



ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ

Филиал Пловдив

Факултет: Електроника и Автоматика

Катедра: Системи за управление

маг. инж. Стела Галинова Стойкова

**ПРИЛОЖЕНИЯ НА ИЗКУСТВЕНИЯ ИНТЕЛЕКТ В
ИНФОРМАЦИОННИТЕ СИСТЕМИ ЗА УПРАВЛЕНИЕ И
БИЗНЕС АНАЛИЗ**

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен
"Доктор"

Област: 5. Технически науки

Професионално направление: 5.2 Електротехника, електроника, автоматика

Научна специалност: Системи с изкуствен интелект

Научен ръководител: доц. д-р инж. Никола Шакев

Пловдив, 2026 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Системи за управление“ към Факултет „Електроника и Автоматика“ на ТУ-София, филиал Пловдив, на редовно заседание, проведено на 02.06.2026 г.

Публичната защита на дисертационния труд ще се състои на 12.10.2026 г. от 13,00 часа, в зала 3206 на Технически университет - София, Филиал Пловдив на открито заседание на научното жури, определено със заповед № ОЖ-5.2-48/ 19.06.2026 г. на Ректора на ТУ-София в състав:

1. Доц. д-р инж. Радослав Николов Хрисчев – председател
2. Доц. д-р инж. Диляна Вълкова Будакова – научен секретар
3. Проф. д-р инж. Коста Петров Бошнаков
4. Проф. д-р инж. Петя Дойчева Копринкова
5. Доц. д-р инж. Денис Сафидинов Чикуртев

Рецензенти:

1. Доц. д-р инж. Радослав Николов Хрисчев
2. Проф. д-р инж. Петя Дойчева Копринкова

Материалите по защитата са на разположение на интересующите се в канцеларията на Факултет „Електроника и Автоматика“ на ТУ-София, филиал Пловдив, корпус IV, кабинет 4242.

Дисертантът е задочен докторант към катедра „Системи за управление“ към Факултет „Електроника и Автоматика“ на ТУ-София, филиал Пловдив. Изследванията по дисертационната разработка са направени от автора, като някои от тях са подкрепени от научноизследователски проект.

Автор: маг. инж. Стела Стойкова

Заглавие: Приложения на изкуствения интелект в информационните системи за управление и бизнес анализ

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

В дисертационния труд са представени разработени иновативни софтуерни решения за приложение на съвременни методи на изкуствения интелект в информационните системи за управление и бизнес анализ в три направления: интелигентна процесна автоматизация на документообработка в ERP системи, класификация по ключови показатели за ефективност при оперативен анализ в CRM системи и когнитивно взаимодействие с интелигентен дигитален асистент в НХМ системи. В дисертационния труд са изследвани съвременни модели на изкуствения интелект и актуални методи за тяхното обучение: дълбоки невронни мрежи тип двупосочен енкодер (BERT), дървета на решенията, ансамблови дървета на решенията (RF, XGB, GBM, LightGBM), KNN алгоритми, машини с опорни вектори (SVC), големи езикови модели (GPT, LLaMA, Qwen, Gemma, Deepseek R1), обучение по механизъм на вниманието и трансфер на знания, параметрично-ефикасно обучение (LoRA, QLoRA, LLaMa Adapter) и контекстуализация (RAG). Разработването и внедряването на софтуерни решения, базирани на съвременни методи на изкуствения интелект, в интегрираните информационни системи за управление в индустрията е приоритетно направление за осъществяване на научноизследователска дейност за взаимосвързаните научно-приложни области на изкуствения интелект и информационни системи за управление.

Цел и основни задачи на дисертационния труд

Целта на дисертационния труд е разработване на иновативни софтуерни решения за приложение на изкуствен интелект в информационните системи за управление и бизнес анализ в три направления, генериращи добавена стойност:

- интелигентна процесна автоматизация на документооборота и документообработка за бизнес процеси в информационни системи за управление на ресурсите на предприятието;
- класификация по ключови показатели за ефективност при оперативен анализ в CRM информационни системи за управление на взаимодействията с клиенти;
- когнитивно взаимодействие (човек-машина) в информационни системи за управление на човешки ресурси и интегрирани информационни системи в индустрията.

За постигане на поставената цел са конкретизирани следните задачи за изпълнение във всяко от направленията:

А) Интелигентна процесна автоматизация на документооборота и документообработка:

- Да се изследват възможностите за последващо обучение на дълбоки невронни мрежи тип двупосочен енкодер за целите на разпознаване на наименувани обекти (NER) и класификация на документи на база на разпознатите обекти.
- Да се изследва възможността за внасяне на подобрения в точността на разпознаване на наименувани обекти от тип организация на български език чрез настройка на хиперпараметри и последващо обучение на дълбоки невронни мрежи тип двупосочен енкодер.

Б) Класификация по ключови показатели за ефективност при оперативен анализ:

- Да се изследва възможността за прилагане на модели за машинно обучение (дървета на решенията, ансамблови дървета на решенията, KNN алгоритми, машини с опорни вектори) за решаване на задачи за двоична класификация по ключови показатели за ефективност при оперативен анализ в три приложни направления: продажби, маркетинг и обслужване на клиенти.

- Да се определи подход към процеса за предварителна обработка на данните и да се изследват фактори, повлияващи върху точността на класификация и засягащи подходите за предварителна обработка на данни (като алгоритми за импутация и коригиране на класов дисбаланс в данните) и селекцията на характеристики за обучение.

В) Когнитивно взаимодействие в информационни системи за управление на човешки ресурси:

- Да бъде разработен интелигентен дигитален асистент, базиран на големи езикови модели, за подпомагане на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubischi RH-6CH7020-D в научноизследователския и учебния процес в следните етапи на изпълнение: формиране на база данни за обучение на големи езикови модели за генериране на код в програмен език MELFA Basic V; формиране на корпус от обучителни документи за работа със съответния стенд (технически ръководства, инструкции за работа); подбор на големи езикови модели за обучение и обучаване на моделите върху формираните база данни и корпус от документи; свързване на обучения интелигентен дигитален асистент с интегрирана информационна система за управление SAP чрез уеб-клиент.
- Да се изследват подходи за параметрично-ефикасно последващо обучение на големи езикови модели с цел разширяване на обхвата на интелигентния дигитален асистент до платформа за обучение на големи езикови модели върху бази данни и корпуси от документи от специализирани научни, приложни и научно-приложни области с възможност за локално обучение и съхранение на данни за запазване на тяхната поверителност.

Научна новост

Представените в дисертационния труд резултати от научноизследователска и експериментална дейност включват научна новост в три приложни направления, засягащи съвременните интегрирани информационни системи за управление в индустрията.

В направление интелигентна процесна автоматизация е внесено подобрение в точността на разпознаване на наименувани обекти от тип организация в бизнес документи на български език, което се явява слабо място на съществуващи широкоразпространени в практиката модели за разпознаване на наименувани обекти. Подобрението е постигнато посредством предлагане на иновативен хибриден подход към разпознаването на наименувани обекти, включващ техники от РПА (регулярни изрази) и иРПА (последващо обучение с учител на дълбоки невронни модели тип двупосочен енкодер по механизъм на вниманието и трансфер на знания). Хибридният подход позволява постигане на по-висока степен на проследимост на постигнатите резултати при извличане на информация в сравнение с разпространени в практиката модели тип „черна кутия“ като SAP DocX.

В направление оперативен анализ са изследвани възможности за прилагане на съвременни модели за машинно обучение (дървета на решенията, ансамблови дървета на решенията, KNN алгоритми, машини с опорни вектори) за решаване на задачи за двоична класификация като в направление маркетинг по ключов показател за ефективност доживотна стойност на клиента (CLV) е постигнато подобрение в точността на класификация чрез предлагане на нов метод за селекция на характеристики при двоична класификация, който комбинира оригинални за набора от данни характеристики с такива извлечени чрез прилагане на вероятностно моделиране.

В направление когнитивно взаимодействие са разработени иновативни решения за интелигентна автоматизация на процесите по онбординг и обучение на персонал в специализирани технически и приложни области за целите на НХМ системи - *интелигентен дигитален асистент RTB*, базиран на голям езиков модел GPT 2, за

подпомагне на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubishi RH-6CH7020-D в научноизследователския и учебния процес и **платформа PrivateEdgeLLMs** за параметрично-ефикасно последващо обучение и контекстуализация на големи езикови модели, представлява значителни подобрения по отношение на установени предизвикателства в областта: риск от нарушаване на конфиденциалността на обработвани данни и неточност на генерирано съдържание, липса на проследимост, както и ресурсоемкост на необходимото последващо обучение, особено в тясноспециализирани приложни области. Разработен е нов **бенчмарк MBV-Bench** за оценяване на способността на обучени големи езикови модели да генерират код по потребителска заявка на специализиран програмен език MELFA Basic V.

Практическа приложимост

Разработеното софтуерно решение Hybrid DIE представлява съвременно и конкурентоспособно решение за интелигентна роботизирана процесна автоматизация на документообработка и извличане на данни от документи за бизнес процеси, приложимо в информационните системи за управление на ресурсите на предприятието (ERP системи). Представеният в Глава 3 на дисертационния труд формиран инструментариум за осъществяване на оперативен анализ на базата на съвременни методи на машинно обучение е подходящ за обучителни цели в специалности като „Индустриален мениджмънт“ и „Индустриално инженерство“ и като инструмент със свободен достъп за малкия и среден бизнес. Изследваните решения за двоична класификация на база на модели за машинно обучение са актуални и приложими за целите на осъществяване на оперативен анализ в интегрирани информационни системи за управление в индустрията (ERP системи) и по-специфично системи за управление на взаимодействията с клиенти (CRM системи). Разработеният интелигентен дигитален асистент RTB, базиран на голям езиков модел GPT 2 е приложим за подпомагне на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubishi RH-6CH7020-D в научноизследователския и учебния процес. Разработената платформа Private Edge LLMs за обучение на големи езикови модели върху бази данни и корпуси от документи от специализирани научни, приложни и научно-приложни области с възможност за локално обучение и локално съхранение на данни за запазване на тяхната конфиденциалност може да бъде внедрена като иновативно решение към информационни системи за управление в сектори, за които конфиденциалността на данните за обучение и полученото генерирано съдържание е от ключово значение като научноизследователски институции. RTB и Private Edge LLMs са приложими като съвременни софтуерни решения за автоматизация на когнитивни взаимодействия в ERP системи и HXM системи.

Апробация

Основните резултати от проведените изследвания в дисертационния труд са докладвани и обсъдени на международна научна конференция International Conference Automatics and Informatics (ICAI, 2024), международна научна конференция 10th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE, 2025) и международен научен симпозиум 33th International Symposium „Control of Energy, Industrial and Ecological Systems“ (2025), организиран от Съюз „Автоматика и Информатика“ (САИ) към ФНТС .

Публикации

Основни постижения и резултати от дисертационния труд са публикувани в 4 научни статии, 1 в научно списание Algorithms (Q2), 2 на международни научни конференции и 1 на международен научен симпозиум като една от публикациите е самостоятелна.

Структура и обем на дисертационния труд

Дисертационният труд включва предговор, една глава, представляваща литературен обзор и анализ по поставената тема на дисертационния труд, три глави за решаване на

формулираните основни задачи, една глава със заключение и обобщени изводи, списък на основните приноси, списък на публикациите по дисертацията, три приложения с програмен код по осъществените разработки и библиография, с общ обем на същинските глави (без приложения) от **197** страници. Цитирани са общо **(187)** литературни източници, като **(179)** са на латиница и **(8)** на кирилица. Работата включва общо **(55)** фигури и **(45)** таблици.

II. ОСНОВНО СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

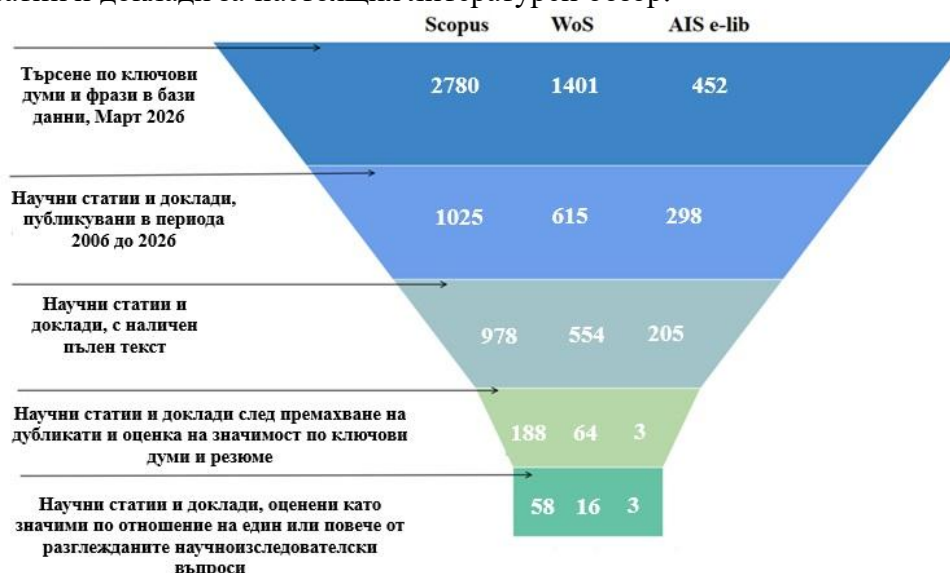
ГЛАВА 1. Литературен обзор и анализ по темата.

