Funded by the European Union
NextGenerationEU

SUMMIT
SOFIA UNIVERSITY
MARKING MOMENTUM
FOR INNOVATION AND
TECHNOLOGICAL TRANSFER

SOFIA UNIVERSITY -
MARKING MOMENTUM
FOR INNOVATION AND TECHNOLOGICAL TRANSFER

National Recovery and Resilience Plan
OF THE REPUBLIC OF BULGARIA

BEST

Research Area
Informatics and Computer science

Artificial Intelligence and Big data for Education, Software and information Technologies

INTRODUCTION

In recent years, the fields of artificial intelligence (AI) and big data (BD) have gained paramount importance in scientific research, both in Bulgaria and Europe, and worldwide. Development of areas such as data management, analysis and visualization, neural networks, machine learning, knowledge management, etc. is observed. This is a prerequisite for their application in increasing the potential for implementing innovations in a number of subfields of science and industry - including economics, medicine, education, etc.

Main goal of the project is to research and develop approaches and technologies for applying artificial intelligence and big data methods in various fields of informatics and computer science.

PROJECT ACTIVITIES

Project activities are divided into three main work-packages:

1. Application of AI and Big Data in information technologies - progress in these areas allows for increasing the efficiency of automatic detection of patterns and deviations in the data. This allows for the development of more accessible tools for automatic analysis, as well as architectures for activating specific behavior in relation to large volumes of incoming information. This work package aims to analyze opportunities for using AI in such systems, and to propose methods for overcoming the challenges associated with their construction.
2. Application of AI and DG in software technologies - increasingly, software systems that are being developed include elements and components with artificial intelligence that have specific characteristics. They pose new challenges to software engineers, compared to the development of traditional software. The work package aims to analyze and categorize these challenges and propose methods for solving them.
3. Application of AI and DG in educational technologies - this work package aims to analyze the opportunities and limitations that achievements pose for the development of innovative approaches to training. This includes the use of large language, model, serious and adaptive games, including for the application of the competency-based approach in education, assessment of competencies and literacy levels.

PUBLICATIONS

Members of the project team has taken part in 4 published papers, indexed in Web of Science as listed below (team members are in bold):

1) Secondary use of data for data analysis: A case of modeling medical data for treatment analysis and assessment, **Kaloyanova Kalinka,** and K. Kaloyanov, Procedia Computer Science, vol. 234, pp. 461-458, 2024.
2) Three-body periodic collisionless equal-mass free-fall orbits revisited, Hristov, I., **Radoslava Hristova**, Dmitrašinović V. and Tanikawa, K, Celestial Mechanics and Dynamical Astronomy, 136(7), 2024.
3) Lexical Representation of Dense Numerical Vectors: Introducing LangVec, **Simeon Emanuilov and Aleksandar Dimov**, Mathematics and Informatics, 67(3), 2024.
4) Billion-Scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering, **Simeon Emanuilov, Aleksandar Dimov**, CYBERNETICS AND INFORMATION TECHNOLOGIES, Vol. 24(4), 2024, pp.45-58.

Additionally, the following papers has been published and indexed in Scopus:

1) Extending Primary Analysis: Exploring Clinical Data Modeling for Secondary Applications, Kalinka Kaloyanova, Elitsa Kaloyanova, ICTO 2024, Париж, Франция, 27-28.06.2024
2) Cross-Continental Insights: Comparative Analysis of Using AI for Information System Stakeholder Analysis in Undergraduate Courses in the EU and USA, Vijay Kanabar and Kalinka Kaloyanova, CSECS 2024, София, 28-30.06.2024

Our team is currently working on many aspects of AI and Big Data, among them:

1. Visualization techniques
2. Virtual robotics
3. Analysis of medical data
4. Hallucinations of Large Language Models (LLMs)
5. Methods for development of data-intensive software systems
6. Development of efficient algorithms for HPC calculations on large dataset.s

Three more papers on those topics have been submitted for publication in WoS indexed journals.

