УДК:651.1:005.932

Швець Олена Вікторівна
(наук. керівник – канд. екон. наук, доцент Прігунов О. В.)
Донецький національний університет імені Василя Стуса, м. Вінниця

ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В ЦИФРОВІЗАЦІЇ ТА АНАЛІЗІ ДОКУМЕНТНИХ МАСИВІВ

Анотація. Розглянуто сутність документних масивів як відносно статичних і впорядкованих сукупностей документів, що формуються відповідно до визначеного призначення та мають специфічні ознаки. Висвітлено значення аналізу документів як якісного методу дослідження інформаційних ресурсів. Окрему увагу приділено процесам цифровізації та оцифрування документів, їх перевагам і труднощам. Розглянуто сучасні підходи до вирішення проблем оцифрування, зокрема використання технологій OCR та штучного інтелекту, а також подано алгоритм автоматизованої обробки документів.

Ключові слова: документні масиви, аналіз документів, цифровізація, оцифровування, OKR.

Документні масиви – відносно статичні, більш упорядковані, створюються через добір, оскільки формуються відповідно до конкретного призначення, мають профіль і сукупність специфічних ознак документів, що обмежують формування масиву [**Помилка! Джерело посилання не знайдено.,** с. 3].

Масивам документів притаманні специфічні властивості:

- термінальність – призначеність для відбору та кумулювання профільних видів документів з метою організації їх функціонування в каналах документної комунікації;
- скінченність – обмеженість функціональним призначенням та іншими ознаками;
- стабільність – пристосованість для введення й виведення окремих видів документів з метою їх упорядкування й організації розповсюджування.

Для якісного опрацювання інформаційних ресурсів виникає потреба в аналізі, вивченні та обробці документних масивів.

Аналіз документів – це якісний метод дослідження письмових документів, візуальних ресурсів або інших фізичних матеріалів з метою інтерпретації значення, виявлення закономірностей або відстеження змін. Він може включати аналіз протоколів зустрічей, особистих листів, урядових звітів, фотографій або рукописних нотаток – будь-якого артефакту, яка відображає інформацію, що стосується дослідницького питання.

Аналіз документів – це практичний та гнучкий метод оцінки документів та інших матеріалів. Він широко використовується в тематичних дослідженнях, історичних дослідженнях, оцінці політики та якісних дослідженнях, що вивчають комунікацію, значення або організаційні процеси [6].

Цифровізація – це процес впровадження та використання цифрових технологій і методів у різні аспекти життя та діяльності, включно з бізнесом, управлінням, освітою та повсякденним побутом. У межах цифровізації відбувається переведення інформації та процесів у цифровий вигляд, що дає змогу автомати-

зувати завдання, підвищити ефективність, покращити комунікацію та розширити можливості [**Помилка! Джерело посилання не знайдено.**; 4].

Оцифрування – це переведення друкованих видань і документів у якісний електронний формат. Оцифрування документів є сучасним та прогресивним процесом, проте має певні складнощі (рис. 1).

