METHODOLOGY FOR BUILDING A BENCHMARK DATASET OF BULGARIAN INVOICES FOR AUTOMATED INFORMATION EXTRACTION

STOYAN K. KONSTANTINOV **HRISTO PARASKEVOV**

ABSTRACT: *The paper proposes a methodology for preparing a benchmark dataset of Bulgarian invoices for evaluating automated information extraction systems. The approach is based on an existing collection of approximately 5,000 real invoices with manually verified structured data. The methodology includes data quality assessment, duplicate detection, identification of document templates and template families, annotation normalization, and dataset partitioning to prevent information leakage. Particular attention is paid to generalization to previously unseen document layouts and the specific challenges of recognizing Bulgarian text. Preliminary OCR observations illustrate the importance of language-specific evaluation. The proposed framework provides a basis for reproducible benchmarking and subsequent experimental comparison of artificial intelligence models.*

KEYWORDS: *document information extraction, benchmark dataset, invoice processing, document templates, OCR, Bulgarian language, artificial intelligence.*

DOI: https://doi.org/10.46687/MQCO1569

МЕТОДИКА ЗА ИЗГРАЖДАНЕ НА ЕТАЛОНЕН НАБОР ОТ БЪЛГАРСКИ ФАКТУРИ ЗА АВТОМАТИЗИРАНО ИЗВЛИЧАНЕ НА ДАННИ*

СТОЯН К. КОНСТАНТИНОВ **ХРИСТО ПАРАСКЕВОВ**

АБСТРАКТ: *В статията е предложена методика за подготовка на еталонен набор от български фактури, предназначен за оценяване на системи за автоматизирано извличане на данни. Изходната основа е съществуващ набор от приблизително 5 000 реални фактури с ръчно проверени структурирани данни. Разгледани са процедурите за проверка на качеството, откриване на дубликати, идентифициране на шаблони и семейства шаблони, нормализация на извлечените стойности и разделяне на данните с цел предотвратяване на изтичане на информация между обучаващите и тестовите извадки. Специално внимание е отделено на оценяването върху непознати оформления и на особеностите при разпознаване на български текст. Представени са предварителни наблюдения от проверка на система за оптично разпознаване. Предложената методика създава основа за възпроизводима оценка и последващо експериментално сравнение на модели с изкуствен интелект.*

1 Въведение

Автоматизираната обработка на документи намира все по-широко приложение в информационните системи за управление на бизнеса. Една от характерните задачи в тази област е извличането на структурирани данни от фактури, които се различават по оформление, съдържание, качество на изображението и използвани текстови формати. В практиката тази задача включва разпознаване на номера и датата на документа, данни за доставчик и получател, данъчни идентификатори, стойности, валута и информация за отделните артикули.

Традиционните подходи за автоматизация често използват предварително дефинирани шаблони, чрез които се определя местоположението на необходимите полета. Подобни решения могат да постигат висока точност при документи с позната структура, но

* Настоящата статия е частично финансирана по проект № РД-08-88/02.02.2026

изискват допълнително конфигуриране при появата на нови оформления. Развитието на машинното обучение, компютърното зрение и големите езикови модели предоставя възможности за изграждане на по-гъвкави системи, които извличат информация чрез съвместен анализ на текста, неговото пространствено разположение и визуалните характеристики на документа.

Независимо от постигнатия напредък, обобщаването към непознати шаблони остава съществен проблем. Изследванията VRDU [1] и DocILE [2] показват необходимостта системите да се оценяват не само върху документи, подобни на използваните при обучението, но и върху оформления, които моделите не са срещали предварително. Подобни ограничения са установени и при практически системи за обработка на фактури [3].

Качеството на оценяването зависи в значителна степен от използваните набори от данни. При случайно разделяне на голям корпус е възможно документи с еднакво или почти еднакво оформление да попаднат едновременно в обучаващата и тестовата извадка. В този случай измерената точност може да отразява способността на модела да разпознава вече познати структури, без да дава достатъчно надеждна оценка за поведението му при нови доставчици.

Допълнително предизвикателство възниква при обработката на документи на български език. Голяма част от използваните публични набори от бизнес документи са създадени за други езици, а наличието на поддръжка на кирилица в дадена система не гарантира еднакво качество при разпознаване на специфични за българските фактури данни.

В настоящото изследване се разглежда възможността за изграждане на еталонен набор от данни въз основа на съществуващ корпус от приблизително 5 000 реални български фактури, за които са налични ръчно проверени структурирани стойности. Наличието на тези данни позволява основните усилия да бъдат насочени към проверка на тяхното качество, идентифициране на шаблоните, стандартизиране на представянето и формиране на независими извадки за оценяване.

