RIMAP - Advanced NLP-Based Matchmaking and Reporting Platform

1st Damir Medved, M.Sc.
EDIH ADRIA
University of Rijeka
Rijeka, Croatia
damir.medved@uniri.hr

2nd Assist. Prof. Benedikt Perak, Ph.D.
FFRI - CCL
University of Rijeka
Rijeka, Croatia
bperak@ffri.uniri.hr

***Abstract*—This paper presents the development and implementation of RIMAP 2.0, an advanced NLP-based matchmaking and reporting platform designed to facilitate efficient and effective collaboration among diverse stakeholders. Leveraging cutting-edge AI technologies, RIMAP 2.0 aims to enhance resource utilization, accelerate innovation, and improve decision-making processes. The platform integrates various AI methodologies, including collaborative filtering, content-based filtering, and network analysis, to provide accurate and reliable matchmaking. The paper discusses the platform's architecture, methodologies, implementation challenges, and results, highlighting its potential impact on research and business communities.**

Keywords—matchmaking, NLP, LLM, AI.

I. INTRODUCTION

During the realization of the EDIH ADRIA project in interaction with small and medium-sized enterprises, a serious lack of their capacities for connecting with scientific institutions was identified to improve existing or create new products and services, and in forming a consortium for applying to national or European funds that finance these activities.

Matchmaking platforms can play a crucial role in connecting researchers, institutions, and businesses for collaborative projects. In this context, a few years ago, as a part of University of Rijeka's cooperation with the Regional Development Agency of Primorsko-Goranska county (PRIGODA), the RIMAP (Regional Innovation Matchmaking Platform) platform was created, which tried to create this connectivity function. But it turned out that traditional keyword-based database methods often fall short in efficiency and accuracy, needing the development of more advanced solutions.

One of the main disadvantages of the keyword-based approach in creating a matchmaking platform include limited context understanding. Keyword-based approaches often fail to understand the context behind the search terms. This limitation can lead to irrelevant matches where the keywords might be present, but the actual content or intent does not align with what the user is looking for. On the same issue, keywords can have multiple meanings (homonyms), or different words can have the same meaning (synonyms) which can lead to inaccurate matches. For example, a search for "bank" could refer to a financial institution or the side of a river, causing confusion and inefficiency in the matchmaking process. The underlying issue being that effectiveness of keyword searches heavily relies on the rigid specific words used which can hinder the platform's ability to match and connect suitable partners effectively. Moreover, keywords alone do not capture the semantic meaning of the search query. For instance, a search for "AI in healthcare" should ideally return results that discuss artificial intelligence applications in the medical field, but keyword-based systems might not fully understand this intent, leading to less relevant matches.

Addressing these disadvantages often requires the integration of more advanced techniques such as natural language processing (NLP), machine learning algorithms, and semantic search capabilities. These technologies can better understand user intent, context, and the nuanced meanings behind search queries, leading to more accurate and relevant matchmaking results.

Therefore, in context of EDIH ADRIA with accordance with the concepts of open science, and with the help of CRORIS, Open AIRE and other public platforms, a technological demonstrator was prepared that indicates the possibility of improving the existing RIMAP 1.0 platform with the help of advanced LLM and NLP-based technologies.

An analysis of the use of the RIMAP 1.0 platform so far revealed several shortcomings, namely:

- Data collection – the platform did not possess any automation for collecting or updating data but depended entirely on manual data entry.
- Data update – given that the process of updating the data was also manual, it turned out that users did not do it, which very quickly led to inadequacy and inaccuracies of data in the system.
- Data cleanup – a system for checking or cleaning data was not implemented, and users were not stimulated to participate in it.
- Data enrichment – since no links to external systems were implemented, there was no possibility to enrich information with external data sources.
- Data ranking – no ranking or classification of data was implemented within the platform, both in the context of

the time of creation of the content and in terms of their value

- Data search – data search was only possible through keyword search; more complex queries were not possible.
- Data analytics – only rudimental analytics and query statistics in the database are implemented.

All these shortcomings meant that the platform was poorly used and that it is necessary to do its complete reengineering if it is to be useful to all stakeholders in the innovation process.