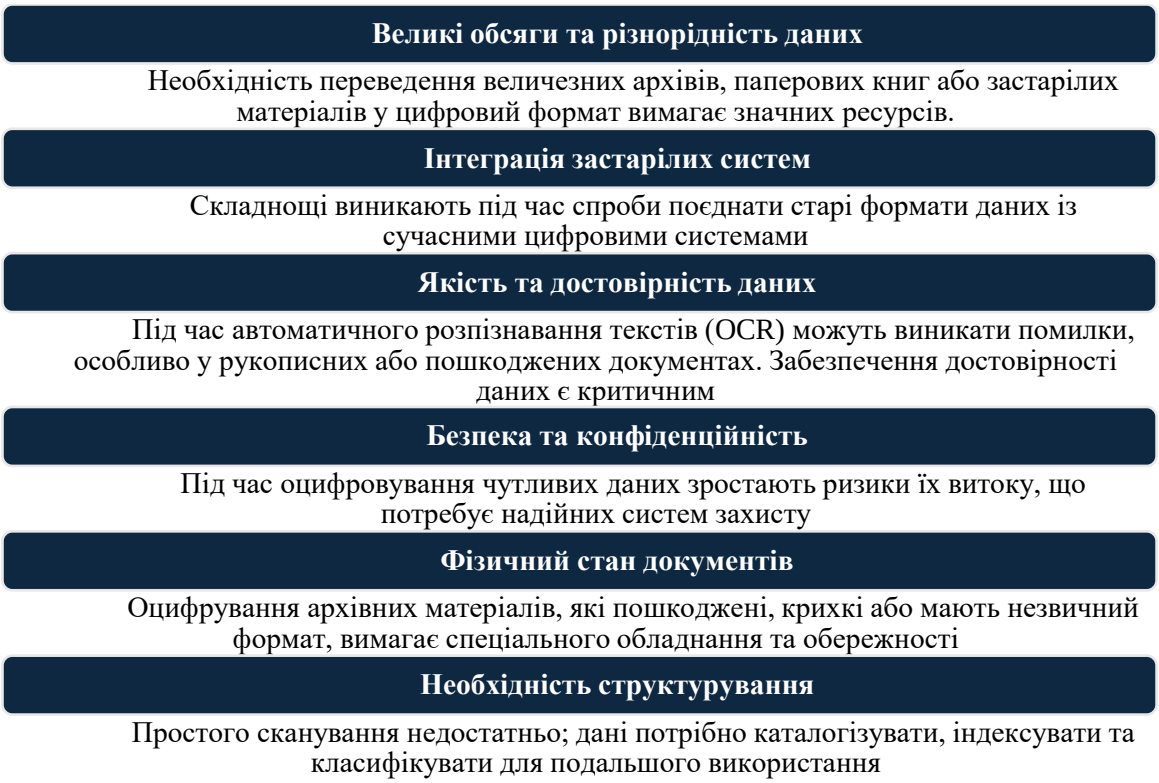

Рисунок 1 – Складнощі під час процесу оцифрування

Для вирішення проблем оцифрування документних масивів пропоновано такі рішення (рис. 2).

Рисунок 2 – Рішення проблеми оцифрування

OCR (оптичне розпізнавання символів) – це технологія, яка перетворює зображення тексту, як-от скановані документи чи фотографії, у цифровий текст. Це дає змогу редагувати, шукати та зберігати текст в електронному вигляді, полегшуючи роботу з документами та керування ними [**Помилка! Джерело посилання не знайдено.**].

Інтелектуальні системи оптичного розпізнавання символів особливо ефективні в галузях із великою кількістю документів, автоматизуючи процес зчитування, розуміння та обробки документів. Ці системи дотримуються структурованого, покращеного штучним інтелектом конвеєра, який починається зі введення документів і закінчується виведенням структурованих даних.

Алгоритм вирішення проблем та аналіз документних масивів шляхом оптичного розпізнавання символів за допомогою штучного інтелекту (рис. 3).

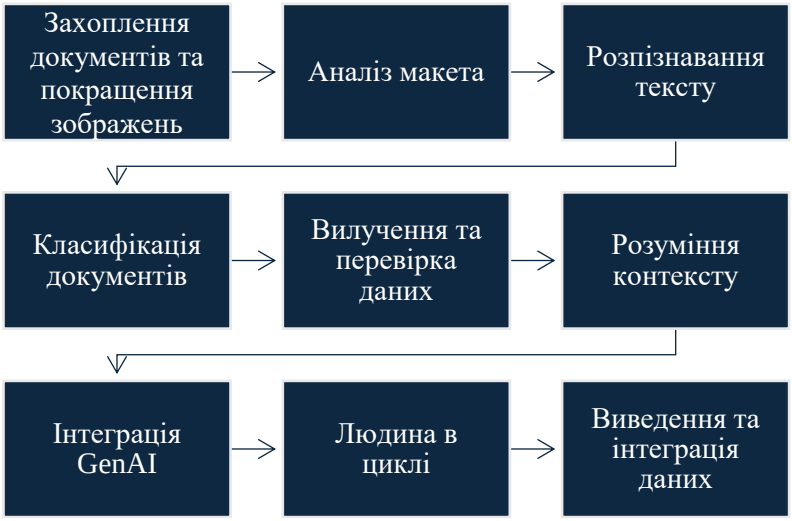

Рисунок 3 – Алгоритм OCR та ШІ

1. Захоплення документів та покращення зображень: процес починається із захоплення документа, який може бути будь-чим: від сканованої форми до PDF-файла та фотографії зі смартфона. Документи можуть бути отримані з мобільних пристроїв, електронної пошти, спільних папок, мережевих сканерів та прямих підключень до бізнес-систем через API або попередньо створені конектори.

