розпізнавання іменованих сутностей (Named Entity Recognition – NER) у середині 2010-х років, демонструючи високу здатність до узагальнення та стійкість до шуму в даних [6]. Ця комбінація дозволяє ідентифікувати сутності на основі контексту та семантичних зв’язків у тексті. Методи NER, які використовують рекурентні нейронні мережі, умовні випадкові поля та трансформерні архітектури, продемонстрували значне підвищення точності виявлення сутностей у порівнянні з rule-based підходами [5], [7]. Систематичні огляди підтверджують ефективність NER-моделей у задачах деідентифікації клінічних записів і фінансових документів [8], [9].

Сучасний етап характеризується домінуванням трансформерних моделей (BERT, RoBERTa), які використовують механізми уваги для моделювання глобальних залежностей у тексті. Як зазначають F. Dernoncourt та J. Y. Lee, такі моделі значно ефективніші за попередні аналоги при роботі з довгими документами, оскільки здатні враховувати контекст всього речення або навіть абзацу, проте вони вимагають значних обчислювальних ресурсів і специфічних підходів до токенізації [7]. Попри наявність ефективних алгоритмічних підходів, більшість наявних досліджень зосереджується або на вузькоспеціалізованих медичних застосуваннях (зокрема деідентифікації клінічних записів) [3], [8], [9] або на ізольованих академічних завданнях обробки природної мови, наприклад [10]. Водночас проблема проєктування комплексної архітектури для корпоративного середовища, здатної поєднувати різні методи аналізу та адаптуватися до специфіки бізнес-лексики, залишається недостатньо опрацьованою. Разом із тим, огляд сучасних датасетів і методик NER свідчить про домінування англомовних ресурсів і відсутність універсальних рішень для мов із розвиненою флективною морфологією [11]. Ефективність rule-based та NER-підходів у контексті опрацювання саме україномовних корпоративних документів досі вивчена обмежено. Додаткові методологічні труднощі зумовлені морфологічною складністю української мови, браком масштабних відкритих розмічених корпусів, а також особливостями оформлення ділової документації українською мовою. Такі тексти вирізняються специфічними структурними, стилістичними й лінгвістичними характеристиками та потребують окремого аналізу в межах автоматизованого опрацювання текстової інформації [2]. Зазначені чинники обумовлюють необхідність адаптації як алгоритмічних підходів, так і архітектурних рішень для забезпечення коректної обробки україномовних текстів у корпоративному середовищі. Таким чином, актуальною є задача проєктування програмної системи, здатної інтегрувати різні методи виявлення чутливої інформації та забезпечувати їх систематичний аналіз у контексті україномовних корпоративних документів.

Формулювання цілей статті

Метою роботи є: проєктування модульної програмної архітектури експериментальної інформаційної системи, призначеної для коректного та відтворюваного порівняльного аналізу rule-based і контекстних (NER) методів виявлення чутливої інформації в україномовних корпоративних документах, з урахуванням вимог системного проєктування, інформаційної безпеки та специфіки дослідницьких DLP-сценаріїв.

Виклад основного матеріалу

Проєктування програмного забезпечення експериментальної системи здійснювалося з урахуванням специфіки досліджуваної задачі, а саме – необхідності порівняльного аналізу методів виявлення чутливої інформації в україномовних корпоративних документах. Система розглядається не як промисловий DLP-продукт, а як дослідницька платформа, призначена для контрольованого тестування різних підходів до аналізу тексту в єдиному обчислювальному середовищі. Основною метою проєктування є створення програмної архітектури, яка забезпечує відтворюваність експериментів, порівнюваність результатів і можливість незалежного розвитку окремих функціональних компонентів. У межах такої архітектури передбачається підтримка двох принципово різних підходів до виявлення чутливої інформації – rule-based та контекстного (NER), а також уніфікований механізм оцінювання їх ефективності за спільними критеріями: якість на рівні знайдених токенів (Precision/Recall/F1), продуктивність (час обробки документа або набору документів), покриття класів сутностей (кількість типів, які виявляються з прийнятною точністю), стійкість до обфускацій даних, інтерпретованість (причини спрацювання/помилки).

Загальна структура експериментальної системи орієнтована на поетапну обробку документів і включає модулі зчитування даних, аналізу тексту, агрегації результатів та формування звітності. Такий підхід дозволяє чітко розмежувати відповідальність між компонентами та мінімізувати взаємні залежності між ними.