Прегледът на литературата показва, че са съставени редица литературни обзори по отношение на внедряването на изкуствен интелект в организационна или бизнес среда като част от съществуващи или новоразработени информационни системи за управление в разнообразие от тематични направления: счетоводство и финансов анализ, управление на информацията и оперативен анализ, управление на човешките ресурси, бизнес анализ и планиране, интелигентни агенти, чатботове и дигитални асистенти. Тематичните обзори са съсредоточени върху тесен сегмент на приложенията на изкуствения интелект в информационните системи за управление, затова информацията, която предоставят, е богата за всяка отделна изследователска тема, но същевременно е ограничена до конкретен сегмент на приложение на ИИ. Това не позволява изследването на възможности, произтичащи от обмена на данни в пресичащи се тематични сегменти като например: изследването на ефектите от внедряването на дигитални асистенти върху вземането на решения в организационната среда или резултатите от приложението на интелигентни агенти в системите за управление на човешките ресурси. В следствие на бързия напредък в развитието на ИИ и нарастването на инвестициите в изследвания на ИИ през последните години, някои от литературните обзори, проведени на база източници публикувани преди 2019 г., вече не представят в пълнота наличните технологии в определен сегмент. Сегментираната и остаряла информация е предпоставка да бъде предпочетен подход на разглеждане на актуални систематични литературни проучвания и обзори, обединяващи множество тематични направления, а не фокусиращи се върху конкретно тематично направление или икономически сектор на приложение на изкуствен интелект. Разгледани са четири систематични литературни проучвания [45, 66, 67, 68 от библиографията на дисертационния труд], обхващащи тематика, свързана с припокриването на научноприложните области изкуствен интелект и информационни системи за управление. На база на установените в разгледаните тематични и систематични литературни обзори неизследвани аспекти по отношение на пресичащи се тематични сегменти на взаимосвързаните научно-приложни области ИИ и ИСУ са дефинирани 6 изследователски въпроса, на които литературният обзор към дисертационния труд се стреми да даде отговор.

Като се има предвид поставената цел за настоящия литературен обзор, а именно: да бъдат установени ключови фактори за прилагането на изкуствен интелект в ИСУ (вид на внедряването, направления за генериране на добавена стойност въз основа на внедряването, предизвикателства и предимства пред реализация на ИИ-базирана функционалност) по методически прозрачен начин от наличните изследователски източници на базата на поставени изследователски въпроси, систематичният подход към съставяне на обзора е най-подходящ.

Търсенето на източници за обзора е проведено в три научноизследователски бази данни: Scopus, Web of Science и електронната библиотека на Асоциацията на информационните системи (AIS). Що се отнася до търсенето в базата данни Scopus, е направена селекция от научни списания, свързани с областта на интегриране на информационни системи за управление в индустрията като изборът на списания е направен въз основа на SJR и квантилното класиране в областите „Информационни

системи за управление“, „Информационни системи“ и „Изкуствен интелект“. Избраните списания имат висок SJR, като най-ниският е 1,753, а най-високият – 5.899 (класиране от 2025г.), и всички те попадат в категорията Q1 за изследвания в областта на информационните системи за управление или информационните системи. Списанията „Knowledge-Based Systems“ и „International Journal of Information Management“ също попадат в категорията Q1 за изследвания в областта на изкуствения интелект. Избраните по тези критерии научни списания са издания, представящи върховните постижения и иновации в презастъпващите се научноприложни области „Информационни системи за управление“ и „Изкуствен интелект“, което ги определя като подходящи за извличане на научни статии и доклади за настоящия литературен обзор.



Фиг. 1 Етапи на извличане на научни публикации за целите на литературния обзор.

За целите на литературния обзор също така са извлечени научни статии и доклади от следните конференции, покриващи двете презастъпващи се научноприложни области на ИИ и ИСУ: Международна конференция за информационни системи (ICIS), Международна конференция автоматика и информатика (ICAI) и Европейска конференция за информационни системи (ECIS). Подборът на доклади от конференциите се основава на комбинация от показатели: процентно отношение на подадени към приети доклади/статии, процент на цитиране, както и видимост и история на научните изследвания на организационния и редакционния екип на конференцията.

Докладваните резултати от настоящия литературен обзор са представени под формата на анализ на 77-те избрани първични изследвания (отбелязани като [75 – 151] в раздела „Библиография“ на дисертационния труд) въз основа на дефинираните научноизследователски въпроси 1-6. Всеки изследователски въпрос е разгледан в отделена подточка.

- **Въпрос 1: Какви методи и модели базирани на ИИ и с каква честота се прилагат в ИСУ?**

Въпрос 1 има за цел да предостави съвременен поглед върху това кои характеристики и функционалности, присъщи за ИИ технологиите, се срещат най-често в научните изследвания в областта на ИИ и по-специално по отношение на неговото внедряване в ИСУ. Видимо е, че научноизследователските усилия са съсредоточени приоритетно върху изследвания, засягащи интелигентна роботизирана автоматизация на процесите (иРПА), машинно обучение и обработка на естествен език. Отбелязва се нарастващ интерес към обработката на естествен език (NLP) и предсказващия анализ в сравнение с резултатите от предишни литературни обзори, които представят NLP като относително неизследвана област и приемат предсказващия анализ за прекалено слабо застъпена област в научните

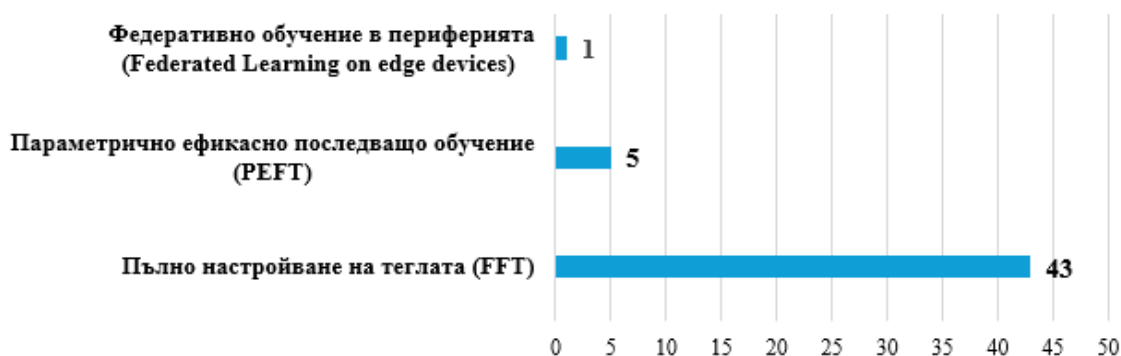
изследвания, че да бъде разглеждана отделно от машинното обучение. Засиленият интерес към областта на обработката на естествен език се дължи на масовата комерсиализация на големи езикови модели като GPT, Copilot, Gemini в последните пет години.



Фиг. 2 Разпределение на първични научни изследвания по ИИ характеристики и функционалности.

- Въпрос 2: Какви стратегии за обучение на модели, базирани на ИИ са използвани?

От разгледаните 77, първични за обзора, научни изследвания около 90% реализират обучение с учител (supervised learning), а останалите 10% са разделени между обучение без учител (unsupervised learning) и обучение с утвърждаване (reinforcement learning). От техниките за обучение с учител при дълбоките невронни мрежи (Фиг. 3) преобладават подходите, изискващи пълна актуализация на тегловете коефициенти (full fine-tuning), което е ресурсоемък процес по отношение на памет, особено при големите езикови модели. Следователно подходяща насока за изследване са техниките за параметрично-ефикасно последващо обучение (като LoRA и QLoRA), с които могат да се постигнат значителни редукции по отношение на необходимия за обучение изчислителен ресурс.



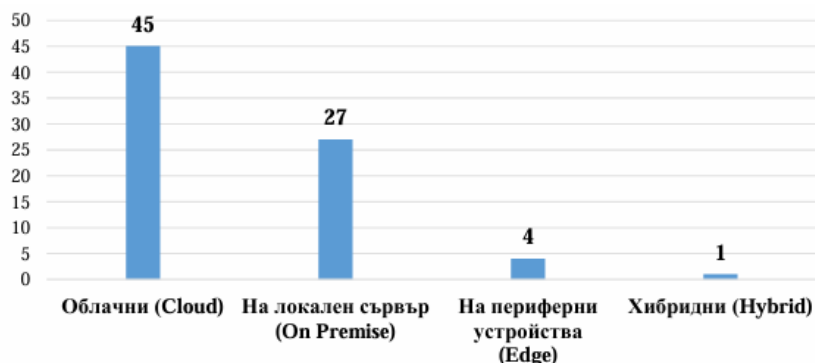
Фиг. 3 Разпределение на първични научни изследвания по техники за обучение с учител при дълбоки невронни мрежи.

- Въпрос 3: Какъв вид внедряване на ИИ в ИСУ е осъществено?

Въпрос 3 е съсредоточен върху идентифицирането на тенденции по отношение на вида платформи, използвани за внедряване на приложения, базирани на ИИ за целите на изследванията в областта на ИСУ, т.е. да се определи дали приложението, базирано на ИИ е внедрено в облачна платформа на ИСУ (SAP Core AI/Conversational AI, Oracle AI services и др.), на локални сървъри (on premise) или на периферни устройства, където се генерират данните (смартфони, таблети, компютри и др.). Фиг. 4 показва, че по-голямата част от научните изследвания са съсредоточени върху облачни или локални внедрявания, докато само четири от първичните за обзора изследвания представят приложения с ИИ, реализирани върху периферни устройства, а едно изследване е фокусирано върху хибридни архитектури.

Получените резултати предразполагат към съсредоточаване на бъдещите научни изследвания върху разработването на приложения, базирани на изкуствен интелект, които частично или изцяло да позволяват локално обучение на модели, базирани на ИИ и

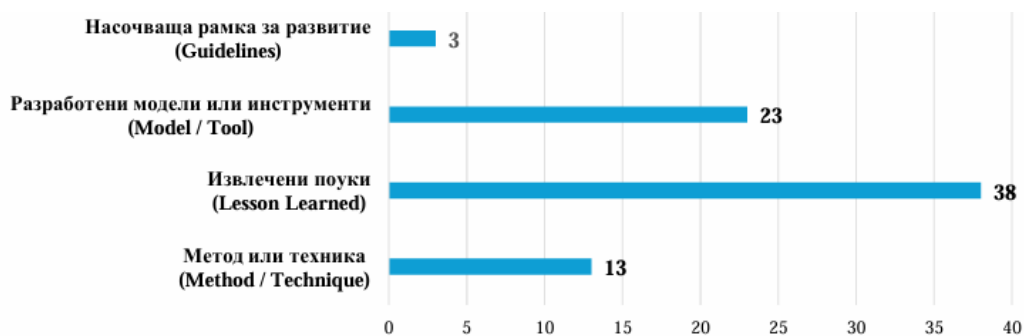
съхранение на данните за обучение върху периферни устройства. Тази насока за развитие е обвързана с изследване на методи за параметрично-ефикасно последващо обучение (LoRA, QLoRA), които редуцират необходимостта от изчислителен ресурс на графичните процесори (VRAM) при обучение. Приложенията, базирани на ИИ и реализирани изцяло или частично върху периферни устройства, се приемат за осигуряващи по-голяма поверителност на обработваните данни в сравнение с изцяло облачните решения.



Фиг. 4 Разпределение на първични научни изследвания по вид внедряване на ИИ в ИСУ.

- **Въпрос 4: Какъв е видът научен принос, реализиран от всяко от разглежданите в обзора научни изследвания, в областта на прилагане на ИИ-базирани технологии в ИСУ?**

Разгледани са четири категории принос, а именно: насочваща рамка за развитие; разработен модел или инструмент; извлечени поуки от практическа реализация или експериментална дейност; метод или техника. Анализът на 77-те първични изследвания част от литературния обзор показва, че приносите са направени до голяма степен като извлечени поуки от практическа реализация и експериментална дейност (38 изследвания) и разработени модели или инструменти (23 изследвания), докато приносите под формата на методи и техники (13 изследвания) и насочващи рамки за развитие (3 изследвания) съставляват по-малка част от изследваната извадка.



Фиг. 5 Разпределение на първични научни изследвания по вид научен принос.

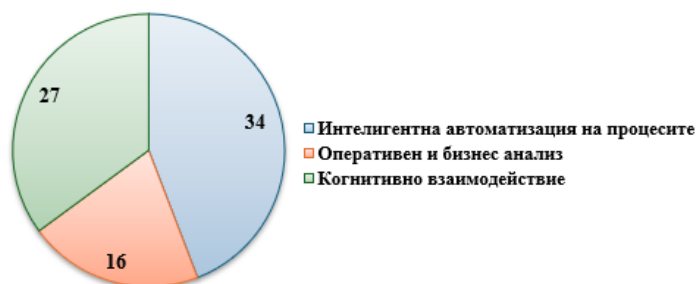
- **Въпрос 5: В какви направления е генерирана добавена стойност въз основа на представените в изследванията внедрявания на ИИ в ИСУ?**

На база анализ на 77-те първични изследвания, разгледани в литературния обзор, могат да бъдат изведени следните три вида генерирана добавена стойност за бизнеса чрез внедряване на ИИ в ИСУ:

Интелигентна автоматизация на процесите, т.е. добавена стойност е генерирана благодарение на автоматизацията на бизнес процесите, елиминиране на задачи с ниска добавена стойност от работни потоци, изпълнявани изцяло от хора и алокирането им към работни потоци, чието изпълнение е подпомагано от изкуствен интелект. В ИСУ това са предимно дейности свързани с документооборот и обработка на бизнес документи, т.е. става въпрос за софтуерни процеси и работни потоци, а не физически такива;

Оперативен и бизнес анализ, т.е. добавена стойност е генерирана благодарение на приложение на методите на машинното обучение за анализ и обработка на данни, предоставени от информационната система на организацията;

Когнитивно взаимодействие, т.е. добавена стойност е генерирана благодарение на взаимодействие между служители и клиенти с интелигентни виртуални агенти, дигитални асистенти, чатботове и други инструменти, базирани на генеративен изкуствен интелект и обработка на естествен език.



Фиг. 6 Разпределение на първични научни изследвания по вид генерирана добавена стойност за бизнеса.

