JobKG: A Knowledge Graph of the Romanian Job Market based on Natural Language Processing

Petre Caraiani[1], Liviu-Adrian Cotfas[2], Flavius Cosmin Darie[1], Camelia Delcea[2], Carlos Angel Iglesias[3] and Radu Prodan[4]

[1]Faculty of Business Administration in Foreign Languages, Bucharest University of Economic Studies, Romania
[2]Department of Economic Cybernetics and Informatics, Bucharest University of Economic Studies, Romania
[3]Departament of Telematic Engineering Systems, Polytechnic University of Madrid, Spain
[4]Institute for Information Technology, University of Klagenfurt, Austria

Abstract

The JobKG project aims to comprehensively analyze the Romanian labor market using data available on online recruiting platforms deploying state-of-art approaches in natural language processing (NLP), semantic web, and agent-based modeling (ABM). For this purpose, it performs an extensive quantitative and qualitative analysis of the Romanian labor market to prospect the opinion of various groups of stakeholders regarding the alignment and calibration of skills demanded by the labor market and those developed and trained in the education system. Given that employment-oriented online services are the main source of information for job seekers willing to search and apply for vacant positions, they represent a rich data source for understanding the occupations in demand and the relevant skills required. The project will create a knowledge graph of the Romanian labor market, starting from existing taxonomies and extracting data using advanced NLP techniques. An ABM approach will analyze the labor market dynamics, considering various scenarios and defining agents with characteristics similar to those of the entities (job seekers and firms), which will simulate the demand and supply.

Keywords

Knowledge graph, job market, natural language processing, agent-based models

1. Introduction

We live in a world profoundly characterized by digitalization that has entered a new phase of Industry 4.0 with technological breakthroughs in many fields: AI, robotics, IoT, biotechnology, healthcare, energy, and blockchain, to name a few. Implementing technological innovation, new business models, changing job requirements, and demographic pressures due to aging and falling birth rates shift employers' required skill structure. While over one-third of the skills needed should have changed by 2020 [1], at least one-fourth of OECD workers report skill mismatches with their current job requirements. Estimates produced by several cross-country analyses indicate that 49 % to 69 % of the currently employed people could lose their jobs in Romania due to automation [2]. Moreover, recent developments indicate substantial changes in the task content of jobs, with a sizable share likely taken over by AI [1]. For developing countries, the readiness of their labor markets and economies to comply with digitalization trends regarding their existing stock of applicable skills and ability to learn them is a significant issue [3]. In the search for solutions, education and training are essential in achieving the right mix of technical skills [4] as a critical element to support skills delivery through tertiary education, integrating skills policies with other labor market policies.

The importance of this topic resides in the significant impact digitalization shall have on the labor market in the future in the composition of the labor force, the creation of second-round impacts on the training needs, skills requirements, and reskilling of existing workers, and the substantial change of the policies of businesses and production processes. The difficulties stem from understanding the evolving skill demands of the job market, which necessitates analyzing job postings over an extended

RuleML+RR 2024: The 8th International Joint Conference on Rules and Reasoning, September 16–18, 2024, Bucharest, Romania
petre.caraiani@gmail.com (P. Caraiani); liviu.cotfas@ase.ro (L. Cotfas); flavius.darie@fabiz.ase.ro (F. C. Darie); camelia.delcea@csie.ase.ro (C. Delcea); carlosangel.iglesias@upm.es (C. A. Iglesias); radu.prodan@aau.at (R. Prodan)


period. However, the laborious nature of manually undertaking this task and the absence of efficient data extraction algorithms for the Romanian language have rendered this endeavor impractical.

Current research limitations exist in *agent-based models* (ABM) for analyzing the job market dynamics, which cannot provide a useful representation of reality [5]. Recent *knowledge graph* (KG) approaches for the job market prove their potential but have a limited scope [6, 7]. Additional limitations exist in the Romanian job market, stemming from the lack of evolution of skills data in the job market and of advanced *natural language processing* NLP algorithms adapted for the Romanian language.

