УДК 004.8:004.932:004.93
DOI https://doi.org/10.32782/IT/2025-1-14

Дмитро КНИШ
здобувач ОС магістр, Державний торговельно економічний університет, вул. Кіото, 19, Київ, Україна, індекс 02156
***ORCID:** 0009-0008-5001-8432*

Наталія КОТЕНКО
кандидат педагогічних наук, доцент, доцент кафедри інженерії програмного забезпечення та кібербезпеки, Державний торговельно економічний університет, вул. Кіото, 19, Київ, Україна, індекс 02156
***ORCID:** 0000-0002-2675-6514*
***Scopus Author ID:** 57223192989*

Бібліографічний опис статті: Книш, Д., Котенко, Н. (2025). Сучасні підходи та виклики у поєднанні технологій OCR та NLP для автоматизованого аналізу друкованого тексту. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 98–104, doi: https://doi.org/10.32782/IT/2025-1-14

СУЧАСНІ ПІДХОДИ ТА ВИКЛИКИ У ПОЄДНАННІ ТЕХНОЛОГІЙ OCR ТА NLP ДЛЯ АВТОМАТИЗОВАНОГО АНАЛІЗУ ДРУКОВАНОГО ТЕКСТУ

У статті розглянуто сучасні підходи до поєднання технологій оптичного OCR та NLP з метою автоматизованого аналізу друкованого тексту. Проведено порівняльний аналіз методів OCR і NLP з акцентом на точність розпізнавання, багатомовну підтримку та контекстуальне розуміння тексту. Особливу увагу приділено нейронним мережам, трансформерним моделям (зокрема TrOCR, BERT, GPT) та алгоритмам глибокого навчання, що забезпечують високу ефективність обробки текстових даних. Запропоновано новий підхід до інтеграції OCR та NLP, який дозволяє підвищити точність і швидкість аналізу, а також адаптувати системи до специфіки різних форматів тексту. Практичне значення дослідження полягає у його застосуванні в галузях освіти, медицини, права та логістики. Визначено основні переваги і виклики таких інтегрованих систем, включаючи обчислювальну складність, чутливість до якості зображень та потребу в якісних навчальних даних.

***Метою дослідження** є розгляд сучасних підходів до інтеграції технологій OCR та NLP для автоматизованого аналізу друкованого тексту. Метою є підвищення точності, ефективності та швидкості обробки таких систем шляхом використання нейронних мереж, трансформерів та алгоритмів машинного навчання.*

***Методологія.** У статті проведено порівняльний аналіз існуючих методів OCR і NLP, зосереджений на точності розпізнавання, підтримці багатомовності та контекстному розумінні. У дослідженні оцінюється продуктивність різних підходів залежно від швидкості обробки та адаптивності до різних форматів тексту.*

***Наукова новизна:** запропоновано новий підхід до інтеграції OCR-NLP, який оптимізує як точність, так і швидкість обробки. На відміну від традиційних методів, це дослідження акцентує увагу на синергії між передовими технологіями глибокого навчання та звичайними стратегіями розпізнавання тексту.*

***Висновки.** Інтеграція технологій OCR і NLP відкриває нові можливості для автоматизованого аналізу друкованого тексту, значно покращуючи точність і ефективність обробки даних. Подальші дослідження мають зосередитися на підвищенні швидкості роботи алгоритмів та їх адаптації до рукописного і багатомовного тексту, що розширить сферу їх застосування та ефективність.*

***Ключові слова:** OCR, NLP, глибоке навчання, трансформери, автоматизований аналіз тексту, багатомовність.*

Dmytro KNYSH
Master's Student, State University of Trade and Economics, 19, Kyoto Str., Kyiv, Ukraine, 02156, dymaknysh@gmail.com
***ORCID:** 0009-0008-5001-8432*

Nataliia KOTENKO
Candidate of Pedagogical Sciences, Associate Professor, Associate Professor at the Department of Software Engineering and Cybersecurity, State University of Trade and Economics, 19, Kyoto Str., Kyiv, Ukraine, 02156, kotenkono@outlook.com
***ORCID:** 0000-0002-2675-6514*
***Scopus Author ID:** 57223192989*