Преобладават изследванията, при които се генерира добавена стойност на база на интелигентна процесна автоматизация или когнитивно взаимодействие. Задачите, основаващи се на когнитивно взаимодействие представляват значителен дял от ежедневните операции, осъществявани в информационните системи за управление в индустрията, което ги превръща във фокус на редица научни изследвания и разработки както в държавния, така и в частния сектор. Развитието на тази област е драстично повлияно в последните няколко години от приоритизираното разработване и масово интегриране на технологии, базирани на методите на генеративен изкуствен интелект като големите езикови модели (LLMs). Редица доставчици на интегрирани информационни системи за управление като SAP, Microsoft и Oracle обявиха в периода 2022-2025, че преориентират предоставяните в платформите им софтуерни услуги към пазара на изкуствения интелект като услуга (AIaaS) в две основни направления: обработка на данни с цел оперативен анализ и разработване на интелигентни виртуални агенти, базирани на генеративен изкуствен интелект.

- Въпрос 6: Какви са предизвикателствата и ползите от внедряването на ИИ в ИСУ?

Идентифицираните в 77-те първични за обзора източници предизвикателства и ползи от внедряването на ИИ в ИСУ според вида генерирана добавена стойност са представени в Таблица 1.5 и Таблица 1.6 на дисертационния труд.

Въз основа на идентифицираните в 77-те първични за обзора източници предизвикателства и ползи от внедряването на ИИ в ИСУ според вида генерирана добавена стойност са формулирани следните насоки за изследване и развитие за всеки от видовете генерирана добавена стойност посредством внедряване на ИИ в ИСУ.

- **Интелигентна автоматизация на процесите:** изследване на иновативни софтуерни решения за иРПА, базирани на дълбоки невронни мрежи, за документообработка и извличане на информация от документи в интегрирани информационни системи за управление в индустрията (ERP системи) с цел внасяне на подобрения в разпознаването на наименувани обекти на български език и класификацията на документи на база на разпознатите обекти;
- **Оперативен и бизнес анализ:** изследване на съвременни подходи за машинно обучение за осъществяване на оперативен анализ в информационни системи за управление на взаимодействията с клиенти (CRM системи) по ключови показатели

за ефективност в три функционални направления: продажби, маркетинг и обслужване на клиенти;

- **Когнитивно взаимодействие:** изследване на съвременни подходи за параметрично-ефикасно последващо обучение и техники за контекстуализация при последващо обучение (RAG), които да позволят локално обучение и обучение в периферни устройства на големи езикови модели за целите на разработване на интелигентни дигитални асистенти за подпомагане на процесите по обучение и онбординг на персонал в информационни системи за управление на човешките ресурси (НХМ системи).

Конкретизирана формулировка на поставените цел и задачи за дисертационния труд е представена в настоящия автореферат в секция „Цел и основни задачи на дисертационния труд“ на раздел „I Обща характеристика на дисертационния труд“. Формулираните цел и задачи отговарят на идентифицираните направления и насоки за бъдеща научноизследователска дейност в 6-те изследователски въпроса, разгледани в литературния обзор.

ГЛАВА 2. Интелигентна автоматизация на документооборота и документообработка за бизнес процеси в информационни системи за управление на ресурсите на предприятието.

С внедряването на технологии, базирани на ИИ в своите информационни системи, организациите генерират стойност благодарение на интелигентна процесна автоматизация, елиминираща участието на персонал в повтаряеми дейности с ниска добавена стойност. В информационните системи за управление това са предимно дейности свързани с документооборота и обработка на бизнес документи, извличане на информация от документи и нейното преобразуване за целите на производственото планиране т.е става въпрос за софтуерни процеси и работни потоци, засягащи обработката на информация, а не физически такива. Обработката на често чувствителни данни, засягащи множество процеси, носи редица предизвикателства по отношение както на проследимост и прозрачност на операциите, извършвани върху данните, така и на подсигуряване на точност и верификация на извлечената информация в процеса на обработка.

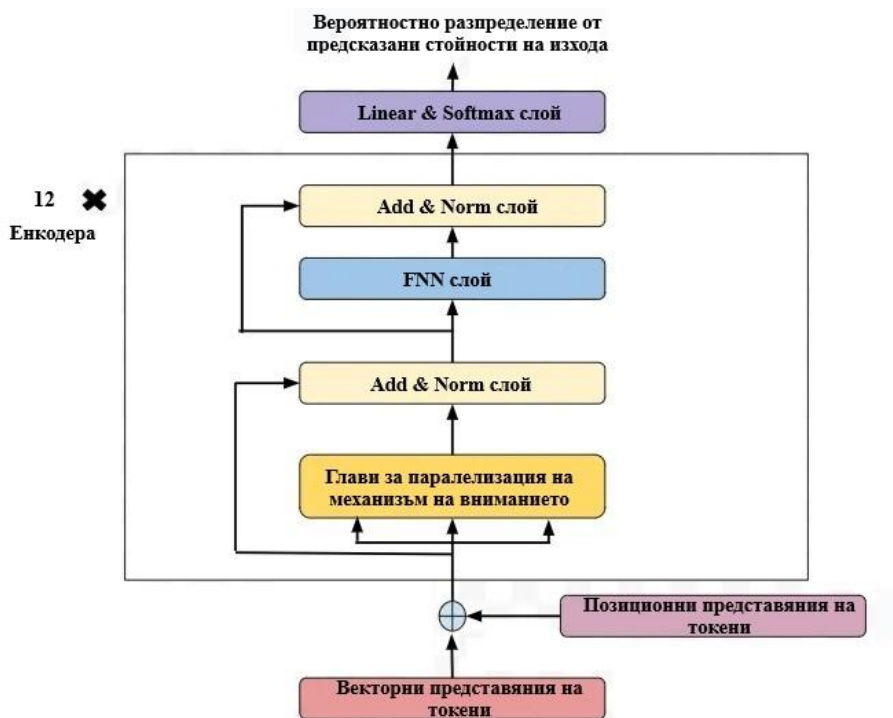
Основна функция на информационните системи за управление е документооборотът. Документооборот е процесът на създаване, обработка, съхранение и разпределение на документи в една организация, който обхваща всички етапи от тяхното създаване до архивиране или унищожаване. Той е от ключово значение за управлението на информацията и повишаване на ефективността на бизнес процесите. Тази функционалност може да бъде обособена в отделен модул (такова е решението на популярната българска ИСУ с отворен код bgERP, т.нар модул е-Документация) или да бъде реализирана като отделна подсистема, взаимодействаща с останалите функционални модули на ИСУ (такова решение предоставят SAP в техните ERP системи, т.нар Document Management Subsystem). Внедряването на процесна автоматизация в областта на документооборота и документообработката се осъществява в две основни направления от решения: роботизирана процесна автоматизация (РПА) и интелигентна роботизирана процесна автоматизация (иРПА).

иРПА се явява приоритетно направление за инвестиции от страна на бизнеса и научноизследователска дейност по отношение на внасяне на подобрения, базирани на методи на изкуствения интелект особено в областта на документооборота, обработка и извличане на информация от документи, като обработката на текст често се основава на

обучение с учител по механизъм на вниманието на специализирани за областта на лингвистиката дълбоки невронни мрежи и архитектури тип двупосочен енкодер.

Таблица 1. Сравнение на основни характеристики на РПА и иРПА.

Характеристика	РПА	иРПА
Вид на обработваните данни	структурирани	структурирани (таблицы, бази данни) и неструктурирани (документи)
Подход към вземане на решения	бази с логически правила, шаблони	базиран на модели за машинно обучение и обработка на естествен език, хибридни решения
Подход към внедряване в ERP системите	отделни локални функции и скриптове (автоматизация на ниво отделен потребител)	ИИ-базирани услуги, внедрявани на организационно ниво чрез APIs
Приложения	въвеждане и консолидиране на данни	Извличане и анализ на информация от документи
Потребителски взаимодействия	пасивно и активно (преобладава); Тъй като автоматизацията е ограничена до отделни подзадачи, асоциирани с конкретен потребител, потребителите имат активна роля в прехвърлянето на данни между отделни модули или процеси в системата, както и за създаването на асоциирани с данните документи.	пасивно (преобладава) и активно; Тъй като повечето задачи в рамките на даден процес се осъществяват от ИИ-базирани софтуерни ботове, потребителите изпълняват предимно функции на верифициране на получени резултати.



Фиг. 7 Архитектура на дълбока невронна мрежа BERT_Base, тип двупосочен енкодер.

Всяка потребителска заявка се токенизира (разбива на най-малки, носещи собствен смисъл съставни единици), напр. „Идентификационен код на доставчик Мелтрейд ЕООД: MEL01“ се превръща в последователност от токени „Идентификационен“, „код“, „на“, „доставчик“, „Мелтрейд“, „ЕООД“, „:“, „MEL01“. За всеки от тези токени се извежда векторно представяне и представяне на позицията им в последователността като векторното и позиционното представяне на всеки token от заявката представляват входната последователност за енкодерите. В структурата на самите енкодер слоеве главите за паралелизация по механизъм на вниманието осъществяват операция по

определяне на контекстовия смисъл на всеки от елементите от входната последователност, съответстващи на токени от заявката.

Математическото изражение на тази операция за едномерни данни е: при дадени сет от n вектора, представляващи заявки (query vectors) $q_1, q_2, \dots, q_n \in R^d$, сет от m векторни ключа (key vectors) $k_1, k_2, \dots, k_m \in R^d$, и сет от m вектора със стойности (value vectors) $v_1, v_2, \dots, v_m \in R^p$ механизмът на вниманието изчислява сет от изходни вектори $o_1, o_2, \dots, o_n \in R^q$ като осъществява линейна комбинация на g -трансформирани вектори със стойности $g(v_i) \in R^q$, използвайки взаимовръзката между векторите на потребителските заявки и съответстващите им векторни ключове за определяне на коефициентите, използвани в операцията. Формално:

$$o_j = \frac{1}{C} \sum_{i=1}^m f(q_j, k_i) g(v_i) \quad (2.4)$$

където $f(q_j, k_i)$ е функция, характеризираща сходството между q_j и k_i , а $g(\cdot)$ най-често е линейно преобразуване от вида:

$$g(v_i) = W_v v_i \in R^q \quad (2.5)$$

където $W_v \in R^{q \times p}$, а $C = \sum_{i=1}^m f(q_j, k_i)$ е коефициент за нормализация.

Често използвана в практиката функция за определяне на сходство (подобие) е Гаусово разпределение с вградени семантични или функционални характеристики, дефинирано, както следва:

$$f(q_j, k_i) = e^{\theta(q_j)^T \varphi(k_i)} \quad (2.6)$$

като $\theta(\cdot)$ и $\varphi(\cdot)$ обикновено представляват линейни преобразувания от вида: $\theta(q_j) = W_q q_j$ и $\varphi(k_i) = W_k k_i$. Да се отбележи, че ако приемем векторите със стойности за входни вектори, всеки изходен вектор o_j е зависим от всички входни вектори. Прилагането на сходство по Гаус и линейни преобразувания (2.6), позволява тези изчисления да бъдат представени в компактна матрична форма:

$$O = W_v V \times \text{softmax}((W_k K)^T W_q Q) \quad (2.7)$$

където: $Q = [q_1, q_2, \dots, q_n] \in R^{d \times n}$, $K = [k_1, k_2, \dots, k_m] \in R^{d \times m}$, $V = [v_1, v_2, \dots, v_m] \in R^{p \times m}$, $O = [o_1, o_2, \dots, o_n] \in R^{q \times n}$ и $\text{softmax}(\cdot)$ изчислява нормализиран вариант на матрицата с входни вектори, където всяка колона на матрицата е нормализирана да се сумира до нула. Важно е да се отбележи, че броят на изходните вектори е равен на броя на векторите за заявки. Коефициент на сходство се изчислява между всеки вектор на заявка като колона от матрицата Q и всеки вектор ключ като колона от матрицата K . $\text{Softmax}(\cdot)$ нормализира изчислените коефициенти на сходство и прави сумите по колони равни на 1. Умножение между нормализираните коефициенти и матрицата V дава съответния изходен вектор. Механизмът на самовниманието улавя вътрешната корелация на дадена матрица с входни вектори $X = [x_1, x_2, \dots, x_n] \in R^{d \times n}$, като се допуска, че $Q = K = V = X$. Тогава формулата (2.7) добива следния вид за механизма на самовнимание:

$$O = W_v X \times \text{softmax}((W_k X)^T W_q X) \quad (2.8)$$

като в този случай броя изходни вектори е определян от броя на входните вектори. При механизма на самовнимание с обучаемата заявка принципът $K = V = X$ се запазва, но матрицата на заявката $Q \in R^{d \times n}$ не се подава на входа и е независима от другите входни

матрици, вместо това тя се получава от обучение върху данните. Тогава формула (2.7) добива следния вид за механизма на самовнимание с обучаемая заявка:

$$O = W_v X \times \text{softmax}((W_k X)^T Q) \quad (2.9)$$

Този механизъм на обучение е разпространен при решаването на задачи за естествена обработка на език (NLP), тъй като позволява на невронните мрежи да уловят общи характеристики от всички входни инстанции по време на обучение, понеже заявката е независима от входа и се споделя от всички входни инстанции. Важно е да се отбележи, че понеже броят на изходните вектори се определя от броя на векторите на заявката, размерът на матрицата на изхода при обучение по механизъм за самовнимание с обучаемая заявка е фиксиран и вече не е гъвкаво свързан с размера на входните матрици. Паралелизация на обучение по механизъм на вниманието се състои от множество оператори за внимание наричани глави (head) с различни функции за сходство, определени от различни групи тегловни матрици. Формално, изходът H_i на i -тата глава при механизъм с паралелизация по M -глави се изчислява:

$$H_i = W_v^{(i)} V \times \text{softmax}\left(\left(W_k^{(i)} K\right)^T W_q^{(i)} Q\right) \in R^{q_i \times n} \quad (2.10)$$

където $W_q^{(i)}$, $W_k^{(i)}$ и $W_v^{(i)}$ са тегловите матрици, които определят функцията за сходство f_i за i -тата глава. Обобщеният изход за M -глави е:

$$O = W_o \begin{bmatrix} H_1 \\ \vdots \\ H_M \end{bmatrix} \in R^{q \times n} \quad (2.11)$$

където $W_o \in R^{q \times (\sum_i q_i)}$ е научената теглова матрица, която проектира конкатенираните глави в желано измерение. В сравнение с непаралелизираното обучение по механизъм на вниманието паралелизираното обучение по механизъм на вниманието позволява всяка глава да присъства на различни позиции въз основа на сходството в различни подпространства на представяне.