Целта на статията е да се предложи методика за подготовка и структуриране на еталонен набор от български фактури, позволяващ надеждна оценка на системи за автоматизирано извличане на данни, включително при непознати документни оформления.

За постигането на целта се разглеждат основните етапи на подготовката, начините за класифициране на шаблоните, процедурите за контрол на качеството и правилата за изграждане на тестови извадки. Методиката е предназначена да служи като основа за последващо експериментално сравнение на различни модели с изкуствен интелект.

2 Съществуващи подходи и еталонни набори от данни

Съвременните методи за извличане на информация от документи могат условно да бъдат разграничени според начина, по който използват текстовата и визуалната информация. Една група подходи съчетава разпознатия текст с координатите на неговото разположение върху страницата. Такива са моделите от фамилията LayoutLM, включително LayoutLMv3 [4], при които текстовите и визуалните характеристики участват съвместно в представянето на документа.

Друга група модели обработва изображенията непосредствено, без задължително използване на отделен етап за оптично разпознаване. Представител на този подход е Donut [5], който позволява генериране на структуриран резултат въз основа на изображението на документа.

При многоезична обработка особено значение имат архитектури като LayoutXLM [6] и LiLT [7]. Те предоставят възможности за използване на информация за пространственото разположение на текста в различни езикови среди. Въпреки това наличието на многоезична архитектура не отменя необходимостта от експериментална оценка върху конкретния език и тип документи.

Съществено значение за развитието на тази област имат публичните еталонни набори от данни. VRDU [1] въвежда протоколи, чрез които може да се изследва поведението на моделите при обучение с един, множество или непознати шаблони. DocILE [2] обхваща реални бизнес документи и предлага оценяване според наличието на съответните оформления в обучаващите данни. Това позволява да се разграничат резултатите върху познати и непознати структури.

Специализираният набор FATURA [8] включва синтетично генерирани фактури с различни оформления. Подобни ресурси са полезни за разработване и предварителна проверка на алгоритми, но синтетичният характер на документите ограничава възможността всички резултати да се пренасят непосредствено към реалната практика. При реалните фактури се срещат допълнителни различия, свързани с използвания счетоводен софтуер, качеството на сканиране, допълнителните обозначения и структурата на отделните полета.

Значението на разнообразието от оформления е отчетено и при DocLayNet [9], предназначен за анализ на разположението на елементи в документи. От друга страна, изследвания върху съществуващи документни корпуси показват, че наличието на голям брой документи само по себе си не гарантира високо качество на еталонните данни. Larson и съавтори [10] обръщат внимание на проблеми с качеството на етикетите и разделянето на данните, които могат да повлияят върху достоверността на експерименталните резултати.

Отделен въпрос представлява поддръжката на български език. Системата Tesseract разполага със специализиран езиков модел за български (bul) [11]. При PaddleOCR е налична поддръжка на български чрез общ модел за разпознаване на кирилица [12]. Тези възможности обаче следва да се разграничават от действително измерената точност при обработка на български фактури, съдържащи едновременно кирилица, латиница, числови стойности, валутни знаци и специфични идентификатори.

Проблемите при изграждането на реалистични еталонни набори от бизнес документи са разгледани и от Skalický и съавтори [13], които подчертават практическите ограничения при достъпа до реални документи и необходимостта от подходящи процедури за оценяване.

Прегледът на разгледаните източници очертава няколко изисквания към изграждането на еталонен набор от български фактури: използване на реални документи с разнообразно оформление; наличие на проверени структурирани стойности; идентифициране на доставчиците и използваните шаблони; предотвратяване на припокриването между обучаващите и тестовите данни; прилагане на единни правила за сравняване на извлечените стойности; отделна оценка на качеството при обработка на български текст. Тези изисквания определят основните елементи на предложената методика.

3 Изходни данни и методика за подготовка

3.1 Характеристика на изходния корпус

Основата за изграждане на предложения еталонен набор е съществуваща колекция от приблизително 10 000 реални български фактури, съхранявани във формат PDF или като изображения. За всеки документ е наличен съответстващ JSON файл, съдържащ предварително извлечени и ръчно проверени стойности.

Тази особеност представлява важно предимство при подготовката на данните. За разлика от случаите, при които създаването на еталонен набор изисква първоначално ръчно извличане и означаване на информацията, тук основната задача е свързана с проверката на вече наличните стойности, тяхното уеднаквяване и осигуряването на подходяща организация за последващо оценяване.