To cite this article: Knysh, D., Kotenko, N. (2025). Suchasni pidkhody ta vyklyky u poiednanni tekhnolohii OCR ta NLP dlia avtomatyzovanoho analizu drukovanoho tekstu [Modern approaches and challenges in combining OCR and NLP technologies for automated printed text analysis]. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 98–104, doi: https://doi.org/10.32782/IT/2025-1-14

MODERN APPROACHES AND CHALLENGES IN COMBINING OCR AND NLP TECHNOLOGIES FOR AUTOMATED PRINTED TEXT ANALYSIS

The article examines modern approaches to the integration of OCR and NLP technologies for automated analysis of printed texts. A comparative analysis of OCR and NLP methods is presented, focusing on recognition accuracy, multilingual support, and contextual understanding. Particular attention is paid to neural networks, transformer-based models (such as TrOCR, BERT, GPT), and deep learning algorithms that ensure high efficiency in text processing. A new approach to OCR-NLP integration is proposed, which enhances both the accuracy and processing speed of such systems and enables their adaptation to various text formats. The practical value of the study lies in its applicability across domains such as education, medicine, law, and logistics. The main advantages and challenges of integrated systems are outlined, including computational complexity, sensitivity to image quality, and the need for high-quality training data.

***The objective** of the study aims to explore current approaches to integrating OCR and NLP technologies for automated printed text analysis. The objective is to increase the accuracy, efficiency, and processing speed of such systems through the use of neural networks, transformers, and machine learning algorithms.*

***Methodology.** The article presents a comparative analysis of existing OCR and NLP methods, focusing on recognition accuracy, multilingual support, and contextual understanding. The performance of various approaches is evaluated based on processing speed and adaptability to different text formats.*

***The novelty** of this study a new approach to OCR-NLP integration is proposed, optimizing both recognition accuracy and processing speed. Unlike traditional methods, the study emphasizes the synergy between cutting-edge deep learning techniques and conventional text recognition strategies.*

***The results.** The integration of OCR and NLP technologies opens new opportunities for automated printed text analysis, significantly improving data processing accuracy and efficiency. Further research should focus on enhancing algorithm speed and adapting to handwritten and multilingual texts to expand the scope and effectiveness of such systems.*

***Key words:** OCR, NLP, deep learning, transformers, automated text analysis, multilingualism.*

Актуальність проблеми. Нині OCR (Optical Character Recognition) та NLP (Natural Language Processing) перебувають на етапі стрімкого розвитку. Це пов'язано зі всебічним впровадженням глибокого навчання і збільшенням доступності цифрових технологій. Значного прогресу досягнуто у обробці багатомовних текстів, рукописів та документів з застарілими шрифтами (Smith, 2007).

Окрім того, нові алгоритми, такі як трансформери, значно покращили точність розпізнавання тексту в різноманітних умовах. Вони здатні враховувати контекст, що дозволяє ефективно обробляти навіть складні чи неструктуровані дані. Розпізнавання тексту в рукописах стало набагато точнішим завдяки використанню нейронних мереж, які можуть вивчати індивідуальні стилі письма та адаптуватися до них.

Водночас NLP дозволяє глибше розуміти значення тексту, не лише обробляючи слова, а й виявляючи зв'язки між ними. Це відкриває нові можливості для аналізу документів, автоматичного перекладу, а також для створення систем підтримки прийняття рішень в реальному часі. Ці досягнення також стимулюють розвиток нових додатків у галузях від юриспруденції до медицини, де точність та швидкість обробки інформації є критичними.

Аналіз останніх досліджень і публікацій. У сучасному інформаційному середовищі інтелектуальна обробка тексту набуває стратегічного значення, особливо в контексті цифровізації та автоматизації роботи з великими масивами текстових даних. Комбінація технологій OCR та NLP стала перспективним напрямом наукових досліджень, що знаходить застосування в багатьох галузях.

Одним із найбільш відомих та стабільних інструментів OCR є Tesseract, який детально описано у роботі R. Smith (Smith, 2007). Автор акцентує увагу на архітектурі системи та її здатності адаптуватися до багатьох мов, що є важливою передумовою для реалізації багатомовних OCR-NLP систем. Smith підкреслює значення попередньої обробки зображення для підвищення точності розпізнавання, що підтверджується у подальших публікаціях.

Низка досліджень зосереджуються на розширенні функціональності NLP моделей шляхом інтеграції статистичних та логічних структур. Наприклад, у публікації H. Tao та ін. (Tao, 2022) запропоновано підхід до навчання моделей класифікації тексту за допомогою запитань і відповідей, що покращує їхню інтерпретованість та загальну ефективність.