За всеки токен се изчислява оценка по механизма на самовниманието, която зависи от векторното представяне на заявката на настоящия токен и векторните представяния на ключовете за всички останали токени:

$$A_{score} = Q_{current} + K_{other\ tokens} \quad (2.12)$$

Тегловите коефициенти, актуализирани при обучение по механизма на самовниманието се обобщават, както следва за непарализирано прилагане:

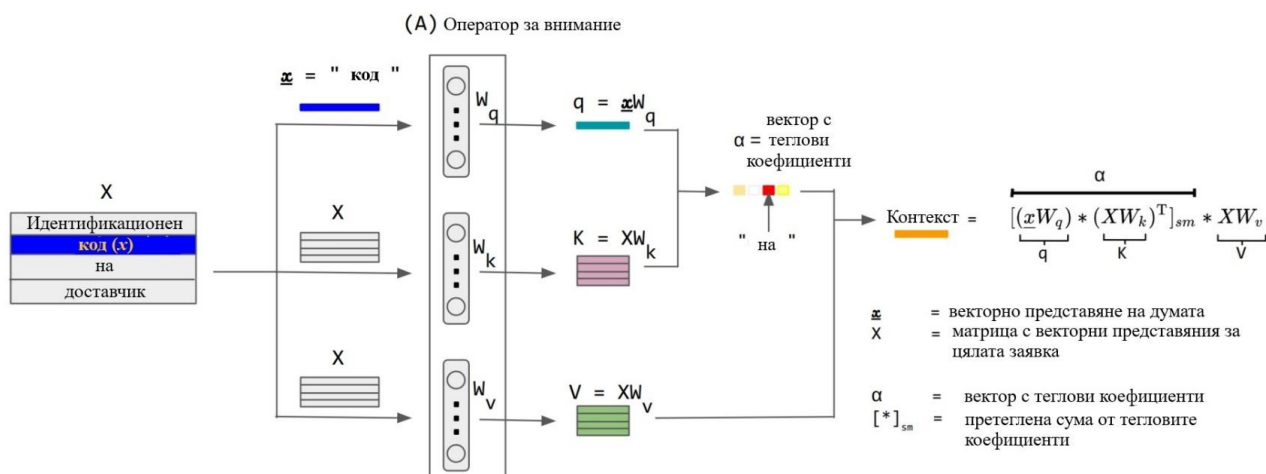
$$\alpha = \text{softmax}\left(\frac{Q \cdot K}{\sqrt{n}}\right) \quad (2.13)$$

Претеглената сума от тегловите коефициенти, актуализирани при обучение по механизма на самовниманието, се прилага, както следва за парализиране с i -глави:

$$\alpha_i = \sum_{i=1}^r \text{softmax}\left(\frac{Q \cdot K}{\sqrt{n}}\right) \times V_i \quad (2.14)$$

Изследвани са възможностите за последващо обучение на дълбоки невронни мрежи тип двупосочен енкодер BERT за целите на разпознаване на наименувани обекти (NER) и класификация на документи на база на разпознатите обекти. Осъществено е обучение на варианти на модела BERT за разпознаване на наименувани обекти (градове, организации, продукти, и д.р.) върху корпус от бизнес документи, съставен от Hasso-Platner Intitut и предоставен под лиценз за свободно ползване: CORPUS_COMP_NER. Този корпус от

документи е стандарт в обучението на модели за разпознаване на наименувани обекти в приложения на иРПА за документообработка в ERP системи и е използван при обучението на модела за разпознаване на наименувани обекти, предлаган от SAP: SAP DocX NER. Корпусът се състои от 1000 документа, анотирани в JSON формат. Обучението на моделите е осъществено на Nvidia A100 (40 GB HBM2e) и H100 (80 GB HBM3) GPU, експериментите са осъществени на Nvidia RTX 2060 (6 GB GDDR6, 32 GB споделена памет) и RTX 4060 (8 GB GDDR6, 24 GB споделена памет).



Фиг. 8 Представяне на механизма на самовнимание.

Таблица 2. Хиперпараметри за обучение на BERT модели.

Хиперпараметър		BERT_Base	BERT_Large	DistilBERT
Скорост на обучение (Learning Rate)	Допустима	2e-5 – 5e-5	1e-5 – 3e-5	3e-5 – 7e-5
	Избрана	3e-5	2e-5	5e-5
Размер на партида (Batch Size)	Допустима	16 – 32	8 – 16	32 – 64
	Избрана	16	16	32
Епохи на обучение (Epochs)	Допустима	3 – 5	2 – 4	3 – 6
	Избрана	4	3	5
Максимален размер на входна последователност в бр. токени	Допустима	128–512	128–512	128–512
	Избрана	512	512	512
ε (за Adam оптимизация)	-	1.00e-08	1.00e-08	1.00e-08

От дълбоките невронни модели (BERT_Base, BERT_Large, DistilBERT), обучени на основния корпус от документи, BERT_Base (110 млн. параметъра) показва най-добро отношение между показателите за точност на разпознаване (Таблица 3) и време за обучение (Таблица 4), т.е. има най-добро отношение точност – бързина при изпълнение на процеса на обучение, следователно този модел е избран за допълнително последващо обучение по механизъм на самовниманието с обучаемая заявка и паралелизация върху данни извлечени от търговския регистър на Република България (1675 наименования на активни вписани организации) с цел подобряване на разпознаването на наименованията на регистрирани организации (със стопанска и нестопанска цел) в бизнес документи на български език, което се явява слабо място на съществуващи широко-разпространени модели за разпознаване на наименувани обекти като SAP DocX NER. В резултат на осъщественото последващо обучение е получен подобрен модел за разпознаване на наименувани обекти от тип организация на български език BERT_bg_bizorg.

Таблица 3. Оценяване на резултатите от обучението за разпознаване на наименувани обекти по модели за CORPUS_COMP_NER.

CORPUS_COMP_NER			
Модели	P	R	F1
BERT_Base	0.755	0.773	0.764
BERT_Large	0.819	0.831	0.825
DistilBERT	0.689	0.708	0.698
SAP DocX NER	0.834	0.852	0.843
CORPUS_COMP_NER и търговски регистър			
BERT_bg_bizorg	0.896	0.915	0.905

Таблица 4. Време за обучение по модели в секунди за епоха от обучението.

CORPUS_COMP_NER	
Модели	T_{tr} , s/epoch
BERT_Base	143.15
BERT_Large	478.25
DistilBERT	75.21
SAP DocX NER	325.55
CORPUS_COMP_NER и търговски регистър	
BERT_bg_bizorg	165.45

Таблица 5. Резултати от обучението за разпознаване на наименувани обекти за извадка от 6 основни наименувани обекта с модел SAP DocX NER.

CORPUS_COMP_NER, модел SAP DocX NER			
NER обекти	P	R	F ₁
локация	0.814	0.848	0.831
организация: доставчик	0.643	0.689	0.665
организация: клиент	0.642	0.674	0.658
продуктов код	0.877	0.908	0.892
код на материал	0.864	0.893	0.878
номер на документ	0.786	0.828	0.806

Таблица 6. Резултати от обучението за разпознаване на наименувани обекти за извадка от 6 основни наименувани обекта с модел BERT_bg_bizorg.

CORPUS_COMP_NER + TP, модел BERT_bg_bizorg			
NER обекти	P	R	F1
локация	0.794	0.816	0.805
организация: доставчик	0.778	0.813	0.795
организация: клиент	0.768	0.798	0.783
продуктов код	0.875	0.904	0.889
код на материал	0.884	0.917	0.900
номер на документ	0.828	0.855	0.841

Обученият модел BERT_bg_bizorg постига 13% по-висока точност на разпознаване на наименуван обект от тип „организация: доставчик“ по критерий F_1 и 12.5% по-висока точност на разпознаване на наименуван обект от тип „организация: клиент“ по критерий F_1 в сравнение със съществуващо решение за разпознаване на наименувани обекти SAP DocX NER, предлагано от един от най-мощните доставчици на ERP софтуер. Моделът BERT_bg_bizorg поддържа сравнимо висока точност на разпознаване по другите видове наименувани обекти в сравнение с SAP DocX NER, т.е. разработеният модел за разпознаване на наименувани обекти BERT_bg_bizorg представлява конкурентоспособно решение по отношение на точност на разпознаване на наименувани обекти.

Създаден е модифициран вариант на BERT архитектурата (BERT_mod) за класификация на документите по тип документ на база на извлечените наименувани обекти като неговата ефективност при класификация е сравнена със съществуващи алтернативи (SVM, XML-CNN, HAN). Обучението на BERT_mod следва математично представяне на обучение по механизъм на вниманието като е използван подход за трансфер на знания от голям обучен модел (модел-учител BERT_bg_bizorg за разпознаване на наименувани обекти) към по-малък модел за обучение (модел-ученик за класификация на документи по разпознати наименувани обекти), наречен „дестилиране на знания“.

Таблица 7. Резултати от обучението за класификация на документи по F_1 при валидиране и тест: модел BERT_mod сравнен с алтернативни модели за класификация.

Модели	F_1 , валидиране	F_1 , тест
SVM	85.8	82.8
XML-CNN	86.4±0.3	84.1±0.4
HAN	89.4±0.3	87.5±0.2
BERT_Base	90.7	89.2
BERT_Large	92.6	91.0
BERT_mod	91.3	89.8

Разработеният модел BERT_mod постига до 5.5% подобрене в точността на класификация по критерий F_1 при валидиране и до 7.7% при тестване в сравнение с разгледаните алтернативни решения (SVM, XML-CNN, HAN).

Таблица 8. Латентност при инференция за BERT модели за класификация.

Модели	L инференция, s
BERT_Base	45.2
BERT_Large	115.8
BERT_mod	14.2

Основно предимство на създадения чрез подход за трансфер на знания модел BERT_mod е неговата в пъти по-ниска латентност при инференция в сравнение с по-големите модели, базирани на дълбоки невронни мрежи тип двупосочен енкодер: 3 пъти по-ниска латентност в сравнение с BERT_Base и 8 пъти по-ниска латентност и в сравнение с BERT_Large. Латентността е осреднена за 100 заявки за класификация на документ в сет от 1000 документа, включващи поръчки за закупуване на материали, поръчки за продажба, фактури.

Създадено е иновативно софтуерно решение Hybrid DIE за интелигентна роботизирана процесна автоматизация на документообработка и извличане на данни от документи за бизнес процеси в информационните системи за управление на ресурсите на предприятието (ERP системи), като част от идентифицираното в литературния обзор направление за генериране на добавена стойност, чрез внедряване на инструменти за иРПА, базирани на изкуствен интелект в ИСУ. Разработеното софтуерно решение се базира на два дълбоки невронни модела с архитектура тип двупосочен енкодер: BERT_bg_bizorg за разпознаване на наименувани обекти и BERT_mod за класификация на документи.

ГЛАВА 3. Класификация по ключови показатели за ефективност при оперативен анализ в CRM информационни системи за управление на взаимодействията с клиенти.

Изследвана е възможността за прилагане на съвременни модели за машинно обучение (дървета на решенията CART; ансамблови дървета на решенията: GBM, LightGBM, XGB, RF; KNN алгоритми и машини с опорни вектори SVC) за решаване на задачи за двоична класификация по ключови показатели за ефективност при оперативен анализ в информационни системи за управление на взаимодействията с клиенти (CRM системи) в три направления: продажби, маркетинг и обслужване на клиенти. Предложен е универсален 5-етапен подход към процеса за предварителна обработка на данните (Фиг. 9).



Фиг. 9 Предложен подход за предварителна подготовка на данните за обработка.

Изследвани са методи за импутация (hot deck, mean, MICE) и методи за корекция на класов дисбаланс (SMOTE, Tomek, OSS) в набора данни като фактори, повлияващи върху точността на класификация за класификационни задачи в направления продажби и обслужване на клиенти.

Предложен е нов метод за селекция на характеристики при двоична класификация на клиенти спрямо ключов показател за ефективност доживотна стойност (CLV), който комбинира оригинални за набора от данни характеристики (актуалност, продължителност, честота на повтарящи се покупки, парична стойност) с такива извлечени чрез прилагане на стандартния вероятностен подход (брой очаквани покупки на седмица, брой очаквани покупки на месец, очаквана осреднена реализирана печалба). Предложеният нов подход при селекция на характеристики за обучение на моделите за класификация балансира влиянието на характеристиките, които произлизат от събраните данни и тези, които са изчислени на база вероятностен модел, което води до по-голяма точност на предсказване на принадлежност на клиентите към един от двата класа в сравнение със стандартните чисто вероятностни подходи, използвани в практиката.

Направление продажби с ключов показател за ефективност процент на надпродаване (UR): решена е задача за двоична класификация – определяне на две категории потребители с „висока“ или „ниска“ вероятност за надпродаване на услуги в масив от данни, предоставени от телекомуникационен оператор, съдържащ 230 потребителски характеристики (190 числови/количествени характеристики и 40 категориен/качествени характеристики, отговарящи на параметри, по които телекомуникационният оператор събира, съхранява и анализира информация за своята клиентска база) в размер от 50 000 редове с данни, предоставени от CRM системата на телекомуникационен оператор (съответстващи на потребители на услугите предлагани от телекомуникационния оператор);

Таблица 9. Импутация като фактор върху точността на класификация по ключов показател „надпродаване“.