Въпреки наличието на предварително обработени документи, корпусът не може непосредствено да бъде разглеждан като готов еталонен набор. Необходимо е предварително да бъдат установени неговата структура, разпределението на документите по доставчици, наличието на повтарящи се записи, разнообразието от оформления и качеството на съществуващите анотации.

Особено значение има разграничението между цифрово създадени фактури и документи, получени чрез сканиране или фотографиране. При първата група текстът може да бъде извлечен непосредствено от PDF файла, докато при втората е необходимо оптично разпознаване. Различията между тези два вида документи могат да повлияят върху точността на системите и следва да бъдат отчетени при изграждането на еталонния тест.

3.2 Етапи на подготовка

Предложената методика включва последователни процедури за инвентаризация, проверка и структуриране на документите. Основните етапи са представени в таблица 1.

Таблица 1. Основни етапи при подготовката на еталонния набор от данни

Етап	Основни дейности	Очакван резултат
Инвентаризация	Установяване на наличните документи, съответстващите JSON файлове, формати и полета	Пълен опис на корпуса
Проверка за дубликати	Сравняване на файлове, номера на фактури, доставчици и визуално сходни документи	Премахване на повторенията
Проверка на данните	Контрол на идентификатори, дати, парични стойности и логически зависимости	Откриване на съмнителни записи
Проверка на анотациите	Повторен ръчен преглед на представителна извадка	Оценка на качеството на еталонните стойности
Класификация на шаблоните	Групиране на документите по доставчик, оформление и структура	Идентификатори на шаблони и семейства шаблони
Формиране на извадки	Разделяне на данните по предварително определени правила	Обучаващи, валидационни и тестови извадки

Първоначалната инвентаризация има за цел да установи съответствието между документите и съпровождащите ги структурирани данни. Предвижда се автоматизирано извличане на основни характеристики, включително брой страници, файлов формат, наличие на текстов слой, използвани полета и идентификатори на доставчиците.

Проверката за дубликати се извършва на няколко нива. За откриване на напълно еднакви файлове може да се използва криптографска хеш-функция SHA-256. Документи с еднакъв доставчик и номер на фактурата се разглеждат като потенциални повторения, а при визуално сходни документи може да се приложи перцептивно хеширане. Необходимо е резултатите от автоматичните проверки да бъдат потвърдени, тъй като два файла с еднакви основни реквизити могат да представляват различни версии на един документ.

Контролът на структурираното съдържание включва проверки за допустимостта на датите, структурата на идентификаторите и съответствието между данъчната основа, начисления данък и общата стойност. Установените несъответствия не следва автоматично да се приемат за грешки, а да бъдат отделяни за допълнителен преглед.

3.3 Оценка на качеството на анотациите

Наличието на ръчно проверени стойности не изключва възможността в корпуса да съществуват грешки. Те могат да произтичат от неправилно разчетен текст, неточно определено поле, пропусната стойност или разлика в използвания формат.

За предварителна оценка на качеството се предлага повторна проверка на случайна извадка от 300 фактури. Проверката следва да бъде извършена независимо от първоначалната обработка, като за всяка фактура се установява дали има поне една грешка в критично поле.

Делът на документите с установени грешки се изчислява като отношение между броя документи с грешки и общия брой повторно проверени документи. Заедно с оценката се отчита 95% доверителен интервал по метода на Уилсън, който позволява да се представи неопределеността при ограничен обем на извадката.

Резултатите от тази проверка ще определят необходимостта от допълнително коригиране на анотациите. Особено важно е тестовите документи да бъдат подложени на засилен контрол, тъй като грешките в еталонните стойности могат да повлияят непосредствено върху оценката на всички сравнявани модели.

4 Идентифициране на шаблоните и изграждане на еталонен тест

4.1 Определяне на шаблони и семейства шаблони

Едно от основните предизвикателства при подготовката на еталонни данни от фактури е определянето на сходството между отделните документи.

За целите на предложената методика е необходимо да се разграничат три понятия:

Шаблон на документ – съвкупност от характеристики на оформлението, които определят разположението на основните реквизити, използваните надписи и структурата на табличната информация.

Версия на шаблон – разновидност на съществуващ шаблон, при която са настъпили ограничени промени, например в логото, допълнителните обозначения или форматирането, без съществена промяна в общата структура.