Контекстне представлення предметної області відображене у вигляді діаграми прецедентів (рис. 1), яка демонструє функціональні можливості системи та сценарії взаємодії з нею. Єдиним актором виступає дослідник, який здійснює налаштування параметрів експерименту, завантаження тестових документів, запуск процесу аналізу та генерацію підсумкового звіту. Така модель взаємодії відповідає дослідницькому характеру системи та не передбачає інтеграції з зовнішніми користувачами або автоматизованими корпоративними процесами.

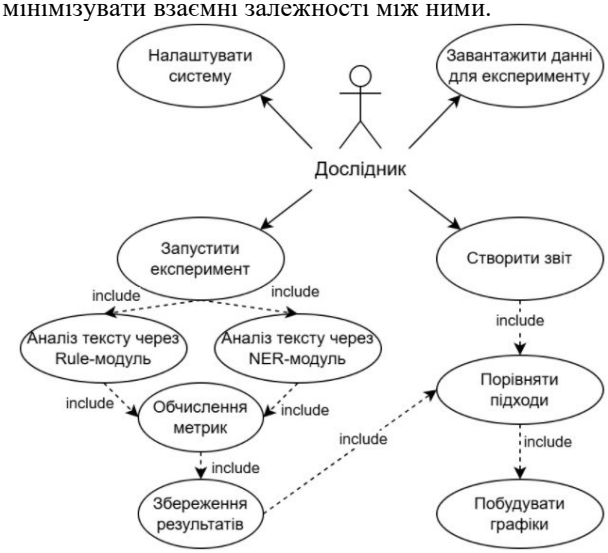

Рис. 1. Діаграма прецедентів експериментальної системи

Архітектурно система побудована за модульним принципом і містить такі групи компонентів: керування експериментом та взаємодія з користувачем, аналіз тексту, обробка та представлення результатів. Ключовим елементом є experiment_manager (рис. 2), який ініціалізує конфігурацію та підсистеми (NER-модель, rule-based сканер, сховища) і керує повним циклом експерименту. Обробка виконується поетапно: документ зчитується блоками, кожен блок аналізується двома методами (NER та rule-based), після чого результати об'єднуються, очищуються від дублікатів і оцінюються за precision/recall/F1 з формулюванням звіту. Це забезпечує коректне порівняння різних підходів до виявлення сутностей.

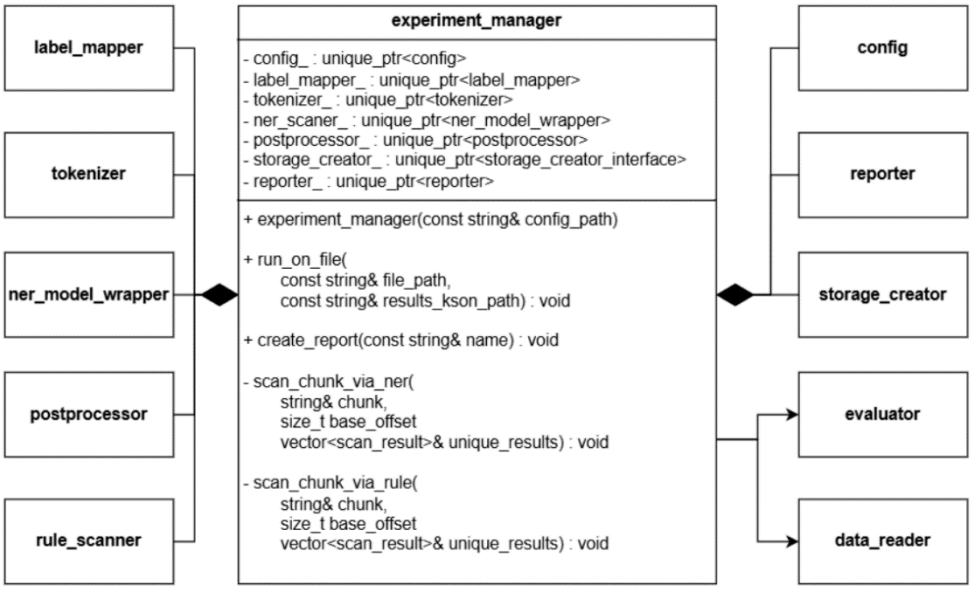

Рис. 2. Діаграма модуля керування експериментом

Параметризація експерименту здійснюється через окремий модуль конфігурації config (рис. 3), який реалізує централізований механізм зчитування та зберігання налаштувань. Під час ініціалізації клас має завантажувати JSON-файл, розбирати його структуру та зберігати дані у внутрішніх полях. Окремі методи