So EDIH ADRIA project team decided to revisit RIMAP and check if following objectives could be met:

- To introduce an AI-based matchmaking platform with enhanced semantic search capabilities.
- To discuss the architecture and methodologies used – scalability, upgradability, ease of use.
- To evaluate the platform's performance and user satisfaction.

II. MATCHMAKING CONCEPTS

Traditional matchmaking systems often rely on relational database queries and manual data handling, which face significant challenges in terms of scalability and accuracy. Determining the optimal matchmaking approach involves considering the specific requirements of the application, the nature of the data, and the desired outcomes. The following approaches are commonly considered:

1. **Traditional Database-Based Queries:**
   - Advantages: Relational databases using SQL queries are well-established, simple to build, super-fast, and cost-effective. These systems can quickly retrieve data using indexed queries [1].
   - Disadvantages: They require extensive manual work for data cleaning and updates and struggle with scalability and complex relationship management [2].
2. **NoSQL Databases:**
   - Advantages: NoSQL databases like MongoDB and Cassandra provide flexibility in handling unstructured data and can scale horizontally, handling large volumes of data efficiently [3]. They support various data models, including document, key-value, column-family, and graph databases [4].
   - Disadvantages: While they offer flexibility and scalability, they often lack the strong consistency guarantees provided by traditional SQL databases and can introduce complexity in data management and query optimization.
3. **Graph-Based Methods:**
   - Advantages: Graph databases, such as Neo4j, excel in handling highly interconnected data, making them suitable for recommendation systems where relationships between entities are complex and dynamic [5]. Graph algorithms can efficiently traverse relationships to provide accurate recommendations [6].
   - Disadvantages: These systems can be complex to implement and require specialized knowledge to design and maintain effectively.
4. **Advanced AI Technologies:**
   - Advantages: Techniques such as collaborative filtering [7], content-based filtering [8], and machine learning algorithms, including deep learning and natural language processing (NLP) [9], offer significant potential for improving matchmaking accuracy and efficiency. Vector databases, used in large language models, enable sophisticated semantic search and recommendation capabilities [10].
   - Disadvantages: These technologies are complex to implement, frequently evolving, and can be expensive. They often involve opaque processes ("black boxes") and can be unpredictable, necessitating robust validation and monitoring frameworks.
5. **Hybrid Models:**
   - Advantages: Combining the strengths of traditional methods, NoSQL, graph databases, and AI technologies can lead to optimized performance and cost-efficiency. Hybrid models can leverage the scalability of NoSQL, the relational insights from graph databases, and the advanced analytical capabilities of AI [11].
   - Disadvantages: While mitigating some issues of individual approaches, hybrid models still require careful integration and management to be effective.

Given the complexity and diverse needs of modern matchmaking platforms, our development team evaluated several trials and decided to use a hybrid approach. This approach leverages multiple AI methodologies, including collaborative filtering and NLP, integrated with NoSQL and graph methods to enhance matchmaking accuracy and efficiency. This hybrid model ensures that the platform’s operating costs remain within acceptable limits while maximizing the effectiveness of the matchmaking process [12][13].

III. METHODOLOGY

RIMAP 2.0's architecture integrates advanced AI components, including an intelligent AI-based harvester for automated data capture, natural language processing (NLP) for metadata enrichment and classification, large language model (LLM) vector databases, and an advanced matchmaking assistant (chatbot) to enhance user experience and facilitate seamless interactions.

A. RIMAP Backend

RIMAP architecture is depicted on (Fig.1) and is aimed to assure simplified user experience on frontend and maximum performance and upgradability on backend.

Fig. 1 RIMAP2.0 architecture

Harvester - The platform utilizes both user-provided data and automatically retrieved information from external databases such as CRORIS [14], OpenAIRE [15], ORCID [16], and other research repositories. The automated data capture ensures comprehensive and up-to-date information. The data is partially stored in the internal data storage, which helps to maintain data upscaling and increases possibility for contextual matchmaking. The harvester dynamically updates metadata and the vector database, ensuring that the platform remains current with the latest research activities and opportunities.