Семейство шаблони – група документи с близки структурни и визуални характеристики, които могат да произхождат от различни доставчици, използващи една и съща счетоводна система или сходен механизъм за генериране на фактури.

Разграничението между доставчик и шаблон е особено важно. Един доставчик може да използва повече от един шаблон, докато различни доставчици могат да генерират фактури с практически еднакво оформление. Следователно броят на доставчиците не може да бъде използван като единствен показател за разнообразието на документите.

Предлага се определянето на шаблоните да се извършва чрез комбиниран автоматизиран и ръчен подход. В началния етап документите се групират по идентификатор на доставчика. След това се анализират пространственото разположение на основните реквизити, текстовите блокове, таблиците и други устойчиви елементи на оформлението.

Сходството между документите може да се оценява чрез характеристиките на разпознатия текст и неговите координати, както и чрез векторни представяния на

изображенията. Получените автоматично групи се подлагат на ръчен контрол с цел установяване на неправилно обединени или разделени шаблони.

За проверка на надеждността на класификацията се предвижда независимо групиране на извадка от документи от двама проверяващи. Степента на съгласие може да бъде оценена чрез коригирания индекс на Ранд (Adjusted Rand Index).

4.2 Разделяне на данните

Разделянето на набора от данни е решаващо условие за достоверността на бъдещите експериментални резултати.

При обичайното случайно разделяне съществува риск фактури с еднакво оформление да попаднат както в обучаващата, така и в тестовата извадка. Това може да доведе до изкуствено повишаване на измерената точност, тъй като моделът вече е обработвал документи със сходна структура.

За ограничаване на този проблем се предлага използване на разделяне, основано на идентификаторите на доставчиците и семействата шаблони.

Основният принцип е документите от едно семейство шаблони да не се разпределят между обучаващата и тестовата извадка, когато се оценява обобщаването към непознати оформления. При теста с непознати доставчици всички документи на съответния доставчик се изключват от обучението.

Когато се изисква едновременно отсъствие на познати доставчици и познати семейства шаблони, следва да се отчетат връзките между тях. Това може да наложи обединяване на свързани групи документи преди окончателното разделяне.

За бъдещото оценяване се предлага формиране на няколко основни части, представени в таблица 2.

Таблица 2. Предложена структура на еталонния набор от данни

Извадка	Предназначение
Обучаваща	Обучение и дообучение на модели
Валидационна	Избор на настройки и контрол на обучението
Тест с непознати шаблони	Оценка върху оформления, които не присъстват в обучението
Допълнителна тестова извадка	Оценка върху документи от познати и непознати доставчици при ясно разграничаване на двете групи

Окончателните размери на извадките следва да бъдат определени след инвентаризацията на корпуса. Те зависят не само от общия брой фактури, но и от разпределението им по доставчици, шаблони и семейства шаблони.

За предотвратяване на изтичане на информация се предвижда допълнителна проверка за напълно еднакви и почти идентични документи между отделните части. Необходимо е също настройките на моделите да бъдат определяни без използване на основната тестова извадка.

4.3 Защита и достъп до данните

Използването на реални фактури налага допълнителни изисквания, свързани със защитата на личните данни и поверителната търговска информация.

Документите могат да съдържат имена на физически лица, адреси, банкови сметки, електронни адреси, подписи и други идентифициращи сведения. Обработването им трябва да бъде съобразено с изискванията на Регламент (ЕС) 2016/679 [14].

Предлага се разграничаване на три нива на достъп: вътрешно ниво с оригиналните документи, изследователско ниво с ограничени идентифициращи данни и публично ниво, включващо само информация, за която са изпълнени необходимите условия за предоставяне.

При подготовката на изследователския набор може да се използва замяна на идентифициращите стойности със синтетични. Псевдонимизирането обаче не трябва да се приема автоматично за анонимизиране, тъй като възможността за повторно идентифициране зависи от наличната допълнителна информация и използваните защитни мерки [15].

Същевременно трябва да се отчете, че замяната на текст в изображенията може да промени тяхното визуално представяне и по този начин да повлияе върху резултатите от автоматизираното извличане. Затова се предлага предварителна проверка на влиянието на тази обработка върху качеството на разпознаване.

5 Правила за оценяване и предварителни наблюдения

5.1 Нормализация на извлечените стойности

При оценяването на системите за извличане на данни е необходимо да се разграничават действителните грешки от разликите във форматирането.

Например една и съща дата може да бъде представена като 01.09.2026, 1 септември 2026 г. или 2026-09-01. По подобен начин паричните стойности могат да съдържат различни разделители за хилядите и десетичната част.