Head of the research group
Assoc. Prof. Aleksandar Dimov, PhD
Members of the group
Prof. Kalinka Kaloyanova, PhD
Prof. Pavel Boychev, PhD
Assoc. Prof. Trifon Trifonov, PhD
Assoc. Prof. Ioannis Patias, PhD
Assoc. Prof. Radoslava Hristova, PhD
Assoc. Prof. Kalin Georgiev, PhD
Assoc. Prof. Atanas Semerdzhiev
Ch. Assist..Tasos Papapostolu, PhD
Ch. Assistant Yavor Dankov, PhD
Ch. Assistant Dafinka Miteva, PhD
Melania Berbatova, PhD student, PhD
Simeon Emanuilov, PhD student, PhD
Yoan Salambashev, MsC Student, PhD
Ivan Makaveev, BsC Student, PhD
Boris Vasilev, BsC Student, PhD

HALLUCINATIONS IN LLMs FOR BULGARIAN LANGUAGE

An important work is on evaluation of the hallucination of large language models for the Bulgarian language. We make a research on what evaluation methods for measuring hallucinations exist. Next, we give an overview of the multilingual evaluation of the latest large language models, focusing on the evaluation of the performance in Bulgarian on tasks, related to hallucination. We also present a method to evaluate the level of hallucination in a given language with no reference data, and provide some initial experiments with this method in Bulgarian.

Another research explores two distinct methodologies for detecting toxic content. The developed methodologies have potential applications across diverse online platforms and content moderation systems. We propose an ontology that models the potentially toxic words in Bulgarian language. Then, we compose a dataset that comprises 4,384 manually annotated sentences from Bulgarian online forums across four categories: toxic language, medical terminology, non-toxic language, and terms related to minority communities. We then train a BERT-based model for toxic language classification, which reaches a 0.89 F1 macro score. The trained model is directly applicable in a real environment and can be integrated as a component of toxic content detection systems.

TSL TEXTURES

Textures are a traditional technology to achieve realistic graphics. They are images applied onto the surface of 3D objects. Although they are widely used, they have intrinsic limitations in respect to memory footprint, 3D surface topology and real-time modification. Three.js Shading Language (TSL) is a new technology still under a rapid development that avoids these restrictions. TSL can be used to effectively generate 3D textures that are calculated in real time by the GPU processor. This generation is available for both desktop and mobile devices. For modern browsers TSL textures are compiled to use the WebGPU API, if this is not supported, TSL falls back to using WebGL2.

Currently over 30 configurable TSL textures are defined, ranging from abstract forms to natural formations and biological patterns. In addition, the application of TSL textures is extended to modify not only colors, but also shapes of objects. For example, a 3D model that is not designed to be animated, can be animated while its surface features are preserved. This help modelling skin and muscle distortion in biological forms.

3D ASSETS

“3D Assets” is a set of generators of 3D object from real life. The goal is threefold: (1) providing ready-to-use assets for multimedia projects and educational games; (2) mass generation of assets to train image recognition systems and Generative Artificial Intelligent systems; (3) modelling 3D environments for robot training. Each asset can be configured via a set of parameters that controls its dimensions, complexity and components. The generators can be reused in various ways: online real-time generators that users can configure an asset and download it as a GLTF file; or a programming API where the asset is generated in real-time as a 3D model. Key features of the asset generators are the intrinsic support for self-adjusting sizes and proportions of assets elements; and the readjustment of texture coordinates depending on the currently selected parameters.

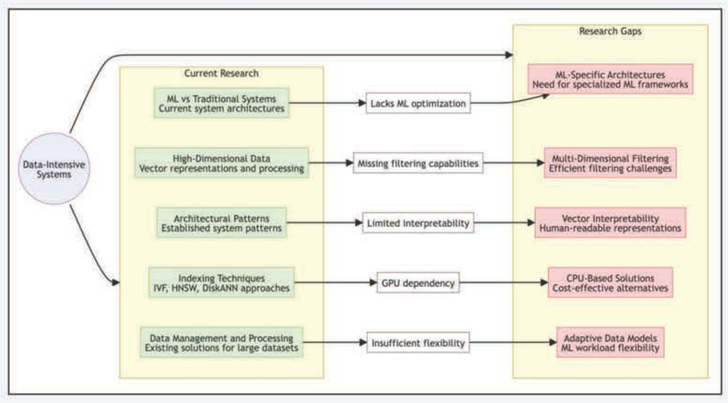
APPLICATION IN MEDICAL DATA

Medical data sets were studied with a focus on their secondary use. New data models were defined, which leveraged these data sets to target new research objectives and address emerging research questions.

A comprehensive series of medical data analyses was conducted, with a focus on a set of analyses related to hospitalization of patients. Applying statistical and machine learning methods to the studied data set, important trends in the patient admissions were identified, enabling researchers to conduct predictive analyses that forecast future trends in hospital admissions. These predictive models can help healthcare to efficiently manage hospital resource allocation and improve the patient care planning.