мають надавати готові значення для компонентів підсистеми, а функції get_rule_scanner_actions() та get_ner_scanner_class_names() будуть формувати списки правил і класів для ініціалізації сканерів. Зчитування та підготовка вхідних документів реалізується модулем data_reader (рис. 4), який абстрагує роботу з файлами різних форматів та забезпечує потокове зчитування тексту фрагментами фіксованого розміру. Під час виклику open_file() модуль буде створювати сховище даних, а потім виконувати попереднє вилучення текстового вмісту з файлу за допомогою docx_text_extractor (для файлів формату DOCX). Після цього буде ініціалізовано об'єкт chunk_reader, який дозволить поступово зчитувати текст. Метод read_next_chunk() має передавати чергову порцію тексту в буфер для подальшого аналізу. Такий підхід є принципово важливим для обробки великих корпоративних документів і дозволяє уникнути надмірного використання оперативної пам'яті.

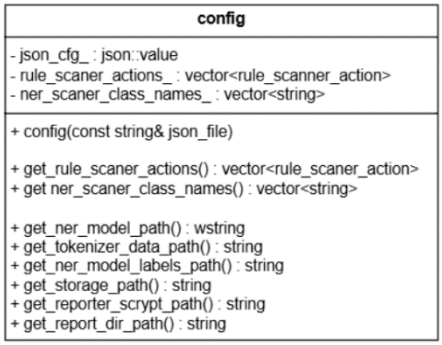

Рис. 3. Діаграма модуля конфігурації

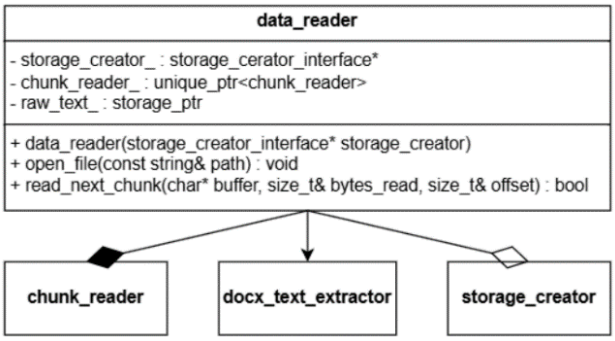

Рис. 4. Діаграма модуля зчитування даних

Для документів формату DOCX передбачено окремий механізм вилучення текстового вмісту, архітектура якого наведена на рис. 5. Документ розглядається як ZIP-контейнер, з якого вилучається XML-ресурс word/document.xml з подальшим синтаксичним розбором. Обробка XML-структури організована через фабрику відвідувачів, що дозволяє масштабувати парсер у разі появи нових типів вузлів.

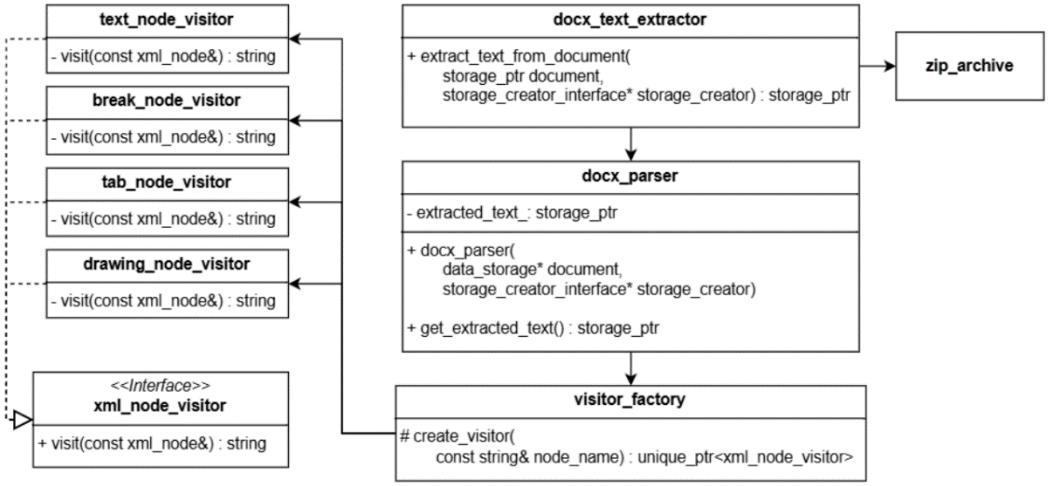

Рис. 5. Діаграма парсера DOCX-документів

Робота з архівами інкапсульована у модулі zip_archive (рис. 6), який використовує бібліотеку zlib та реалізує потокове вилучення даних великими блоками. Метод extract_file() має відкривати архів, зчитувати його метадані, виділяти цільове сховище через storage_creator, а далі потоково копіювати вміст великими блоками.