У сфері глибокого навчання важливе місце посідає архітектура LSTM (Long Short-Term Memory), яка дозволяє моделі зберігати довготривалі залежності в тексті. Розроблена S. Hochreiter та J. Schmidhuber (Hochreiter, 1997), ця модель є базисом для сучасних RNN-систем і використовується у задачах як OCR, так і NLP.

Окрему увагу у дослідженнях приділено аналізу україномовних текстів. У роботі О. Окунькової (Окунькова, 2023) висвітлюється застосування сучасних інформаційних технологій для автоматичного аналізу текстів українською мовою, що є особливо актуальним у контексті створення локалізованих NLP моделей.

Значний внесок у розвиток NLP зробили дослідження векторного представлення слів, зокрема моделі Word2Vec (Mikolov, 2013), які дозволяють кодувати семантичну інформацію у вигляді багатовимірних векторів. Це є критичним етапом у побудові високоточних NLP-систем.

У сучасних публікаціях також активно вивчається тема обробки мови в режимі «end-to-end». У роботі Prabhavalkar R. та ін. (Prabhavalkar, 2023) розглянуто підходи до цілісної обробки мовлення без попередньої сегментації, що має потенціал і в OCR/NLP інтеграції, особливо при роботі з рукописними та аудіотекстами.

Особливу роль відіграють трансформери – архітектура, запропонована A. Vaswani (Vaswani, 2023), стала основою для створення сучасних моделей на кшталт BERT, GPT, TrOCR. Механізм самоуваги дозволяє таким моделям одночасно враховувати як локальний, так і глобальний контекст, що критично важливо для розпізнавання складних або структурно неоднорідних текстів.

Дослідження Hemmer A. та ін. (Hemmer, 2024) розширює тему постобробки результатів OCR, зокрема в аспекті виявлення помилок розпізнавання. Це особливо важливо для побудови стійких до похибок систем, які працюють із неякісними або спотвореними джерелами.

Загальну картину сучасного стану інтеграції OCR та NLP доповнює огляд постобробних підходів, представлений у роботі Nguyen T. та ін. (Nguyen, 2021). Автори систематизують методи виправлення помилок OCR за допомогою лінгвістичних моделей та нейронних мереж, акцентуючи увагу на важливості контекстуального аналізу результатів розпізнавання.

Мета дослідження. У світлі стрімкого зростання обсягів цифрової інформації та необхідності її ефективної обробки, особливо у форматі сканованих документів, друкованих текстів та багатомовних джерел, постає актуальне завдання створення високоточних, адаптивних і масштабованих інструментів для автоматизованого аналізу текстових даних. У цьому контексті мета дослідження полягає в теоретичному обґрунтуванні, технічному аналізі та практичному узагальненні сучасних підходів до інтеграції OCR та NLP з метою формування комплексної моделі для автоматизованої обробки друкованих текстів.

Основна мета полягає у виявленні, систематизації та порівнянні сучасних технологічних рішень, що забезпечують ефективне злиття можливостей OCR, як інструменту переведення візуальної текстової інформації в машиночитану форму, та NLP, як системи, здатної глибоко аналізувати, інтерпретувати та трансформувати текст на основі лінгвістичних і статистичних закономірностей.

Виклад основного матеріалу дослідження. Сучасні підходи до OCR демонструють суттєвий прогрес завдяки активному впровадженню технологій глибокого навчання та розвитку архітектур нейронних мереж. Одним із провідних напрямів є застосування згорткових нейронних мереж (CNN) та рекурентних нейронних мереж (RNN), які зарекомендували себе як високоефективні інструменти для обробки зображень тексту. Зокрема, CNN використовуються для виділення ключових ознак текстових елементів на сканованих документах або фото, тоді як RNN забезпечують здатність моделі враховувати послідовність символів і їхній контекст. Такий підхід дозволяє значно зменшити кількість помилок у процесі розпізнавання. Яскравим прикладом реалізації цього методу є OCR-система Tesseract, яка в останніх версіях інтегрує можливості глибокого навчання, що позитивно впливає

на її точність, зокрема при роботі з друкованими матеріалами, що мають неоднорідну структуру або фонові артефакти (Smith, 2007).