NLP metadata enrichment – Metadata within RIMAP is enriched using NLP techniques, which process and analyse the content of research publications, project proposals, company product and service information, and available funding opportunities. The use of large language models (LLMs) such as BERT [167] and GPT [10], enables the expansion of keywords to encompass semantic similarity through lexical networks and embeddings. This metadata update is reducing overall platform costs and optimising matchmaking searches (consumption of tokens).

Keyword Expansion with LLMs - LLMs leverage vast amounts of textual data to understand and generate human-like text. In RIMAP, these models are used to expand keywords by analysing the semantic context in which terms are used. For example, an initial keyword like "artificial intelligence" can be expanded to include semantically similar terms such as "machine learning," "neural networks," and "deep learning," ensuring a more comprehensive metadata enrichment process. This is achieved through the embeddings generated by the LLMs, which capture the contextual meaning of words [18], [19].

Lexical Networks and Embeddings - Using lexical networks [20], and advanced embeddings, RIMAP can map relationships between words and their meanings. Word embeddings, such as those generated by Word2Vec [9], or GloVe [19], but also OpenAI's text-embedding-3-large models are utilized to create dense vector representations of words. These vectors capture semantic similarity, enabling the platform to understand that terms like "AI" and "machine learning" are related, even if they do not appear in the same context. This process enhances the accuracy of metadata enrichment by ensuring that related concepts are included in the search and matchmaking algorithms.

This enriched metadata enhances the platform's ability to conduct efficient and accurate matchmaking searches while optimizing resource consumption, such as the use of computational tokens. By incorporating sophisticated NLP algorithms, RIMAP can understand and classify complex textual information, leading to more relevant and precise matchmaking results [9], [17].

Private vector database LLM/RAG - A private vector database [21] is constructed using LLMs and retrieval-augmented generation (RAG) techniques. Retrieval-augmented generation (RAG) techniques combine information retrieval methods with generative models, allowing the system to retrieve relevant documents from a large corpus and use this information to generate more accurate and contextually relevant responses. Within a retrieval system there are several databases that contribute to contextualisation, including a network graph of relationships between scientists, laboratories, industry partners, and funding sources. Graph algorithms, such as community detection [22] and centrality measures [23] are applied to identify potential collaborations by detecting clusters of researchers with common interests. This approach enables the transparent identification of key influencers and emerging research trends, facilitating strategic matchmaking that aligns with users' specific needs and objectives.

B. RIMAP Fontend

The frontend of the RIMAP 2.0 platform is designed to be user-friendly and highly functional, utilizing web technologies and AI integration. It is built on WordPress, a popular content management system, with custom widgets and modules to enhance its capabilities. The platform's design is powered by CSS to ensure a responsive and visually appealing interface.

A key feature of the frontend is the AI assistant, which is created using Flask, a lightweight Python web framework. This AI assistant is embedded within the WordPress environment through custom templates, enabling seamless interaction with users.

The AI assistant includes a speech-to-text module, leveraging light Python libraries to convert spoken words into text. This feature enhances user input methods, although it faces challenges with misspellings and inaccuracies. These issues are particularly pronounced for the Croatian language, necessitating the adoption of more efficient and accurate models tailored to Croatian linguistic nuances to improve performance and user experience.

Additionally, the AI assistant incorporates a text-to-speech module, enabling the conversion of text into spoken words. However, generating natural-sounding speech in Croatian presents challenges, such as correctly pronouncing complex words and maintaining a natural intonation. Addressing these problems requires advanced models and further refinement to ensure clarity and naturalness in generated speech, thereby improving accessibility and user satisfaction.

IV. DEVELOPMENT PROCESS

RIMAP 2.0 was developed through a series of phases, including initial metadata collection, algorithm selection, and system integration. Core components was completely redesigned offering unlimited scalability, introducing multilingual capabilities and API interfaces to automate interactions with large users. Advanced communication and

information capabilities in form of specialized thematic newsletters and info dispatch, where also introduced to assure prompt information regarding potential partnerships or available funding is not overlooked (Fig. 2).