The *JobKG project* will be the first comprehensive research carried out in Romania, which will use open data from job postings and job-related information to map the supply and demand for job skills based on NLP. A transformer-based model will enhance the state-of-the-art method of extracting data from Romanian text. Additionally, the project will be the first attempt to reconcile qualitative and quantitative results to cross-validate them. Furthermore, integrating information using KG methodology is a state-of-the-art practice that will enable the integration and visualization of results in a more accessible and complex manner. Using KG to improve the capabilities and insights derived from ABM is another project contribution, facilitating fact-based policymaking on the labor market and adjacent fields. Thus, the project will yield actionable results and research findings, following harmonized Romanian and international occupational and skills classification schemes to increase applicability.

2. Objectives

The JobKG project aims to reduce the gap between job seekers' skills and the skills that employers require in the Romanian labor market. The recommendations made through the project could positively affect the labor market by overcoming skills barriers to employment. The project considers several specific objectives to achieve its overall goal.

O1: Research on multidisciplinary approaches to understand the Romanian job market, combining qualitative and quantitative analysis based on statistical and data analysis;

O2: Research semantic and big data technologies for modeling the Romanian job market as a KG;

O3: Research semantic similarity algorithms for matching job offer and demand;

O4: Research on NLP and deep learning (DL) technologies for extracting and cataloging skills from online data sources;

O5: Research data-driven what-if prediction models to simulate the Romanian job market;

O6: Create an online platform as a support tool for job seekers, companies, and learning providers;

O7: Propose policy recommendations based on ABM-simulated scenarios on the specific Romanian job market conditions.

3. Methods

The JobKG project will commence with quantitative and qualitative analyses to gain a comprehensive understanding of the labor market in Romania. Subsequently, it will employ NLP and KGs to analyze data from major employment-oriented online services. Afterward, it will develop a policy decision support tool based on agent-based modeling. Figure 1 illustrates the project's conceptual architecture.

3.1. Quantitative and Qualitative Analysis

To better understand the labor market in Romania, the project will conduct a qualitative (literature review, semi-structured interviews, focus groups) and a quantitative (statistical analysis of macroeconomic indicators and questionnaire-based survey) analysis.

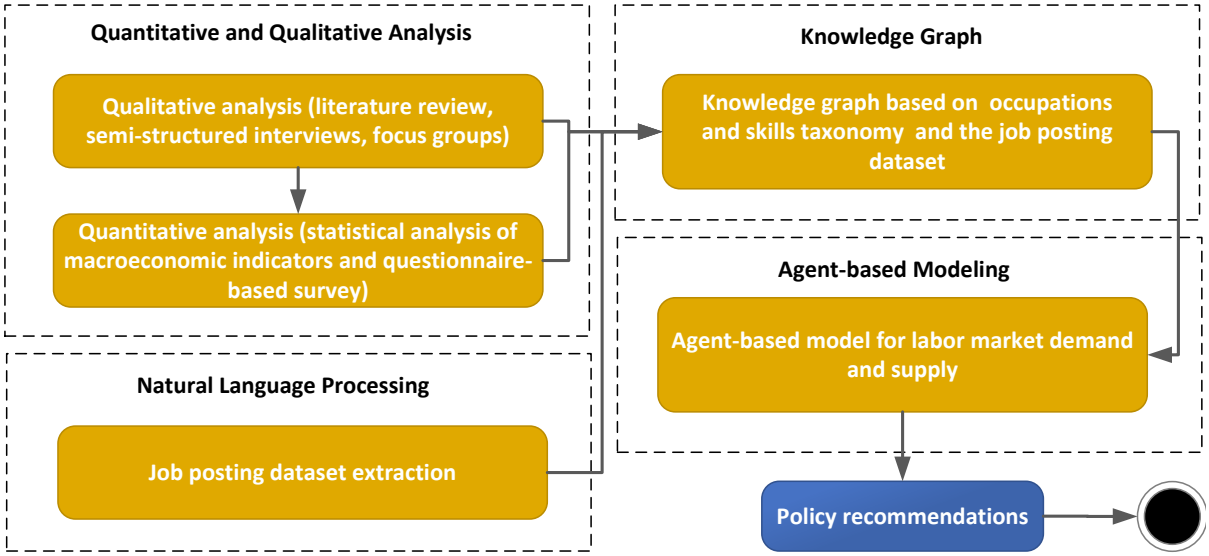

Figure 1: JobKG conceptual architecture.