Окрім класичних нейронних мереж, в OCR дедалі активніше впроваджуються моделі на базі трансформерної архітектури, зокрема TrOCR (Transformer Optical Character Recognition). Ці моделі використовують механізм самоуваги, який дозволяє ефективно враховувати взаємозв'язки між окремими частинами зображення тексту. Завдяки цьому TrOCR демонструє високу точність при розпізнаванні текстів різних стилів, шрифтів, мов і структур. Цей підхід особливо ефективний для складних і багатомовних документів, які раніше створювали труднощі для традиційних OCR-систем.

Ще одним важливим напрямом розвитку OCR є підтримка багатомовності за рахунок використання моделей, орієнтованих на Unicode. Такі системи здатні одночасно працювати з текстами, написаними різними мовами, що має критичне значення для глобальних інформаційних проєктів, багатокультурних архівів, міжнародних юридичних документів та інших сфер, де в одному документі може міститися кілька мовних кодів. Таким чином, сучасні OCR-технології не лише підвищують точність і адаптивність до різних типів текстів, а й забезпечують високу гнучкість при обробці мультикультурного цифрового контенту.

Кожна з розглянутих систем OCR складається з кількох ключових етапів (Martin, 2008), таких як попередня обробка зображення, сегментація, розпізнавання тексту, постобробка.

Попередня обробка зображення включає: перетворення кольорового або градаційного зображення у двоколірне для полегшення виділення тексту; усунення артефактів, таких як плями або зайві лінії, які можуть заважати розпізнаванню, корекція нахилу чи викривлення тексту для більш точного аналізу.

Сегментація включає виділення окремих блоків тексту, рядків, слів або символів із загального зображення, а також використання методів, таких як адаптивні алгоритми порогової обробки або CNN.

Розпізнавання тексту поділяють на шаблонний підхід (порівняння символів із попередньо збереженими шаблонами) та моделі глибокого навчання (використання нейронних мереж, що аналізують особливості тексту для точнішого розпізнавання).

Постобробка включає корекцію помилок розпізнавання на основі мовних моделей або словників, а також виправлення формату, структури тексту та інтеграція в кінцеву систему.

Розглядаючи методи CNN, RNN, TrOCR та багатомовні моделі можна зазначити, що кожен має свої переваги та недоліки. Нижче наведемо особливості роботи кожного з методів, головні переваги та обмеження.

Методи оптичного розпізнання символів з використанням глибокого навчання, а саме з використання CNN мають певні переваги порівняно з іншими підходами. Однією з головних є навчання нейронної мережі, яке дозволяє працювати з ширшим спектром вхідних даних, що дозволяє адаптуватися під різні розміри шрифтів, макетів тексту та інших особливостей вхідних даних. Проте через це потребується велика кількість обчислювальних ресурсів, особливо у процесі навчання, а також довший час на створення, налаштування та навчання моделі. Також важливим є набір даних, який використовується у процесі навчання моделі, оскільки саме від нього залежить точність розпізнання тексту.

Трансформери, це сучасна архітектура глибокого навчання, що здійснила революцію у сфері NLP і дедалі більше застосовується у задачах OCR. Їхньою основною інновацією є механізм самоуваги (self-attention) (Tao, 2022), який дозволяє ефективно аналізувати кожен елемент вхідних даних у контексті всіх інших елементів послідовності. У випадку OCR це забезпечує високу точність розпізнавання тексту, навіть у документах зі складною та неоднорідною структурою.

Сучасні системи OCR стрімко розвиваються завдяки впровадженню трансформерних архітектур, які значно перевершують традиційні моделі в точності та адаптивності. Архітектура трансформерів базується на двох ключових компонентах: енкодері та декодері. Енкодер здійснює обробку вхідних даних, формуючи їх приховане представлення, тоді як декодер на основі цього представлення генерує текстовий результат. Ключовим елементом є механізм самоуваги (self-attention), який дозволяє кожному елементу вхідної послідовності враховувати інші елементи, що дає змогу моделі працювати з глобальним контекстом та розв'язувати проблеми довгих залежностей, властиві архітектурам RNN (Tao, 2022).