Модел за класификация	AUC при различни подходи за импутация		
	Mean	Hot-Deck	MICE
CART (DT classifier)	0.6201	0.6357	0.6365
RF	0.8121	0.8328	0.8387
GBM	0.8331	0.8440	0.8689

Таблица 10. Методи за коригиране на дисбаланс в класово разпределение на данните като фактор върху точността на класификация по ключов показател „надпродаване“.

Модел за класификация	AUC при различни подходи за коригиране на дисбаланс		
	SMOTE	TOMEK	OSS
RF	0.9165	0.8168	0.8395
GBM	0.9428	0.8318	0.8617

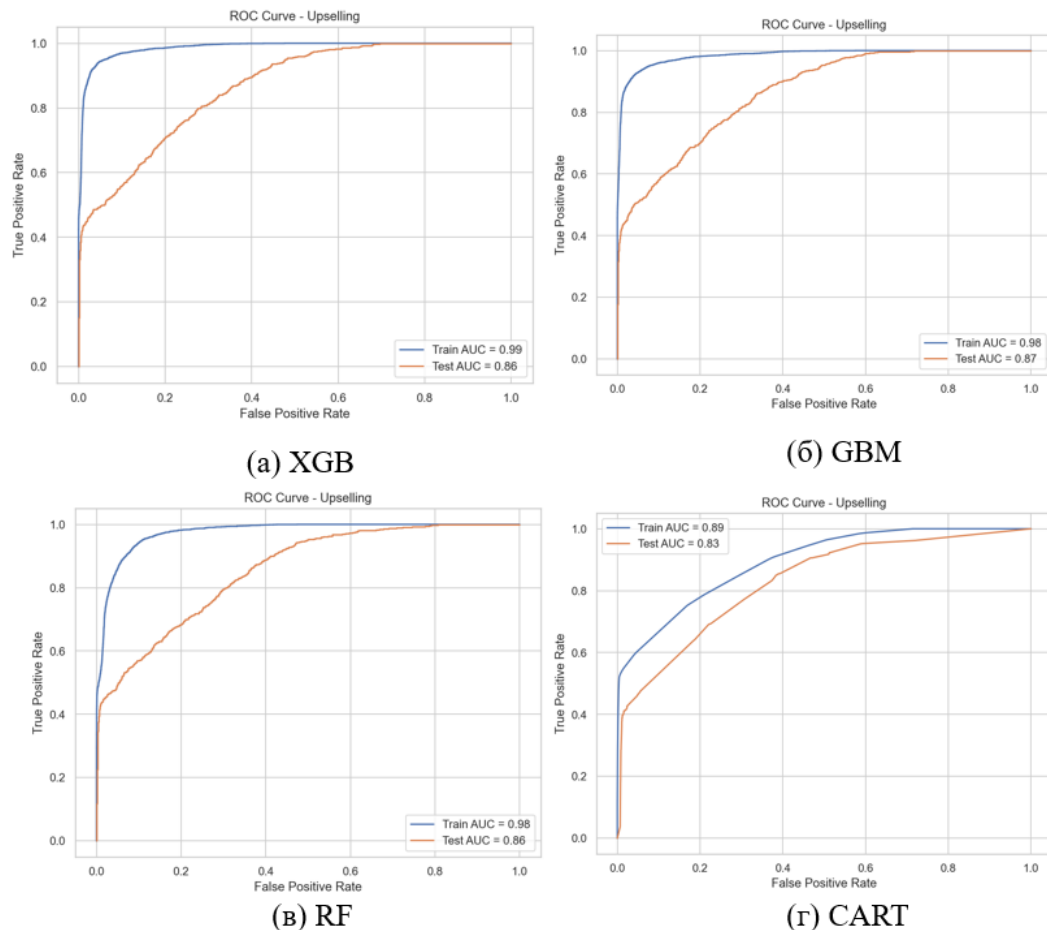
Въз основа на получените резултати са избрани подход за множествена импутация (MICE) и подход за едностранна селекция (OSS) при корекция на класов дисбаланс, която постига обяснимо ниво на подобрение в точността на класификация спрямо първоначалната AUC стойност за всеки от класификаторите.

Таблица 11. Резултати от класификация по ключов показател „надпродаване“ за оперативен анализ на данни в направление продажби в рамките на CRM система с дървета не решенията, ансамблови алгоритми, KNN, SVC.

CART	AUC		По класове	P		R		F1	
	Обучение	Тест		Обучение	Тест	Обучение	Тест	Обучение	Тест
	0.89	0.83	Кл. „-1“:	0.90	0.85	0.88	0.75	0.88	0.79
			Кл. „1“:	0.86	0.81	0.72	0.69	0.78	0.74

Модел за класификация	AUC		
	Обучение	Тест	Базови модели
XGB	0.99	0.86	0.83
GBM	0.98	0.87	0.83
LightGBM	0.94	0.81	0.78
RF	0.98	0.86	0.81
KNN	0.80	0.75	0.72
SVC	0.86	0.79	0.75

Постигнато е подобрение между 3-5% по отношение на AUC показателя за коректност на класификация по ключов показател „надпродаване“ между AUC стойностите за базовите модели (преди прилагане на методи за импутация, корекция на дисбаланс и настройка на хиперпараметри за моделите за класификация) и тези получени при класификация върху тестовите данни след настройка на хиперпараметри и обучение върху набора данни за всички модели за класификация (CART, GBM, LightGBM, XGB, RF, KNN, SVC), базирани на машинно обучение.



Фиг. 10 ROC криви за класификатори с най-висок AUC при класификация по ключов показател „надпродаване“.

Направление маркетинг с ключов показател за ефективност доживотна стойност на клиента (CLV): решена е задача за двоична класификация – определяне на две категории потребители, генериращи „значителна“ стойност за бранда в определения времеви хоризонт или такива, чиито взаимодействия с бранда генерират „незначителна“ стойност за масив от 9300 редове с данни, съответстващи на потребители, закупили продукти от електронния магазин, внедрен към CRM системата на телекомуникационния оператор (напр. зарядно, кейс и протектор за телефон, слушалки, смарт часовник и т.н.) в рамките на едногодишен период;

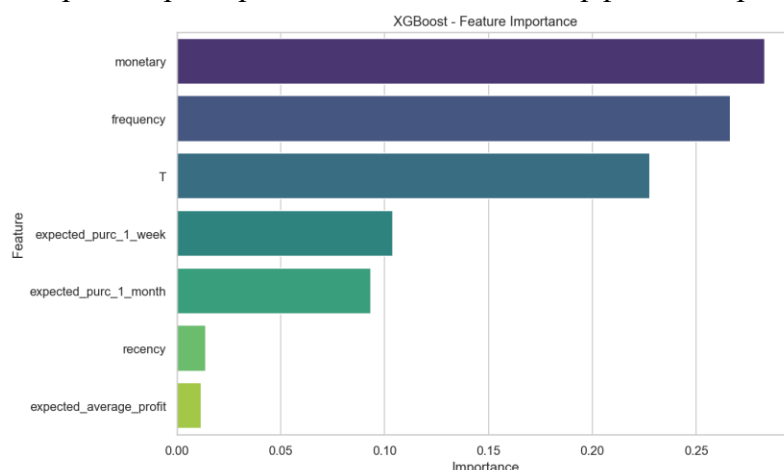
Предложен е нов метод за селекция на характеристики при двоична класификация на клиенти спрямо ключов показател за ефективност доживотна стойност (CLV), който комбинира оригинални за набора от данни характеристики с такива извлечени чрез прилагане на стандартния вероятностен подход на изчисляване на доживотна стойност на клиента. Приложен е пермутационен подход за определяне на влиянието на характеристиките върху осъществената класификация, което е важен елемент за постигане на обяснимост на получените резултати от обучение и класификация с моделите, базирани на машинно обучение.

На Фиг. 12 и 13 са представени селектираните характеристики, сортирани по важност, за осъществяване на класификация по ключов показател за ефективност доживотна стойност (CLV) при оперативен анализ на данни в направление маркетинг в рамките на CRM система с XGB класификатор и LightGBM класификатор.

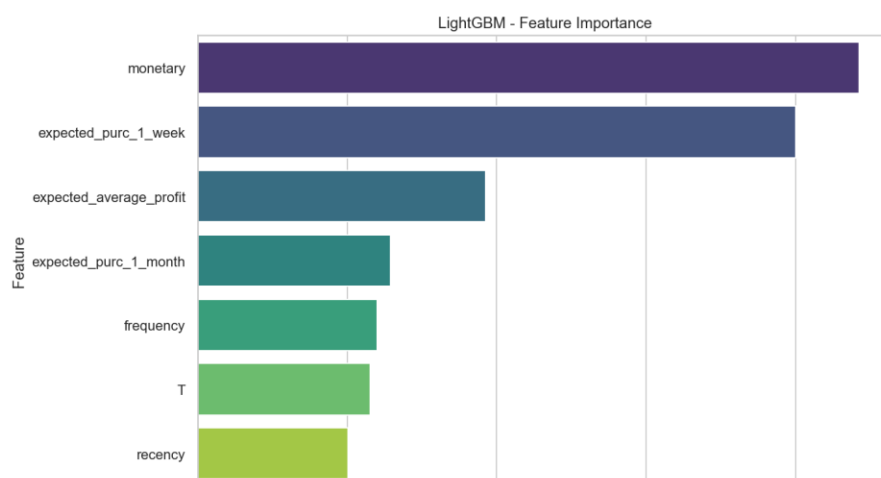
В Таблица 12 са представени резултати от класификация по ключов показател доживотна стойност за оперативен анализ на данни в направление маркетинг в рамките на CRM система по вероятностен подход (BUYD) и с ансамблови модели за класификация (XGB, RF, LightGBM).

	recency	T	frequency	monetary	expected_purc_1_week	expected_purc_1_month
Customer ID						
12748.0	53.142857	53.428571	210	154.192429	3.265158	13.025670
14911.0	53.142857	53.428571	201	691.710100	3.126645	12.473095
17841.0	53.000000	53.428571	124	330.134355	1.940290	7.740345
13089.0	52.285714	52.857143	97	606.362474	1.537528	6.133456
14606.0	53.142857	53.428571	93	130.139032	1.463999	5.840269
15311.0	53.285714	53.428571	91	667.779121	1.433717	5.719467
12971.0	52.571429	53.285714	86	127.485872	1.357024	5.413481
14646.0	50.428571	50.714286	74	3596.804392	1.222517	4.876360
13408.0	53.000000	53.428571	62	453.500645	0.986249	3.934373
18102.0	52.285714	52.571429	60	3859.739083	0.968607	3.863856

Фиг. 11 Селектирани характеристики: извадка от интерфейс за обработка в Python.



Фиг. 12 Селектирани характеристики за класификация по ключов показател за ефективност CLV с XGB класификатор.



Фиг. 13 Селектирани характеристики за класификация по ключов показател за ефективност CLV с LightGBM класификатор.

Таблица 12. Резултати от класификация по ключов показател доживотна стойност с вероятностен подход (BUYD) и с ансамбови модели за класификация (RF, XGB, LightGBM).

Модел за класификация	ERR, %
BUYD (вероятностен)	9.57
RF	8.71
LightGBM	8.56
XGB	8.62
RF + BUYD (hybrid)	4.53
LGBM + BUYD (hybrid)	4.41
XGB + BUYD (hybrid)	4.48

Постигнато е подобрение под формата на редукция от приблизително 4% по отношение на ERR показателя за обща грешка при класификация при сравнение на ансамбовите модели тип дървета на решенията в комбинация с хибридна селекция на характеристики за обучение (RF + BUYD hybrid, XGB + BUYD hybrid и LightGBM + BUYD hybrid) с тези без хибридна селекция на характеристики за обучение (RF, XGB и LightGBM), т.е. предложеният подход за селекция и класификация представлява значимо подобрение в двоичната класификация по ключов показател „доживотна стойност“. Постигнато е подобрение под формата на редукция от приблизително 5% по отношение на ERR показателя за обща грешка при класификация в сравнение със стандартния вероятностен подход.

Направление обслужване на клиенти с ключов показател за ефективност процент на изоставяне (CR): решена е задача за двоична класификация – определяне на две категории потребители с „ниска“ и „висока“ вероятност за изоставяне на ползваните услуги от потребителя (преминаване на потребителски план към друг оператор, закриване на услугата и т.н.) в масив от данни, съдържащ 230 потребителски характеристики за клиентска база от 50 000 потребителя, предоставени от CRM системата на телекомуникационен оператор.

Таблица 13. Импутация като фактор върху точността на класификация по ключов показател „изоставяне“.

Модел за класификация	AUC при различни подходи за импутация		
	Mean	Hot-Deck	MICE
CART (DT classifier)	0.6563	0.6932	0.6368
RF	0.6708	0.7488	0.6899
GBM	0.6784	0.7507	0.6714

Таблица 14. Методи за коригиране на дисбаланс в класово разпределение на данните като фактор върху точността на класификация по ключов показател „изоставяне“.

Модел за класификация	AUC при различни подходи за коригиране на дисбаланс		
	SMOTE	TOMEK	OSS
RF	0.7985	0.6801	0.7199
GBM	0.8269	0.7100	0.7406

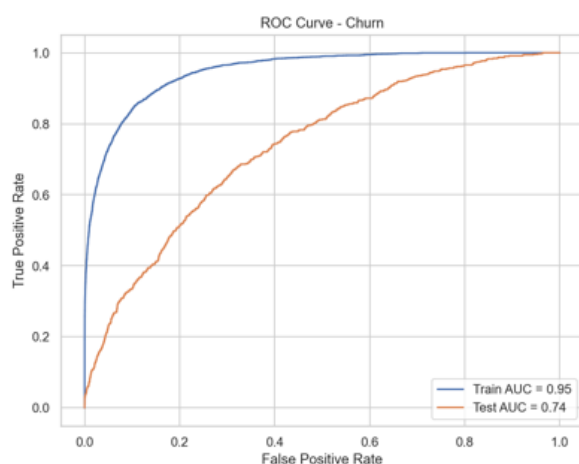
Въз основа на получените резултати са избрани подход за единична кълъстерна импутация (hot deck) и подход за едностранна селекция (OSS) при корекция на класов

дисбаланс, която постига обяснимо ниво на подобрение в точността на класификация спрямо първоначалната AUC стойност за всеки от класификаторите.