За да бъдат резултатите съпоставими, се предлага използване на правила за нормализация според типа на полето.

Таблица 3. Основни правила за нормализация на извлечените стойности

Тип данни	Предложена нормализация
Дата	Единен формат ISO 8601
Парична стойност	Десетично представяне с точно определена валута
Валута	Код по ISO 4217
ЕИК, ДДС номер, IBAN	Премахване на несъществени интервали и проверка на допустимата структура
Наименование на организация	Уеднаквяване на регистъра, интервалите и типографските знаци
Свободен текст	Нормализация на Unicode и интервалите; отделно оценяване на текстовото сходство

При числовите стойности сравнението трябва да запазва точността на оригиналните данни. Разлики в паричните суми не следва да се пренебрегват само защото стойностите са визуално близки.

Особено внимание е необходимо при документи с обозначения в различни валути. За всяко поле трябва да бъде ясно определено коя стойност се извлича, без моделът самостоятелно да преобразува сумата в друга валута.

При идентификаторите нормализацията трябва да бъде достатъчно ограничена, за да не превръща неправилно разпознати символи в привидно верни стойности.

5.2 Критерии за оценяване

Основната мярка за оценяване на качеството на извлечените структурирани данни се предлага да бъде F1 на ниво поле.

За всяко поле се определят верните положителни резултати (TP), лъжливите положителни резултати (FP) и лъжливите отрицателни резултати (FN). Мярката се изчислява по формулата:

$$(1) \qquad F_1 = \frac{2TP}{2TP + FP + FN}.$$

При неправилна непразна стойност се отчита както пропуснатата вярна стойност, така и неправилно предсказаната. Сравняването се извършва след прилагане на правилата за нормализация.

За ограничаване на влиянието на доставчици с голям брой фактури се предлага резултатите да се обобщават и на ниво доставчик. При този подход първоначално се изчислява качеството за отделните доставчици, след което се определя средната стойност за съответната група.

Допълнителни показатели могат да бъдат: дял на документите, при които всички критични полета са извлечени правилно; точност на извличането на редовете с артикули; дял на неправилно генерираните или липсващите стойности; време за обработка на документ; разлика между точността върху познати и непознати оформления.

При последващите експериментални сравнения доверителните интервали следва да отчитат зависимостта между фактурите на един доставчик. Подходяща възможност е използването на клъстерна повторна извадка с връщане (cluster bootstrap), при която единицата на повторната извадка е доставчикът. Необходимостта от подходяща статистическа проверка при сравняване на модели е разгледана и в [16].

5.3 Предварителна проверка на оптичното разпознаване на български текст

Една от особеностите на разглежданата задача е необходимостта от надеждно разпознаване на документи, съдържащи едновременно български текст, числови стойности, латински идентификатори и специални символи.

В рамките на предварителна технологична проверка са използвани 24 характерни текстови реда от български фактури, представени с три различни шрифта. Получените 72 синтетични изображения са обработени чрез Tesseract 5.3.4 с различни езикови настройки.

За оценяване е използван делът на неправилно разпознатите символи (Character Error Rate, CER). Получените в предварителната проверка стойности са представени в таблица 4.

Таблица 4. Предварителни резултати от разпознаването на български текст със Tesseract

Езиков модел	CER
Български (bul)	2,73%
Български и английски (bul+eng)	1,27%
Руски (rus)	1,94%

Най-нисък дял на неправилно разпознатите символи в тази ограничена проверка е получен при комбинираното използване на български и английски езиков модел.

Наблюдавани са обаче характерни грешки при символи със съществено значение за фактурите. Сред тях са неправилното разпознаване на знака за номер, процента и валутните обозначения, както и смесването на визуално сходни кирилски и латински букви.

При използването на руския езиков модел се наблюдават и замени на символи, характерни за българския език. Това показва, че общата поддръжка на кирилица не е достатъчна предпоставка за надеждно извличане на информация от български бизнес документи.

Необходимо е да се подчертае, че представените резултати са получени върху малка извадка от синтетично генериран текст, без характерните за реалните документи шум, геометрични изкривявания и разнообразие на оформленията. Поради това те не могат да се разглеждат като окончателна оценка на качеството на Tesseract върху реални български фактури.

Тяхното значение е преди всичко методическо: те илюстрират необходимостта оценяването на системите за разпознаване да се извършва върху документи от конкретната езикова и приложна област.