Така схема розмежовує відповідальність між розпаковкою та зберіганням даних, а використання принципів проєктування PIMPL та RAII дозволять приховати деталі реалізації та поліпшити керування ресурсами.

Модуль storage (рис. 7) буде реалізовувати універсальну підсистему зберігання даних, яка має абстрагувати джерело – оперативну пам'ять або диск через спільний інтерфейс data_storage. Основна ідея полягає в тому, щоб уніфікувати доступ до потоку даних незалежно від того, де він фізично знаходиться, і надати можливість ефективної роботи з великими файлами. Класи memory_storage та file_storage будуть реалізовувати відповідно зберігання даних в оперативній пам'яті та на диску. Перший має використовуватись для невеликих файлів і буде забезпечувати швидкий доступ до інформації, тоді як другий має створювати тимчасові файли на диску й слідкувати за їх часом життя.

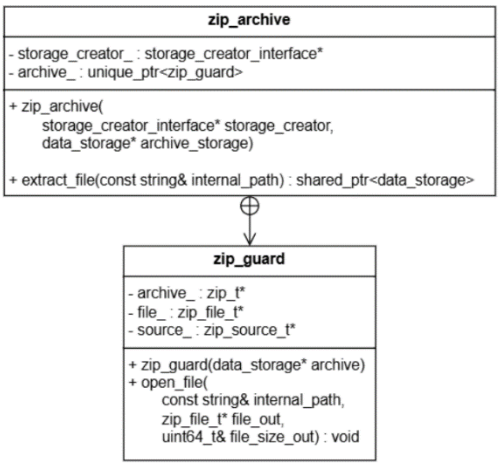

Рис. 6. Діаграма модуля розпакування ZIP-

контейнерів

Клас chunk_reader надасть змогу зчитувати дані частинами із перекриттям, що зручно для потокового аналізу великих документів. Компонент storage_creator буде реалізовувати фабрику сховищ, яка автоматично обирає тип зберігання залежно від розміру файлу. Така архітектура забезпечить гнучке керування ресурсами та спростить роботу з даними для інших модулів системи.