Однією з найуспішніших реалізацій трансформерів у задачах OCR є модель TrOCR, яка поєднує візуальний енкодер, побудований на основі Vision Transformer (ViT), та текстовий GPT-подібний декодер. Алгоритм її роботи складається з декількох етапів. Спочатку зображення тексту розбивається на невеликі фрагменти (патчі), що конвертуються у вектори

ознак. Потім Vision Transformer аналізує ці фрагменти з урахуванням їхньої взаємодії за допомогою самоуваги, формуючи приховане представлення зображення. На наступному етапі текстовий декодер генерує текст, послідовно відтворюючи символи на основі цього прихованого представлення, а також уже згенерованих елементів. У результаті формується повноцінна текстова послідовність, що відтворює зміст вхідного зображення.

Серед головних переваг трансформерів в OCR можна виокремити контекстуальне розуміння – здатність моделі враховувати не лише локальні, а й глобальні особливості тексту; гнучкість – адаптацію до різних типів даних (друковані документи, рукописи, тексти зі спотвореннями); а також цілісну обробку послідовностей, що усуває потребу в додаткових етапах сегментації.

Водночас трансформери мають певні обмеження. По-перше, їхня обчислювальна складність є квадратичною відносно довжини послідовності:

$$O(n^2 d), \quad (1)$$

де n – довжина послідовності; d – розмірність векторного представлення токена.

Через це час розпізнавання тексту збільшується із збільшенням кількості вхідних токенів.

Також до недоліків можна віднести високі вимоги до пам'яті. Матриця уваги має розміри R^{n*n}, що потребує значного обсягу пам'яті для роботи алгоритму. Однак деякі методи, такі як, Sparse Attention, можуть знижувати складність до логарифмічної або навіть лінійної.

Ще одна важлива проблема – висока залежність від великої кількості навчальних даних. Через відсутність вбудованої пам'яті, як у LSTM (Hochreiter, 1997), трансформери потребують великих і якісно збалансованих корпусів для адекватного навчання. Інакше модель не здатна сформувати стійке узагальнення і має схильність до overfitting – запам'ятовування навчальної вибірки без перенесення знань на нові дані (Hochreiter, 1997). Формально це проявляється у зниженні точності на тестових даних при збереженні низької функції втрат на тренуванні:

$$I = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \quad (2)$$

де y_i – правильна ймовірність; $\hat{y}_i$ – прогнозована ймовірність.

Одним із суттєвих обмежень сучасних OCR-систем, зокрема побудованих на трансформерних моделях, є висока залежність від великих навчальних вибірок. Якщо обсяг тренувальних даних недостатній, модель не здатна сформувати коректні статистичні узагальнення. У таких випадках вона схильна до оверфітингу – тобто запам'ятовування навчальної вибірки без здатності до узагальнення. Це призводить до низьких втрат (loss) на тренувальних даних і водночас – до значних похибок на нових, раніше не бачених текстах. Таким чином, модель демонструє хороші результати лише в межах того, чого «навчилась дослівно», і не здатна коректно працювати з незнайомими прикладами.

Ще однією актуальною темою є підтримка багатомовного OCR, яка реалізується через інтеграцію моделей з підтримкою Unicode. Це дозволяє обробляти документи, що містять текст кількома мовами одночасно. Подібні системи базуються на глибоких нейронних мережах, які комбінують переваги традиційних CNN/RNN підходів із трансформерною архітектурою, здатною опрацьовувати широкий мовний спектр. Універсальність таких моделей відкриває перспективи для реалізації глобальних цифрових платформ, але водночас ускладнює навчання та оптимізацію, особливо з огляду на значну варіативність мов, алфавітів і граматичних структур.

Загальними проблемами як для OCR, так і для моделей NLP залишаються:

- низька якість вхідних зображень (розмитість, шуми, викривлення);
- нестандартні шрифти та рідкісні мови, які знижують точність розпізнавання;
- висока обчислювальна складність, що робить моделі важкими для використання на пристроях із обмеженими ресурсами або в умовах великого навантаження;
- семантична неоднозначність, яку сучасні системи не завжди можуть врахувати, що призводить до помилок у розпізнаванні контексту.

Попри ці обмеження, технології OCR залишаються одним із найбільш потужних інструментів цифровізації. Сфери їхнього застосування постійно розширюються – від сканування архівів до автоматизованого аналізу зображень у соціальних мережах. Наприклад, у текстовій аналітиці OCR використовується для отримання статистичних характеристик: довжини тексту, кількості слів, частоти ключових слів, водності, визначення тематики тощо. Одним із таких прикладів є сервіс Istio, що дозволяє аналізувати як текстові файли, так і фото з текстовим наповненням (Окунькова, 2023, с. 75).