2. Аналіз макета: система виконує аналіз макета, щоб виявити структурні елементи, як-от таблиці, текстові блоки, зображення, штрих-коди, галочки та підписи. Цей крок зберігає логічну структуру документа під час обробки.

3. Розпізнавання тексту: потім система використовує OCR та ICR для оцифрування друкованого та рукописного тексту, готуючи його до подальшої обробки. Ці технології здатні розпізнавати логічну структуру всього документа, що дає змогу класифікувати документи, витягувати дані та високоякісний експорт у цифрові формати.

4. Класифікація документів: моделі класифікації на основі штучного інтелекту аналізують як текстові, так і графічні характеристики, щоб розпізнавати та впорядковувати документи, класифікуючи кожен документ за типом. Отже, кожен документ може бути спрямований через відповідний робочий процес обробки.

5. Вилучення та перевірка даних: тепер дані можна точно витягувати зі структурованих, напівструктурованих та неструктурованих документів. Ключові дані, як-от імена, дати та номери довідок, витягуються з документа за допомогою передового штучного інтелекту та машинного навчання, що імітує людське розуміння. Потім витягнуті дані перевіряються на відповідність бізнес-правилам або системам компанії для впевненості, що все відповідає дійсності.

6. Розуміння контексту: для інтерпретації значення та контексту вилученої інформації використовується обробка природної мови (NLP). Наприклад, система може визначити, чи слово «Меркурій» стосується хімічного елементу, плане-

ти чи марки автомобіля, а «Рахунок» – це ім’я людини чи рахунок-фактура, який потрібно оплатити.

7. Інтеграція GenAI: після того, як дані будуть надійно вилучені з документа, відповідні фрагменти даних можна надіслати до LLM для виконання конкретних завдань, наприклад, класифікації типу контракту та підсумовування його ключових зобов’язань простою мовою для швидшого перегляду.

8. Людина в циклі: якщо щось виглядає не так або відсутнє, система надсилає це людині на перевірку – цей процес називається перевіркою «людина в циклі» (HITL). Щоразу, коли вноситься виправлення, моделі ШІ вдосконалюються завдяки постійному навчанню та стають точнішими. Цей крок є вирішальним, коли потрібна 100 %-ва точність або коли документ не відповідає конкретним правилам перевірки, встановленим для кожної моделі ШІ.

9. Виведення та інтеграція даних: зрештою, чисті, структуровані дані можна експортувати у відповідний файл – JSON, CSV, XML чи інші [5].

Отже, ефективне опрацювання документних масивів є важливим складником управління інформаційними ресурсами. Незважаючи на наявні труднощі цифровізації та оцифрування документів, застосування сучасних технологій, як-от OCR і штучний інтелект, відкриває широкі можливості для автоматизації, підвищення точності та оптимізації роботи з документами. Це сприяє покращенню доступу до інформації, її структуризації та подальшому ефективному використанню.

Список використаних джерел

1. Парафійник Н. І. Документно-інформаційні комунікації: навч. посіб. Харків: Нац. аерокосм. ун-т ім. М. Є. Жуковського «ХАІ», 2011. Ч. ІІ. 144 с.

2. Роль OCR у оцифровці документів. *Shaip*. 2023. URL: https://uk.shaip.com/blog/ocr-in-document-digitization/ (дата звернення: 03.04.2026).

3. Цифровізація (Діджиталізація) – це що таке, визначення. *Karapuziki*. 2025. URL: https://karapuziki.com.ua/tsyfrovizatsiia-didzhytalizatsiia-tse-shcho-take-vyznachennia/ (дата звернення: 16.03.2026).

4. Прігунов О. В., Яворська Ю. Л. Цифрова трансформація національних архівів і бібліотек на основі інформаційно-аналітичних систем. *Культура, інформація, комунікація: міждисциплінарний діалог:* матеріали Всеукр. наук. конф. Київ, 2025. С. 121.

5. Що таке оптичне розпізнавання символів (OCR) за допомогою штучного інтелекту та чому це важливо. *ABBYY*. 2025. URL: https://www.abbyy.com/blog/ai-ocr/ (дата звернення: 13.04.2026).

6. Як аналіз документів сприяє кращим дослідницьким рішенням. *LinkedIn*. 2025. URL: https://www.linkedin.com/pulse/what-document-analysis-how-drives-better-research-decisions-fkpqc/ (дата звернення: 11.03.2026).