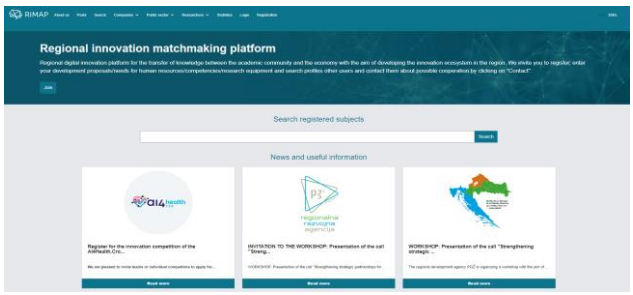


Fig. 2 RIMAP platform, user experience

A. *Implementation*

AI portion of project was completed using multiple open-source components, based on platform usage final decision regarding appropriate licencing will be adopted in 2024.

For Data Augmentation and Enrichment Langchain was used to augment and enrich the existing dataset by generating synthetic data, expanding the variety of potential matches, and improving the diversity and quality of recommendations.

For Text Analysis and Understanding Open AI was used to analyse and understand the textual descriptions of scientists' research areas, laboratory capabilities, industry needs, and funding opportunities. This understanding was crucial for metadata updates and accurately match entities based on their characteristics and requirements.

Langchain is powering recommendation systems by generating personalized suggestions for potential collaborations, funding opportunities, and industry partnerships based on the input provided by users and historical data.

For Chatbots and Conversational Interfaces Open AI assistants were used to build conversational interfaces to interact with users, gather requirements, aid, and answer queries related to matchmaking, collaborations, and funding opportunities.

This methodology improves the precision and relevance of matchmaking ensuring that the platform operates efficiently while maintaining cost-effectiveness by minimizing unnecessary data storage and processing. The combination of these technologies positions RIMAP 2.0 as a leading solution for fostering collaborations and advancing research and innovation efforts.

B. *Challenges*

As usually happens most of the problem on the project stemmed from the context of the data and not so much from the context of the process or the design of the platform itself. Key challenges included data preprocessing, algorithm tuning, and ensuring user privacy and data security. There are three areas that spend the most project team time and energy:

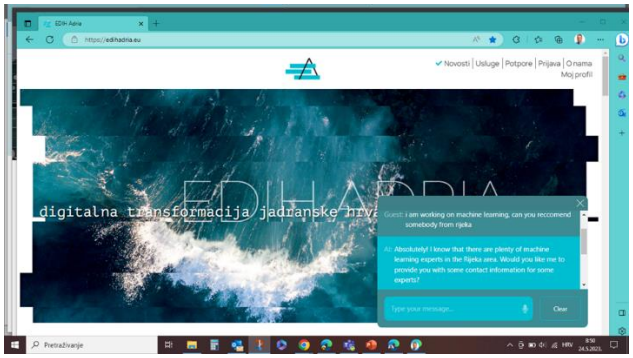
Data Quality and Preprocessing - Most of the time was spent to assure that data used by the matchmaking system is accurate, relevant, and up to date. Team conducted thorough data preprocessing, including data cleaning, normalization, and feature engineering, to enhance the quality of input data and remove bias, noise, or inconsistencies.

Feature Selection and Engineering - Second most demanding effort was to identify and select the most relevant features that contribute to matchmaking success. Team explored advanced feature engineering techniques to create new features or representations that capture important patterns and relationships in the data. Key motivation was to increase speed of the platform and reduce operational costs.

Algorithm Selection and Optimization - Third most demanding component was evaluation and comparison of different matchmaking algorithms to identify the most suitable ones for the specific context and objectives of the system. Hyperparameters where fine-tuned through techniques like cross-validation or grid search to optimize performance.

V. RESULTS

The modular approach used within RIMAP 2.0 leverages state-of-the-art technologies to create a robust and efficient matchmaking platform. The AI-based harvester ensures continuous data updates, while NLP enrichment processes enhance the semantic understanding of metadata, leading to more accurate matchmaking outcomes. The use of a private vector database supported by LLMs and advanced graph algorithms allows for the dynamic visualization and analysis of complex relationships within the research and innovation ecosystem. Special attention was paid to the creation of integrable interfaces (chatbots and assistants) that can be ebmed into partner websites (Fig. 3).



Fig, 3 RIMAP chatbot embedded in EDIH ADRIA WEB page.

Performance Metrics:

The platform's performance was evaluated using metrics such as matchmaking accuracy, user satisfaction, and processing speed.