Завдяки активному розвитку глибокого навчання, трансформерів і комп'ютерного зору OCR поступово долає свої обмеження. Інтеграція з NLP забезпечує не лише корекцію

результатів, а й семантичний аналіз, класифікацію та витяг ключових даних, що робить технологію універсальним інструментом для обробки інформації у цифровій екосистемі.

Ключовим компонентом NLP-моделей є векторні представлення слів, які дозволяють комп'ютерам опрацьовувати слова як точки в багатовимірному просторі:

$$\vartheta(w) \in R^d, \quad (3)$$

де d – розмірність векторного простору, $\vartheta(w)$ – векторне представлення слова, R^d – d-вимірний евклідовий простір дійсних чисел, де кожен вимір відповідає одній координаті у векторному поданні слова. Такі моделі як Word2Vec, GloVe, FastText дозволяють моделювати семантичну подібність і контекстні зв'язки між словами, що є основою сучасної текстової аналітики.

NLP активно застосовується в багатьох сферах, зокрема:

- генерація тексту (ChatGPT, Jasper);
- голосові помічники (Siri, Google Assistant);
- резюмування;
- аналіз настроїв і тональності;
- пошукові системи з семантичним розпізнаванням;
- персоналізовані рекомендації (Netflix, YouTube).

Моделі NLP можуть суттєво відрізнятись за архітектурою та призначенням.

BERT – модель заснована на трансформерній архітектурі, яка використовує лише енкодер. Основна особливість BERT – двонаправлений контекст, що дозволяє аналізувати слова у взаємозв'язку з усім реченням [6]. Завдяки цьому модель досягає високої точності в завданнях класифікації, аналізу тональності, виявлення іменованих сутностей (NER) та інших завданнях розуміння тексту.

GPT – модель на основі трансформерів, яка, на відміну від BERT, використовує лише декодер і працює в режимі авторегресії (тобто передбачає наступне слово на основі попередніх). Завдяки цьому GPT є однією з найкращих моделей для генерації тексту (Prabhavalkar, 2023), перекладу, створення чат-ботів та креативного письма. Водночас вона схильна до «галюцинацій» – генерації неправдивої або непослідовної інформації (Vaswani, 2023), також не має двонаправленого контексту, що може знижувати точність в аналітичних NLP-завданнях.

spaCY – високопродуктивна бібліотека для обробки природної мови, яка оптимізована для швидкості та ефективності. Вона забезпечує токенізацію, частиномовний аналіз, лематизацію, синтаксичний аналіз та розпізнавання іменованих сутностей. Завдяки добре оптимізованому коду spaCy ідеально підходить для виробничих NLP-систем, де важлива продуктивність. Однак вона менш гнучка в навчанні нових моделей у порівнянні з BERT або GPT і може поступатися точністю у складних NLP-завданнях.

NLTK – потужний інструмент для обробки природної мови, який містить широкий набір методів, зокрема токенізацію, морфологічний аналіз, частиномовне тегування, стемінг, роботу з мовними корпусами та аналіз синтаксичних дерев. NLTK широко використовується в академічних дослідженнях і навчальних цілях завдяки своїй гнучкості. Водночас він значно поступається spaCy за швидкістю та вимагає більше коду для реалізації базових NLP-завдань.

Перевагами які характерні для всіх моделей є автоматизація обробки тексту, що значно спрощує пошук та генерацію контенту. Як зазначалося NLP моделі мають глибоке розуміння тексту, що може допомагати у задачах визначення настроїв текстів та розуміння залежності у реченнях.

Спільне використання технологій OCR та NLP дозволяє мінімізувати недоліки обох підходів та значно розширити можливості використання. Поєднання технологій дозволяє отримати інтелектуальні системи обробки тексту, які не лише розпізнають, а й розуміють зміст документа.

Сумісне використання OCR і NLP успішно використовується у багатьох реальних задачах, таких як: автоматизація обробки документів; аналіз історичних архівів і рукописних текстів; розпізнавання тексту на зображеннях із соціальних медіа.