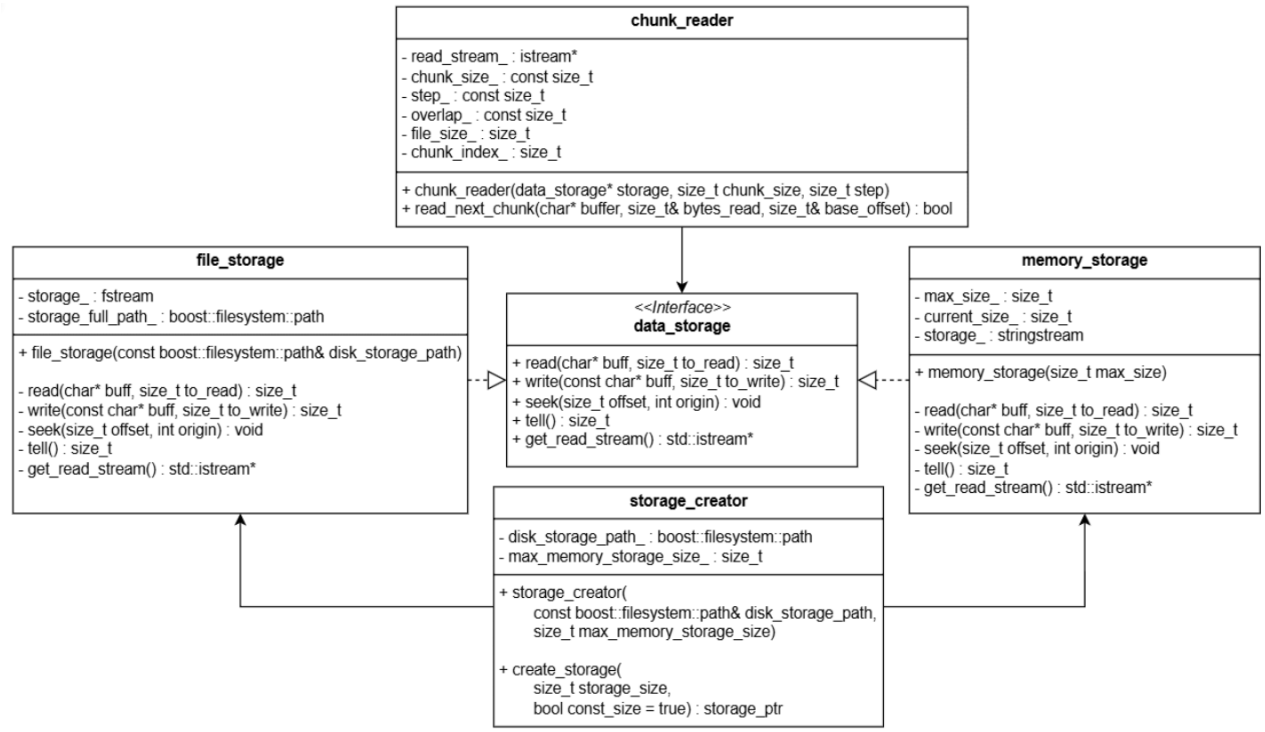

Рис. 7. Діаграма модуля зберігання даних

Контекстний аналіз тексту реалізується NER-модулем (рис. 8), який інкапсулює повний конвеєр контекстного виявлення сутностей. Клас tokenizer має використовувати SentencePiece для стабільного розбиття на субтокени та формування вхідних тензорів. Клас label_mapper буде завантажувати і зберігати дані про BIO мітки та нормалізуватиме їх. Основним класом для аналізу тексту буде виступати ner_model_wrapper, який інкапсулює використання моделі XLM-RoBERTa через бібліотеку ONNX Runtime. На самому початку він має валідувати сигнатуру моделі, створити тензори представлення й вирахувати прогін. Після інференсу має здійснюватися сортування і вибірка по виміру класів для кожного знайденого токена та зібратися вектор token_info, де буде збережено ID класу, бал, розмічений токен і його офсети. Реалізація postprocessor має групувати сусідні токени однієї сутності, ігноруючи клас O та корегуючи можливі помилки послідовностей. Для кожної групи будуть склеюватись текст і офсети в результаті формуючи scan_result з глобальними офсетами, результуючим текстом і назвою класу.

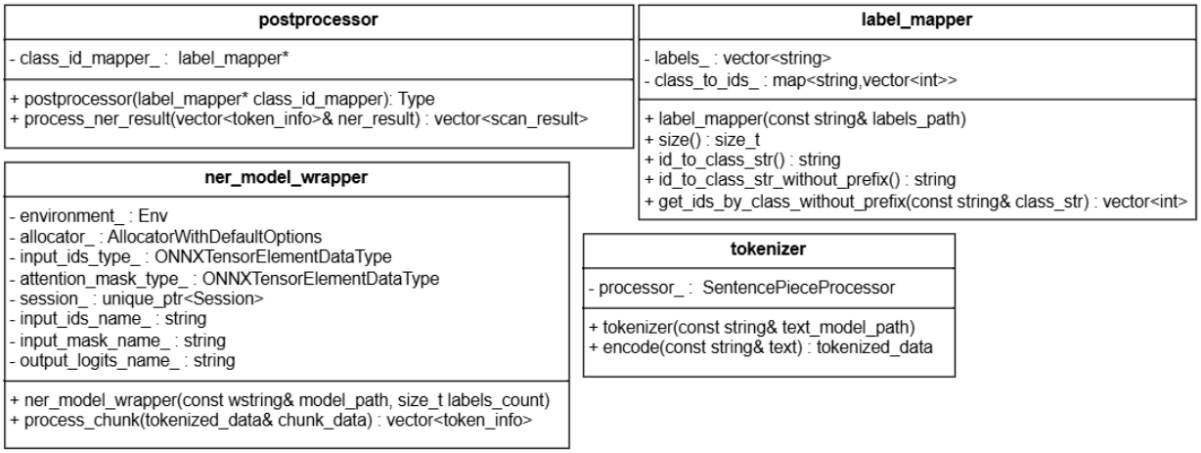

Рис. 8. Діаграми класів NER-модуля

Rule-based сканування, а саме виявлення сутностей у тексті на основі регулярних виразів, реалізовано у вигляді окремого модуля сканування правилами (рис. 9). На етапі ініціалізації він має приймати набір правил і компілювати регулярні вирази. Під час сканування, метод scan_chunk() буде нормалізувати фрагмент тексту, а далі виконається перевірка регулярним виразом. За наявності збігу модуль зобов'язаний сформулювати результат у вигляді вектору scan_result і розрахувати оригінальні офсети для знайдених фрагментів.Архітектурне відокремлення цього компонента дозволяє чітко простежити його внесок у загальні результати аналізу та порівняти його ефективність із контекстним підходом.