Initially, it will perform extensive literature research on the taxonomy of skills and their role in employability, focusing on the Romanian labor market. It will analyze several indicators related to the labor market and education to evaluate the mismatch between the demand and supply of skills in the Romanian labor market. Comparative analyses with similar indicators exhibited by other countries in the region might also be relevant to placing the Romanian labor market in the European context.

Next, the project will conduct qualitative research based on in-depth semi-structured interviews and focus groups. The qualitative research aims to prospect the opinion of various groups of stakeholders (e.g., representatives of employers, representatives of educational institutions, students, and decision-makers in the field of education and the labor market) regarding the alignment and calibration of skills demanded by the labor market and trained in the education system. Specialized software, such as NVivo, ATLAS.ti, and MAXQDA, will process the obtained data.

3.2. Natural Language Processing

In addition to the traditional qualitative and quantitative state-of-the-art approaches, JobKG will rely on extracting and analyzing large volumes of data using NLP techniques and semantic web technologies to understand the labor market in Romania fully. Employment-oriented online services are the primary way job seekers search and apply for vacant positions. Thus, such platforms represent a rich data source for understanding the demanded occupations and the required relevant skills. However, the data on these platforms is available in a format that can facilitate data analysis; hence, extracting relevant information using NLP and representing the data using the semantic web is necessary.

JobKG will further use NLP techniques to analyze the job posting dataset data to extract the soft and hard skills mentioned, with benefit from the quantitative and qualitative analysis results. As job markets change, new skills will emerge, uncovered by existing taxonomies [8].

To detect such skills, after an initial cleanup phase, the project will compare the performance of various machine learning and DL algorithms based on F-Score to determine the best-performing algorithm. In the category of DL algorithms, transformers derived from BERT [9] will receive special attention, providing state-of-the-art results in many NLP tasks, including skills extraction from English texts [10]. JobKG will use a pre-trained BERT model for the Romanian language [11], fine-tuned on job postings annotated by two experienced experts with a level of agreement evaluated using Cohen's Kappa. Semantic similarity techniques will evaluate the skills matching, and an n-grams Jaccard similarity compared the best approach with more advanced techniques.

3.3. Knowledge Graph

The JobKG project will create a KG graph [6] of the Romanian labor market by combining the occupations and skills taxonomy with the data extracted from the job posting dataset. Recent KG approaches for the job market prove the approach's potential but have a limited scope [6, 7]. Additional limitations exist in the context of the Romanian job market, stemming from the lack of data regarding the evolution of skills in the job market and the lack of advanced NLP algorithms adapted for the Romanian language.

The JobKG project addresses such challenges through advanced methodologies by using two web-scraping tools, Scrapy and Selenium, targeting the most popular Romanian job platforms to revolutionize its labor market, ejobs.ro and bestjobs.ro, which will keep the KG graph up-to-date with the real-time market dynamics. On top of it, the project will accurately classify soft and hard skills from job postings by leveraging NLP and machine learning and deep learning models such as BERT to identify new emerging competencies not included in current static taxonomies and enrich the existing KG graph with more relevant information. The integration of these technologies will ultimately enable JobKG to offer a more accurate and real-time depiction of the Romanian labor market landscape.

3.4. Agent-based Modeling

An ABM will define agents mirroring job seekers and firms to simulate labor market demand and supply, thanks to its ability to handle complexity through heterogeneous agents [12], representing dynamic interactions. We selected ABM, as it can easily handle complexity through heterogeneous agents, dynamically reflecting people and firms and their interactions, and works well when modeling various situations at the microeconomic level and observing the results at a macro scale [12].

To gain a comprehensive grasp of ABM, it is imperative to establish a nuanced understanding of its foundational principles. Within the ABM paradigm, a system assembles autonomous decision-making entities known as agents, operating within a defined set of rules, guiding its decision-making processes and allowing it to independently assess the prevailing circumstances. The behavioral responses of agents are contingent upon the specific system to which they belong, resulting in varied manifestations.