The key advantage has been shown to be the specialization of the matchmaking platform and AI functionality provide unified proposals on all dimensions of searching a particular topic (Fig. 4). And the search for such information through internet search engines takes 5-7 times longer and does not give an appropriate result.

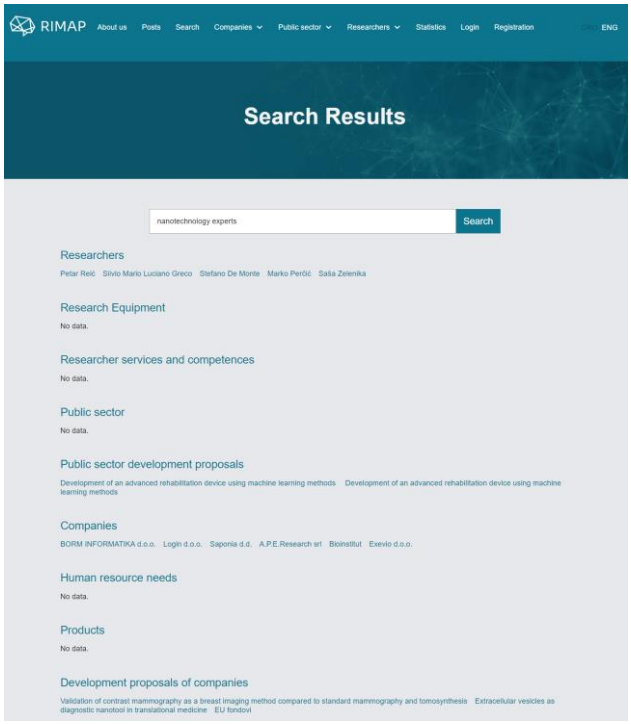


Fig. 4 RIMAP comprehensive matchmaking proposal

Another popular way to interact with RIMAP platform is use of embedded chatbot interface. Here interaction is executed in more fluid and natural conversation with AI assistant. In (Fig. 5) is visible typical interaction and comprehensive platform answer identifying appropriate subject expert and available laboratory equipment for given input (piezo technology in this case).

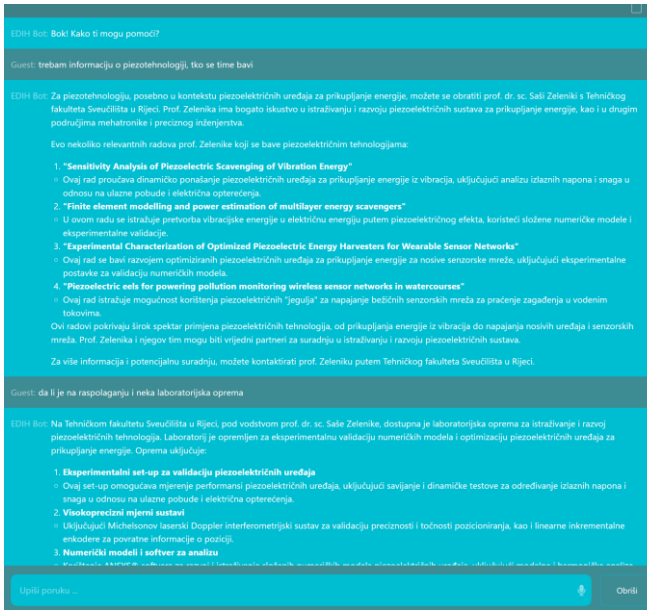


Fig. 5 Interaction with RIMAP chatbot

Of course, the current platform limitation is that matchmaking is possible only for organizations that are members and researchers who have shared their data on the platform, so in the coming period intensive work will be done on its popularization and increase of user base.

VI. Discussion

The results indicate that integrating multiple AI methodologies significantly improves matchmaking accuracy and user satisfaction, however, despite its success, RIMAP 2.0 faces number of limitations mostly related to GDPR and IP rights.

Ensuring fairness, transparency, and bias mitigation are critical for the platform's ethical deployment. As EDIH ADRIA project has ethical committee as part of project, and University of Rijeka adopted Artificial Intelligence Tools Usage Policy [24], project team had lot of attention to ethical considerations such as fairness, transparency, and bias mitigation throughout the development and deployment of the matchmaking system. Measures to detect and mitigate biases in data, algorithms, and recommendations were implemented to ensure fair and unbiased outcomes.