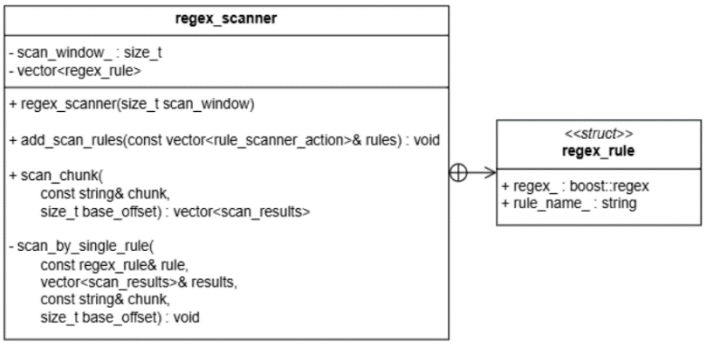

Рис. 9. Діаграма модуля rule-based сканування

Оцінювання результатів і формування звітності здійснюється спеціалізованими модулями – модулем оцінки та модулем підготовки результатів (рис. 10). Оцінювання буде реалізовано зіставленням знайдених сутностей з еталонною розміткою, після чого відбудеться підрахунок True Positive/False Positive/False Negative випадків та розрахунок метрик precision, recall та F1-score. Клас оформлення результатів reporter буде накопичувати результати NER і правил та формувати графіки по кожному класу сутностей.

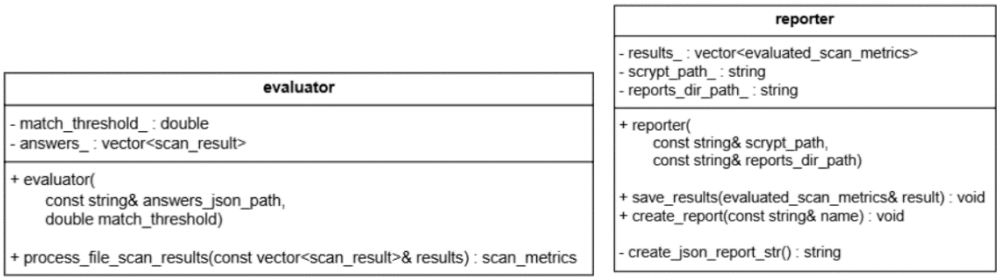

Рис. 10. Діаграми модулів оцінювання та звітності

Потоки даних в експериментальній системі організовані у вигляді послідовного конвеєра, що охоплює етапи завантаження, аналізу, оцінювання та звітності (рис. 11).

1. Етап завантаження: data_loader буде зчитувати документ і отримувати сирий текст з нього для подальшого аналізу.
2. Етап аналізу: кожен сканер почне аналіз тексту. regex_scanner почне пошук конфіденційних даних за правилами, які були надані в конфігурації. Для NER методу сканування буде викликано tokenizer, який розіб'є текст на токени. Далі токенізований текст буде передано в ner_model_wrapper, після чого результат буде оброблено через postprocessor для формування фінального результату пошуку та виправлення можливих помилок моделі.
3. Етап оцінювання: evaluator розрахує необхідні метрики та порівняє результати аналізу тексту. 4. Етап звітності: reporter сформує таблиці з результатами експерименту.

Така організація забезпечує прозорість обробки даних і дозволяє відтворювати експериментальні результати.

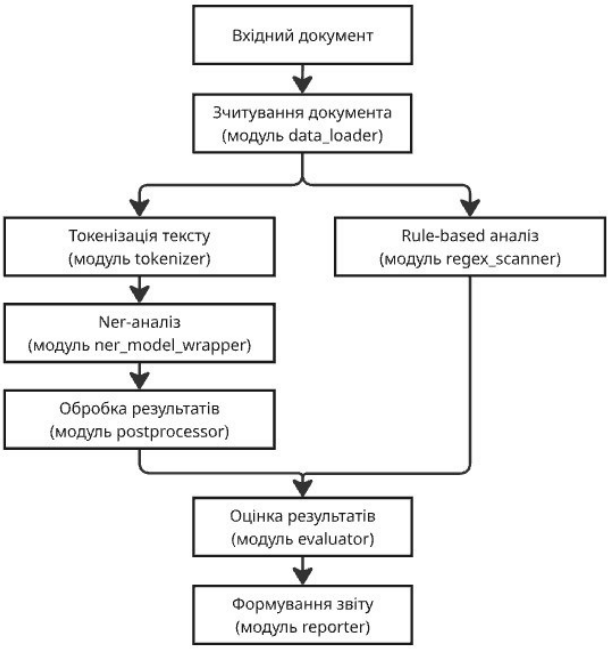

Рис. 11. Потоки даних в системі

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Наукова новизна роботи полягає у системному проєктуванні експериментальної архітектури для порівняльного аналізу методів виявлення чутливої інформації в україномовних корпоративних документах, орієнтованої на потреби дослідження DLP-платформ.