Предложеният еталонен набор ще предостави възможност за подобна проверка при по-голямо разнообразие от реални документи и при използване на различни системи за разпознаване.

6 Заключение

Изграждането на надежден еталонен набор от данни е важна предпоставка за обективно оценяване на системите за автоматизирано извличане на информация от документи. При обработката на реални фактури качеството на резултатите зависи не само от използваните алгоритми, но и от характеристиките на документите, достоверността на еталонните стойности и начина на организиране на тестовите извадки.

В статията е предложена методика за подготовка на еталонен набор от български фактури въз основа на съществуващ корпус от приблизително 10 000 документа с предварително проверени структурирани данни.

Методиката обхваща последователни процедури за инвентаризация и проверка на данните, идентифициране на шаблони и семейства шаблони, нормализация на стойностите и формиране на независими извадки за оценяване. Особено внимание е отделено на ограничаването на изтичането на информация между обучаващите и тестовите данни, което е съществено условие за достоверна оценка на поведението на моделите при непознати документни оформления.

Предварителната проверка на оптичното разпознаване показва характерни затруднения при обработката на български текст и специфични за фактурите символи. Макар резултатите да са ограничени до синтетична извадка, те подкрепят необходимостта от специализирано оценяване върху реални български документи.

Основният методически принос на разработката е систематизирането на процедури за подготовка и оценяване на реални български фактури, при които разнообразието на оформленията и независимостта на тестовите данни се разглеждат като основни фактори за достоверността на резултатите.

Следващият етап на изследването предвижда прилагане на предложената методика върху целия наличен корпус, установяване на действителното разпределение на шаблоните и окончателно изграждане на еталонните извадки.

Подготвеният набор ще бъде използван за последващо експериментално изследване на влиянието на разнообразието от шаблони върху точността на различни модели за

автоматизирано извличане на данни. По този начин методическата разработка създава основа за отделно експериментално изследване, насочено към сравнение на моделите и определяне на условията за надеждното им приложение.

ЛИТЕРАТУРА

[1] Wang, Z., Zhou, Y., Wei, W., Lee, C.-Y., Tata, S. VRDU: A Benchmark for Visually-rich Document Understanding. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2023). DOI: 10.1145/3580305.3599929.

[2] Šimsa, Š., Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Skalický, M., Matas, J., Doucet, A., Coustaty, M., Karatzas, D. DocILE Benchmark for Document Information Localization and Extraction. Document Analysis and Recognition – ICDAR 2023 (2023). arXiv:2302.05658.

[3] Palm, R. B., Winther, O., Laws, F. CloudScan – A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. 14th IAPR International Conference on Document Analysis and Recognition (2017). arXiv:1708.07403.

[4] Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. Proceedings of the 30th ACM International Conference on Multimedia (2022).

[5] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S. OCR-free Document Understanding Transformer. European Conference on Computer Vision (2022).

[6] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding (2021). arXiv:2104.08836.

[7] Wang, J., Jin, L., Ding, K. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022).

[8] Limam, M., Dhiaf, M., Kessentini, Y. FATURA: A Multi-Layout Invoice Image Dataset for Document Analysis and Understanding (2023). arXiv:2311.11856.

[9] Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., Staar, P. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022). DOI: 10.1145/3534678.3539043.

[10] Larson, S., Lim, G., Leach, K. On Evaluation of Document Classification using RVL-CDIP. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (2023). arXiv:2306.12550.

[11] Tesseract OCR. Tessdata: Trained Models for Tesseract. https://github.com/tesseract-ocr/tessdata.

[12] PaddlePaddle. PP-OCRv5 Multilingual Text Recognition. https://www.paddleocr.ai/main/en/version3.x/algorithm/PP-OCRv5/PP-OCRv5_multi_languages.html.

[13] Skalický, M., Šimsa, Š., Uřičář, M., Šulc, M. Business Document Information Extraction: Towards Practical Benchmarks. Experimental IR Meets Multilinguality, Multimodality, and Interaction – CLEF 2022 (2022). arXiv:2206.11229.

[14] European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union, L 119 (2016).

[15] European Data Protection Board. Guidelines 01/2025 on Pseudonymisation. Version 1.0 for public consultation (2025).

[16] Dror, R., Baumer, G., Shlomov, S., Reichart, R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018).

Стоян К. Константинов
Месторабота: Direct Service Ltd.
E-mail: s.konstantinov@shu.bg

Христо Параскевов
ШУ „Епископ Константин Преславски“
ORCID ID: 0000-0001-8097-6131
e-mail: h.paraskevov@shu.bg