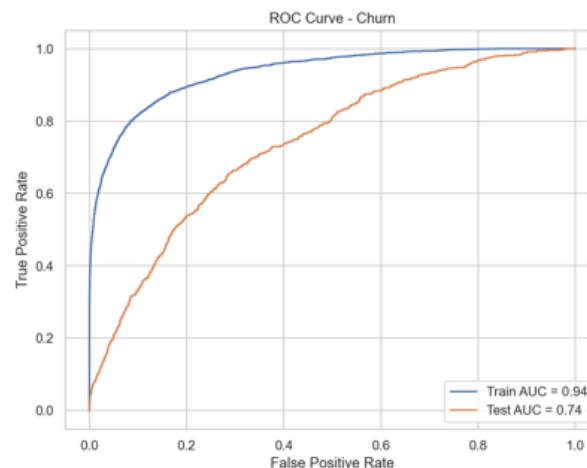
Таблица 15. Резултати от класификация по ключов показател „изоставяне“ за оперативен анализ на данни в направление обслужване на клиенти в рамките на CRM система с дървета на решенията, ансамблови алгоритми, KNN, SVC.

CART	AUC		По класове	P		R		F1	
	Обучение	Тест		Обучение	Тест	Обучение	Тест	Обучение	Тест
	0.78	0.67		0.99	0.93	1	0.99	0.99	0.96
			Кл. „1“	0.85	0.73	0.74	0.67	0.79	0.7

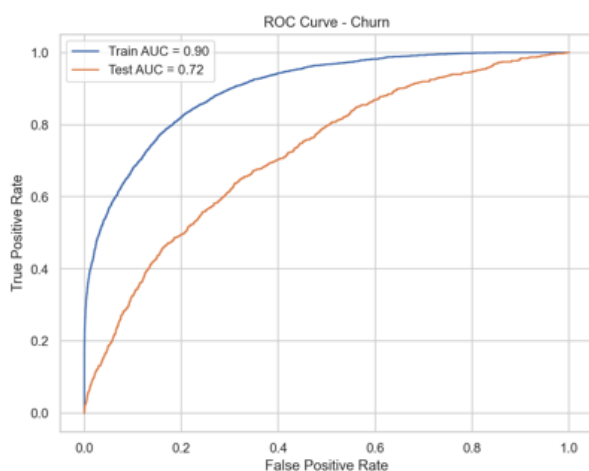
Модел за класификация	AUC		
	Обучение	Тест	Базови модели
XGB	0.95	0.74	0.70
GBM	0.94	0.74	0.70
LightGBM	0.92	0.71	0.67
RF	0.90	0.72	0.68
KNN	0.81	0.63	0.58
SVC	0.86	0.67	0.63



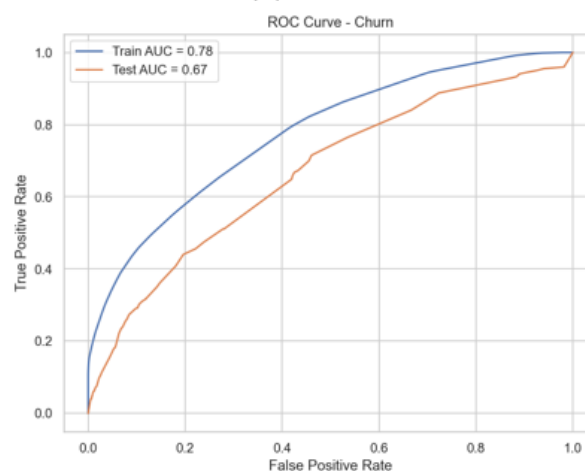
(a) XGB



(б) GBM



(в) RF



(г) CART

Фиг. 14 ROC криви за класификатори с най-висок AUC при класификация по ключов показател „изоставяне“.

Постигнато е подобрение от 4-5% по отношение на AUC показателя за коректност на класификация по ключов показател „изоставяне“ между AUC стойностите за базовите модели (преди прилагане на методи за импутация, корекция на дисбаланс и настройка на хиперпараметри за моделите за класификация) и тези получени при класификация върху тестовите данни след настройка на хиперпараметри и обучение върху набора данни за всички модели за класификация (CART, GBM, LightGBM, XGB, RF, KNN, SVC), базирани на машинно обучение.

ГЛАВА 4. Когнитивно взаимодействие в информационни системи за управление на човешки ресурси и интегрирани информационни системи в индустрията.

Задачите, основаващи се на когнитивно взаимодействие представляват значителен дял от ежедневните операции, осъществявани в информационните системи за управление в индустрията, което ги превръща във фокус на редица научни изследвания и разработки както в държавния, така и в частния сектор. Тези задачи включват динамичен обмен на информация между потребители/служители и машинен интерфейс, софтуерни платформи и интелигентни дигитални агенти за целите на извличане на познание, систематизиране и обобщаване на информация, генериране на съдържание и подпомагане на процеса по вземане на решения. Това означава, че е необходим интердисциплинарен подход в разработването на модели и изследването на възможни решения в областта. Въпреки че взаимодействията човек-машина не са нова област на изследване, а всъщност са обект на засилен научноизследователски интерес още от зората на компютърните технологии и информационните науки. Развитието на тази област е драстично повлияно в последните няколко години от приоритизираното разработване и масово интегриране на технологии, базирани на методите на изкуствен интелект, специфично тези с генеративен характер като големите езикови модели (LLMs).

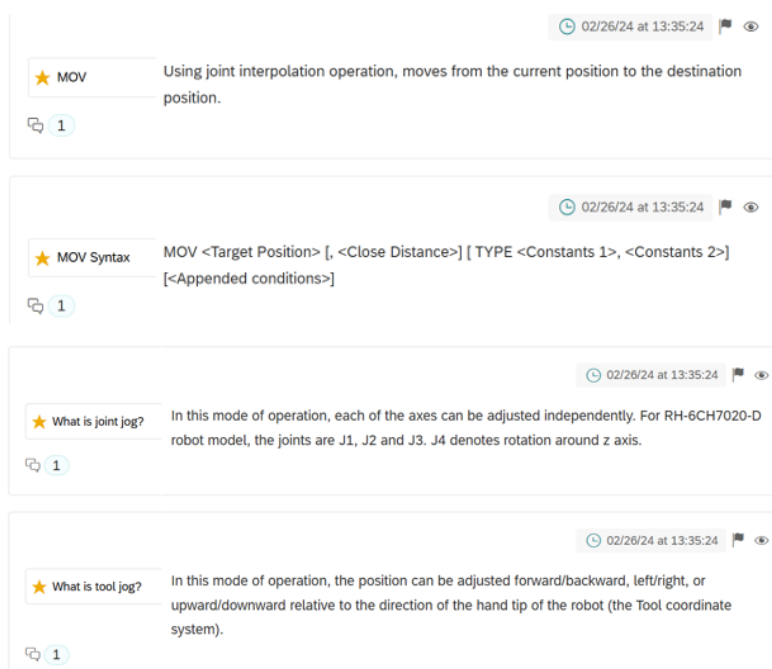
В осъществения литературен обзор задачите за когнитивно взаимодействие са идентифицирани не само като основно направление, носещо стойност за бизнеса, но и като основна тенденция в развитието на информационните системи за управление в индустрията. Този клас задачи се характеризират с висока актуалност и приложимост като същевременно са обременени от множество проблеми, свързани с прозрачност, проследимост, конфиденциалност на обработваните данни и истинност на генерираното съдържание, произтичащи от същността на предпочитания подход за разработване на решения, базирани на генеративен изкуствен интелект и по-специфично на големи езикови модели.

Внедряването на софтуерни решения: интелигентни дигитални асистиенти/агенти, базирани на големи езикови модели, в информационни системи за управление в индустрията като ERP системи, CRM системи и HXM системи в работни потоци, включващи задачи за когнитивно взаимодействие като комуникация с клиенти или онбординг и обучение на персонал създава както възможности за технологичното обезпечаване на тези процеси, така и предизвикателства пред функционалната реализация на такива решения.

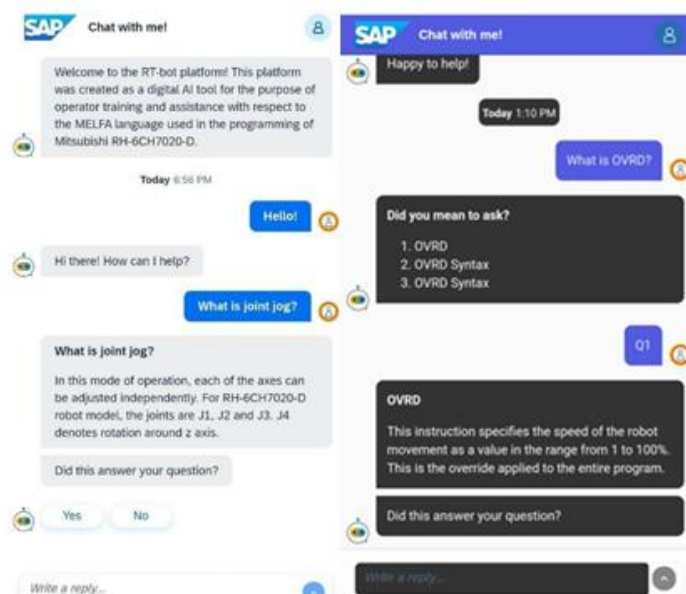
Разработен е интелигентен дигитален асистент RTB, базиран на голям езиков модел GPT 2, за подпомагане на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubischi RH-6CH7020-D в научноизследователския и учебния процес, в следните етапи на изпълнение:

- Формирана е база данни за обучение на големи езикови модели за генериране на код в програмен език MELFA Basic V. Формиран е корпус от учебни документи: техническа документация за робот Mitsubischi RH-6CH7020-D и ръководства за работа със софтуерната платформа на робота RT Toolbox 3;

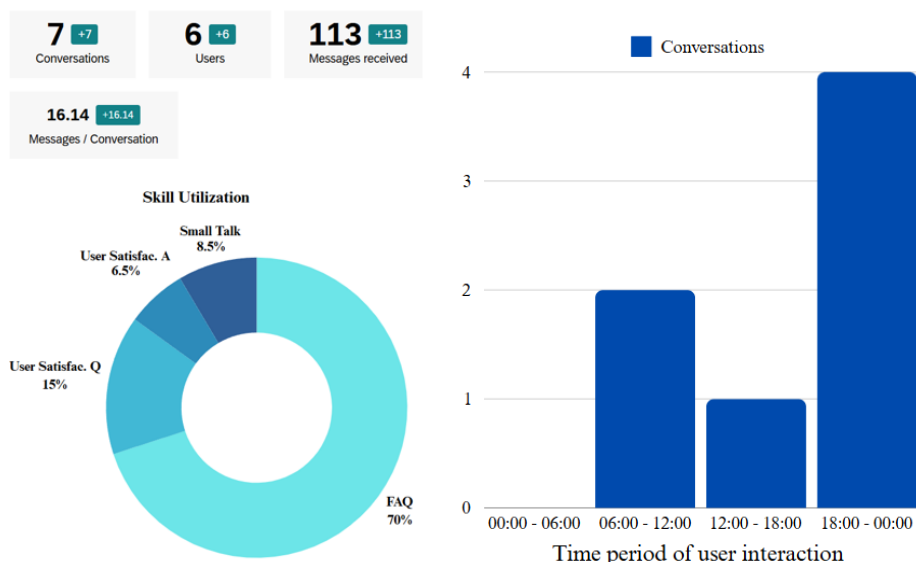
- Разработен е бенчмарк MBV-Bench за оценяване на способността на обучени големи езикови модели да генерират код по потребителска заявка на специализиран програмен език MELFA Basic V;
- Подбрани са големи езикови модели за обучение и е осъществено обучение на моделите върху формираните база данни и корпус от документи. Избран е подход за последващо обучение и контекстуализация на познание от специализирана област LoRA, тъй като при тази техника се актуализира само малък набор от параметри, значително намалявайки разходите за памет и изчисления;
- Обученият интелигентен дигитален асистент е свързан с интегрирана информационна система за управление SAP чрез уеб-клиент.



Фиг. 15 Извадка от формираната база данни за обучение на програмен език MELFA Basic V.



Фиг. 16 Извадки от потребителски взаимодействия с разработения дигитален асистент, реализирани в SAP BTP чрез API.



Фиг. 17 Визуализации на статистически данни за употребата на дигитален асистент RTB, достъпни в платформата за управление на дигиталния асистент.

Обхватът на интелигентния дигитален асистент (RTB) е разширен до разработване на платформа Private Edge LLMs за последващо обучение на големи езикови модели (LLaMA, Gemma, Qwen, Deepseek R1) върху бази данни и корпуси от документи от специализирани научни, приложни и научно-приложни области с възможност за локално обучение чрез реализация на техники за параметрично-ефикасно последващо обучение (LoRA, QLoRA, LLaMA Adapter), контекстуализация (RAG) и локално съхранение на данни за запазване на тяхната конфиденциалност. Разработената платформа Private Edge LLMs може да бъде внедрена като иновативно решение към информационни системи за управление в сектори, за които конфиденциалността на данните за обучение и полученото генерирано съдържание е от ключово значение (като научноизследователски институции) или в легаси системи, които не са обезпечени с облачно внедряване.