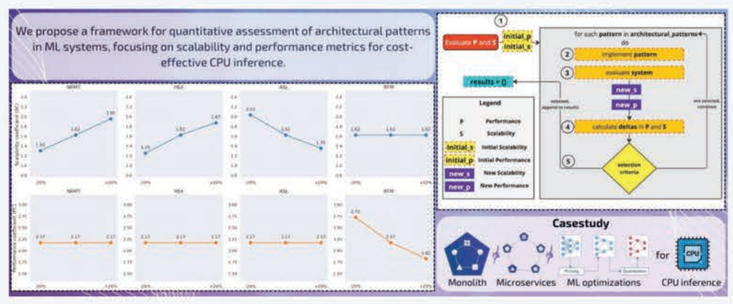
Additionally, systematic efforts are being made to build a logical data warehouse model on clinical data. The model is designed to be suitable for multidimensional analysis, that facilitates efficient data querying and analysis, and supports more effective decision-making in clinical and research settings.

METHODS FOR DEVELOPMENT OF DATA-INTENSIVE SYSTEMS



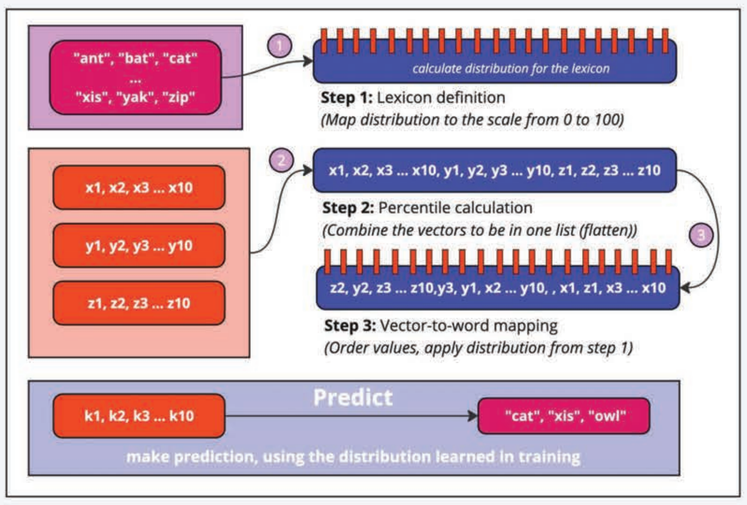
Traditionally, data-intensive systems have emerged from the notion of data-intensive computing, which was inspired by the requirements of traditional science disciplines while moving to the so-called eScience – a research perspective that involves extensive usage of Information and Communication Technologies in all research initiatives.

Nowadays, most computing systems collect and process data from various sources, ranges, and applications. As a result, there is a shift in the area of software engineering from software-intensive systems towards data-intensive systems. More challenges reside in size (e.g., amount of data), its complexity, heterogeneity, and velocity. This requires activities that differ from solving traditional software development problems. For example, in data-intensive systems, the machine's computing power is not the main limiting factor but the I/O and additional quality characteristics like reliability, scalability, maintainability, etc.



LEXICAL REPRESENTATION OF DENSE NUMERICAL VECTORS

High-dimensional numerical vectors are widely used in machine learning for searching and indexing data. However, it is often difficult for users to interpret their meaning. To address this, we introduce a novel approach that transforms dense vectors into human-readable lexical representations using a percentile-based mapping approach. The essence of the approach is a mapping of words from a predefined/custom lexicon to vectors based on their relative local magnitudes. This way, it enables intuitive visualization of the semantic similarities and differences between complex data points and allows for domain-specific interpretability. It provides an easy way to deduplicate dense vectors (even near-duplicates) and can generate locality-aware hash-like representations, which can be used for efficient indexing and retrieval in various applications.



AI-BEST: Artificial Intelligence and Big data for Education, Software and information Technologies

Contact information:
ai-best@fmi.uni-sofia.bg

Contract BG-RRP-2.004-0008 for the financing of project "Sofia University - Marking Momentum for Innovation and Technological Transfer" under pillar 2 "Establishing a network of research higher education institutions in Bulgaria", component "Innovative Bulgaria" from National Recovery and Resilience Plan part of the program to accelerate economic recovery and transformation through science and innovation