На відміну від наявних досліджень, у яких rule-based та NER-підходи зазвичай аналізуються ізольовано або в різних програмних середовищах, у цій роботі запропоновано єдину модульну архітектуру, що забезпечує:

– однакові умови обробки вхідних даних;
– уніфіковані механізми зчитування, нормалізації та потокового аналізу тексту;
– спільний контур оцінювання результатів за стандартизованими метриками якості.

Новизна архітектурного підходу полягає також у поєднанні потокової обробки великих корпоративних документів із контекстним NER-аналізом, що дозволяє масштабувати експериментальні дослідження без втрати відтворюваності результатів. Запропонована структура системи створює основу для коректного дослідження впливу мовної специфіки українських ділових текстів на ефективність методів виявлення чутливої інформації, що раніше залишалося поза увагою більшості робіт.

У роботі розглянуто підхід до проєктування експериментальної програмної системи, призначеної для порівняльного аналізу методів виявлення чутливої інформації в україномовних корпоративних документах у контексті DLP-платформ. Основну увагу зосереджено на архітектурних рішеннях, які забезпечують коректність, відтворюваність і порівнюваність результатів дослідження.

Запропонована модульна архітектура дозволяє інтегрувати rule-based і контекстні NER-методи в єдине середовище обробки текстових даних, не порушуючи принципу розмежування відповідальності між компонентами. Реалізація потокового зчитування документів, уніфікованої підсистеми зберігання даних та централізованого керування експериментом створює передумови для аналізу великих масивів корпоративних текстів з урахуванням їх структурної та мовної специфіки.

Використання трансформерної моделі XLM-RoBERTa у складі контекстного NER-сканера забезпечує можливість урахування семантичних і контекстних залежностей, характерних для україномовних ділових документів, тоді як rule-based модуль слугує базовою точкою порівняння. Така організація системи дозволяє надалі досліджувати сильні та слабкі сторони кожного підходу в умовах реальних корпоративних текстів.

Отримані архітектурні рішення можуть бути використані як методологічна основа для подальших наукових досліджень у сфері захисту текстових даних, а також для проєктування експериментальних компонентів у складі DLP-систем, орієнтованих на обробку україномовного контенту.

Література

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (GDPR) [Електронний ресурс] / European Parliament, Council of the European Union. – Режим доступу : https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng – (Дата звернення : 10.10.2025).

2. Kabak R. A. Виявлення чутливих сутностей у корпоративних документах для систем Data Loss Prevention: виклики, методи та напрями дослідження / R. A. Kabak, O. A. Liashenko, O. A. Radul // Комп'ютерне моделювання та оптимізація складних систем (КМОСС-2025) : матеріали IX Міжнар. наук.-техн. конф., м. Дніпро. – Дніпро : ДВНЗ УДХТУ, 2025. – С. 187–189. – Режим доступу : https://udhtu.edu.ua/wp-content/uploads/2025/11/zbirnyk-kmoss-2025_compressed.pdf.

3. Neamatullah I. Automated de-identification of free-text medical records / I. Neamatullah, M. M. Douglass, L.-W. H. Lehman [et al.] // BMC Medical Informatics and Decision Making. – 2008. – Vol. 8. – Art. 32. – DOI: https://doi.org/10.1186/1472-6947-8-32.

4. Zhao Z. Re-examination of rule-based methods in de-identification of electronic health records: Algorithm development and validation / Z. Zhao, M. Yang, B. Tang, T. Zhao // JMIR Medical Informatics. – 2020. – Vol. 8, № 4. – e17622. – DOI: https://doi.org/10.2196/17622.

5. Liu Z. De-identification of clinical notes via recurrent neural network and conditional random field / Z. Liu, B. Tang, X. Wang, Q. Chen // Journal of Biomedical Informatics. – 2017. – Vol. 75. – P. S34–S42. – DOI: https://doi.org/10.1016/j.jbi.2017.05.023.