Фиг. 18 Модулна архитектура на разработената платформа Private Edge LLMs за допълнително обучение на големи езикови модели.

Осъществено е изследване на влиянието на настройвани хиперпараметри на големите езикови модели при различни техники за параметрично-ефикасно последващо обучение в рамките на разработената платформа за последващо обучение на големи езикови модели Private Edge LLMs.

Таблица 16. Хиперпараметри за обучение на LLM модели по LoRA.

Хиперпараметър		GPT2 Large (1.5B) RTB	Deepseek R1 (1.5 B)	LLaMA 3.1 (8B)	Gemma 1 (7B)	Qwen 2 (7B)
Скорост на обучение (Learning Rate)		1e-4	1e-4	5e-5	3e-5	5e-5
Епохи на обучение (Epochs)	Допустима Избрана	3 – 5 4	3 – 5 4	2 – 3 3	2 – 3 3	2 – 3 3
Размер на блокове текст (Block size)	Допустима Избрана	256–512 512	256–512 512	128–256 256	128–256 256	128–256 256
LoRA-ранг	Допустима Избрана	16-32 16	16-32 16	8-16 16	8-16 16	8-16 16
LoRA- α		32	32	32	32	32
LoRA-dropout		0.1	0.1	0.05	0.05	0.05

Таблица 17. Хиперпараметри за обучение на LLM модели по QLoRA.

Хиперпараметър		Deepseek R1 (1.5 B)	LLaMA 3.1 (8B)	Gemma 1 (7B)	Qwen 2 (7B)
Скорост на обучение (Learning Rate)		2e-4	1e-5	3e-5	3e-5
Епохи на обучение (Epochs)	Допустима Избрана	3 – 5 4	2 – 3 3	2 – 3 3	2 – 3 3
Размер на блокове текст (Block size)	Допустима Избрана	64–128 128	64–128 128	64–128 128	64–128 128
QLoRA-ранг	Допустима Избрана	32-64 32	16-32 16	16-32 16	32-64 32
QLoRA- α		64	32	32	64
QLoRA-dropout		0.1	0.05	0.1	0.05

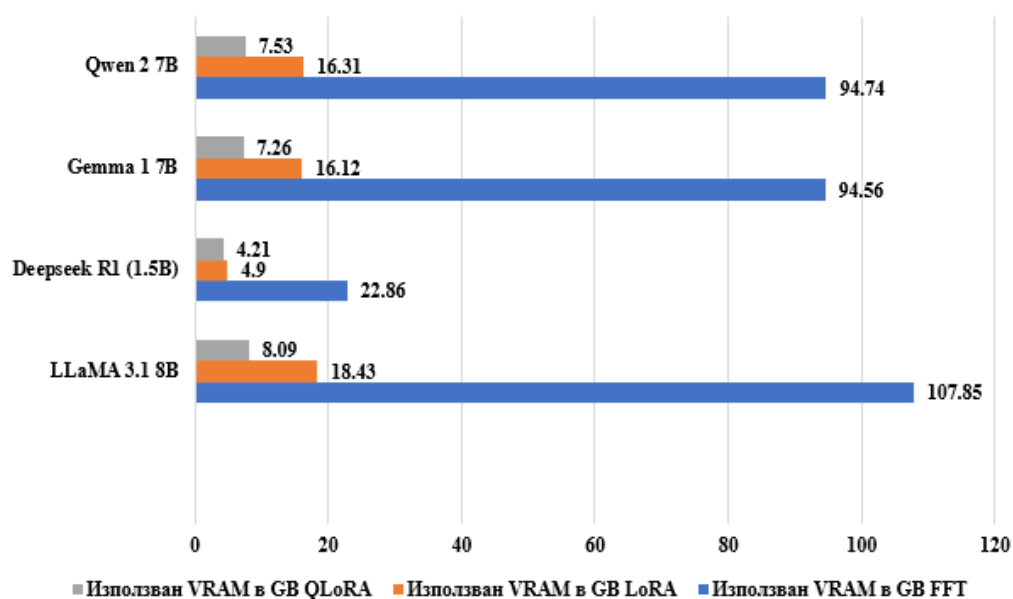
Таблица 18. Оценка на резултати от последващо параметрично-ефикасно обучение на големи езикови модели в рамките на RTB и PrivateEdgeLLMs при генериране на текстови отговор.

Големи езикови модели + вид последващо обучение	F1		L, ms/t
	Преди последващо обучение	След последващо обучение	
GPT2 (1.5B) + LoRA	0.67	0.80	20 ± 5
LLaMA 3.1 8B + LoRA	0.77	0.91	25 ± 3
LLaMA 3.1 8B + QLoRA	0.77	0.80	
LLaMA 3.1 8B + Adapter	0.77	0.89	
Deepseek R1 (1.5B) + LoRA	0.70	0.81	15 ± 5
Deepseek R1 (1.5B) + QLoRA	0.70	0.73	
Deepseek R1 (1.5B) + Adapter	0.70	0.77	
Gemma1 7B + LoRA	0.73	0.88	40 ± 6
Gemma1 7B + QLoRA	0.73	0.77	
Gemma1 7B + Adapter	0.73	0.82	
Qwen2 7B + LoRA	0.72	0.86	22 ± 8
Qwen2 7B + QLoRA	0.72	0.76	
Qwen2 7B + Adapter	0.72	0.79	

В резултат от последващо параметрично-ефикасно обучение на модели с архитектура от тип трансформатор (GPT2, Deepseek R1) са постигнати 11-13% подобрене (при подход LoRA) и 3% подобрене (при подход с квантуване на тегловите коефициенти QLoRA) в точността на генерирано съдържание по параметър F1 с максимално средно увеличение в латентността преди сливане на тегловите коефициенти на LoRA адапторите от 5 ms/t, сравнено с базовите модели.

В резултат от последващо обучение LoRA на модели с архитектура от тип авторегресивен декодер (LLaMA, Gemma, Qwen) са постигнати 14-15% подобрене (при подход LoRA) и 3-4% подобрене (при подход с квантуване на тегловите коефициенти QLoRA) в точността на генерирано съдържание по параметър F1 с максимално средно увеличение в латентността преди сливане на тегловите коефициенти на LoRA адапторите от 8 ms/t, сравнено с базовите модели

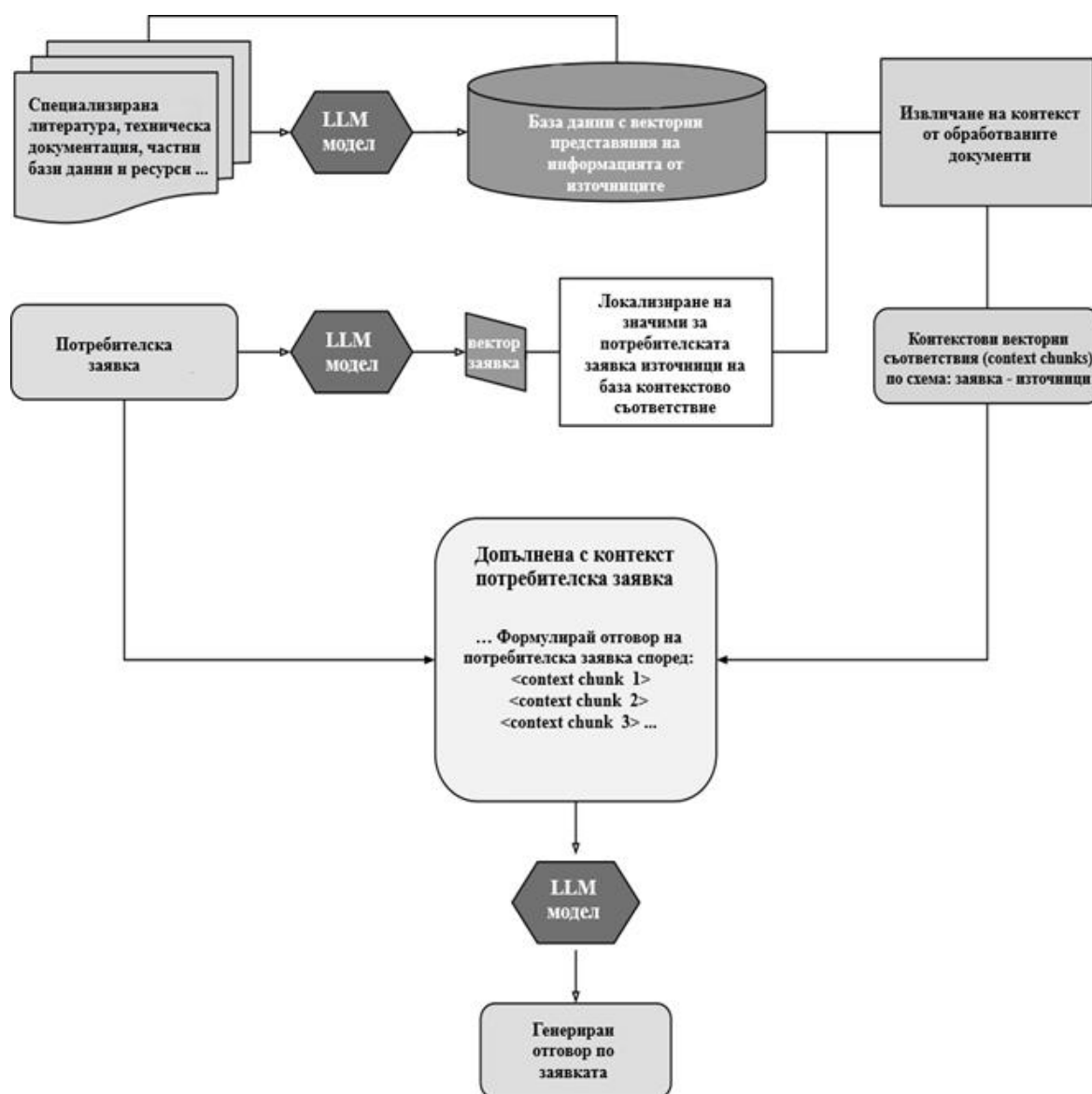
Подхода на обучение с добавяне на адаптационни промптове LLaMA Adapter постига 7-12% подобрене по F1 спрямо базовите модели при архитектури тип авторегресивен декодер и до 7% подобрене при архитектури тип трансформатор.



Фиг. 19 Утилизация на VRAM при последващо параметрично-ефикасно обучение на големи езикови модели в рамките на PrivateEdgeLLMs.

С прилагането на настройка на хиперпараметри и подходи за параметрично-ефикасно последващо обучение на големи езикови модели в рамките на разработената платформа Private Edge LLMs необходимият за обучение VRAM на графичния процесор е редуциран: с между 81-92.5% при LoRA подход и с между 78.5-83% при QLoRA подход спрямо VRAM-а необходим за пълна актуализация на теглата за различните обучавани големи езикови модели. Постигнатите редукции по отношение на необходим изчислителен ресурс (VRAM) на графичния процесор позволяват последващото обучение на отворени големи езикови модели в рамките на разработената платформа Private Edge LLMs да бъде осъществено на графични процесори от потребителско ниво като Nvidia RTX 3090 и Nvidia RTX 4090.

В рамките на разработената платформа Private Edge LLMs е реализирана схема за допълнителна контекстуализация RAG. Прилагането на подход за контекстуализация при обучение RAG помага да се преодолеят някои от ограниченията на големи езикови модели като остаряваща информация, неточности в базата знания, използвана за предварителното обучение на модели или неточности в генерираното съдържание т.нар. „халюцинации“ и „кофабулации“ като позволява на големите езикови модели да имат достъп и да включват информация от конкретни документи или бази данни при генериране на отговор.



Фиг. 20 Представяне на внедрената схема RAG за контекстуализация на познание при последващо обучение на големи езикови модели.

Постигнато е подобрене 4-7% на индекса за обобщаване на съдържание ROUGE-L според вида голям езиков модел при контекстуализацията с RAG върху корпус от техническа документация за работа с робот Mitsubishi RH-6CH7020-D и среда RT Toolbox 3, както и редуциране на индекса за несигурност PPL за всички обучавани модели.

Таблица 19. Резултати от последваща контекстуализация с RAG на големи езикови модели в рамките на PrivateEdgeLLMs в задачи за генериране на текстови отговор и обобщаване.

Модел + избран подход за обучение	Параметри за оценяване			
	PPL, без RAG	PPL с RAG	ROUGE-L без RAG	ROUGE-L с RAG
LLaMA 3.1 8B	8	6	0.86	0.92
Deepseek R1 (1.5B)	18	11	0.78	0.82
Gemma1 7B	12	8	0.81	0.85
Qwen2 7B	10	6	0.84	0.91

Таблица 20. Резултати от последващо параметрично-ефикасно обучение на големи езикови модели в рамките на RTB и PrivateEdgeLLMs в задачи за генериране на код.

Модел + избран подход за последващо обучение	Параметри за оценяване		
	Pass@10 оценка върху разработен бенчмарк <i>MBV-Bench</i>		L, ms/t
	Преди последващо обучение	След последващо обучение	
GPT2 + LoRA	58%	66%	18± 3
LLaMA + LoRA	74%	83%	20±5
LLaMA + QLoRA	74%	77%	
Deepseek R1 + LoRA	80%	92%	12±5
Deepseek R1 + QLoRA	80%	84%	
Gemma + LoRA	67%	75%	35± 3
Gemma + QLoRA	67%	71%	
Qwen + LoRA	87%	95%	20 ± 6
Qwen + QLoRA	87%	90%	

В резултат от последващо обучение на модели с архитектура от тип авторегресивен декодер (LLaMA, Gemma, Qwen) са постигнати 8-9% (за LoRA) и 3-4% (за QLoRA) подобрение в точността на генерирани програмни решения по параметър Pass@10 с максимално средно увеличение в латентността преди сливане на тегловите коефициенти на LoRA адапторите от 6 ms/t, сравнено с базовите модели.

