Extracting Competencies from Job Adverts and Academic Syllabi: A Large Language Model Approach

Viktor Andonovikj[1,2], **Pavle Boškoski**[2,3]

[1]*International Postgraduate School Jožef Stefan*
[2]*Jožef Stefan Institute*
[3]*Faculty of Information Studies in Novo Mesto*
E-mail: viktor.andonovikj@ijs.si

Abstract

This paper proposes a method for automatically extracting competencies and skills from syllabi of job advertisements and higher education institutions. The proposed approach leverages sentence transformers to bridge the gap between the natural language used in syllabi and the formal descriptions found in the ESCO (European Skills, Competences, Qualifications and Occupations) ontology. We utilise the pre-trained sentence transformer (SBERT) for generating sentence embeddings and employ cosine similarity for identifying skills within syllabi that closely align with the ESCO framework. We outline the methodology, discuss the potential challenges, and explore the benefits of utilising this approach for educational institutions and students.

1 Introduction

The landscape of higher education is constantly evolving, with a growing emphasis on equipping students with the necessary skills to succeed in the dynamic job market. At the same time, enforced by the rapid technological advancements and dynamic economic shifts, the competencies required by employers are continually evolving. Academic syllabi play a crucial role in this context, outlining the learning objectives and expected outcomes of a particular course. Thus, there is a pressing need for methodologies that can systematically identify and bridge the gap between the skills taught in academic syllabi and those sought by industry. However, manually identifying the specific competencies and skills targeted within a syllabus can be a time-consuming and laborious task. This paper proposes a novel approach for automatically extracting skills from job advertisements and syllabi of higher education institutions.

Our proposed method utilises the power of Natural Language Processing (NLP) techniques, specifically Sentence Transformers (SBERT) (Reimers and Gurevych, 2019), to bridge the gap between the natural language used in syllabi and the formal descriptions found in standardised skill ontologies. By embedding academic syllabi into a high-dimensional latent space, we can effectively utilise them to identify the most relevant competencies for the labour market. The flowchart of the methodology is given in Figure 1.

The results of our methodology can be used by educational institutions to refine their curricula, ensuring that they remain responsive to the evolving demands of the labor market. This approach allows for a data-driven analysis of the alignment between educational content and labor market requirements. The alignment of academic offerings with industry needs is crucial for enhancing the employability of graduates and ensuring that educational institutions produce a workforce that meets the current demands of the labor market. Moreover, this approach is versatile and can be applied across various fields, making it a valuable tool for continuous improvement in higher education.

In the following sections, we detail our data collection process for the experiments, describe the SBERT-based methodology for competence extraction, and present examples from our dataset to illustrate the practical applications of this approach. By leveraging machine learning techniques, this paper contributes to the ongoing efforts to enhance the relevance and impact of academic programs in preparing students for successful careers.

2 Related Work

The relationship between educational attainment, mismatched education levels (both over- and under-education), and the field of education in the context of job search and employment quality has been extensively studied in the literature (Bauer, 2002; Hartog, 2000). While these concepts are not new, previous studies have predominantly relied on qualitative data or registry-based datasets, which often lack granularity or are limited to broad categories such as years of education, fields of study, or specific segments of educational programs.

Prior studies, such as those by (Mardis et al., 2018) and (Alanazi and Benlaria, 2023), have highlighted the challenges and opportunities in aligning academic programs with the evolving demands of the labor market. The methodologies employed in these studies range from qualitative analyses to survey-based approaches (Anastasiu et al., 2017; Terblanche, 2011). More recent approaches have begun to explore employers' needs using data mining techniques, including textual analysis (Maer Matei and Aldea, 2019).

The use of advanced language models for competence analysis has become increasingly common in recent re-