Серед сучасних систем де використовується комбінація методів OCR та NLP для структурованого аналізу, класифікації та вилучення ключових даних з текстів можна виділити такі: Google Document AI, Amazon Textract, ABBYY FlexiCapture, Microsoft Azure Form Recognizer.

Інтеграція технологій OCR та NLP відкриває нові можливості для автоматизованого аналізу друкованого тексту, значно підвищуючи точність та ефективність обробки. OCR дозволяє перетворювати друковані документи у цифровий формат, тоді як NLP допомагає розпізнавати контекст та виправляти помилки (Nguyen, 2021).

Висновки і перспективи подальших досліджень. Аналіз існуючих підходів показав, що сучасні методи, такі як глибоке навчання та трансформерні моделі, забезпечують найкращі результати у розпізнаванні та подальшому аналізі тексту. Наступні дослідження у цій

сфері можуть бути спрямовані на покращення швидкості алгоритмів та адаптацію їх до рукописних документів та багатомовних текстів. Сумісне використання технологій OCR та NLP є перспективним і вже є сервіси які демонструють їх успішну роботу. Проте, є ряд проблем, які все ще є складними для вирішення через певні обмеження.

ЛІТЕРАТУРА:

1. Smith R. An Overview of the Tesseract OCR Engine // Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, Curitiba, Parana, Brazil, 23–26 September 2007. P 1–5. https://doi.org/10.1109/icdar.2007.4376991

2. Martin J. H., Jurafsky D. Speech and Language Processing. 2nd ed. Prentice Hall, 2008. P 1–29.

3. Teaching Text Classification Models Some Common Sense via Q &A Statistics: A Light and Transplantable Approach / H. Tao et al. Natural Language Processing and Chinese Computing. Cham. *Springer International Publishing*. 2022. P. 593–605. https://doi.org/10.1007/978-3-031-17120-8_46

4. Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Computation*. 1997. Vol. 9, no. 8. P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

5. Окунькова О. Сучасні інформаційні технології аналізу україномовних текстів. *Вісник Кременчуцького національного університету імені Михайла Остроградського*. 2023. No. 1. P. 1–7. https://doi.org/10.32782/1995-0519.2023.1.10

6. Efficient Estimation of Word Representations in Vector Space / Mikolov T., Chen K., Corrado G., Dean J. 2013. P 1–12.

7. End-to-End speech recognition: a survey / R. Prabhavalkar та ін. 2023. С. 1–27.

8. Attention is all you need / A. Vaswani та ін. 2023. С. 1–15.

9. Confidence-Aware document OCR error detection / A. Hemmer та ін. 2024.

10. Survey of Post-OCR Processing Approaches / T. T. H. Nguyen et al. *ACM Computing Surveys*. 2021. Vol. 54, no. 6. P. 1–37. https://doi.org/10.1145/3453476

REFERNCES:

1. Smith, R. (2007). An overview of the tesseract OCR engine. U Ninth international conference on document analysis and recognition (ICDAR 2007) vol 2. IEEE. https://doi.org/10.1109/icdar.2007.4376991

2. Martin,J. H., Jurafsky, D. (2008). Speech & language processing. Dorling Kindersley India. P 1–29.

3. Tao, H., Zhu, G., Xu, T., Liu, Q., & Chen, E. (2022). Teaching text classification models some common sense via Q &a statistics: A light and transplantable approach. U Natural language processing and chinese computing (s. 593–605). *Springer International Publishing*. https://doi.org/10.1007/978-3-031-17120-8_46

4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

5. Okunkova, O. (2023). Suchasni informatsiini tekhnolohii analizu ukrainomovnykh tekstiv [Modern information technologies for analyzing Ukrainian-language texts] *Visnyk Kremenchutskoho natsionalnho universytetu imeni Mykhaila Ostrohradskoh*o, (1), 73–79. https://doi.org/10.32782/1995-0519.2023.1.10

6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1–12.

7. Prabhavalkar, R., Hori, T., Sainath, T. N., Schluter, R., & Watanabe, S. (2023). End-to-End speech recognition: A survey. 1–27.

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is all you need. 1–15.

9. Hemmer, A., Coustaty, M., Bartolo, N., & Ogier, J.-M. (2024). Confidence-Awфare document OCR error detection.

10. Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021). Survey of post-ocr processing approaches. *ACM Computing Surveys*, 54 (6), 1–37. https://doi.org/10.1145/3453476