

Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0826U003470

Особливі позначки: відкрита

Дата реєстрації: 13-07-2026

Статус: Захищена

Реквізити наказу МОН / наказу закладу:



II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Божко Олександр Юрійович

2. Oleksandr Bozhko

Кваліфікація:

Ідентифікатор ORCID ID: Не застосовується

Вид дисертації: доктор філософії

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 122

Назва наукової спеціальності: Комп'ютерні науки

Галузь / галузі знань: інформаційні технології

Освітньо-наукова програма зі спеціальності: Комп'ютерні науки

Дата захисту: 31-07-2026

Спеціальність за освітою: Автоматизовані системи обробки інформації та управління

Місце роботи здобувача: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): PhD 14141

Повне найменування юридичної особи: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

V. Відомості про дисертацію

Мова дисертації: Українська

Коди тематичних рубрик: 20.17.15, 20.53.19, 28.23.24, 20.54.06

Тема дисертації:

1. Моделі та методи екстракції слабоструктурованих даних з PDF-документів
2. Models and methods for semi-structured data extraction from PDF documents.

Реферат/анотація:

1. Божко О. Ю. Моделі та методи екстракції слабоструктурованих даних з PDF-документів. – Кваліфікаційна наукова праця на правах рукопису. Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 Комп'ютерні науки. (галузь знань 12 Інформаційні технології) – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2026. Актуальність дослідження зумовлена протиріччям між практичною потребою в надійній автоматизованій екстракції слабоструктурованих даних з різномірних PDF-документів та відсутністю науково обґрунтованого інструментарію для прогнозування якості такої екстракції та обґрунтованого вибору стратегії їх обробки. Метою дисертаційного дослідження є розробка моделей та методів, які б забезпечили реалізацію процедури автоматизованої адаптивної екстракції слабоструктурованих даних з PDF-документів на основі обґрунтованого вибору стратегії їх обробки з урахуванням прогнозованої якості екстракції та економічних критеріїв. Об'єктом дослідження є процес автоматизованої екстракції слабоструктурованих даних з PDF-документів. Предметом дослідження є моделі та методи адаптивної екстракції слабоструктурованих даних з

PDF-документів. Наукова новизна проведеного дослідження полягає в наступному: 1. Уперше розроблено метод прогнозування якості екстракції слабоструктурованих даних з PDF-документів на основі регресійного аналізу залежності міри F1 від системи бінарних ознак складності документа, який, на відміну від існуючих методів, що використовують характеристики запиту або рівень впевненості моделі, дозволяє апіорно оцінити очікуваний результат обробки до її виконання. 2. Уперше розроблено метод автоматичної анотації ознак складності PDF-документів з використанням мультимодальних великих мовних моделей, який, на відміну від ручного маркування, забезпечує масштабованість формування навчальних вибірок без залучення людини. 3. Уперше розроблено метод вибору стратегії екстракції слабоструктурованих даних на основі порогової політики вибору, що, на відміну від підходів з пробною обробкою, інтегрує класифікацію рівня складності документа та прогнозування якості екстракції з урахуванням обмежень на вартість і мінімально допустиму якість без попередньої обробки документа. 4. Удосконалено модель представлення слабоструктурованих даних у PDF-документах шляхом введення профілю складності на основі системи бінарних ознак структурних характеристик, що, на відміну від існуючих моделей, дозволяє формально описати властивості документа, релевантні для прогнозування якості автоматизованої екстракції слабоструктурованих даних. 5. Удосконалено модель оцінки якості екстракції слабоструктурованих даних шляхом введення багаторівневої системи метрик на основі декомпозиції за семантичними групами бізнес-сутностей документа, що, на відміну від інтегральних показників якості, дозволяє диференційовано оцінювати результати екстракції на рівні метаданих, записів та окремих полів і локалізувати джерела помилок. 6. Удосконалено метод класифікації документів за рівнем їх складності шляхом застосування багатокласової логістичної регресії на системі бінарних ознак структурної складності документа, що, на відміну від існуючих підходів до класифікації документів за типом або тематикою, дозволяє стратифікувати документи за прогнозованим рівнем якості автоматизованої екстракції за допомогою великих мовних моделей.

2. Bozhko O. Yu. Models and methods for semi-structured data extraction from PDF documents. – Qualifying scientific work as a manuscript. Dissertation for the degree of Doctor of Philosophy in specialty 122 Computer Science (field of knowledge 12 Information Technologies) – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2026. The relevance of this study stems from the contradiction between the practical need for reliable automated extraction of semi-structured data from heterogeneous PDF documents and the lack of scientifically sound tools for predicting the quality of such extraction and for making an informed choice regarding the processing strategy. The purpose of the dissertation research is to develop models and methods that would ensure the implementation of an automated adaptive extraction procedure for semi-structured data from PDF documents based on an informed selection of a processing strategy taking into account the predicted extraction quality and economic criteria. The object of research is the process of automated extraction of semi-structured data from PDF documents. The subject of research is models and methods of adaptive extraction of semi-structured data from PDF documents. The scientific novelty of the conducted research is as follows: 1. For the first time, a method for predicting the quality of semi-structured data extraction from PDF documents has been developed based on regression analysis of the dependence of the F1 measure on a system of binary document complexity features, which, unlike existing methods that use query characteristics or model confidence level, allows a priori estimation of the expected processing result before its execution. 2. For the first time, a method for automatic annotation of complexity features of PDF documents using multimodal large language models has been developed, which, unlike manual labeling, ensures scalability of training sample formation without human involvement. 3. For the first time, a method for selecting a semi-structured data extraction strategy based on a threshold selection policy has been developed, which, unlike approaches with trial processing, integrates document complexity level classification and extraction quality prediction taking into account cost constraints and minimum acceptable quality without prior document processing. 4. The model for representing semi-structured data in PDF documents has been improved by introducing a complexity profile based on a system of binary structural characteristic features, which, unlike existing models, allows formally describing document properties relevant to predicting the quality of automated