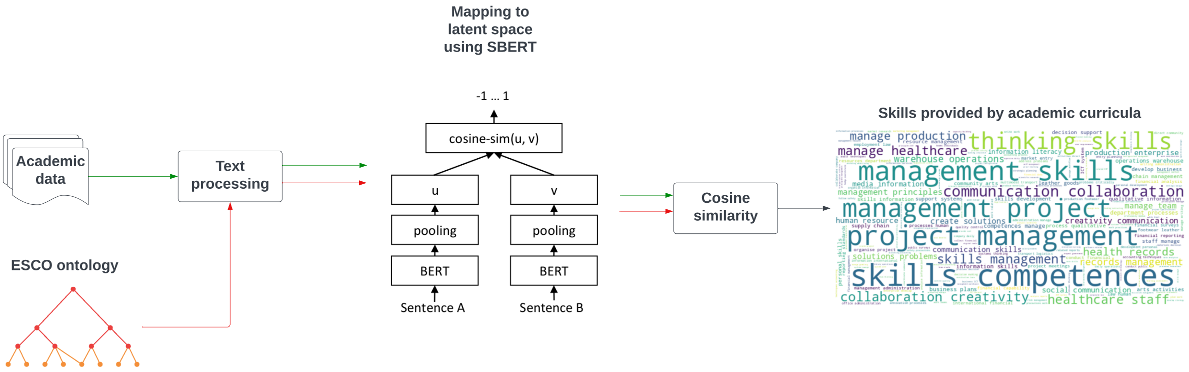

Figure 1: Flowchart of the methodology for competency extraction from academic curricula

search. (Debortoli et al., 2014) applied latent semantic analysis to develop a competency taxonomy for big data and business intelligence, which assists individuals, organizations, and academic institutions in evaluating and enhancing their skills. Additionally, (Schedlbauer et al., 2021) employed text mining of job advertisements to categorise the expected skills for standardised medical informatics, which serves as a foundation for identifying educational goals.

The rapid advancements in statistical techniques, particularly those involving artificial intelligence and large language models, have opened new avenues for research (Andonovikj et al., 2024). For example, (Föll and Thiesse, 2021) used text mining to analyse the content of informatics curricula in Germany. Other researchers, such as (Bommarito et al., 2018), have demonstrated the effectiveness of language models in extracting detailed information from academic datasets. This methodology aligns closely with our approach, where we utilise a language model to extract competencies from academic study plans.

The integration of language models, such as SBERT, with the comparative analysis provided by the European standard classification of occupations (ESCO) framework contributes significantly to ongoing research on decision support in academic planning and the dynamic interplay between academia and industry.

3 Methodology

The methodology involves a systematic approach to extracting competencies from job advertisements and academic syllabi. This approach leverages advanced NLP techniques to facilitate a thorough analysis of the alignment between educational offerings and industry needs. The overall framework is designed to be robust, scalable, and adaptable to various fields of study. It can be divided into three key components:

- data collection and preprocessing;
- text encoding using SBERT;
- extracting competencies from the ESCO ontology.

Data collection and preprocessing In the data collection phase, job advertisements are usually sourced from online job portals, providing a rich dataset of the skills and competencies currently in demand by employers. Job advertisements are selected based on ISCO classifications to ensure relevance to specific areas, such as aligning with the academic syllabi we aim to analyze. For example, for Business and Economics curricula, we choose job adverts under relevant ISCO codes. This ensures that the competencies in the job adverts are directly comparable to those in the academic syllabi, facilitating focused analysis of alignment with labor market needs. Concurrently, academic syllabi are gathered from educational institutions, representing a wide array of courses. These syllabi detail the competencies that students are expected to acquire through their coursework.

After the data is collected, it undergoes a preprocessing stage. This involves normalising the text data to ensure consistency across different sources. Text normalisation includes converting all text to lowercase, removing non-ASCII characters, and eliminating unnecessary whitespace and newline characters. This step is crucial for standardising the textual data, making it suitable for further analysis.

Text encoding using SBERT Our methodology centers on using SBERT to encode text from job advertisements and academic syllabi into 512-dimensional embeddings, which accurately capture the essence of competencies. SBERT handles entire texts as single inputs, generating unified embeddings that represent competencies in a high-dimensional space. We parsed the ESCO ontology to extract and encode both competency labels and descriptions, enabling detailed semantic analysis.

Extracting competencies from the ESCO ontology The ESCO (European Skills, Competences, Qualifications, and Occupations) (Commission et al., 2017) ontology provides a standardised framework for categorising and describing skills, competences, qualifications, and occupations within the European Union. It serves as a bridge between the labor market and education systems by offering a common language that facilitates the alignment of educational outcomes with industry needs. In our methodology, ESCO plays a crucial role by providing a structured reference for mapping extracted competencies from job advertisements and academic syllabi, enabling precise and meaningful comparisons that enhance the relevance and applicability of our analysis.