RIMAP is fully compliant with GDPR requirements. Users have full control over their personal data, including the permission to access, rectify, and erase their data. RIMAP is providing mechanisms for users to exercise these rights and ensure that their personal data is accurate, up-to-date, and securely managed.

RIMAP implemented feedback mechanisms that allow users to provide feedback on suggested matches or recommendations. The team is using this feedback to continuously update and refine the matchmaking algorithms, improving their accuracy over time.

One of the problems found is the lack of domain expertise and contextual information for improvement matchmaking process. Factors such as domain-specific knowledge, business rules, constraints, and user preferences must be used to a greater extent to tailor the matchmaking algorithm to the specific needs of the domain.

VII. Conclusion

RIMAP 2.0 represents a significant advancement in matchmaking technology, combining various AI techniques to enhance collaboration.

Accurate matchmaking ensures that resources such as research funding, laboratory facilities, and expertise are efficiently allocated to projects and collaborations with the highest likelihood of success. This leads to better resource utilization and maximizes the impact of investments. On the other hand, by connecting researchers, laboratories, and industry partners more effectively, accurate matchmaking accelerates the pace of innovation and progress in various fields. This results in faster development of new technologies, products, and solutions that benefit society and drive economic growth.

And last, but not least, accurate matchmaking expands collaboration opportunities by identifying compatible partners with complementary skills, expertise, and resources. This fosters interdisciplinary collaboration and knowledge exchange, leading to more impactful research outcomes and breakthrough discoveries.

Future developments will focus on expanding the platform's capabilities, incorporating more diverse data sources, and refining algorithms to further improve accuracy and efficiency. **Data Quality** will be ensured through thorough preprocessing and feature engineering. And **Algorithm Optimization** will be permanently improved using cross-validation and parameter tuning techniques.

ACKNOWLEDGMENT

This work has been financially supported by the University of Rijeka EDIH ADRIA project and this support is gratefully acknowledged.

REFERENCES

[1] Stonebraker, M., et al. (1986). The design of the Postgres storage system. VLDB '86: Proceedings of the Twelfth International Conference on Very Large Data Bases.

[2] Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, 13(6), 377-387.

[3] Stonebraker, M. (2010). SQL databases v. NoSQL databases. Communications of the ACM, 53(4), 10-11.

[4] Han, J., et al. (2011). Survey on NoSQL database. Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE.

[5] Angles, R., & Gutierrez, C. (2008). Survey of graph database models. ACM Computing Surveys (CSUR), 40(1), 1-39.

[6] Gupta, M., et al. (2014). Wtf: The who-to-follow system at twitter. Proceedings of the 22nd international conference on World Wide Web, 505-514.

[7] Sarwar, B., et al. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th international conference on World Wide Web.

[8] Lops, P., et al. (2011). Content-based recommender systems: State of the art and trends. Recommender systems handbook, 73-105.

[9] Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[10] Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[11] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, 12(4), 331-370.

[12] Resnick, P., et al. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work.

[13] Schafer, J. B., et al. (1999). Recommender systems in e-commerce. Proceedings of the 1st ACM conference on Electronic commerce.

[14] Končić, I., Konjević, S., Hoić, M., & Macan, B. (2024). The role of CroRIS in promoting Open Science in Croatia.

[15] Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., ... & De Bonis, M. (2019). The OpenAIRE research graph data model. Zenodo.

[16] Baessa, M., Lery, T., Grenz, D., & Vijayakumar, J. K. (2015). Connecting the pieces: Using ORCIDs to improve research impact and repositories. F1000Research, 4.

[17] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[18] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[19] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

[20] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[21] Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. arXiv preprint arXiv:2310.11703.

[22] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

[23] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, 40(1), 35-41.

[24] Zelenika, S., Meštrović, A., Pošćić, A., Lerga, J., Medved, D., Lazzarich, L., Jelenić, G., University of Rijeka. (2024). "Artificial Intelligence Tools Usage Policy at the University of Rijeka", Croatia. Zenodo.