semi-structured data extraction. 5. The model for evaluating the quality of semi-structured data extraction has been improved by introducing a multi-level metric system based on decomposition by semantic groups of business entities of the document, which, unlike integral quality indicators, allows differentiated evaluation of extraction results at the level of metadata, records, and individual fields and localization of error sources. 6. The method for classifying documents by their complexity level has been improved by applying multiclass logistic regression to a system of binary features of document structural complexity, which, unlike existing approaches to classifying documents by type or subject matter, allows stratifying documents by the predicted quality level of automated extraction using large language models.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки: Інформаційні та комунікаційні технології

Стратегічний пріоритетний напрям інноваційної діяльності: Впровадження нових технологій та обладнання для якісного медичного обслуговування, лікування, фармацевтики

Підсумки дослідження: Теоретичне узагальнення і вирішення важливої наукової проблеми

Публікації:

- 1. Божко О. Ю. Використання великих мовних моделей для розпізнавання інформації в нумізматичних описах. Наука і техніка сьогодні. 2024. № 1(29). С. 615–625.
- 2. Божко О. Ю. Розробка ітеративного методу екстракції даних з неструктурованих документів на основі використання великих мовних моделей. Вісник Кременчуцького національного університету імені Михайла Остроградського. 2025. № 1. С. 119–124.
- 3. Божко О. Ю. Метод класифікації документів за складністю екстракції даних великими мовними моделями. Системи та технології. 2025. Вип. 2(70). С. 84–92.
- 4. Петров К. Е., Божко О. Ю. Метод автоматичної розмітки ознак структурної складності документів із використанням великих мовних моделей. Автоматизовані системи управління та прилади автоматики. 2026. Вип. 188. С. 54–68.

Наукова (науково-технічна) продукція: програмні продукти, програмно-технологічна документація

Соціально-економічна спрямованість: підвищення продуктивності праці

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами:

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Петров Костянтин Едуардович
2. Костянтин Е. Петров

Кваліфікація: д.т.н., професор, 01.05.04

Ідентифікатор ORCID ID: 0000-0003-1973-711X

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VII. Відомості про офіційних опонентів та рецензентів

Офіційні опоненти

Власне Прізвище Ім'я По-батькові:

1. Турута Олексій Петрович

2. Oleksii Turuta

Кваліфікація: к. т. н., доцент, 05.13.06

Ідентифікатор ORCID ID: 0000-0002-0970-8617

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, Харків, Харківський р-н., 61022, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Ізонін Іван Вікторович

2. Ivan V. Izonin

Кваліфікація: д. т. н., доц., 05.13.23

Ідентифікатор ORCID ID: 0000-0002-9761-0096

Додаткова інформація:

Повне найменування юридичної особи: Національний університет "Львівська політехніка"

Код за ЄДРПОУ: 02071010

Місцезнаходження: вул. Степана Бандери, Львів, 79013, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Рецензенти

Власне Прізвище Ім'я По-батькові:

1. Чалий Сергій Федорович

2. Serhii F. Chalyi

Кваліфікація: д.т.н., професор, 05.13.06

Ідентифікатор ORCID ID: 0000-0002-9982-9091

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Золотухін Олег Вікторович

2. Oleg V. Zolotukhin

Кваліфікація: к. т. н., доцент, 05.13.23

Ідентифікатор ORCID ID: 0000-0002-0152-7600

Додаткова інформація: <https://www.webofscience.com/wos/author/record/E-5814-2018>;

<https://www.scopus.com/authid/detail.uri?authorId=57207774022>;

<https://scholar.google.com.ua/citations?user=B1sT8KsAAAAJ>

Повне найменування юридичної особи: Харківський національний університет радіоелектроніки

Код за ЄДРПОУ: 02071197

Місцезнаходження: проспект Науки, Харків, Харківський р-н., 61166, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові
голови ради

Євланов Максим Вікторович

Власне Прізвище Ім'я По-батькові
головуючого на засіданні

Євланов Максим Вікторович

Відповідальний за підготовку
облікових документів

Іванова Олена Олександрівна

Реєстратор

Юрченко Тетяна Анатоліївна

**Керівник відділу УкрІНТЕІ, що є
відповідальним за реєстрацію наукової
діяльності**



Юрченко Тетяна Анатоліївна