In the second step of the competence extraction we use cosine similarity to find the 10 most relevant competencies from the ESCO ontology. This metric measures the similarity between two vectors by calculating the cosine of the angle between them. The expression for the cosine similarity is given in Equation 1.

$$\text{cos\_sim}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|}, \qquad (1)$$

The cosine similarity score ranges from -1 to 1, with higher scores indicating greater similarity. The SBERT embedding from the source text is compared to the embeddings of competencies extracted from ESCO. The top 10 most similar ESCO competencies are identified based on their cosine similarity scores, ensuring a robust alignment with standardised competency frameworks.

4 Results and discussion

Using SBERT, we encoded competencies mentioned in job advertisements, gathered through PES (public employment services of Slovenia) and mapped them to the ESCO framework through cosine similarity.

Figure 2 illustrates the most relevant competencies extracted from job advertisements in Slovenia, specifically related to the Business and Economics field. These competencies include a range of management skills, analytical and thinking competences, as well as social and communication skills. The competencies with the highest frequency in job advertisements were management skills, thinking skills and competences, and communication, collaboration, and creativity skills. Specific competencies such as project management, accounting, and human resource management were also prominently featured, indicating their high demand in the Slovenian job market.

Aligning the extracted competencies with the ESCO ontology allowed us to standardise and categorise these competencies, facilitating a more structured comparison. The SBERT embeddings allowed us to perform semantic comparisons, indicating potential alignments between identified competencies from job advertisements and relevant ESCO categories. This alignment process not only validated the competencies but also provided a clearer understanding of the specific skills required in the job market.

Implications for Academic Syllabi While the primary focus of this paper is on the methodology for extracting competencies, the implications of these findings are significant for aligning academic offerings with industry needs. By understanding the competencies in demand, educational institutions can adjust their curricula to better prepare students for the labor market. The detailed extraction and alignment process ensure that the insights derived are actionable and relevant for academic program development.

Discussion The use of SBERT for encoding both job advertisements and ESCO competencies proved to be effective in capturing the semantic nuances of the competencies described in various texts. This approach offers several advantages, including scalability and adaptability to different fields of study. The high-dimensional embeddings generated by SBERT provide a robust foundation for comparing and aligning competencies, making this methodology a valuable tool for both academic and industry stakeholders.

Moreover, the alignment with the ESCO framework ensures that the competencies are categorised according to a standardised European classification system, enhancing the applicability of the findings across different contexts and regions. This standardisation is crucial for facilitating a common understanding of competencies and supporting the alignment of educational outcomes with labor market demands.

5 Conclusion

This paper presents a novel methodology for extracting competencies from job advertisements and academic syllabi using Sentence-BERT (SBERT) embeddings, and the ESCO ontology. We provide a robust framework for identifying and analysing competencies across various textual data sources.

The core focus of this methodology is the extraction of competencies through SBERT and cosine similarity, enabling precise semantic comparisons and mapping to the ESCO framework. This approach is versatile and can be applied across different fields of study, providing detailed insights into the competencies described in both job adverts and academic syllabi.

While the primary contribution is the methodological advancement, its implications extend to aligning academic offerings with industry needs. By systematically identifying areas of alignment and gaps, educational institutions can use these insights to make data-driven adjustments to their curricula, ensuring that graduates are better prepared for the evolving demands of the labor market. The methodology offers a powerful tool for enhancing the relevance and responsiveness of educational programs, contributing to improved alignment between education and employment, and ultimately benefiting both job seekers and employers.

References

Alanazi, A. S. and Benlaria, H. (2023). Bridging higher education outcomes and labour market needs: A study of jouf university graduates in the context of vision 2030. *Social Sciences*, 12(6):360.