6. Liu Z. Entity recognition from clinical texts via recurrent neural network / Z. Liu, M. Yang, X. Wang, Q. Chen // BMC Medical Informatics and Decision Making. – 2017. – Vol. 17, Suppl. 2. – P. 67. – DOI: https://doi.org/10.1186/s12911-017-0468-7.

7. Dernoncourt F. De-identification of patient notes with recurrent neural networks / F. Dernoncourt, J. Y. Lee, O. Uzuner, P. Szolovits // Journal of the American Medical Informatics Association. – 2017. – Vol. 24, № 3. – P. 596–606. – DOI: https://doi.org/10.1093/jamia/ocw156.

8. Negash B. De-identification of free text data containing personal health information: A scoping review of reviews / B. Negash, A. Katz, C. J. Neilson [et al.] // International Journal of Population Data Science. – 2023. – Vol. 8, № 1. – Art. 2153. – DOI: https://doi.org/10.23889/ijpds.v8i1.2153.

9. Kovačević A. De-identification of clinical free text using natural language processing: A systematic review of current approaches / A. Kovačević, B. Bašaragin, N. Milošević, G. Nenadić // Artificial Intelligence in Medicine. – 2024. – Vol. 151. – Art. 102845. – DOI: https://doi.org/10.1016/j.artmed.2024.102845.

10. Au T. W. T. E-NER – An annotated named entity recognition corpus of legal text [Електронний ресурс] / T. W. T. Au [et al.]. – 2022. – Режим доступу : https://arxiv.org/abs/2212.09306.

11. Jehangir B. A survey on Named Entity Recognition – datasets, tools, and methodologies / B. Jehangir, S. Radhakrishnan, R. Agarwal // Natural Language Processing Journal. – 2023. – Vol. 3. – Art. 100017. – DOI: https://doi.org/10.1016/j.nlp.2023.100017.

References

1. European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed October 10, 2025).

2. Kabak, R. A., Liashenko, O. A., Radul, O. A. Detection of sensitive entities in corporate documents for Data Loss Prevention systems: challenges, methods, and research directions. In: *Computer Modeling and Optimization of Complex Systems (KMOSS-2025): Proceedings of the 9th International Scientific and Technical Conference*. Dnipro: Ukrainian State University of Chemical Technology, 2025, pp. 187–189. Available at: https://udhtu.edu.ua/wp-content/uploads/2025/11/zbirnyk-kmoss-2025_compressed.pdf.

3. Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., Clifford, G. D. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 2008, vol. 8, art. 32. https://doi.org/10.1186/1472-6947-8-32.

4. Zhao, Z., Yang, M., Tang, B., Zhao, T. Re-examination of rule-based methods in de-identification of electronic health records: algorithm development and validation. *JMIR Medical Informatics*, 2020, vol. 8, no. 4, e17622. https://doi.org/10.2196/17622.

5. Liu, Z., Tang, B., Wang, X., Chen, Q. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 2017, vol. 75, pp. S34–S42. https://doi.org/10.1016/j.jbi.2017.05.023.

6. Liu, Z., Yang, M., Wang, X., Chen, Q. Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, 2017, vol. 17, Suppl. 2, p. 67. https://doi.org/10.1186/s12911-017-0468-7.

7. Dernoncourt, F., Lee, J. Y., Uzuner, O., Szolovits, P. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 2017, vol. 24, no. 3, pp. 596–606. https://doi.org/10.1093/jamia/ocw156.

8. Negash, B., Katz, A., Neilson, C. J., Moni, M., Nesca, M., Singer, A., Enns, J. E. De-identification of free text data containing personal health information: a scoping review of reviews. *International Journal of Population Data Science*, 2023, vol. 8, no. 1, art. 2153. https://doi.org/10.23889/ijpds.v8i1.2153.

9. Kovačević, A., Bašaragin, B., Milošević, N., Nenadić, G. De-identification of clinical free text using natural language processing: a systematic review of current approaches. *Artificial Intelligence in Medicine*, 2024, vol. 151, art. 102845. https://doi.org/10.1016/j.artmed.2024.102845.

10. Au, T. W. T., et al. *E-NER: An annotated named entity recognition corpus of legal text*. arXiv preprint, 2022. Available at: https://arxiv.org/abs/2212.09306.

11. Jehangir, B., Radhakrishnan, S., Agarwal, R. A survey on named entity recognition: datasets, tools, and methodologies. *Natural Language Processing Journal*, 2023, vol. 3, art. 100017. https://doi.org/10.1016/j.nlp.2023.100017.