Разработената платформа Private Edge LLMs позволява комбиниране на локален изчислителен ресурс и облачен такъв при последващо обучение, както и достъпване на облачно-внедряеми големи езикови модели като GPT (т.нар хибридно обучение), което в комбинация с разнообразието от поддържани техники за последващо обучение и контекстуализация я превръща в конкурентоспособно решение за последващо обучение на големи езикови модели. PrivateEdgeLLMs позволява и обособяване на обучението на отворени големи езикови модели върху изцяло локален изчислителен ресурс с изцяло локално съхранение на обработваните данни, актуализираните теглови коефициенти и полученото генерирано съдържание, което я превръща в атрактивно софтуерно решение за приложни области, изискващи поверителност на обработваната информация като научноизследователски институции и развойни звена.

ГЛАВА 5. ИЗВОДИ

От Глава 2:

Създадено е иновативно софтуерно решение Hybrid DIE за интелигентна роботизирана процесна автоматизация на документообработка и извличане на данни от документи за бизнес процеси приложимо в информационните системи за управление на ресурсите на предприятието (ERP системи), базирано на два дълбоки невронни модела с архитектура тип двупосочен енкодер: BERT_bg_bizorg за разпознаване на наименувани обекти и BERT_mod за класификация на документи.

- Изследвани са възможностите за последващо обучение с учител на дълбоки невронни мрежи тип двупосочен енкодер BERT по механизъм на вниманието за целите на разпознаване на наименувани обекти (NER) в бизнес документи с настройка на хиперпараметри за разпознаване на наименувани обекти (организации, продукти, и д.р.) върху корпус от бизнес документи, съставен от Hasso-Platner Intitut и предоставен под лиценз за свободно ползване.

- Последващо допълнително обучение на BERT_Base модела е осъществено по механизъм на самовниманието с обучаемая заявка и паралелизация върху данни извлечени от търговския регистър на Република България с цел подобряване на разпознаването на наименованията на регистрирани организации в бизнес документи на български език, което се явява слабо място на съществуващи широко-разпространени модели за разпознаване на наименувани обекти като SAP DocX NER. В резултат на осъщественото последващо обучение е получен подобрен модел за разпознаване на наименувани обекти от тип организация на български език BERT_bg_bizorg. Обученият модел BERT_bg_bizorg постига 13% по-висока точност на разпознаване на наименуван обект от тип „организация: доставчик“ по критерий F1 и 12.5% по-висока точност на разпознаване на наименуван обект от тип „организация: клиент“ по критерий F1 в сравнение със съществуващо решение за разпознаване на наименувани обекти SAP DocX NER, предлагано от един от най-мощните доставчици на ERP софтуер.
- Моделът BERT_bg_bizorg поддържа сравнимо висока точност на разпознаване по другите видове наименувани обекти в сравнение с SAP DocX NER, т.е. разработеният модел за разпознаване на наименувани обекти BERT_bg_bizorg представлява конкурентоспособно решение по отношение на точност на разпознаване на наименувани обекти.
- Обучен е модел с модифициран вариант на BERT архитектурата BERT_mod за класификация на документите по тип документ на база на извлечените наименувани обекти. Разработеният модел BERT_mod постига до 5.5% подобрение в точността на класификация по критерий F1 при валидиране и до 7.7% при тестване в сравнение с разглежданите алтернативни решения SVM, XML-CNN, HAN. Основно предимство на създадения чрез подход за трансфер на знания модел BERT_mod е неговата в пъти по-ниска латентност при инференция в сравнение с други BERT модели: 3 пъти по-ниска латентност в сравнение с BERT_Base и 8 пъти по-ниска латентност и в сравнение с BERT_Large.

От Глава 3:

Осъществено е изследване на приложимостта на модели за машинно обучение (дървета на решенията CART; ансамблови дървета на решенията GBM, LightGBM, XGB, RF; KNN алгоритми и машини с опорни вектори SVC) за решаване на задачи за двоична класификация по ключови показатели за ефективност при оперативен анализ в информационни системи за управление на взаимодействията с клиенти (CRM системи) в три направления: продажби, маркетинг и обслужване на клиенти.

- Предложен е универсален 5-етапен подход към процеса за предварителна обработка на данните при решаване на задачи за двоична класификация в CRM системи.
- Изследвани са методи за импутация (hot deck, mean, MICE) и методи за корекция на класов дисбаланс (SMOTE, Tomek, OSS) в набора данни като фактори, повлияващи върху точността на класификация за класификационни задачи в направления продажби и обслужване на клиенти.
- Предложен е нов метод за селекция на характеристики при двоична класификация на клиенти спрямо ключов показател за ефективност доживотна стойност (CLV), който комбинира оригинални за набора от данни характеристики с такива извлечени чрез прилагане на стандартния вероятностен подход, което води до 5% по-висока точност на предсказване на принадлежност на клиентите към един от

двата класа в сравнение със стандартните чисто вероятностни подходи, използвани в практиката.

От Глава 4:

Разработени са съвременни софтуерни решения за автоматизация на когнитивни взаимодействия в процесите по онбординг и обучение на служители, базирани на допълнително обучение на големи езикови модели с техники за параметрично-ефикасно последващо обучение, които са приложими в ERP системи и НХМ системи.

- Разработен е интелигентен дигитален асистент RTB, базиран на голям езиков модел GPT 2, за подпомагане на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubischi RH-6CH7020-D в научноизследователския и учебния процес: формиран е корпус от обучителни документи за работа със съответния стенд; подбрани са големи езикови модели за обучение и са обучени върху формираната обучителна извадка; обученият интелигентен дигитален асистент е свързан с интегрирана информационна система за управление SAP чрез уеб-клиент.
- Разработен е бенчмарк MBV-Bench за оценяване на способността на обучени големи езикови модели да генерират код по потребителска заявка на специализиран програмен език MELFA Basic V.
- Обхватът на интелигентния дигитален асистент е разширен до разработване на платформа Private Edge LLMs за обучение на големи езикови модели (LLaMA, Gemma, Qwen, Deepseek R1) върху бази данни и корпуси от документи от специализирани научни, приложни и научно-приложни области с възможност за локално обучение чрез реализация на техники за параметрично-ефикасно последващо обучение (LoRA, QLoRA, LLaMA Adapter), контекстуализация (RAG) и локално съхранение на данни за запазване на тяхната поверителност.
- Изследвано е влиянието на прилагане на техники за параметрично-ефикасно последващо обучение LoRA, QLoRA, LLaMA Adapter върху точността на генериране на съдържание от избраните големи езикови модели и върху необходимия изчислителен ресурс (VRAM) на графичния процесор.
- Изследвано е влиянието на прилагане на схема за допълнителна контекстуализация RAG върху индекса за обобщаване на съдържание и индекса за несигурност за всички обучавани големи езикови модели.

НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

Резултатите от изследванията в настоящия дисертационен труд могат да се резюмират в следните научно-приложни и приложни приноси:

Научно-приложни приноси:

- Предложени са и са експериментално тествани дълбок невронен модел тип двупосочен енкодер за разпознаване на наименувани обекти (BERT_bg_bizorg) и дълбок невронен модел тип двупосочен енкодер за класификация на документи (BERT_mod) по разпознатите наименувани обекти, както и софтуерно решение, базирано на моделите (Hybrid DIE) за интелигентна роботизирана процесна автоматизация на документообработка и извличане на данни от документи за бизнес процеси в информационните системи за управление на ресурсите на предприятието (ERP системи).
- Предложен е нов метод за селекция на характеристики при двочна класификация на клиенти спрямо ключов показател за ефективност доживотна стойност (CLV) по данни от CRM системата на телекомуникационен оператор. Новият предложен

метод комбинира оригинални за набора от данни характеристики (актуалност, продължителност, честота на повтарящи се покупки, парична стойност) с такива извлечени чрез прилагане на стандартния вероятностен подход (брой очаквани покупки на седмица, брой очаквани покупки на месец, очаквана осреднена реализирана печалба). Така се балансира влиянието на характеристиките, които произлизат от събраните данни и тези, които са изчислени на база вероятностен модел, което води до по-голяма точност на предсказване на принадлежност на клиентите към един от два класа в сравнение със стандартните чисто вероятностни подходи, използвани в практиката.

- Разработен е бенчмарк MBV-Bench за оценяване на способността на обучени големи езикови модели да генерират код по потребителска заявка на специализиран програмен език MELFA Basic V.

Приложни приноси:

- Осъществено е изследване на приложимостта на модели за машинно обучение (дървета на решенията CART; ансамблови дървета на решенията GBM, LightGBM, XGB, RF; KNN алгоритми и машини с опорни вектори SVC) за решаване на задачи за двоична класификация по ключови показатели за ефективност при оперативен анализ в информационни системи за управление на взаимодействията с клиенти (CRM системи) в три направления: продажби, маркетинг и обслужване на клиенти.
- Разработен е интелигентен дигитален асистент RTB, базиран на голям езиков модел GPT 2, за подпомагане на обучението по програмиране и използване на роботизиран лабораторен стенд Mitsubischi RH-6CH7020-D в научноизследователския и учебния процес, приложим в интегрирани информационни системи за управление в индустрията (ERP системи) и информационни системи за управление на човешки ресурси (HXM системи).
- Разработена е платформа Private Edge LLMs за обучение на големи езикови модели (LLaMA, Gemma, Qwen, Deepseek R1) върху бази данни и корпуси от документи от специализирани научни, приложни и научно-приложни области с възможност за локално обучение чрез реализация на техники за параметрично-ефикасно последващо обучение (LoRA, QLoRA, LLaMA Adapter) и локално съхранение на данни за запазване на тяхната поверителност. Изследвано е влиянието на прилагането на техники за параметрично-ефикасно последващо обучение LoRA, QLoRA, LLaMA Adapter върху точността на генериране на съдържание и програмен код при последващо обучение върху данни от специализирана техническа приложна област на избраните големи езикови модели и върху необходимия за осъществяване на обучението изчислителен ресурс (VRAM) на графичния процесор.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

- Stoykova S, Shakev N. (2023). Artificial Intelligence for Management Information Systems: Opportunities, Challenges, and Future Directions., Algorithms (Q2); 16, 357, <https://doi.org/10.3390/a16080357>
- S. Stoykova and N. Shakev, (2024). Generative AI-Driven Personnel Training in Industrial Robotics through Intelligent HXM, International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, pp. 228-232, DOI: [10.1109/ICAI63388.2024.10851547](https://doi.org/10.1109/ICAI63388.2024.10851547).
- Stoykova, S. (2025). Platform for Local Fine-tuning of Large Language Models, 33th International Symposium "Control of Energy, Industrial and Ecological Systems", https://e-university.tu-sofia.bg/e-publ/files/20100_SAI_symposium_paper.pdf
- Stoykova, S., Shakev, N., Popov, V., Ahmed Shieva, S. (2025). Energy consumption evaluation for large language models. In 2025 10th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE) (pp. 1–6). IEEE, <https://doi.org/10.1109/EEAE65901.2025.11273449>

ABSTRACT of “Application of artificial intelligence in management information systems and business analysis “

Authored by: Stela Stoykova

The aim of the dissertation is the development of innovative software solutions for the application of artificial intelligence in management information systems and business analysis in three directions, generating added value:

- intelligent process automation of document flow and document processing for ERP systems;
- classification by KPIs in operational analysis for CRM systems;
- cognitive interaction (human - AI) for HXM and ERP systems.

To achieve the set aim, the following tasks were completed in each of the directions:

A) Intelligent process automation of document flow and document processing:

- Research and facilitation of fine-tuning of bidirectional encoder deep neural networks BERT for the purpose of named object recognition (NER) and document classification based on the recognized objects.
- Study of improvements in the accuracy of NER of organization entities in Bulgarian language by hyperparameter tuning and fine-tuning of bidirectional encoder DNNs resulting in the development of refined models for NER: BERT_bg_bizorg and document classification: BERT_mod.

B) Classification by KPIs in operational analysis:

- Study of the applicability of machine learning models: decision trees: CART; ensemble decision trees: RF, GBM, XGB, LightGBM; KNN algorithms and support vector machines, to solve binary classification tasks by key performance indicators in operational analysis in three application areas: sales, marketing and customer service.
- Formulating an approach to the data pre-processing process and investigation of factors affecting classification accuracy such as imputation algorithms (hot deck, mean, MICE) and approaches for correcting data class imbalance (SMOTE, Tomek, OSS) in pre-processing as well as features selection in training processes.
- Proposal of a new method for feature selection in binary CLV classification tasks, which combines features original to the dataset with those extracted by applying the standard probabilistic approach.

C) Cognitive interaction in human resource management information systems:

- Development of an intelligent digital assistant *RTB*, based on GPT 2, to support human training in programming and utilization of Mitsubishi RH-6CH7020-D factory automation robot in the research and educational process in the following stages of implementation: formation of a database MELFA Basic V commands and a corpus of training documents for working with the robot; development of a benchmark *MBV-Bench* for evaluation of LLM code generation in MELFA Basic V programming language, selection of LLMs for training and fine-tuning the models on the formed database and corpus of documents; connection of the trained intelligent digital assistant with an integrated management information system SAP via a web client.
- Study of approaches for parameter-efficient fine-tuning (LoRA, QLoRA, LLaMA Adapter) of large language models (LLaMA, Gemma, Qwen, Deepseek R1) in order to expand the scope of the intelligent digital assistant to a platform *Private Edge LLMs* for training large language models on databases and document corpora from specialized scientific and applied fields with the possibility of local training and data storage to preserve data confidentiality. Study of the impact of applying PEFT techniques LoRA, QLoRA, LLaMA Adapter on the accuracy of content generation from the selected LLMs and on the required computational resource (VRAM) of the graphics processor.
- Study of the impact of an additional contextualization scheme RAG on the content summarization index and the perplexity index for all trained large language models.