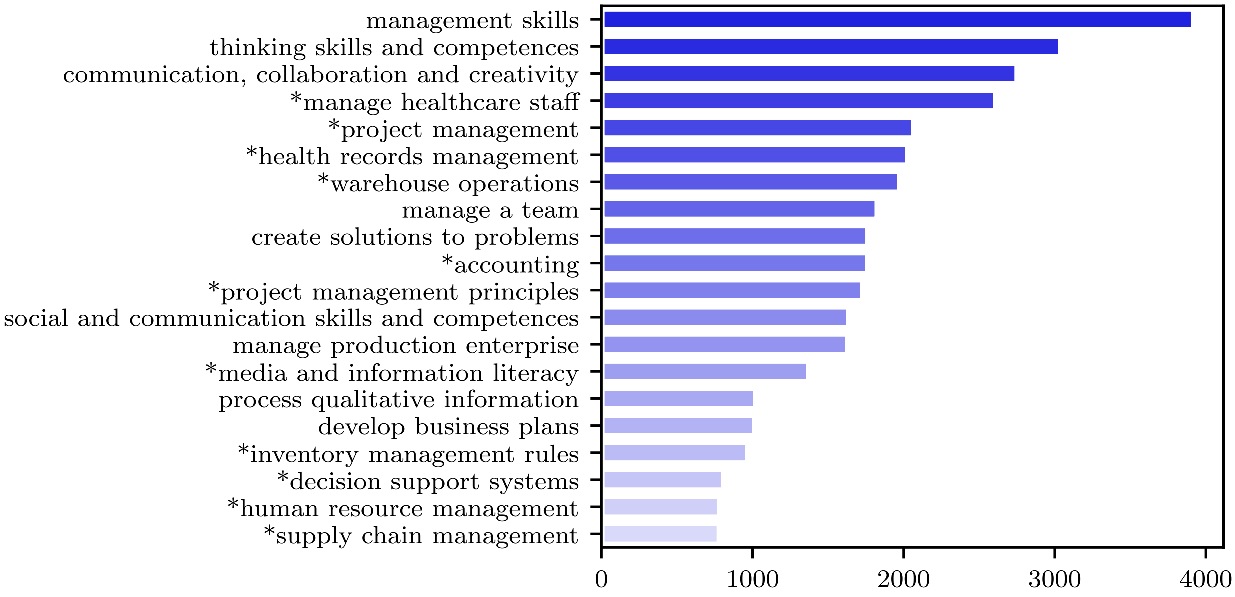

Figure 2: Extracted competencies from job advertisements in Slovenia in the fields of Business and Economics. Competencies marked with * represent ESCO knowledges.

Anastasiu, L., Anastasiu, A., Dumitran, M., Crizboi, C., Holmaghi, A., and Roman, M. (2017). How to align the university curricula with the market demands by developing employability skills in the civil engineering sector. *Education Sciences*, 7(3):74.

Andonovikj, V., Mileva Boshkoska, B., Redek, T., and Boškoski, P. (2024). A data-driven approach to aligning academic offerings with industry needs for business and economy in slovenia. *Journal of Decision Systems*, pages 1–13.

Bauer, T. K. (2002). Educational mismatch and wages: a panel analysis. *Economics of education review*, 21(3):221–229.

Bommarito, M. J., Katz, D. M., and Detterman, E. (2018). Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *SSRN Electronic Journal*.

Commission, E., Directorate-General for Employment, S. A., and Inclusion (2017). *ESCO handbook : European skills, competences, qualifications and occupations*. Publications Office.

Debortoli, S., Müller, O., and Brocke, J. v. (2014). Vergleich von kompetenzanforderungen an business-intelligence- und big-data-spezialisten: Eine text-mining-studie auf basis von stellenausschreibungen. *WIRTSCHAFTSINFORMATIK*, 56(5):315–328.

Föll, P. and Thiesse, F. (2021). Exploring information systems curricula: A text mining approach. *Business & Information Systems Engineering*, 63(6):711–732.

Hartog, J. (2000). Over-education and earnings: where are we, where should we go? *Economics of education review*, 19(2):131–147.

Maer Matei, M. M. and Aldea, A. B. (2019). Employers' requirements for data scientists - an analysis of job posts. *Logos Universality Mentality Education Novelty: Economics & amp; Administrative Sciences*, 4(1):21–32.

Mardis, M. A., Ma, J., Jones, F. R., Ambavarapu, C. R., Kelleher, H. M., Spears, L. I., and McClure, C. R. (2018). Assessing alignment between information technology educational opportunities, professional requirements, and industry demands. *Education and Information Technologies*, 23:1547–1584.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schedlbauer, J., Raptis, G., and Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, 150:104453.

Terblanche, C. (2011). Meeting employers' and students' expectations through the use of employer demand ontology in curriculum development. In *2011 5th IEEE International Conference on E-Learning in Industrial Electronics (ICELIE)*. IEEE.