The model will include skill vectors for job seekers and job offers, reflecting characteristics from the KG and quantitative analysis. This approach allows job seekers to observe real-life skill matches and adaptations. To ensure representativity, the model will also account for new skills emerging from technological changes, simulated in extensive scenarios based on expected market changes. Calibration and validation will use previous results (Section 3.1 and Section 3.2), implemented in NetLogo and following ODD or ODD+D protocols for best practices.

4. Conclusion

The design of the JobKG project addresses the significant changes in the Romanian labor market amidst the digitalization and Industry 4.0 revolution, hoping to achieve unparalleled success in job innovation. The project will harness state-of-the-art technologies such as NLP, KG, and ABM to establish a nuanced and rigorously constructed picture of job demands and the required knowledge components. Building on these foundations, sophisticated new methods will enable the routine extraction and analysis of data from large national job platforms to depict emerging and accurately used skills.

By combining different techniques and taking an interdisciplinary path, the JobKG project aims to significantly advance the understanding of Romania's labor market dynamics. A comprehensive KG will enhance job matching, bridging gaps between job seekers' skills and employers' requirements. Policy recommendations derived using ABM will simulate labor market scenarios, address job market gaps, and inform workforce development strategies. Future research could extend the JobKG approach to multiple countries.

Acknowledgement

The Romanian Ministry of Research, Innovation, and Digitalization, project "JobKG: A Knowledge Graph of the Romanian Job Market based on Natural Language Processing", CF178/31.07.2023, contract number CN760046/ 23.05.2024 financed by Romania's National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8 supported this work.

References

[1] CEDEFOP, Artificial or human intelligence? Digitalisation and the Future of Jobs and Skills: opportunities and risks, Technical Report, CEDEFOP, 2019. URL: http://data.europa.eu/doi/10.2801/164782. doi:doi/10.2801/862703.

[2] World Bank, World Development Report 2016: Digital Dividends, Technical Report, World Bank, 2016. URL: https://www.worldbank.org/en/publication/wdr2016.

[3] Deloitte, Preparing tomorrow's workforce for the Fourth Industrial Revolution, Technical Report, Deloitte, 2018. URL: https://www.unicef.org/rosa/reports/preparing-tomorrows-workforce-fourth-industrial-revolution.

[4] ILO, Changing demand for skills in digital economies and societies: Literature review and case studies from low- and middle-income countries, Technical Report, ILO, 2021. URL: https://www.ilo.org/skills/areas/skills-training-for-poverty-reduction/WCMS_831372/lang--en.

[5] J.-D. Kant, G. Ballot, O. Goudet, WorkSim: An Agent-Based Model of Labor Markets, Journal of Artificial Societies and Social Simulation 23 (2020) 4. URL: http://jasss.soc.surrey.ac.uk/23/4/4.html. doi:10.18564/jasss.4396.

[6] M. de Groot, J. Schutte, D. Graus, Job Posting-Enriched Knowledge Graph for Skills-based Matching, 2021. URL: http://arxiv.org/abs/2109.02554, arXiv:2109.02554 [cs].

[7] K. Yao, J. Zhang, C. Qin, P. Wang, H. Zhu, H. Xiong, Knowledge Enhanced Person-Job Fit for Talent Recruitment, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), 2022, pp. 3467–3480. URL: https://ieeexplore.ieee.org/document/9835552. doi:10.1109/ICDE53745.2022.00325, iSSN: 2375-026X.

[8] I. Khaouja, I. Kassou, M. Ghogho, A Survey on Skill Identification From Online Job Ads, IEEE Access 9 (2021) 118134–118153. doi:10.1109/ACCESS.2021.3106120, conference Name: IEEE Access.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[10] M. Zhang, K. Jensen, S. Sonniks, B. Plank, SkillSpan: Hard and Soft Skill Extraction from English Job Postings, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4962–4984. URL: https://aclanthology.org/2022.naacl-main.366. doi:10.18653/v1/2022.naacl-main.366.

[11] S. Dumitrescu, A.-M. Avram, S. Pyysalo, The birth of Romanian BERT, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4324–4328. URL: https://aclanthology.org/2020.findings-emnlp.387. doi:10.18653/v1/2020.findings-emnlp.387.

[12] L. Hamill, N. Gilbert, Agent-Based Modelling in Economics, 1st edition ed., Wiley, Chichester, UK ; Hoboken, NJ, 2016.